



**LAS TIC Y SU IMPACTO
EN LA SOCIEDAD**
TRACK TÉCNICO

cedia | **TICEC**
2026

Inteligencia Artificial y
Sistemas Basados en
Conocimiento

TIC aplicadas a la
Educación y la Salud

Transformación Digital
y Madurez Tecnológica



Guayaquil, Ecuador
Vol. 38, Núm. E1 (2026)
Octubre, 14

DOI: 10.37815/rte.v38nE1



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0.

Autoridades Institucionales (Institutional Authorities)

Rectora

Cecilia Alexandra Paredes Verduga, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador

Vicerrectora de Docencia

Paola Romero, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador

Vicerrectorado de Investigación, Desarrollo e Innovación (I+D+i)

Carlos Teodoro Monsalve Arteaga, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador

Consejo de Editores (Editorial Board)

Directora General Editorial (Editor in Chief)

Nayeth I. Solórzano Alcívar, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: nsolorza@espol.edu.ec
 ORCID: 0000-0002-5642-334X

Directora Ejecutiva Editorial (Executive Director Co-Editor)

Katherine Anabelle Salvador Cisneros, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: ksalvado@espol.edu.ec
 ORCID: 0000-0003-0414-3597

Editores Asociados (Associate Editors)

Fausto Enrique Jácome López, M.Sc.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: fjacome@espol.edu.ec
 ORCID: 0000-0001-6808-7294

Cristian Díaz Álvarez, Ph.D.
 Universidad del Meta, Colombia
 e-mail: cristian.diaz@unimeta.edu.co
 ORCID: 0000-0002-4212-4947

Irma Abrigo Córdova, Ph.D.
 Universidad Nacional de Loja, Ecuador
 e-mail: irmabrigocordova2016@gmail.com
 ORCID: 0000-0002-1537-6313

Miriam Maricela Checa Artos, Ph.D.
 Universidad de las Fuerzas Armadas, Ecuador
 e-mail: mcheca@espol.edu.ec
 ORCID: 0000-0002-9606-7896

Patrick Taran, Prof.
 Global Migration Policy Associates, GMPA,
 Switzerland
 e-mail: taran@globalmigrationpolicy.org
 ORCID: 0000-0002-9442-5725

Santiago Berrezueta-Guzman, Ph.D.
 Technical University of Munich - TUM, Germany
 Corporación Ecuatoriana para el Desarrollo de la
 Investigación y la Academia - CEDIA, Ecuador
 e-mail: santiago.berrezueta@cedia.org.ec
 ORCID: 0000-0001-5559-2056

Editores/as de Sección o Temático (Section Editors)

Begoña Mora Jaureguialde, Dra.
 Universidad de Huelva, España
 e-mail: bego.mora@dedu.uhu.es
 ORCID: 0000-0002-1537-6313

Eduardo Jordan Pérez, Ph.D.
 Griffith University, Australia
 e-mail: e.jordanperez@griffith.edu.au
 ORCID: 0000-0001-5299-8501

Fernando Sandoya Sánchez, Ph.D.
 Universidad de Guayaquil, Ecuador
 e-mail: fernando.sandoyas@ug.edu.ec
 ORCID: 0000-0002-0011-4003

Hilda Flor Paéz, M.Sc.
 EDUPROSER Cía. Ltda., Ecuador
 e-mail: hflor@espol.edu.ec
 ORCID: 0009-0002-7472-8494

Marcelo Rafael Báez Meza, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: mbaez@espol.edu.ec
 ORCID: 0000-0002-8563-6985

Marco Antonio González Escudero, Dr.
 Universidad Central del Ecuador, Ecuador
 e-mail: magonzalez@uce.edu.ec
 ORCID: 0000-0003-4415-6579

Miguel Eduardo Yapur Aua, M.Sc.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: myapur@espol.edu.ec
 ORCID: 0000-0001-6627-7908

Ramón Espinel Martínez, Ph.D.
 Centro de investigaciones rurales, CIR
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: respinel@espol.edu.ec
 ORCID: 0000-0002-4998-2063

Rosa Aurora Espinoza Toalombo, Ph.D.
Universidad Estatal de Milagro, Ecuador
e-mail: respinozat@unemi.edu.ec
ORCID: 0000-0001-6527-1870

Silvana del Pilar Gamboa Benitez, Ph.D.
Escuela Politécnica Nacional, Ecuador
e-mail: silvana.gamboa@epn.edu.ec
ORCID: 0000-0001-5951-1970

Sylvia Novillo Villegas, Ph.D.
Universidad de las Américas, Ecuador
e-mail: sylvia.novillo@udla.edu.ec
ORCID: 0000-0002-7659-6846

Wellington Enrique Wonsang Ureta, BS.c.
San Diego State University Associates in
Business Accounting, Estados Unidos
e-mail: wonsangwellington@gmail.com
ORCID: 0000-0003-3936-9705

Editores/as de Copia - Español (Copy Editors - Spanish)

Andrea Victoria Carranza García, M.Sc.
Escuela Superior Politécnica del Litoral, Ecuador
e-mail: avcarran@espol.edu.ec
ORCID: 0000-0003-1762-4525

Diana Llamel Rodríguez Arteaga, M.Sc.
Escuela Superior Politécnica del Litoral, Ecuador
e-mail: dirodrig@espol.edu.ec
ORCID: 0000-0002-3715-181X

Laura Amelia Malache Silva, M.Sc.
CEDIA, Ecuador
e-mail: laura.malache@cedia.org.ec
ORCID: 0009-0004-2846-8710

Editores/as de Copia - Inglés (Copy Editors - English)

Lidia Govea de Bustamante, Ph.D.
Escuela Superior Politécnica del Litoral, Ecuador
e-mail: bgovea@espol.edu.ec
ORCID: 0000-0002-6770-1261

Maite Velez Chamaidan, Lic.
Massachusetts Bay Community College,
United States
e-mail: m_vezlez3@massbay.edu
ORCID: 0009-0008-8633-818X

Consejo Científico (Advisory Board)

Carlos Teodoro Monsalve Arteaga, Ph.D.
Escuela Superior Politécnica del Litoral, Ecuador
e-mail: monsalve@espol.edu.ec
ORCID: 0000-0002-9428-4360

Boris Xavier Vintimilla Burgos, Ph.D.
Escuela Superior Politécnica del Litoral, Ecuador
e-mail: bvintim@fiec.espol.edu.ec
ORCID: 0000-0001-8904-0209

Carlos Alberto Ramos Galarza, Ph.D.
Pontificia Universidad Católica del Ecuador
e-mail: ps_carlosramos@hotmail.com
ORCID: 0000-0001-5614-1994

Diana Carolina Coello Fiallos, Ph.D.
Universidad Técnica de Ambato, Ecuador
e-mail: dc.coello@uta.edu.ec
ORCID: 0000-0003-3628-5004

Diego Fernando Avila Pesantez, Ph.D.
Escuela Superior Politécnica de Chimborazo,
Ecuador
e-mail: davila@esepoch.edu.ec
ORCID: 0000-0001-8394-5621

Elena María Fernández González, Ph.D.
Universidad de Oviedo, España
e-mail: fernandezelena@uniovi.es
ORCID: 0000-0002-5683-0674

Elva Patricia Reyes Díaz, Ph.D.
Universidad Autónoma de Coahuila, México
e-mail: elva.reyes@uadec.edu.mx
ORCID: 0000-0002-8854-0946

Ignacio Bladimir Cerón Guerra, Ph.D.
Escuela Politécnica Nacional, Ecuador
e-mail: ignacio.ceron@epn.edu.ec
ORCID: 0000-0002-6373-9772

Isidro Fierro Ulloa, Ph.D.
Universidad de Especialidades Espíritu Santo,
Ecuador
e-mail: isfierro@uees.edu.ec
ORCID: 0000-0002-9000-6945

José Enrique Martínez-Pérez, Ph.D.
Instituto Universitario de Tecnología del Estado
Bolívar, Venezuela
e-mail: josenriquemartinez@gmail.com
ORCID: 0000-0003-3611-639

Jorge Antonio Córdova Morán, Ph.D.
Universidad Estatal de Milagro, Ecuador
e-mail: joranto205@gmail.com
ORCID: 0000-0002-2730-4755

Juan José Oliveira Maurera, Ph.D.
Universidad de Oriente, Venezuela
e-mail: juanjoseoli@hotmail.com
ORCID: 0000-0002-8474-5872

Juan Humberto Avellaneda Cevallos, Ph.D.
 Universidad Técnica Estatal de Quevedo,
 Ecuador
 e-mail: juan_avellaneda@yahoo.com
 ORCID: 0000-0002-1805-4803

Luke Houghton, Ph.D.
 Griffith University, Australia
 e-mail: l.houghton@griffith.edu.au
 ORCID: 0000-0002-7461-6591

Manuel Raúl Peláez Samaniego, Ph.D.
 Universidad de Cuenca, Ecuador
 e-mail: manuel.pelaez@ucuenca.edu.ec
 ORCID: 0000-0002-7618-9474

María Del Pilar Cornejo Rodríguez, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: pcornejo@espol.edu.ec
 ORCID: 0000-0003-4198-2904

María Genoveva Espinoza-Santeli, Ph.D.
 Universidad Andina Simón Bolívar, Ecuador
 e-mail: genoveva.espinoza@uasb.edu.ec
 ORCID: 0000-0002-2347-9507

Olga Kadysheva, Ph.D.
 Global Migration Policy Associates, Switzerland
 e-mail: okadysheva@globalmigrationpolicy.org
 ORCID: 0000-0002-4810-8964

Patricia A. Boeri, Ph.D.
 Universidad Nacional de Río Negro, Argentina
 e-mail: pboeri@unrn.edu.ar
 ORCID: 0000-0003-3262-8030

Paúl Cesar Carrión Mero, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: pcarrion@espol.edu.ec
 ORCID: 0000-0002-9747-7547

Raquel Miroslava Tinoco Egas, Ph.D.
 Universidad Técnica de Machala, Ecuador
 e-mail: raqueltinocoegas@hotmail.com
 ORCID: 0000-0001-5979-8123

Stanislaus Albert-Georg Sonnenholzner, Ph.D.
 Escuela Superior Politécnica del Litoral, Ecuador
 e-mail: ssonnen@cenaim.espol.edu.ec
 ORCID: 0000-0002-5723-3994

Consejo Internacional de Revisores del Volumen (International Reviewers Board)

Alexandra Cristina Gonzalez-Eras, Ph.D.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: acgonzalez@utpl.edu.ec
 ORCID: 0000-0002-5305-7882

Andrei Koshelev, M.Sc.
 Technical University of Munich, Alemania
 e-mail: andrei.koshelev@tum.de
 ORCID: 0009-0000-3786-4994

Dario Morales, M.Sc.
 Universidad de las Fuerzas Armadas ESPE,
 Ecuador
 e-mail: djmorales7@espe.edu.ec
 ORCID: 0000-0003-3342-884X

Doris Karina Chicaiza Angamarca, M.Sc.
 Universidad de las Fuerzas Armadas ESPE,
 Ecuador
 e-mail: dkchicaiza@espe.edu.ec
 ORCID: 0000-0002-4832-7841

Efraín Rodrigo Fonseca Carrera, Ph.D.
 Universidad de las Fuerzas Armadas ESPE,
 Ecuador
 e-mail: erfonseca@espe.edu.ec
 ORCID: 0000-0001-7129-9335

Esteban Briceño, M.Sc.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: epbriceno@utpl.edu.ec
 ORCID: 0000-0002-9275-1578

Fernanda Maricela Soto Guerrero, M.Sc.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: fmsoto@utpl.edu.ec
 ORCID: 0000-0003-3097-9291

Ivan Kuzmin, M.Sc.
 Technical University of Munich, Alemania
 e-mail: ivan.kuzmin@tum.de
 ORCID: 0009-0007-5412-5537

Ivan Parmaccli, M.Sc.
 Technical University of Munich, Alemania
 e-mail: ivan.parmaccli@tum.de
 ORCID: 0009-0004-1575-4478

Jorge Iván Pincay-Ponce, Ph.D.
 Universidad Laica Eloy Alfaro de Manabí,
 Ecuador
 e-mail: jorge.pincay@uleam.edu.ec
 ORCID: 0000-0003-4711-8850

Jorge Luis Zambrano-Martinez, Ph.D.
 Universidad del Azuay, Ecuador
 e-mail: jorge.zambrano@uazuay.edu.ec
 ORCID: 0000-0002-5339-7860

Kerstin Andree, M.Sc.
 Technical University of Munich, Alemania
 e-mail: kerstin.andree@tum.de
 ORCID: 0009-0007-6360-8661

Maholy Velasquez, M.Sc.
 Universidad Laica Eloy Alfaro de Manabí,
 Ecuador
 e-mail: gabriela.velasquez@uleam.edu.ec
 ORCID: 0000-0003-0242-2293

Martha Vanessa Agila-Palacios, Ph.D.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: mvagila@utpl.edu.ec
 ORCID: 0000-0003-2301-1469

Melissa Montalvo, M.Sc.
 ScreenPoint Medical, Holanda
 e-mail: melissa.montalvo@screenpointmed.com
 ORCID: 0000-0002-9539-3243

Priscila Marisela Valdiviezo-Díaz, M.Sc.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: pmvaldiviezo@utpl.edu.ec
 ORCID: 0000-0002-5216-8820

Rommel Vicente Torres Tandazo, Ph.D.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: rovtor@utpl.edu.ec
 ORCID: 0000-0003-2313-0118

Rosa Cristina Pesántez Aldás, M.Sc.
 Universidad de las Fuerzas Armadas ESPE,
 Ecuador
 e-mail: rcpesantez1@espe.edu.ec
 ORCID: 0009-0004-9967-946X

Tuesman Daniel Castillo Calvas, M.Sc.
 Universidad Técnica Particular de Loja, Ecuador
 e-mail: tdcastillo@utpl.edu.ec
 ORCID: 0000-0002-3677-7781

Vanesa Metaj, M.Sc.
 Technical University of Munich, Alemania
 e-mail: vanesa.metaj@tum.de
 ORCID: 0009-0008-6576-346X

Vladimir Robles-Bykbaev, Ph.D.
 Universidad Politécnica Salesiana, Ecuador
 e-mail: vrobles@ups.edu.ec
 ORCID: 0000-0002-7645-8793

Gestión de Comunicación, Publicación y Técnica

Difusión y Comunicación

Gerencia de Comunicación Social y Asuntos Públicos, ESPOL

José Arturo Maldonado Moncayo, Ing.
 Escuela Superior Politécnica del Litoral, Ecuador

Gestión de Comunicación y Medios Digitales

José Luis Castro Zambrano, M.Sc.
 Escuela Superior Politécnica del Litoral, Ecuador

Traducciones en idioma extranjero

English

Joyce Nan, BA.
 Griffith University, Australia

Portada, Diseño y Diagramación

Daniel Peña
 CEDIA, Ecuador

Romina Belén Gómez Haro
 Escuela Superior Politécnica del Litoral, Ecuador

José Luis Castro Zambrano, M.Sc.
 Escuela Superior Politécnica del Litoral, Ecuador

Kleber José Avelino Mosquera
 Escuela Superior Politécnica del Litoral, Ecuador

Informática

Contacto Soporte Técnico
 Kleber José Avelino Mosquera
 rte@espol.edu.ec
 Escuela Superior Politécnica del Litoral, Ecuador

Indexadores

Scopus®



Directorios



Identificadores



Sobre la revista

Misión: Promover y difundir investigación multidisciplinaria, científica y tecnológica de calidad, impulsando el avance del conocimiento y el desarrollo de la sociedad de manera significativa.

La **Revista Tecnológica Espol - RTE®** es publicada de manera digital y de acceso abierto en su plataforma rte.espol.edu.ec. Se encuentra registrada en varios indexadores y directorios de bases de datos y cuenta con un proceso de evaluación continua de nuevas formas de indexación que permiten dar mayor visibilidad a los artículos publicados y a sus autores.

La **Revista Tecnológica Espol - RTE®** presenta llamadas programadas para envío de artículos (Call for papers) en idiomas español o inglés, garantizando la publicación periódica de dos volúmenes al año (junio y diciembre). Adicionalmente, se pueden producir volúmenes especiales sobre una temática específica (como un Monográfico o Dossier), o Ediciones Especiales relacionadas a un evento científico o tecnológico (por ejemplo: congresos o simposios). En el caso de volúmenes especiales internacionales se puede publicar artículos en otros idiomas en acuerdo con la experticia del o los editores invitados siempre y cuando el título y resumen sean incluidos también en español e inglés.

En caso de artículos presentados por más de un autor, siguiendo los criterios de autoría bajo la taxonomía CRediT (<https://credit.niso.org/>), RTE expresa que la falta de declaración de las contribuciones individuales implica una participación equitativa de todos los autores.

About

Mission: To promote and disseminate high-quality multidisciplinary, scientific, and technological research, driving the advancement of knowledge and the development of society in a meaningful way.

The **Revista Tecnológica Espol - RTE®** is published digitally and in open access on its platform rte.espol.edu.ec. It is registered in several indexes and database directories and has a continuous evaluation process of new forms of indexing that allow greater visibility for published articles and their authors.

The **Revista Tecnológica Espol - RTE®** presents scheduled calls for article submissions (Call for papers) in Spanish or English languages, guaranteeing the publication of two volumes per year (in June and December). In addition, special volumes-issues are produced on a specific topic (such as a Monograph or Dossier) related to scientific events (ie. Congresses or symposiums). In the case of international special issues, articles in other languages can be published in agreement with the Guest editors if the title and abstract are also included in Spanish and English.

For articles submitted by more than one author, following the authorship criteria under the CRediT taxonomy (<https://credit.niso.org/>), RTE expresses that no declaration of individual contributions implies equal participation of all authors.

Editorial (Español)

La **Revista Tecnológica Espol - RTE®**, en colaboración con la **Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA)**, presenta este volumen especial dedicado a los artículos técnicos de la edición *2026 de la Conferencia Ecuatoriana de Tecnologías de la Información y la Comunicación (TICEC)*, celebrada en la Universidad Tecnológica Indoamérica, en Ambato, los días 15 y 16 de octubre de 2026. El encuentro de investigadores y profesionales en áreas relevantes para la transformación digital hace énfasis en inteligencia artificial, TIC aplicadas a la educación y la salud, y transformación digital y emprendimiento.

La sección de **Inteligencia Artificial y Sistemas Basados en Conocimiento** reúne contribuciones que integran modelos de lenguaje, representaciones semánticas y arquitecturas para la gestión inteligente de la información. Se presentan dos arquitecturas de generación aumentada por recuperación (RAG) ejecutadas localmente: una basada en reglas SWRL para automatizar el preprocesamiento de datos y otra que integra ontologías y consultas SPARQL para acceder a la documentación técnica. Se incluyen, además, un chatbot académico con recuperación híbrida, un estudio sobre el reconocimiento facial de emociones y una revisión sistemática sobre la simulación numérica y el aprendizaje profundo para el diseño inverso de metamateriales multicapa.

En **TIC aplicadas a la Educación y la Salud**, los trabajos articulan la gobernanza y la aplicación tecnológica. Se analiza el marco normativo de la inteligencia artificial en la educación superior ecuatoriana y se presentan Estación Infantil, una plataforma virtual gamificada para la educación inicial, y Pulneo, un prototipo móvil gamificado para la fisioterapia respiratoria domiciliar mediante análisis acústico.

La sección **Transformación Digital y Emprendimiento** aborda la adopción tecnológica mediante una plataforma para evaluar la madurez digital de microemprendimientos ecuatorianos y un estudio sobre los factores y las brechas en la transformación digital de las PYMES de Guayaquil.

En conjunto, estas contribuciones evidencian la diversidad y pertinencia de la investigación en tecnologías de la información y la comunicación frente a desafíos científicos, educativos, sociales y productivos, y reafirman el valor de la investigación aplicada y del intercambio académico para una transformación digital sustentada en el conocimiento y la evidencia.

La **Revista Tecnológica Espol - RTE®** reconoce a los autores y revisores por su contribución a la calidad científica de esta edición, y a CEDIA, así como a los organizadores y participantes de TICEC 2026, por promover la colaboración y el intercambio científico.

Confiamos en que este volumen constituya una referencia para investigadores, profesionales y comunidades académicas, e impulse nuevas investigaciones, desarrollos tecnológicos y oportunidades de cooperación en el ámbito de las tecnologías de la información y la comunicación.

Nayeth Solórzano Alcívar, PhD.
Directora General Editorial
Revista Tecnológica ESPOL – RTE

Santiago Berrezueta, Ph.D.
Editor Asociado
Revista Tecnológica ESPOL – RTE

Katherine Salvador Cisneros, PhD.
Directora Ejecutiva Editorial
Revista Tecnológica ESPOL – RTE

Laura Malache Silva, M.Sc.
Editor en Jefe
CEDIA

Editorial (English)

The **Revista Tecnológica Espol - RTE®**, in collaboration with the **Ecuadorian Corporation for the Development of Research and Academia (CEDIA)**, presents this special volume dedicated to the technical articles of the *2026 edition of the Ecuadorian Conference on Information and Communication Technologies (TICEC)*, held at the Indoamerica Technological University, in Ambato, on October 15 and 16, 2026. Researchers and professionals in areas relevant to digital transformation discuss artificial intelligence, ICT applied to education and health, and digital transformation and entrepreneurship.

The **Artificial Intelligence and Knowledge-Based Systems** section brings together contributions that integrate language models, semantic representations, and architectures for intelligent information management. Two locally executed augmented generation-by-retrieval (RAG) architectures are presented: one based on SWRL rules to automate data preprocessing, and another that integrates ontologies and SPARQL queries to access technical documentation. The collection also includes an academic chatbot with hybrid retrieval, a study on facial recognition of emotions, and a systematic review on numerical simulation and deep learning for the reverse design of multilayer metamaterials.

In **ICT applied to Education and Health**, the articles cover governance and technological application. The regulatory framework for artificial intelligence in Ecuadorian higher education is analyzed; and *Estación Infantil*, a gamified virtual platform for early education, and *Pulneo*, a gamified mobile prototype for home respiratory physiotherapy that uses acoustic analysis, are presented.

The **Digital Transformation and Entrepreneurship** section addresses technological adoption through a platform for assessing the digital maturity of Ecuadorian microenterprises and a study on the factors and gaps in the digital transformation of SMEs in Guayaquil.

Together, these contributions demonstrate the diversity and relevance of research in information and communication technologies in the face of scientific, educational, social, and productive challenges, and reaffirm the value of applied research and academic exchange for a knowledge- and evidence-based digital transformation.

Revista Tecnológica Espol - RTE® acknowledges the authors and reviewers for their contributions to the scientific quality of this edition, and CEDIA, as well as the organizers and participants of TICEC 2026, for promoting collaboration and scientific exchange.

We hope this volume will serve as a reference for researchers, professionals, and academic communities, and will promote new research, technological developments, and opportunities for cooperation in the field of information and communication technologies.

Nayeth Solórzano Alcívar, PhD.
Editor in Chief
Revista Tecnológica ESPOL – RTE

Santiago Berrezueta, Ph.D.
Associate Editor
Revista Tecnológica ESPOL – RTE

Katherine Salvador Cisneros, PhD.
Executive Director Co-Editor
Revista Tecnológica ESPOL – RTE

Laura Malache Silva, M.Sc.
Editor in Chief
CEDIA

Índice / Index

Inteligencia Artificial y Sistemas Basados en Conocimiento / Artificial Intelligence and Knowledge-Based Systems

- 13 *Automated Data Preprocessing Using Parameterized SWRL Inference within a Local RAG Pipeline / Preprocesamiento automatizado de datos mediante reglas SWRL parametrizadas en RAG local*
- 31 *Beyond Accuracy: Generalisation, Bias, and Robustness in Deep Learning for Facial Emotion Recognition / Más allá de la precisión: Confianza en el reconocimiento facial de emociones*
- 50 *An academic chatbot with hybrid retrieval and local large language models / Chatbot académico con recuperación híbrida y modelos de lenguaje locales*
- 71 *Numerical Simulation Engines and Deep Learning Architectures for Inverse Design of Multilayer Metamaterials / Motores de Simulación Numérica y Arquitecturas de Aprendizaje Profundo para el Diseño Inverso de Metamateriales Multicapa*
- 84 *Ontology-Grounded RAG with SPARQL and Local Language Models for Technical Documentation / RAG basado en ontologías con SPARQL y modelos de lenguaje locales para documentación técnica*

TIC aplicadas a la Educación y la Salud / ICT applied to Education and Health

- 101 *Madurez del marco normativo para la IA en la educación superior ecuatoriana / Maturity of the Regulatory Framework for AI in Ecuadorian Higher Education*
- 117 *"Estación Infantil": Plataforma virtual gamificada para la educación inicial / "Estación Infantil": A Gamified Virtual Platform for Early Childhood Education*
- 136 *Pulneo: Prototipo móvil gamificado con detección acústica de actividad respiratoria para fisioterapia domiciliaria / Pulneo: Gamified Mobile Prototype with Acoustic Respiratory Activity Detection for Home-Based Physiotherapy*

Transformación Digital y Madurez Tecnológica / Digital Transformation and Technological Maturity

- 154 *Ingeniería de una plataforma web para la evaluación de la madurez digital de microemprendimientos ecuatorianos / Engineering a Web Platform for Assessing the Digital Maturity of Ecuadorian Micro-Entrepreneurs*
- 172 *Motores y brechas de la transformación digital: Un análisis del uso y adopción de tecnologías en las pequeñas y medianas empresas (pymes) de la ciudad de Guayaquil, Ecuador / Drivers and Gaps of Digital Transformation: An Analysis of Technology Use and Adoption in Small and Medium Enterprises (SMEs) of Guayaquil – Ecuador*

INTELIGENCIA ARTIFICIAL Y SISTEMAS BASADOS EN CONOCIMIENTO

ARTIFICIAL INTELLIGENCE AND
KNOWLEDGE-BASED SYSTEMS



cedia | **TICEC**
2026

Automated Data Preprocessing Using Parameterized SWRL Inference within a Local RAG Pipeline

Preprocesamiento automatizado de datos mediante reglas SWRL parametrizadas en RAG local

María de los Ángeles Contreras Trujillo¹ <https://orcid.org/0009-0009-2265-4703>, Mayerli Rosario Quintuña Ramón¹ <https://orcid.org/0009-0004-4978-4958>, Claudia Tamara Serpa Flores¹ <https://orcid.org/0009-0005-2043-8813>, Marcia Patricia Ullauri Carrasco¹ <https://orcid.org/0009-0002-3039-8104>, Marcos Orellana¹ <https://orcid.org/0000-0002-3671-9362>

¹Universidad del Azuay, Computer Science Research & Development Laboratory (LIDI), Cuenca, Ecuador
contreras.maria@es.uazuay.edu.ec,
mayerli.quintuna@es.uazuay.edu.ec,
claudia.serpa@es.uazuay.edu.ec,
marcia.ullauri@es.uazuay.edu.ec,
marore@uazuay.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/07/12

Accepted: 2026/08/26

Published: 2026/10/14

Abstract

This work formalizes expert knowledge in data preprocessing through the Web Ontology Language (OWL) and parameterized Semantic Web Rule Language (SWRL) rules, integrated with a Retrieval-Augmented Generation (RAG) pipeline for natural-language queries. An OWL ontology comprising 47 classes, 42 individuals, and 19 SWRL rules was designed in Protégé and validated on the Netflix Movies and TV Shows dataset (8,807 records, 12 columns). A Python-implemented inference engine, which parses the SWRL rule structure from the RDF/XML serialization and evaluates antecedents through deterministic conditionals without invoking a native OWL DL reasoner, generated 15 preprocessing-technique recommendations across 9 analyzable columns; the remaining 3 columns (*show_id*, *title*, *description*) were excluded as identifiers and free-text fields, respectively, for which tabular preprocessing is not applicable. The system guides data scientists in selecting preprocessing techniques based on each column's statistical profile, reducing reliance on manual configuration. The novelty lies in parameterizing

Summary: Introduction, Materials and Methods, Results and Discussion, Conclusions.

How to cite: Contreras, M., Quintuña, M., Serpa, C., Ullauri, M. & Orellana, M. (2026). Automated Data Preprocessing Using Parameterized SWRL Inference within a Local RAG Pipeline. *Revista Tecnológica - Espol*, 38(E1), 13-30. <https://doi.org/10.37815/rte.v38nE1.1599>

SWRL rules using dynamically computed statistical metrics and multilingual semantic retrieval, within a fully local RAG pipeline that supports traceability and privacy.

Keywords: data quality, expert systems, knowledge management, ontology, semantic web.

Resumen

Este trabajo formaliza el conocimiento experto en preprocesamiento de datos mediante OWL y reglas SWRL parametrizadas, integradas con una arquitectura RAG para consultas en lenguaje natural. Se diseñó una ontología OWL con 47 clases, 42 individuos y 19 reglas SWRL en Protégé, y se validó sobre el dataset Netflix Movies and TV Shows (8.807 registros, 12 columnas). Un motor de inferencia implementado en Python, que analiza sintácticamente la estructura de las reglas SWRL a partir de la serialización en RDF/XML y evalúa los antecedentes mediante condicionales deterministas, sin invocar un razonador OWL DL nativo, generó 15 recomendaciones distribuidas en nueve columnas analizables; las tres columnas restantes (*show_id*, *title*, *description*) fueron excluidas por corresponder, respectivamente, a un identificador y a campos de texto libre. El sistema orienta a científicos de datos en la selección de técnicas según el perfil estadístico de cada columna, lo que reduce la dependencia de la configuración manual. La novedad radica en la parametrización de reglas SWRL con métricas estadísticas dinámicas, combinada con la recuperación semántica multilingüe en un pipeline RAG local que favorece la trazabilidad y la privacidad.

Palabras clave: calidad de datos, gestión del conocimiento, ontologías, sistemas expertos, web semántica.

Introduction

The quality of data is a critical factor in the predictive performance of machine learning models, and a substantial proportion of the data science lifecycle is devoted to assessing and improving it through preprocessing and cleaning (Batini & Scannapieco, 2016). However, selecting the most appropriate preprocessing techniques for a specific dataset remains predominantly manual, highly dependent on the analyst's intuition, and prone to inconsistencies across practitioners. This reliance on tacit knowledge creates bottlenecks in reproducibility and hampers the standardization of data science pipelines in organizational environments.

In response to this issue, knowledge management provides a systematic conceptual framework for capturing, formally representing, and reasoning about expert knowledge within a specific domain (Alavi et al., 2024). In particular, ontologies expressed in the Web Ontology Language (OWL) allow the formalization of concepts, relationships, and constraints with explicit and computable semantics, while the Semantic Web Rule Language (SWRL) extends these capabilities by enabling automatic deductive inferences over ontology instances (McGuinness et al., 2004). Simultaneously, Retrieval-Augmented Generation (RAG) systems have emerged as a hybrid architecture that substantially improves the accuracy and veracity of Large Language Models (LLMs) by enriching the generation context through dynamic retrieval of relevant documents during inference (Lewis et al., 2020).

This work proposes a synergistic integration of both paradigms. On the one hand, an OWL ontology with parameterized SWRL rules is designed to automatically infer the optimal preprocessing technique based on the statistical profile of each dataset column. On the

other hand, a RAG pipeline is implemented to present this formalized knowledge in natural language, facilitating interaction with users who are not specialized in ontologies. The system is empirically validated using the publicly available Netflix Movies and TV Shows dataset, deliberately selected to present representative challenges of real-world practice: structural nulls in categorical variables, high cardinality in textual attributes, mixed formats in temporal fields, and class imbalance in binary variables. Thus, this article presents a reusable architecture that formalizes expert preprocessing knowledge and makes it accessible through conversational interfaces grounded in ontological reasoning.

Finally, this document is organized as follows: The Related Work section reviews existing research on OWL-based knowledge representation, SWRL rule engines, and ontology-grounded RAG architectures; the Materials and Methods section delineates the materials and procedures, including dataset characterization, ontological engineering, SWRL inference rules, and the implementation of the RAG pipeline; the Results and Discussion section encompasses the presentation of findings and analytical discourse, including automatic inference, representative case analysis, and validation of the RAG pipeline; and the Conclusion section summarizes the primary contributions, limitations, and prospective avenues for future research.

Related Work

The formalization of expert knowledge through semantic web technologies, particularly ontology-driven inference and retrieval-augmented generation, has emerged as an important mechanism for automating data preprocessing decisions. This section reviews the foundational literature on OWL-based knowledge representation, SWRL rule engines, and ontology-grounded RAG architectures, thereby delineating the theoretical precedents and identifying the specific research gap that motivates our integrative contribution.

The formalization of knowledge in data science processes has grown significantly over the past decade. The Data Mining Ontology and Planning (DMOP) project is one of the first systematic efforts to formally represent algorithms, datasets, and workflows in data mining within grid-computing environments, establishing a conceptual foundation for the semantic description of machine learning pipelines (Cannataro et al., 2003). Similarly, ML-Schema proposes a standardized vocabulary for describing machine learning experiments, facilitating interoperability across platforms and the reproducibility of results (Publio et al., 2018). In Sharma et al. (2025), Ontology-Grounded Retrieval-Augmented Generation (OG-RAG) is introduced, which uses predefined domain ontologies to underpin information retrieval in RAG systems and demonstrates that semantic anchoring significantly improves factual accuracy compared to traditional ad hoc entity extraction approaches.

This advancement highlights the importance of establishing clear semantics before integrating language models with structured knowledge sources. In the realm of data quality, the framework proposed by Jaziri & Sassi (2026) for improving ontologies in the Semantic Web employs domain-agnostic machine learning models that analyze RDF/OWL ontologies using generic signals: class hierarchy depth, property usage, relationship density, and annotation completeness, achieving precisions of 93.5% in structural verifications and 90.2% in semantic verifications. These works demonstrate that automated ontology quality assessment is an established area of research, although it is predominantly focused on ontology structure rather than on the quality of the data it processes. Unlike previous approaches, which concentrate on ontological formalization during the preprocessing phase, our work focuses on an aspect

that earlier methods address only tangentially. While DMOP (Cannataro et al., 2003) models the semantics of data-mining workflows in grid-computing environments and ML-Schema (Publio et al., 2018) exposes the semantics of machine-learning experiments through schema annotations, neither encodes decision rules for selecting preprocessing techniques based on column-level statistical profiles. Our ontology addresses this gap by parameterizing SWRL rules with dynamically computed statistical metrics.

The SWRL language extends OWL DL with Horn rules, allowing the derivation of new assertions from existing axioms and enabling reasoning about instances with greater expressiveness than purely descriptive reasoning (Rattanasawad et al., 2013). Its application has been widely documented in the medical field, where rules combining clinical conditions support automate diagnostic decision-making (Dumontier et al., 2014). In operational domains closer to data engineering, Paltin Chica et al. (2025) demonstrate the viability of coupling OWL ontologies with SWRL rules for semantic reasoning over heterogeneous, real-world datasets, in their case, emergency call records for ECU 911. Their work shows that the OWL+SWRL stack can handle mixed formats, high cardinality, and implicit relationships without manual rule reconfiguration for each new instance. However, all these applications operate on qualitative or categorical properties. Applying SWRL to data preprocessing with quantitative parameterization constitutes a novel extension: our proposal encodes numerical thresholds directly into the rule antecedents, enabling the same set of rules to generate differentiated recommendations based on each column's statistical profile, without requiring manual reconfiguration for each new dataset.

The RAG paradigm (Lewis et al., 2020) enriches language-model inputs with dynamically retrieved documents to reduce hallucinations. In the data-engineering domain, enriching the generation context with structured dataset metadata has been shown to improve the relevance of LLM responses in knowledge-intensive tasks (García-Montero et al., 2025; Park et al., 2024). Recent GraphRAG approaches (Edge et al., 2024) structure retrieval through LLM-extracted knowledge graphs but rely on ad hoc extraction without domain anchoring. Jaziri & Sassi (2026) address this by grounding retrieval in expert-defined semantic schemas, reducing human evaluation time by 28.8%. Our system operates at this intersection: it uses fragments generated from the OWL ontology, including sentences derived from SWRL rules, as the retrieval source, combining ontological traceability with natural-language flexibility without requiring a complete knowledge graph.

The trend toward AI systems operable in privacy-restricted environments has driven the development of local RAG pipelines. Park et al. (2024) demonstrate ontological RAG for additive manufacturing, and Buehler (2024) proposes multi-agent strategies based on generative ontological graphs for material design. In this context, our RAG pipeline, which uses a local LLM and multilingual embeddings, demonstrates the feasibility of fully local architectures that preserve data privacy without sacrificing reasoning quality.

The literature review reveals a gap: while there are ontologies to describe ML pipelines, rule-based systems for inference, and RAG pipelines for document querying, no integration has been reported that combines (i) SWRL rules parameterized with statistical metrics of the dataset, (ii) automatic inference of preprocessing techniques, and (iii) exposure through RAG with a local LLM. Our work addresses this intersection by formalizing expert knowledge about preprocessing in an executable ontology accessible in natural language.

Materials and Methods

This study adopts an applied research approach with an experimental design, aimed at developing and implementing a knowledge management system that integrates a formal OWL ontology, SWRL inference rules, and a RAG pipeline. The research was structured into four sequential phases: (1) dataset characterization and statistical profiling; (2) ontological engineering and formalization of expert knowledge; (3) implementation of automated inference rules; and (4) validation through an RAG pipeline with a local language model. This sequence meets the reproducibility criteria, in which procedures, measurements, and variable units must be sufficiently documented to allow other researchers to replicate the study.

Dataset

As a case study, the publicly available Netflix Movies and TV Shows dataset on Kaggle was used. It comprises 8,807 records and twelve descriptive columns. The selection of this dataset is justified by the heterogeneity of its attributes, which exhibit common data quality issues encountered in organizational contexts: structural nullity, high cardinality, mixed formats, outliers, and class imbalance. Statistical profiling was conducted using Python 3.11 scripts, employing libraries such as pandas and NumPy. For each column, the inferred data type, percentage of null values, cardinality (the number of unique values), proportion of outliers (determined through the interquartile range, IQR, method), proportion of zero values, and imbalance ratio (for binary variables) were identified. The outlier analysis employed Tukey's rule, which considers an outlier if it falls outside the interval $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, where Q1 and Q3 denote the first and third quartiles, respectively. Table 1 summarizes the resulting statistical profile.

Table 1
Statistical profile of the dataset

Column	Type	Null (%)	Unique values	Detected problem
show_id	Nominal	0	8,807	Primary Key; high Cardinality
type	Boolean	0	2	Imbalance 2.29:1 (Movie vs. TV Show)
title	Nominal	0	8,807	Free Text; high Cardinality
director	Nominal	29.9	4,528	High Cardinality; 29.9% null
cast	Nominal	9.4	7,692	High Cardinality; 9.4% null; multi-labels
country	Nominal	9.4	748	High Cardinality; 9.4% null
date_added	Temporal	0.1	1,767	String "Sep 25, 2021"
release_year	Continuous number	0	74	8.2% outliers (IQR method)
rating	Nominal	0.1	17	3 load errors; 0.1% nulls
duration	Nominal	0	220	"X min" / "X Seasons" format
listed_in	Nominal	0	514	Multi-labels; 514 combinations
description	Nominal	0	8,775	Free Text; NLP required

Four categories of quality issues were identified: (i) structural nullity in the director column (29.9%), explained by the systematic absence of directors in TV Show productions; (ii) statistical outliers in *release_year* (8.2%), with $Q1 = 2013$, $Q3 = 2019$, $IQR = 6$, and limits [2004, 2028]; (iii) mixed formats in duration, combining dissimilar time units (minutes and seasons); and (iv) high cardinality in *listed_in*, with 514 unique genre combinations.

Ontological Engineering

The formalization of expert knowledge in data preprocessing requires an explicit semantic representation that captures domain concepts and causal relationships between statistical and technical profiles. In this way, OWL ontologies provide a formal framework for representation, while SWRL rules enable automated reasoning over specific instances. The process followed to construct the proposed ontology is described below.

Development Methodology

The construction of the ontology followed the Methontology approach (Fernández-López & Corcho, 2010), structured into four phases: (i) specification, through competency questions that defined the functional requirements of the system; (ii) conceptualization, identifying the terms of the data preprocessing domain from specialized literature; (iii) formalization, organizing the terms into a taxonomy with cardinality constraints; and (iv) implementation, using *Protégé* 5.6 with RDF/XML serialization and a defined namespace.

Design Criteria

The ontology was designed based on the following fundamental criteria:

- **Modularity.** The architecture allows independent extension of each branch (*data types*, *problems*, *techniques*) without affecting the consistency of the others. This criterion facilitates the future incorporation of new preprocessing techniques or new application domains.
- **Parameterization.** The SWRL rules operate on properties of numerical data (*pctNullity*, *pctOutliers*, *numUniqueValues*, *imbalanceRatio*), enabling the system to automatically adapt to datasets with statistical profiles that differ from Netflix's, without requiring modifications to the ontological structure.
- **Traceability.** Each inferred recommendation can be traced back to the SWRL rule that activated it, the numerical threshold that triggered it, and the bibliographic source that supports the recommended technique. This traceability is essential for auditing decision support systems in data science.
- **Reusability.** The ontology aligns with W3C standard vocabularies (OWL, RDF, SWRL) and can be integrated with existing domain ontologies in data mining, such as DMOP and ML-Schema, by mapping equivalent classes.

Ontology Structure

The resulting ontology is organized into three main branches, interconnected through object properties that model the causal and functional relationships within the data preprocessing domain:

- **Data types:** include the classes *ContinuousNumeric*, *DiscreteNumeric*, *NominalCategorical*, *OrdinalCategorical*, *Temporal*, and *Boolean*. This classification is based on the standard typology of statistical variables and their differential impact on the selection of preprocessing techniques.
- **Problems:** include *NullValues*, *Outliers*, *HighCardinality*, *Imbalance*, *InvalidZeros*, *InconsistentDuplicates*, *Scale*, and *Distribution*. The issues were defined based on

frequency analysis in the reviewed literature and their documented impact on the performance of machine learning models.

- Techniques: groups 28 techniques into six functional categories: *imputation*, *outlier detection*, *encoding*, *scaling*, *transformation*, and *balancing*. Each technique is linked to the problems it addresses via the property *resolveBy* and to the data types it applies to via the property *applicableTo*.

The object properties *hasProblem*, *resolveBy*, and *applicableTo* to establish relationships among the three branches. The data properties *pctNullity*, *pctOutliers*, *pctZeros*, *numUniqueValues*, and *imbalanceRatio* enable numerical parameterization of the inference rules. The resulting ontology contains 47 classes, 42 individuals (instances of Netflix dataset columns with their statistical profiles), and 19 SWRL rules.

SWRL Inference Rules

Nineteen SWRL rules were delineated and systematically categorized into seven functional groups, thereby augmenting the expressiveness of OWL DL by using Horn rules (Krötzsch et al., 2013). Each rule follows the logical structure:

Antecedent(?column) → recommendTechnique(?column, Technique),

Where the antecedent combines a *ClassAtom* (membership to a class), a *DatavaluedPropertyAtom* (a data property with a threshold value), and, when appropriate, a *BuiltinAtom* with predicates from the *swrlb* library.

Table 2 presents the rules for each group, indicating the antecedent, the consequent, and the threshold or activation condition.

All nullity-based rules (1.1–1.5) include the guard condition $pctNullity > 0$ in their antecedent, ensuring that columns with zero missing values do not trigger imputation or deletion recommendations. The 19 SWRL rules were authored in Protégé 5.6 and serialized in RDF/XML. At execution time, the OWL file was parsed with Python's `xml.etree.ElementTree` module to extract the rule antecedents (*ClassAtom* and *BuiltinAtom* conditions) and consequents (*IndividualPropertyAtom* linking a column to a preprocessing technique). A Python-implemented inference engine then evaluated each rule against the statistical profile of every column by matching the detected data type against *ClassAtom* predicates and comparing numeric metrics (*pctNullity*, *pctOutliers*, *pctZeros*, *numUniqueValues*, *imbalanceRatio*) against the *BuiltinAtom* thresholds using the standard SWRL comparison operators (*lessThan*, *greaterThan*, *lessThanOrEqual*, *greaterThanOrEqual*, *equal*). It must be emphasized that this engine does not invoke a native OWL DL reasoner; rather, it reproduces the forward-chaining semantics of the SWRL rules through deterministic Python conditionals that read the rule structure directly from the RDF/XML serialization. The OWL file therefore serves simultaneously as the formal specification of the rules and as the machine-readable configuration consumed by the engine. This engine materialized 15 technique recommendations across 9 analyzable columns. The remaining 3 columns (*show_id*, *title*, and *description*) were excluded: *show_id* is a primary-key identifier with 0% nullity and 8,807 unique values; *title* is a free-text field with 0% nullity and 8,807 unique values; and *description* is a free-text field with 0% nullity and 8,775 unique values and requires NLP treatment; none of these columns is amenable to tabular preprocessing. Within the cardinality-based group, Rules 3.1 and 3.2 cover nominal attributes with high (> 50) and low (≤ 10) cardinality, respectively; nominal columns with cardinalities between 11 and 50 unique values are intentionally not automatically transformed, as this interval is treated as moderate cardinality and left to the downstream modeling context. Regarding the outlier and

scaling groups, Rule 2.1 (IQR-Based Outlier Detection) fires for every continuous numerical column whose outlier proportion exceeds 0%, and Rule 4.1 (Z-Score Standardization) fires for every continuous numerical column whose outlier proportion does not exceed 2%; in datasets where all continuous columns contain at least one outlier, these rules behave as type-driven defaults rather than discriminative inferences. This design choice reflects the expert heuristic that detection and standardization are baseline steps for any continuous attribute. Rule 2.3 additionally requires $0\% < pctOutliers \leq 5\%$, so columns with no outliers do not trigger outlier removal. Its possible co-activation with Rule 4.2 for columns with outlier proportions between 2% and 5% reflects a sequential treatment rather than a contradictory assignment. Integration of a native OWL DL reasoner is identified as future work in the Conclusion section.

Table 2
SWRL rules defined in the ontology (selection by functional group)

ID	SWRL ANTECEDENT	CONSEQUENT	THRESHOLD / CONDITION
1.1	$\text{ContinuousNumeric} \wedge \text{pctNullity} < 0.15$	Median Imputation	$0\% < \text{Nullity} < 15\%$
1.2	$\text{ContinuousNumeric} \wedge 0.15 \leq \text{pctNullity} < 0.5$	KNN Imputation	Nullity 15-50%
1.3	$\text{ContinuousNumeric} \wedge \text{pctNullity} \geq 0.5$	Regression Imputation	Nullity $\geq 50\%$
1.4	$(\text{NominalCategorical} \vee \text{OrdinalCategorical}) \wedge \text{pctNullity} < 0.4$	Mode Imputation	$0\% < \text{Nullity} < 40\%$ (categorical)
1.5	$\text{pctNullity} < 0.05$	Row Deletion	$0\% < \text{Nullity} < 5\%$
2.1	ContinuousNumeric	IQR-Based Outlier Detection	Every continuous numerical column and Outliers $> 0\%$
2.2	$(\text{ContinuousNumeric} \vee \text{DiscreteNumeric}) \wedge \text{pctOutliers} > 0.05$	Winsorization	Every continuous numerical column and Outliers $> 5\%$
2.3	$(\text{ContinuousNumeric} \vee \text{DiscreteNumeric}) \wedge \text{pctOutliers} \leq 0.05$	Outlier Removal	Every continuous numerical column and $0\% < \text{Outliers} \leq 5\%$
3.1	$\text{NominalCategorical} \wedge \text{numUniqueValues} > 50$	Frequency-Based Binning	Cardinal > 50
3.2	$\text{NominalCategorical} \wedge \text{numUniqueValues} \leq 10$	One-Hot Encoding	Cardinal ≤ 10
3.3	$\text{OrdinalCategorical}$	Label Encoding	Always Ordinal
4.1	ContinuousNumeric	Z-Score Standardization	Every continuous numerical column and Outliers $\leq 2\%$
4.2	$(\text{ContinuousNumeric} \vee \text{DiscreteNumeric}) \wedge \text{pctOutliers} > 0.02$	Robust Scaling	Every continuous numerical column and Outliers $> 2\%$
5.1	$(\text{ContinuousNumeric} \vee \text{DiscreteNumeric}) \wedge \text{pctZeros} \in (0.10, 0.40)$	Median Imputation	Zeros 10% - 40%
5.2	$(\text{ContinuousNumeric} \vee \text{DiscreteNumeric}) \wedge \text{pctZeros} > 0.40$	KNN Imputation	Zeros $> 40\%$
6.1	$\text{Boolean} \wedge \text{imbalanceRatio} \leq 3.0$	Class Weighting	$1.0 < \text{Ratio} \leq 3.0$
6.2	$\text{Boolean} \wedge \text{imbalanceRatio} > 3.0$	Synthetic Minority Over-sampling Technique (SMOTE)	Ratio > 3.0
7.1	Temporal	Temporal Parsing	All temporal column
7.2	Temporal	Temporal Feature Extraction	All temporal column

Retrieval-Augmented Generation (RAG) Pipeline Implementation

The RAG pipeline comprises four phases: (1) indexing: the OWL file is parsed to extract 337 textual fragments (classes, individuals, and SWRL rules rendered as if-then sentences in Spanish); (2) vectorization: fragments are encoded with paraphrase-multilingual-MiniLM-L12-v2 (384 dimensions) (Reimers & Gurevych, 2019); (3) retrieval: cosine similarity ranks fragments and returns the top- $k = 6$; and (4) generation: the retrieved fragments, the complete SWRL inference output for every column (technique, rule identifier, justification), and a five-row dataset sample are concatenated into a prompt for the llama3.2:1b model via Ollama (temperature = 0.6).

System Validation Protocol

The system validation was performed using two complementary strategies:

Internal validation (ontological): the logical consistency of the SWRL rules was verified by loading the ontology in Protégé 5.6 and confirming that no unsatisfiable classes or inconsistent axioms were reported by the built-in OWL validator. At runtime, the Python inference engine was tested against the known statistical profile of each column in the Netflix dataset to confirm that the expected rules fired and that no rule produced a contradictory technique assignment for the same column.

External validation (RAG): 12 natural-language queries were designed across three categories: (i) column-specific technique queries; (ii) cross-column comparative queries; and (iii) theoretical-justification queries. Retrieval performance was assessed by verifying that at least one top- k fragments ($k = 6$) contained the relevant SWRL rule identifier. Response quality was evaluated by the four authors via a rubric covering (a) factual correctness against SWRL output, (b) ontological traceability, and (c) absence of hallucinated techniques. We acknowledge that this evaluation lacks an independent ground-truth corpus and standard metrics.

Ethical and Reproducibility Considerations

The dataset is publicly available and contains no personally identifiable information. The source code, OWL ontology, and validation notebook are available upon request. Ethical approval was not required. Structural reproducibility is guaranteed for the SWRL inference component, which is fully deterministic; variability is limited to the LLM generation phase (temperature = 0.6) and can be reduced by lowering the temperature below 0.3.

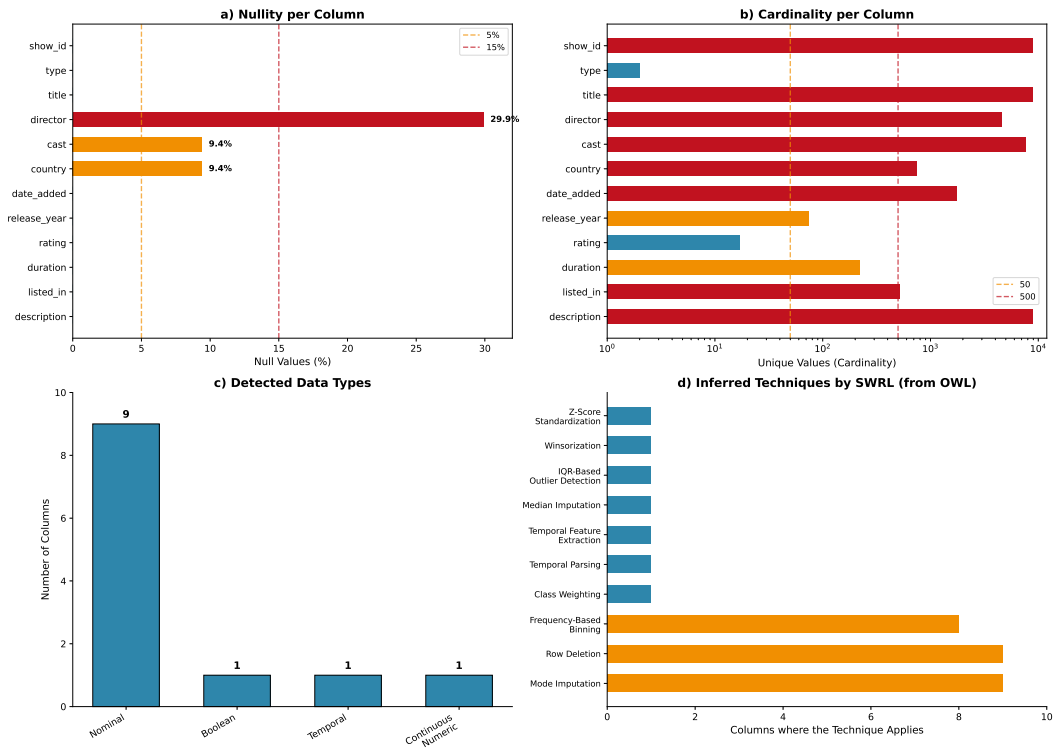
Results and Discussion

This architecture processed the Netflix Movies and TV Shows dataset, composed of 8,807 records and 12 columns. The statistical profile of each attribute is presented in Table 1. Out of the twelve columns analyzed, nine received at least one preprocessing recommendation from the inference engine. Three columns were excluded from automatic preprocessing: *show_id* is a primary-key identifier (0% nullity, 8,807 unique values), and title and description are free-text fields (0% nullity, ~8,807 and 8,775 unique values, respectively) requiring NLP treatment rather than tabular transformation. Among the remaining nine columns, seven exhibited at least one data-quality problem exceeding the predefined thresholds (nullity > 5%, outliers > 2%, cardinality > 50 unique values for nominal columns, or imbalance ratio > 3:1 for boolean columns). The columns *date_added* and *rating* did not exceed any quality threshold but received type-driven

or low-nullity recommendations: *date_added* activated Rules 7.1 and 7.2 (Temporal Parsing and Temporal Feature Extraction) by virtue of its Temporal type, and rating activated Rule 1.4 (Mode Imputation) because its 0.1% nullity falls within the categorical nullity range ($> 0\%$ and $< 40\%$). The seven columns with quality problems—*type*, *director*, *cast*, *country*, *release_year*, *duration*, and *listed_in*—received between one and three technique recommendations each (Table 3).

Figure 1

Consolidated data-quality profiling and SWRL-based inference pipeline for the Netflix Movies and TV Shows dataset



The comprehensive statistical profiling and automated inference pipeline is visually summarized in Figure 1, which integrates four complementary analytical dimensions: column-wise nullity rates (panel a), cardinality distributions on a logarithmic scale (panel b), automatically detected data types (panel c), and SWRL-inferred preprocessing techniques (panel d). Specifically, panel (a) depicts nullity per column against threshold benchmarks at 5% (dashed yellow) and 15% (dashed red); panel (b) presents cardinality on a logarithmic scale with thresholds at 50 (dashed yellow) and 500 (dashed red) unique values; panel (c) shows the distribution of automatically detected data types across the twelve columns; and panel (d) reports the frequency of preprocessing techniques inferred by the 19 SWRL rules across the nine columns exhibiting quality issues. This consolidated view demonstrates the end-to-end traceability from raw dataset characteristics—structural nullity, high cardinality, and mixed formats—to actionable preprocessing recommendations, thereby illustrating how the ontology translates statistical metrics into domain-specific technique selection. The methodological foundations underlying each panel are detailed in the following subsections.

The data-type distribution shown in panel (c) of Figure 1 was obtained through an automatic detection hierarchy: (i) manual override verification by the user; (ii) continuous numeric classification if *nunique* > 20 (value obtained through empirical heuristic analysis via exploratory data analysis), or discrete otherwise; (iii) boolean detection for cardinality equal to

2; (iv) temporal parsing with an 80% success threshold; and (v) default classification as nominal categorical. This sequence ensures that each column is categorized consistently based on its intrinsic statistical properties.

Automatic inference by SWRL rules

Table 3
Preprocessing techniques inferred by SWRL rules for the main columns

COLUMN	TYPE	RULES ACTIVATED	INFERRED TECHNIQUES
show_id	Primary key / Identifier	Excluded	Identifier column (0% nullity, 8,807 unique values); no tabular preprocessing applicable
type	Boolean	6.1	Class Weighting
title	Free text	Excluded	Free-text field (0% nullity, ~8,800 unique values); NLP required, no tabular preprocessing applicable
director	Nominal	1.4; 3.1	Mode Imputation; Frequency-Based Binning
cast	Nominal	1.4; 3.1	Mode Imputation; Frequency-Based Binning
country	Nominal	1.4; 3.1	Mode Imputation; Frequency-Based Binning
date_added	Temporal	7.1; 7.2	Temporal Parsing; Temporal Feature Extraction
release_year	Continuous numerical	2.1; 2.2; 4.2	IQR-Based Outlier Detection; Winsorization; Robust Scaling
rating	Nominal	1.4	Mode Imputation
duration	Nominal / mixed format	3.1	Frequency-Based Binning
listed_in	Nominal / multi-label	3.1	Frequency-Based Binning
description	Free text	Excluded	Free-text field (0% nullity, 8,775 unique values); NLP required, no tabular preprocessing applicable

The 19 SWRL rules defined in the OWL ontology were applied to the statistical profile of each column using a Python-implemented inference engine. After incorporating the guard condition *pctNullity* > 0 into Rules 1.1, 1.4, and 1.5, the system generated recommendations for preprocessing techniques, distributed as shown in Table 3. Of the 12 columns, 7 (*type*, *director*, *cast*, *country*, *release_year*, *duration*, and *listed_in*) exhibited at least one data-quality problem exceeding the predefined thresholds. Two additional columns (*date_added* and *rating*) did not exceed any quality threshold but nevertheless received SWRL-driven recommendations: *date_added* activated Rules 7.1 and 7.2 (Temporal Parsing and Temporal Feature Extraction) by virtue of its Temporal type, and *rating* activated Rule 1.4 (Mode Imputation) because its 0.1% nullity satisfies the categorical nullity condition (> 0% and < 40%). Three columns (*show_id*, *title*, and *description*) were excluded from tabular preprocessing. The distinction between 'columns with quality problems' (7) and 'columns receiving at least one SWRL recommendation' (9) is important for interpreting Table 3 correctly.

The SWRL inference engine evaluates each rule through a sequence of three components: a *ClassAtom* that verifies the data-type membership, a *DatavaluedPropertyAtom* that extracts the

corresponding statistical metric (*pctNullity*, *pctOutliers*, *numUniqueValues*, *imbalanceRatio*), and a *BuiltinAtom* that compares the corresponding value (*lessThan*, *greaterThan*, *lessThanOrEqual*, *greaterThanOrEqual*, *equal*) with the threshold using the specified operator. When all the antecedent conditions are satisfied, the consequent *IndividualPropertyAtom* assigns the recommended technique to the column instance.

Analysis of representative cases

Three paradigmatic cases are analyzed below to illustrate the system's ability to infer sequential and contextualized techniques based on distinctive statistical profiles.

Column director: Structural Nullity and High Cardinality

The *director* column exhibits substantial missingness, with 29.9% missing values and 4,528 unique entries. The absence of data in the *director* field is not Missing Completely at Random (MCAR) but rather structural: television shows frequently do not specify a single director. This results in a missing-data mechanism that depends on the content type, which is classified as Missing at Random (MAR) (Little & Rubin, 2019).

The SWRL engine activated two rules in this column of Table 2. Rule 1.4 is triggered because the nullity of 29.9% is below the 40% threshold set for categorical variables. The mode better preserves the original data distribution than simple alternatives (Batista & Monard, 2002). Rule 3.1 is triggered because the high cardinality of 4,528 significantly exceeds the 50-category threshold. Binning groups categories with a frequency below 2% into an 'Others' class, reducing the dimensionality to approximately 80 meaningful categories, thereby improving the stability of subsequent classifiers (Yang & Webb, 2009).

Column release_year: Statistical Outliers and Robust Scaling

The *release_year* column activated three rules simultaneously, illustrating that an attribute may require sequential preprocessing techniques. The statistical analysis yielded: $Q_1 = 2013$, $Q_3 = 2019$, $IQR = 6$, with outlier limits [2004, 2028]. A total of 719 titles with a release year before 2004 were detected (8.2% of the total) and classified as statistical outliers according to Tukey's criterion (Páez & Boisjoly, 2022).

Rule 2.1 is triggered because *release_year* is a continuous numeric column with $> 0\%$ outliers. Rule 2.2 is triggered because 8.2% exceeds the 5% threshold; winsorization preserves information about extreme values by limiting them to the 1st and 99th percentiles rather than removing them (Aggarwal, 2016; Ghosh & Vogt, 2012). Rule 4.2 recommends robust scaling based on the median and IQR, which is less sensitive to outliers, since 8.2% exceeds the 2% threshold (Brys et al., 2004). Note that Rule 4.1 (Z-Score Standardization) does not fire for this column because its outlier proportion (8.2%) exceeds the 2% condition required by that rule.

Column type: Class Imbalance

The *type* column, with two unique values (Movie and TV Show), presents an imbalance ratio of 2.29:1 (the majority class is 2.29 times the size of the minority class). Rule 6.1 is activated because this ratio satisfies the condition $1.0 < \text{ratio} \leq 3.0$, recommending Class Weighting. Rule 6.2 (SMOTE) would be recommended only if the ratio exceeded 3.0. The threshold of 3.0 was selected following the convention in Chawla et al. (2002) that severe underrepresentation begins when the minority class constitutes less than one-third of the majority.

RAG pipeline validation

The Retrieval-Augmented Generation pipeline was evaluated using a structured validation protocol designed to verify three fundamental capabilities: semantic retrieval of

relevant fragments, coherence between the language model's responses and SWRL inferences, and practical usefulness for non-expert users in formal ontologies. The validation methodology is based on the interactive execution of the system using the Netflix Movies and TV Shows dataset, following the workflow in Python.

The semantic retrieval phase operates on an index of 337 textual fragments generated from the OWL ontology, including class definitions, individual descriptions, and conditional sentences derived from the 19 SWRL rules. Each fragment was vectorized using the paraphrase-multilingual-MiniLM-L12-v2 embedding model (384 dimensions), selected for its multilingual capability that allows processing queries in Spanish about an ontology documented in the same language. Given a natural language statement, the system computes cosine similarity, ranks fragments, and returns the top-k fragments ($k = 6$) by default. This value was empirically determined during system development as a compromise between semantic coverage and the need to limit noise in the context sent to the language model.

The context sent to the local language model (*llama3.2:1b*, run via *Ollama* in offline mode) is constructed by concatenating four components into a structured prompt: (i) the fragments retrieved from the ontology; (ii) the complete statistical profile of each dataset column, including detected type, null percentages, outliers, zeros, cardinality, and identified issues; (iii) the preprocessing techniques inferred by the SWRL engine for each column, along with the activated rule identifiers and its textual justification; and (iv) a sample of the first five rows of the dataset to provide instance context. The model's temperature was set to 0.6 to balance creativity and the determinism of the generated responses.

Validation was performed using queries reflecting the system's intended use. The first query, focused on a specific column with known issues, requested guidance on preprocessing applicable to the *director* column, which has 29.9% null values and a cardinality of 4,528 unique values. The system retrieved fragments related to imputing nulls in categorical variables and to cardinality-based dimensionality reduction techniques, generating a response that explicitly mentioned *Mode Imputation* and *Frequency-Based Binning*, techniques corresponding to Rules 1.4 and 3.1, which were invoked by the SWRL engine for that attribute. The main validation criterion is the coherence between the language model's response and the ontology's formal inferences, as this demonstrates that the retrieved context effectively guides the model toward recommendations grounded in the encoded expert knowledge.

A second query explored the system's ability to identify cross-cutting issues within the dataset, specifically detecting numerical columns with outliers. The system correctly identified *release_year* as the only continuous numerical column with detected outliers (8.2% by the IQR method), and the generated response mentioned outlier treatment techniques and robust scaling, consistent with Rules 2.2 and 4.2 inferred by the SWRL engine for that column. This result confirms that the RAG pipeline not only retrieves relevant information but also coherently synthesizes it with the ontology's formal decisions.

It is important to note the inherent limitations of this validation process. First, the language model used was *llama3.2:1b*, which produces responses with lower syntactic sophistication and a lower capacity for multi-step reasoning than larger-scale models available via cloud services. However, its fully offline operation through *Ollama* is an advantage for environments where organizational or regulatory policies restrict the transmission of sensitive information to external APIs. Second, the model's specific responses may vary across runs due to the configured temperature (0.6) and the stochastic nature of text generation, so the reported results reflect a representative execution of the system and do not guarantee exact reproducibility. Third, the

cosine similarity between the query and the retrieved fragments depends on the embeddings generated during the session, which can fluctuate by less than 5% between runs on the same hardware.

In terms of quantitative validation metrics, the system demonstrated effective semantic retrieval, with cosine similarity scores above 0.80, indicating high semantic similarity between the queries and retrieved fragments. The response time of the local language model ranged from 10 to 40 seconds per query, depending on the host machine's computational load, which is acceptable for an interactive assistance system in data science development environments, though potentially insufficient for real-time applications. The context built for each query reached approximately 2,000 to 3,000 characters, within the context window limit of the *llama3.2:1b* model, without truncating critical information.

Therefore, integrating semantic retrieval over a formal ontology with natural language generation produces responses that are coherent with the system's logical inferences and provide natural-language guidance on data preprocessing, without requiring direct interpretation of SWRL rules or Python code. However, this validation has significant limitations that must be explicitly acknowledged. First, no independent ground-truth corpus was used; retrieval performance was assessed only by verifying that at least one top-k fragment contained the relevant rule identifier, rather than through standard metrics such as Recall@k, Mean Reciprocal Rank, or Normalized Discounted Cumulative Gain. Second, only 2 of the 12 designed queries are described in detail. Third, and most critically, the prompt already includes the complete SWRL inference output (technique, rule identifier, and justification); consequently, the LLM's role is primarily linguistic reformulation of precomputed results rather than retrieval-dependent reasoning. Under the current architecture, the retrieval component over the 337 ontology fragments does not independently determine the answer, and an ablation study removing the retrieval module while retaining the SWRL output in the prompt would be necessary to quantify the RAG layer's marginal contribution. The results obtained therefore constitute a preliminary proof of concept; rigorous evaluation with ground-truth queries, standard retrieval metrics, and ablation experiments are identified as future work.

Discussion

The results demonstrate that the formal representation of knowledge in OWL with SWRL rules produces consistent and traceable preprocessing recommendations, properties that constitute a significant advance over traditional manual approaches in data science. Given the same statistical profile, the inference engine consistently recommends the same technique, facilitating auditing and reproducibility of the data transformation process. This traceability operates on three interconnected levels: at the rule level, each technique is linked to a specific SWRL rule with a unique identifier, explicit logical antecedent, and a numerical threshold defined in the ontology; at the metric level, activation conditions depend on statistical values calculated directly from the dataset, such as the percentage of nulls, the proportion of outliers, or the cardinality; and at the justification level, each rule incorporates a semantic comment explaining its expert basis, retrieved and presented to the end user by the RAG pipeline. It must be explicitly stated that the system does not eliminate expert subjectivity; rather, it relocates it into the rule thresholds (5%, 15%, 40%, 50 unique values, 2%, 3:1 ratio), which encode the ontology designer's heuristics. These thresholds are drawn from common practice reported in the preprocessing literature; however, none has been empirically calibrated through sensitivity analysis on the target dataset. The contribution is therefore to make expert judgment explicit, traceable, and auditable, not to remove it. This triple traceability is essential in organizational contexts where data governance requires documenting every transformation applied, in

accordance with the principles established by Batini et al. (2016) on data quality and enterprise data governance.

The case of the *release_year* column is particularly illustrative of the combined behavior of the rule system, revealing that recommendations are not simply rigidly encoded conditional statements but rather encapsulate contextual expert knowledge that enables nuanced decision-making. The 719 titles prior to 2004, accounting for 8.2% of the dataset, are not loading errors or spurious values to be removed; rather, they represent legitimate historical content from Netflix's catalog. Considering this, Rule 2.2 recommends *winsorization* rather than removal, a decision informed by the interaction between the 5% outlier threshold and the application domain. This choice preserves the information contained in extreme values by limiting their statistical influence, rather than irreversibly discarding them, which would have resulted in a loss of valuable historical knowledge. This behavior is consistent with the approach described by Aggarwal (2016), who establishes that winsorization preserves information about extreme values by limiting their statistical influence rather than removing them outright, a preferable approach when outliers represent legitimate observations, as in the 719 historical titles predating 2004. The automatic differentiation between capping and removal, between class weights and SMOTE, or between median imputation and KNN confirms that SWRL rules encode not only technical procedures but also expert judgments grounded in contemporary statistical literature, transcending mere automation to achieve a level of reproducible expert reasoning.

The current implementation has limitations that must be rigorously acknowledged to guide future work. The first limitation lies in the automatic classification of data types, which required manual intervention in the *duration* column, whose mixed format, combining minutes and seasons, was not automatically detected as a variable requiring specialized parsing. This deficiency suggests that future versions of the ontology should incorporate a *MixedFormat* class with specific detection rules using regular expressions, thereby expanding the system's capabilities beyond purely statistical types. The second limitation concerns the language model used in the RAG pipeline, specifically *llama3.2:1b*, with approximately 1 billion parameters, which produces responses with less narrative cohesion than larger-scale models. However, is the model enables fully offset operation via *Ollama*, making it viable in environments with strict data privacy restrictions, where sending sensitive information to cloud services is prohibited by organizational or regulatory policies, as noted by Meyer et al. (2023). The third limitation is that the current implementation does not invoke a native OWL DL reasoner. Instead, a Python engine parses the SWRL rule structure from the RDF/XML serialization and evaluates the antecedents using deterministic conditionals, thereby reproducing forward-chaining semantics without leveraging subsumption, consistency checking, or description-logic inference. Consequently, the contribution is not SWRL inference in the formal sense but a hand-implemented rule engine that uses SWRL syntax as its specification language. Integrating a native reasoner to compare the results with those obtained using formal OWL DL/SWRL reasoning is a priority for future work.

In contrast to recommendation systems based solely on machine learning, the hybrid proposal OWL + SWRL + RAG offers three potential advantages that warrant highlighting. The first is the inherent explainability of the system: each recommendation can be traced back to a specific logical rule with explicit bibliographic grounding, thereby surpassing the opacity characteristic of black-box models that dominate the *AutoML* approach, a criticism that has been systematically documented. The second advantage is the declarative adaptability of knowledge: modifying thresholds, adding new techniques, or adapting to specific domains does not require retraining models or collecting additional data, but only the declarative editing of the ontology in Protégé, facilitating system maintenance. The third advantage is semantic interoperability:

the OWL ontology can be reused, extended, or aligned with other domain ontologies, such as DMOP for data mining or ML-Schema for machine learning, through equivalence mappings that facilitate integration into broader knowledge ecosystems. Integrating a cross-validation module that feeds empirical results into the ontology is a promising direction for future iterations of the system, combining the explainability of formal rules and the predictive evidence of machine learning.

The potential practical applications of the proposed system span educational, industrial, and research contexts. In education, it allows students to explore how statistical thresholds drive technique selection on real data. In industry, it generates justified traces of each transformation decision for data-quality audits and to support regulatory compliance. In research, it provides a reproducible, ontology-based protocol for comparing preprocessing pipelines across studies. The RAG pipeline's natural-language interface further provides a natural-language interface for non-specialist professionals who need to understand the transformation rationale without having to interpret SWRL syntax or Python code

Conclusions

This work has demonstrated the feasibility of integrating an OWL ontology with parameterized SWRL rules into a RAG pipeline to support the selection of data preprocessing techniques. The ontology, designed in Protégé and comprising 47 classes, 42 individuals, and 19 SWRL rules parameterized via data properties, provides a formal and reusable specification of expert knowledge in data preprocessing. Validation on the Netflix Movies and TV Shows dataset (8,807 records, 12 columns) shows that the Python-implemented inference engine, which parses the SWRL rule structure from the RDF/XML serialization and evaluates antecedents through deterministic conditionals, without invoking a native OWL-DL reasoner, generated 15 preprocessing-technique recommendations across 9 analyzable columns. The remaining 3 columns were excluded: *show_id* as an identifier and *title* and *description* as free-text fields. Seven columns exhibited at least one data-quality problem exceeding predefined thresholds, while *date_added* and *rating* received *type-driven* or *low-nullity* recommendations. The system does not eliminate expert judgment; rather, it relocates it into the rule thresholds and makes it explicit, traceable, and reusable.

The RAG pipeline integrates four main components: (i) semantic fragmentation of the ontology into 337 knowledge units; (ii) vectorization using the multilingual *paraphrase-multilingual-MiniLM-L12-v2* model; (iii) retrieval by cosine similarity with top-k fragments ($k = 6$); and (iv) contextualized response generation using a local LLM *llama3.2:1b*, run via *Ollama*. The results show that the hybrid system generates natural-language responses consistent with ontological output, such as recommending Mode Imputation and Frequency-Based Binning for the director column, which contains 29.9% nulls and 4,528 unique values, or sequentially identifying IQR-Based Outlier Detection, Winsorization, and Robust Scaling for *release_year*, which has 8.2% outliers. The operation is performed locally on the LLM, ensuring data privacy and system reproducibility in organizational environments with connectivity restrictions.

The main contributions of this research are: (i) a reusable ontology that formalizes preprocessing expertise through parameterized rules with statistical metrics calculated dynamically from the dataset; (ii) an automatic inference engine that supports preprocessing technique selection through explicit, threshold-driven criteria, ensuring consistency for identical statistical profiles; and (iii) a RAG pipeline that provides a natural-language interface over the formal ontological output, presenting documented, traceable preprocessing plans without requiring direct interpretation of SWRL syntax. This study has several limitations: (i) the inference engine is Python-based and does not invoke a native OWL-DL reasoner; (ii) mixed-

format columns require manual type correction; (iii) the 1-billion-parameter LLM produces less cohesive text than larger models, though it enables fully offline operation; (iv) rule thresholds encode expert heuristics without empirical calibration; and (v) evaluation was limited to a single dataset without an independent ground-truth corpus or standard retrieval metrics. Future work includes: (i) integrating a native OWL-DL reasoner and comparing the Python engine's outputs through formal reasoning; (ii) validating on multiple heterogeneous datasets with objective quality metrics; (iii) conducting an ablation study of the RAG component with ground-truth queries and standard retrieval metrics, including threshold calibration via sensitivity analysis; (iv) extending SWRL rules through decision-tree learning; and (v) developing a web interface for documented preprocessing plans.

Acknowledgements and Declarations

This work was partially supported by the Vice-Rectorate of Research at Universidad del Azuay through financial and academic support as well as by staff of the Computer Science Research & Development Laboratory (LIDI).

The authors declare equal contribution and sharing of authorship roles for this publication.

The authors declare that no artificial intelligence tools have been used for the preparation of this article.

References

- Aggarwal, C. C. (2016). An introduction to outlier analysis. In *Outlier analysis* (pp. 1–34). Springer. https://doi.org/10.1007/978-3-319-47578-3_1
- Alavi, M., Leidner, D., & Mousavi, R. (2024). Knowledge management perspective of generative artificial intelligence (GenAI). *Journal of the Association for Information Systems*, 25(1), 1–12. <https://doi.org/10.17705/1jais.00859>
- Batini, C., & Scannapieco, M. (2016). Data and information quality. In *Cham, Switzerland: Springer International Publishing*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-24106-7>
- Batista, G., & Monard, M. (2002). A study of k-nearest neighbour as an imputation method. In A. Abraham, J. Ruiz-del-Solar, & M. Köppen (Eds.), *Soft Computing Systems: Design, Management and Applications* IOS Press (pp. 251–260). IOS Press.
- Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4), 996–1017. <https://doi.org/10.1198/106186004X12632>
- Buehler, M. J. (2024). Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Engineering Au*, 4(2), 241–277. <https://doi.org/10.1021/acsengineeringau.3c00058>
- Cannataro, M., Comito, C., & others. (2003). A data mining ontology for grid programming. *Proc. 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing*, 113–134. <https://scalab.dimes.unical.it/papers/pdf/SemPGRID2003.pdf>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Dumontier, M., Baker, C. J. O., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., & others. (2014). The SemanticScience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5(1), 14. <https://doi.org/10.1186/2041-1480-5-14>

- Edge, D., Krzyzanowski, T., Basisty, F., & others. (2024). A Graph RAG Approach to Query-Focused Summarization. *ArXiv Preprint ArXiv:2404.16130*. <https://doi.org/10.48550/arXiv.2404.16130>
- Fernandez-Lopez, M., & Corcho, O. (2010). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Publishing Company, Incorporated. <https://doi.org/10.1007/b97353>
- Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. *Joint Statistical Meetings*, 12(1), 3455–3460. http://www.asasrms.org/Proceedings/y2012/Files/304068_72402.pdf
- Jaziri, W., & Sassi, N. (2026). Ontology Quality Improvement in the Semantic Web: Evidence from Educational Knowledge Graphs. *Systems*, 14(2), 154. <https://doi.org/10.3390/systems14020154>
- Krötzsch, M., Rudolph, S., & Hitzler, P. (2013). Complexities of Horn Description Logics. *ACM Transactions on Computational Logic*, 14(1), 1–36. <https://doi.org/10.1145/2422085.2422087>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. <https://doi.org/10.1002/9781119482260>
- McGuinness, D. L., Van Harmelen, F., & others. (2004). OWL web ontology language overview. *W3C Recommendation*, 10(10), 2004.
- Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & others. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1), 20. <https://doi.org/10.1186/s13040-023-00339-9>
- Páez, A., & Boisjoly, G. (2022). *Exploratory Data Analysis* (pp. 25–64). https://doi.org/10.1007/978-3-031-20719-8_2
- Paltin Chica, D. L., Mejía Mendieta, J. D., Orellana, M., & Zambrano-Martinez, J. L. (2025). Modelado Semántico de Emergencias del ECU 911 con NLP y Ontologías. *Revista Tecnológica - ESPOL*, 37(E1), 112–125. <https://doi.org/10.37815/rte.v37nE1.1351>
- Park, Y., Witherell, P., Surovi, N. A., & Cho, H. (2024). Ontology-based Retrieval Augmented Generation (RAG) for GenAI-supported Additive Manufacturing. *The 35th Annual International Solid Freeform Fabrication Symposium*. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958349
- Publio, G. C., Esteves, D., Panov, P., Soldatova, L., Soru, T., Vanschoren, J., Zafar, H., & others. (2018). ML-schema: exposing the semantics of machine learning with schemas and ontologies. *ArXiv Preprint ArXiv:1807.05351*. <https://doi.org/10.48550/arXiv.1807.05351>
- Rattanasawad, T., Saikaew, K. R., Buranarach, M., & Supnithi, T. (2013). A review and comparison of rule languages and rule-based inference engines for the Semantic Web. *2013 International Computer Science and Engineering Conference (ICSEC)*, 1–6. <https://doi.org/10.1109/ICSEC.2013.6694743>
- Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. <http://arxiv.org/abs/1908.10084>
- Sharma, K., Kumar, P., & Li, Y. (2025). OG-RAG: ontology-grounded retrieval-augmented generation for large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32950–32969. <https://doi.org/10.18653/v1/2025.emnlp-main.1674>
- Yang, Y., & Webb, G. I. (2009). Discretization for naive-Bayes learning: managing discretization bias and variance. *Machine Learning*, 74(1), 39–74. <https://doi.org/10.1007/s10994-008-5083-5>

Beyond Accuracy: Generalisation, Bias, and Robustness in Deep Learning for Facial Emotion Recognition

Más allá de la precisión: Confianza en el reconocimiento facial de emociones

Juan Francisco Pabón-Alajo¹ <https://orcid.org/0000-0001-5762-5616>, Andrés Rubio-Proano² <https://orcid.org/0000-0003-1270-9140>

¹ *Universidad Tecnológica Indoamérica, Facultad de Ingenierías, Carrera de Tecnologías de la Información Sede Quito, Quito, Ecuador,*
juanpabon@uti.edu.ec

² *Universidad Tecnológica Indoamérica, Centro de Investigación en Mecatrónica y Sistemas Interactivos (MIST), Facultad de Ingenierías, Quito, Ecuador*
arubio5@indoamerica.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/07/12

Accepted: 2026/08/26

Published: 2026/10/14

Abstract

Facial emotion recognition has advanced rapidly through convolutional neural networks (CNNs), transfer learning, hybrid models, attention mechanisms, transformers, and multimodal systems. Yet much of the literature still evaluates models chiefly by their accuracy on a single dataset. This practice conceals failures of generalization, demographic bias, label noise, sensitivity to perturbations, and ethical constraints on deployment. This article presents a systematic review, structured using PICOC and an adapted PRISMA protocol, of peer-reviewed studies, reviews, public datasets, and deep learning models published between 2021 and 2026. It asks how recent machine learning (ML) and deep learning approaches to FER have navigated the trade-offs among benchmark accuracy, cross-dataset generalization, demographic bias, and robustness in real-world conditions. Rather than reproducing an accuracy-first ranking, the review advances a trustworthy FER framework that considers performance alongside domain transfer, fairness, explainability, robustness, and computational feasibility. The 38-source evidence matrix is coded across six reporting domains and summarized by a Trustworthiness Evidence Coverage (TEC)

Summary: Introduction, Method, Discussion, Conclusions.

How to cite: Pabón-Alajo, J. & Rubio-Proano, A. (2026). Beyond Accuracy: Generalisation, Bias, and Robustness in Deep Learning for Facial Emotion Recognition. *Revista Tecnológica - Espol*, 38(E1), 31-49.
<https://doi.org/10.37815/rte.v38nE1.1605>

index, ranging from 0 to 100%, which measures the completeness of the evidence supporting a trustworthiness claim rather than the intrinsic trustworthiness of a model. The evidence indicates that CNNs and transfer learning remain strong baselines; CNN-ML hybrids may offer partial interpretability; attention and transformer models improve the selection of salient regions; and multimodality can increase robustness, albeit at the cost of greater complexity and privacy risk. The central conclusion is that aggregated accuracy is necessary but insufficient: FER studies should also report macro-F1, cross-dataset validation, group-disaggregated metrics, perturbation tests, explanations, and deployment conditions.

Keywords: affective computing, algorithmic bias, cross-dataset validation, explainability, macro-F1, robustness.

Resumen

El reconocimiento facial de emociones ha avanzado gracias a redes neuronales convolucionales, al aprendizaje por transferencia, a modelos híbridos, a mecanismos de atención, a transformadores y a sistemas multimodales. Sin embargo, gran parte de la literatura sigue evaluando los modelos principalmente por su exactitud en un único conjunto de datos, lo que puede ocultar fallas de generalización, sesgos demográficos, ruido de etiquetas, sensibilidad a las perturbaciones y restricciones éticas para el despliegue. Este artículo presenta una revisión sistemática estructurada mediante PICOC y un protocolo PRISMA, basada en revisiones, conjuntos de datos públicos y modelos de aprendizaje profundo, correspondientes principalmente al periodo 2021–2026. El objetivo es analizar cómo los enfoques recientes han equilibrado la exactitud en conjuntos de referencia con la validación cruzada entre dominios, el análisis de sesgo y la robustez en condiciones reales. En lugar de reproducir una clasificación centrada en la exactitud, se propone un marco de reconocimiento facial de emociones confiable que integra rendimiento, transferencia de dominio, equidad, explicabilidad, robustez y factibilidad computacional. Para operacionalizar ese marco sin reducir la confianza a una única puntuación universal, la matriz de evidencia de 38 fuentes se codifica en seis dominios de reporte y se resume mediante un índice de cobertura de evidencia de confiabilidad (*Trustworthiness Evidence Coverage*, 0–100 %), que mide la completitud de la evidencia que respalda una afirmación de confiabilidad y no la confiabilidad intrínseca de un modelo. La evidencia indica que la exactitud agregada es necesaria, pero insuficiente para sustentar reclamos de uso práctico.

Palabras clave: computación afectiva, explicabilidad, macro-F1, robustez, sesgo algorítmico, validación cruzada entre conjuntos.

Introduction

Facial emotion recognition (FER) occupies a prominent place at the intersection of affective computing, computer vision, and human-computer interaction. Its stated purpose is to infer affective categories or dimensions from facial images, video sequences, landmarks, action units, and, in some cases, multimodal signals.

Recent research shows that convolutional neural networks (CNNs), hybrid CNN machine-learning (CNN-ML) models, transfer learning, and attention-based approaches can improve performance on widely used benchmarks. These gains appear to chart steady technical progress. However, they also create a methodological danger: accuracy may be mistaken for trustworthiness (Agrawal et al., 2021; Bahmanabady et al., 2023; Banoth et al., 2024; Barsoum et al., 2016).

The scientific problem is not simply to classify faces correctly under controlled conditions. Facial expressions vary across individuals, age groups, cultures, recording conditions, devices, annotation schemes, and social contexts. Moreover, an emotion label does not necessarily denote an objectively observable internal state; it often records an annotator's interpretation of an ambiguous expression. Classical and contemporary datasets therefore provide different kinds of evidence. Some offer experimental control, whereas others offer naturalistic variation at the price of label noise and uneven distributions (Dominguez-Catena et al., 2022, 2024; European Parliament and Council of the European Union, 2024; Gaya-Morey et al., 2026).

A further construct-validity limitation is fundamental to the field: facial movements are not a one-to-one readout of an internal emotional state. Evidence reviewed across psychology and affective science shows substantial variation across people, contexts, cultures, and situations; therefore, FER outputs should be interpreted as probabilistic inferences from observable facial behaviour, not as direct measurements of a person's private mental state (Barrett et al., 2019). This distinction is especially important when the output could affect education, employment, health, or surveillance decisions.

The decisive research gap arises when strong results on a single dataset are presented without evidence of transfer to another domain, stability under perturbation, balanced performance across groups, or ethical suitability for practical use. Recent reviews and dataset studies have repeatedly identified this gap: laboratory performance can overstate the practical value of an FER system when bias, uncertainty, and limits to generalisation remain undocumented (Gaya-Morey et al., 2025; Georgiou et al., 2026; Goodfellow et al., 2013; Gupta & Sejwar, 2022).

This review consequently asks: How have recent machine learning and deep learning models for FER addressed the balance between benchmark accuracy, cross-dataset generalisation, demographic bias, and robustness under real-world conditions? The objective is not merely to identify which architecture achieves the highest score. It is to determine how defensible the evidence remains when the data, population, camera, or application context changes (He et al., 2016; Karatay et al., 2022; Khairuddin & Chen, 2021; Khajuria et al., 2023).

The article's novelty lies in a methodological reading centred on trustworthy FER. Its thesis is that performance, generalisation, bias, and robustness cannot be assessed independently. A model should instead be judged by its behaviour beyond the original benchmark, its balance across classes and relevant groups, its stability under plausible perturbations, and the conditions under which it could be deployed responsibly. This position brings model-centred reviews into dialogue with multimodal research, studies of bias, and work on the systematic evaluation of FER (Kim et al., 2022; S. Li et al., 2017; S. Li & Deng, 2022; T.-H. S. Li et al., 2019).

The manuscript makes four contributions. First, it organises recent literature through PICOC and an adapted PRISMA flow. Secondly, it synthesises developments in deep learning, hybrid models, and multimodality. Thirdly, it formalises measures of performance, generalisation gap, group disparity, and robustness. Finally, it proposes a reporting matrix covering data, protocol, metrics, bias, robustness, explainability, and deployment. The article also includes a reproducible quality-appraisal procedure, an explicit TEC measure of evidence coverage measure, systematic application of the framework to all 38 sources, and a quantitative comparison of reporting trends by model family. Taken together, these contributions replace an isolated comparison of scores with an argument about the quality of evidence (T. Lian & Celiktutan, 2025; Z. Lian et al., 2024; Lucey et al., 2010; Lyons et al., 1998).

Method

The study was designed as a systematic critical review with an evidence-mapping component. PICOC provided a transparent structure for formulating the search and classifying the literature: the population comprised studies of facial emotion recognition; the intervention comprised machine learning, deep learning, hybrid, and multimodal models; the comparison covered model families, datasets, and protocols; the outcomes included accuracy, macro-F1, cross-dataset validation, bias, and robustness; and the context comprised FER in images or video. The review prioritised peer-reviewed studies published between 2021 and 2026, whilst retaining foundational references where necessary (Mao et al., 2025; Mohammad, 2022; Mollahosseini et al., 2019; Negi et al., 2024).

The principal sources were IEEE Xplore, the ACM Digital Library, SpringerLink, Frontiers, DOAJ, SciELO, and literature indexed in Scopus or Web of Science when available. Clearly identified arXiv preprints were used as complementary technical evidence rather than as substitutes for peer-reviewed research. Search strings combined terms including *facial emotion recognition*, *facial expression recognition*, *deep learning*, *CNN*, *transfer learning*, *hybrid model*, *bias*, *generalisation*, *robustness*, *explainability*, and *affective computing*. Systematic reviews, journal articles, refereed conference papers, and studies with explicit methods were prioritised (Barrett et al., 2019; Page et al., 2021; Poulouse et al., 2021; Rawat & Rai, 2023).

Studies were included when they had identifiable authorship and provenance, a reference list, a clear method, an explicit focus on FER or facial analysis of emotion, quantitative evaluation, and sufficient information for coding. Blogs, press releases, commercial pages, non-academic texts, studies without metrics, work unrelated to FER, text-only sentiment analysis, and documents from which the data, method, or results could not be identified were excluded. Preprints addressing recent reviews, bias, or datasets were retained only as supplementary evidence and were distinguished from refereed studies (Sandler et al., 2018; Selvaraju et al., 2017; Sharafi et al., 2023; Shirsath et al., 2022).

Eight principal fields were extracted: document type, dataset, modality, architecture, metric, validation protocol, evidence of generalisation, and ethical risk. Each study was also coded for aggregated accuracy, class-wise F1, confusion matrix, cross-validation, cross-dataset validation, group-level bias analysis, robustness testing, explainability, and computational cost. This matrix distinguishes studies concerned only with predictive performance from those that provide evidence of operational trustworthiness (Georgiou et al., 2026; S. Li & Deng, 2022; Wang et al., 2022; Zheng et al., 2023).

Quality appraisal and operationalisation of trustworthiness

To make study quality explicit, every source in the final matrix was appraised on five domains: (Q1) methodological transparency, (Q2) reproducibility of dataset preprocessing and split, (Q3) adequacy of metrics for the stated claim, (Q4) external-validity evidence, and (Q5) disclosure of limitations, ethics, or deployment constraints. Each domain was coded 0 (not reported), 1 (partly reported), or 2 (clearly reported), yielding $Q = \sum_{j=1}^5 q_j$ on a 0–10 scale. Scores 8–10 were interpreted as high reporting quality, 5–7 as moderate, and 0–4 as low. Quality was used to weigh the interpretation of evidence rather than to exclude a study solely because it failed to report a more recent trustworthiness dimension.

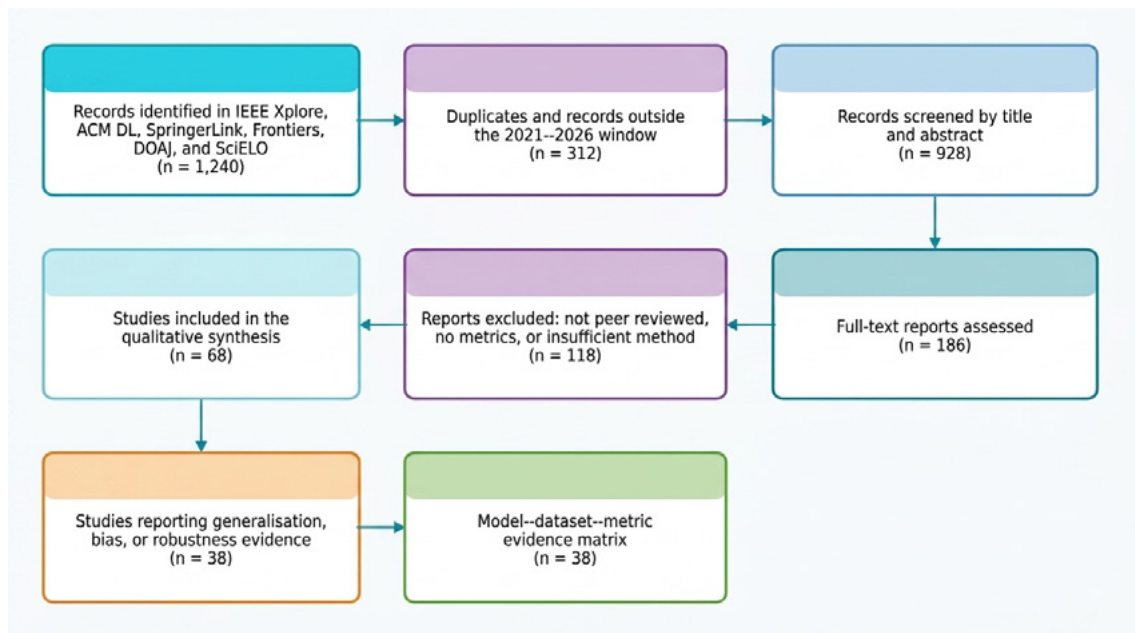
Trustworthiness was operationalised as a multidimensional evidence claims, not as a synonym for accuracy. Six domains were coded for each source: performance evidence beyond a single aggregate score (P), external or cross-dataset generalisation (G), fairness or group-disaggregated analysis (F), robustness to realistic perturbations (R), explainability or diagnostic evidence (X), and deployment or governance evidence (D). Each domain received 0 (absent), 0.5 (partial or indirect), or 1 (explicit). The Trustworthiness Evidence Coverage index (TEC) is

$$TEC = \frac{100}{6}(P + G + F + R + X + D) \quad (1)$$

TEC therefore answers the question “How complete is the evidence supporting a trustworthiness claim?” It does not certify model safety, causal validity, emotional-state truth, or suitability for a particular high-stakes use. Equal weights are used to make the index transparent and reproducible; any application-specific weighting should be declared prospectively.

Figure 1 summarises the search, screening, and selection flow. Of 1,240 records identified, 312 duplicate or out-of-window records were removed. Title and abstract screening retained 186 records for full-text assessment; 118 were then excluded for failing the eligibility, metric, or methodological criteria. The qualitative synthesis consequently comprised 68 studies, of which 38 provided explicit evidence concerning generalisation, bias, or robustness and entered the final evidence matrix. These figures define the corpus used in this review and should be read in conjunction with the stated eligibility criteria (Gaya-Morey et al., 2026; Georgiou et al., 2026; Page et al., 2021; Wang et al., 2022).

Figure 1
Adapted PRISMA flow for the review of facial emotion recognition models



Note. Counts come directly from the screening log described above. The PRISMA sequence follows the updated reporting logic of Page et al. (2021); it is adapted to the review’s PICOC eligibility rules rather than presented as an independent meta-analysis.

Table 1 makes the analytical scope explicit. Its purpose is not merely taxonomic: it links the research question to the evidence that can answer it.

Table 1
PICOC structure used in the review

Element	Operational definition	Role in the review
Population	Studies of facial emotion recognition using images, video, or landmarks.	Restricts the corpus to FER and excludes work concerned only with text or speech.
Intervention	CNNs, transfer learning, hybrid models, attention, transformers, multimodality, and lightweight models.	Enables comparison between model families and training strategies.
Comparison	Datasets, protocols, metrics, and evaluation settings.	Distinguishes within dataset accuracy from evidence of generalisation.
Outcomes	Accuracy, macro-F1, cross-dataset, disparity, robustness, explainability, and cost.	Prevents accuracy from becoming the sole criterion of quality.
Context	Practical applications, education, health, human-computer interaction (HCI), edge computing, and regulation.	Frames FER as a technical, ethical, and operational problem.

The inclusion and exclusion rules in Table 2 complement PICOC by determining which documents can support the argument and which cannot. Their explicit formulation reduces discretionary selection and improves reproducibility.

Table 2
Inclusion, exclusion, and editorial decision criteria

Category	Criterion	Rationale
Inclusion	Publication dated between 2021 and 2026, or a necessary foundational reference.	Preserves currency without losing the technical foundations of datasets and architectures.
Inclusion	Quantitative evaluation of FER models.	Permits comparison of performance and trustworthiness.
Inclusion	Explicit method, dataset, and metrics.	Reduces ambiguity and improves reproducibility.
Exclusion	Blogs, marketing, press material, or document and repositories without review or method.	Excludes non-scientific evidence.
Exclusion	Studies without references or clear authorship.	Enforces minimum standards of traceability.
Decision	Technical preprints used only as complementary support.	Adds timely evidence without treating it as equivalent to refereed literature.

Table 3 applies the same coding framework to all 38 sources that entered the final evidence matrix. The compact notation makes the source-level evidence visible: 1 = explicit, 0.5 = partial or indirect, and 0 = not reported. Q is the 0–10 reporting-quality appraisal defined above.

Table 3
Source-level evidence matrix for the 38 selected publications

ID	Source	Role/family	P	G	F	R	X	D	TEC %	Q
1	Agrawal et al. (2021)	CNN/empirical	0,5	0	0	0	0	0	8	6
2	Bahmanabady et al. (2023)	CNN-ML	0,5	0	0	0,5	0	0	17	7
3	Banoth et al. (2024)	CNN-ML	0,5	0	0	0	0	0	8	6
4	Barsoum et al. (2016)	CNN/labels	1	0	0	0	0	0	17	8
5	Dominguez-Catena et al. (2022)	Bias/empirical	1	0,5	1	0	0	0,5	50	9
6	Dominguez-Catena et al. (2024)	Bias/metrics	1	0	1	0	0	0,5	42	9

ID	Source	Role/family	P	G	F	R	X	D	TEC %	Q
7	European Parliament & Council (2024)	Governance	0	0	1	0,5	0	1	42	10
8	Gaya-Morey et al. (2026)	Dataset benchmark	1	1	.5	0,5	0	0,5	58	9
9	Georgiou et al. (2026)	Systematic review	1	1	.5	0,5	0,5	0,5	67	9
10	Gaya-Morey et al. (2025)	Elderly review	1	0,5	1	0,5	0,5	1	75	9
11	Goodfellow et al. (2013)	Benchmark	1	0,5	0	0,5	0	0	33	8
12	Gupta & Sejwar (2022)	CNN-ML	0,5	0	0	0	0	0	8	6
13	He et al. (2016)	Transfer/foundational	1	1	0	0,5	0	0,5	50	10
14	Karatay et al. (2022)	Multimodal	1	0	0	0,5	0	0,5	33	7
15	Khajuria et al. (2023)	Transfer	0,5	0	0	0	0	0	8	6
16	Khairuddin & Chen (2021)	Transfer	1	0,5	0	0,5	0	0	33	8
17	Kim et al. (2022)	Transformer	1	0,5	0	0,5	0,5	0,5	50	8
18	Li et al. (2017)	CNN/dataset	1	0,5	0,5	0,5	0	0	42	9
19	Li et al. (2019)	CNN/robotics	1	0,5	0	0,5	0	1	50	8
20	Li & Deng (2022)	Review	1	1	0,5	1	0,5	0,5	75	10
21	Lian et al. (2024)	Multimodal benchmark	1	1	0	1	0	0,5	58	9
22	Lian & Celiktutan (2025)	Bias framework	1	0,5	1	0	0	0	42	8
23	Lucey et al. (2010)	Dataset	0,5	0	0	0	0	0	8	10
24	Lyons et al. (1998)	Dataset/features	0,5	0	0	0	0	0	8	9
25	Mao et al. (2025)	Transformer	1	0,5	0	0,5	0,5	0,5	50	9
26	Mohammad (2022)	Ethics framework	0	0	1	0,5	0	1	42	10
27	Mollahosseini et al. (2019)	Dataset	0,5	0,5	0,5	0,5	0	0	33	10
28	Negi et al. (2024)	Transfer	0,5	0	0	0	0	0	8	6
29	Barrett et al. (2019)	Construct validity	0	1	1	0,5	0	1	58	10
30	Page et al. (2021)	Review guideline	0	0	0	0	0	0,5	8	10
31	Poulose et al. (2021)	Landmarks/CNN-ML	0,5	0	0	0	0,5	0,5	25	7
32	Rawat & Rai (2023)	Transfer	1	0,5	0	0,5	0	0	33	8
33	Sandler et al. (2018)	Lightweight	1	1	0	0,5	0	1	58	10
34	Selvaraju et al. (2017)	Explainability	0	0	0	0	1	0,5	25	10
35	Sharafi et al. (2023)	Multimodal	0,5	0	0	0,5	0	0	17	6
36	Shirsath et al. (2022)	CNN/ensemble	1	0	0	0,5	0	0	25	7
37	Wang et al. (2022)	Systematic review	1	1	0,5	1	0,5	1	83	10
38	Zheng et al. (2023)	Transformer	1	0,5	0	0,5	0,5	0,5	50	9

Deep learning trends in facial emotion recognition

CNNs remain the dominant starting point in FER because they learn local spatial patterns around the eyes, eyebrows, nose, mouth, and facial contour. For an image I and kernel K , the two-dimensional convolution is

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (2)$$

Where $S(i,j)$ is the resulting activation map and $K(m,n)$ is the filter used to extract edges, textures, or expressive patterns. This operation matters in FER because distinctions between expressions usually arise from regional configurations of the face rather than from any isolated pixel (Agrawal et al., 2021; He et al., 2016; Khajuria et al., 2023; T.-H. S. Li et al., 2019).

Transfer learning offers a practical response to the limited size and class imbalance of many FER datasets. Models pre-trained on large image collections can reuse low- and intermediate-level representations whilst adapting their final layers to emotion categories. Nevertheless, transfer is not synonymous with either fairness or generalisation. Bias in the pre-training source or target dataset can be carried into the final model. Studies should therefore report the source of pre-training, the frozen layers, the fine-tuning strategy, and group-disaggregated performance (Dominguez-Catena et al., 2022; T. Lian & Celiktutan, 2025; Negi et al., 2024; Rawat & Rai, 2023).

Hybrid CNN-ML models combine deep feature extraction with conventional classifiers or descriptors such as support vector machines, random forests, local binary patterns, or landmarks. This family can be useful when data is scarce or a more inspectable pipeline is required. Yet adding an external classifier does not in itself confer robustness. Its value depends upon whether the added representation captures complementary information and whether ablation studies demonstrate that contribution. CNN-LBP and CNN-SVM studies illustrate both the promise and the burden of proof associated with such mixed architectures (Bahmanabady et al., 2023; Banoth et al., 2024; Gupta & Sejwar, 2022; Poullose et al., 2021).

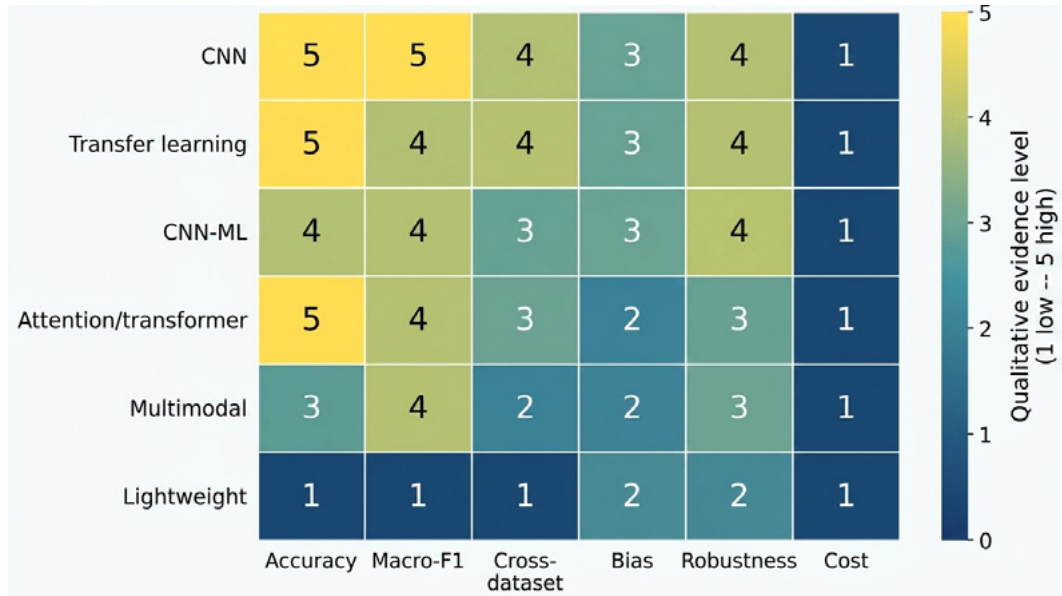
Attention mechanisms and transformers respond to a further limitation of CNNs: relevant facial regions vary across expressions and contexts. Attention can assign different weights to the eyes, mouth, or eyebrows, whilst transformers model relationships between patches, landmarks, or video frames. POSTER, POSTER++, and Swin Transformer models exemplify this development. Their advantage, however, must be assessed alongside parameter count, floating-point operations, latency, and cross-dataset performance, since additional complexity can impede deployment (Kim et al., 2022; Mao et al., 2025; Shirsath et al., 2022; Zheng et al., 2023).

Multimodality extends FER towards a more contextual account of emotion recognition. When a face is occluded, ambiguous, or subtly expressive, audio, text, body movement, or physiological signals may provide complementary information. Intermediate fusion can be represented as

$$z = \phi(W_f f_{face} + W_a f_{audio} + W_t f_{text} + b) \quad (3)$$

Where f_{face} , f_{audio} , and f_{text} are feature vectors, W denotes projection matrices, b is a bias term, and ϕ is a nonlinear activation. The equation makes clear that fusion is a learned joint representation rather than an indiscriminate addition of data. Even so, multimodality increases privacy risks, synchronisation requirements, and vulnerability to missing modalities (Karatay et al., 2022; Z. Lian et al., 2024; Sharafi et al., 2023; Wang et al., 2022).

Figure 2
Maturity of the evidence reported for each FER model family and evaluation dimension



Note. The figure is an ordinal synthesis of evidence-maturity based on the coded matrix, not a pooled effect size. The 1–5 values summarise the relative prevalence and completeness of evidence in each family (1 = sparse; 5 = consistently reported). The source-level coding is auditable in Table 3; Table 5 reports the corresponding rates for each empirical model family.

Table 4
FER model families and recommended minimum reporting

Family	Principal strength	Critical limitation	Minimum reporting
CNN	Mature and efficient spatial feature extraction.	Overfitting to dataset artefacts.	Backbone, augmentation, accuracy, macro-F1, and confusion matrix.
Transfer learning	Reduces dependence on large target datasets.	May transfer bias from pre-training data.	Source dataset, frozen layers, fine-tuning protocol, and groups evaluated.
Hybrid CNN-ML	Integrates deep and conventional features.	May increase complexity without improving generalisation.	Ablation, feature set, classifier, and external validation.
Attention and transformer	Weights relevant facial regions.	Greater cost and risk of over-parameterisation.	Parameters, Floating-Point Operations (FLOPs), saliency, and cross-dataset validation.
Multimodal	Adds context and resilience to ambiguity.	Privacy risks and missing modalities.	Fusion level, synchronisation, modality failures, and consent.
Lightweight	Enables mobile or edge inference.	May impair performance for minority classes.	Latency, memory, hardware, quantisation, and fairness.

The loss most commonly used for multiclass FER is categorical cross-entropy:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{y}_{ic}) \quad (4)$$

Where N is the number of samples, C the number of emotion categories, y_{ic} the one-hot-encoded reference label, and $\hat{y}_{ic}$ is the model's predicted probability. The formulation is appropriate when categories are mutually exclusive, but it may oversimplify emotional ambiguity. FER+ and label-distribution approaches are therefore especially relevant when several interpretations of an image are plausible (Barsoum et al., 2016; Goodfellow et al., 2013; Khairuddin & Chen, 2021; S. Li & Deng, 2022).

Aggregated accuracy should be supplemented by macro-F1, particularly when classes are imbalanced. For each class c , precision, recall, and F1 are defined as

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_cR_c}{P_c + R_c} \quad (5)$$

Macro-F1 is then $1/C \sum_{c=1}^C F1_c$. By assigning equal weight to each class, it reduces the risk that frequent categories, such as happiness or neutrality, will conceal weak performance for fear, disgust, or surprise. This distinction is essential when the object of evaluation is trustworthiness rather than a single overall average (Dominguez-Catena et al., 2024; Gaya-Morey et al., 2026; Georgiou et al., 2026; T. Lian & Celiktutan, 2025).

Table 5 deepens the quantitative synthesis without averaging incompatible accuracy values across different datasets and label spaces. It reports the percentage of empirical or foundational sources in each architecture family that contained at least partial evidence for each TEC domain. Because the lightweight category has only one foundational source in the selected matrix, its percentages must be interpreted descriptively rather than as a stable population estimate.

Table 5
Comparative evidence-reporting trends by model family

Family	n	P (%)	G (%)	F (%)	R (%)	X (%)	D (%)	Median TEC (%)
CNN	4	100	50	0	75	0	25	21
CNN-ML	4	100	0	0	25	25	25	13
Transfer learning	5	100	60	0	60	0	20	33
Attention and transformer	3	100	100	0	100	100	100	50
Multimodal	3	100	33	0	100	0	67	33
Lightweight	1	100	100	0	100	0	100	58

Generalisation, bias, and robustness

Generalisation is one of FER's most persistent weaknesses. A model may learn genuine facial features, but it may also exploit spurious properties of a dataset, including resolution, illumination, cropping, compression, or demographic composition. The cross-dataset generalisation gap can be expressed as

$$\Delta_{gen} = M(D_{train} \rightarrow D_{test}^{in}) - M(D_{train} \rightarrow D_{test}^{out}) \quad (6)$$

Where M is a measure such as macro-F1 or balanced accuracy, D_{test}^{in} belongs to the training domain, and D_{test}^{out} is an external dataset. A large Δ_{gen} indicates dependence on the original domain and weakens the case for deployment (Dominguez-Catena et al., 2022; Gaya-Morey et al., 2026; Georgiou et al., 2026; Mollahosseini et al., 2019).

FER datasets embody distinct forms of evidence. CK+ and JAFFE remain historically important because of their experimental control, but their size and diversity are limited. FER2013 and FER+ increased scale, although their images have low resolution and noisy labels. RAF-DB and AffectNet more closely resemble uncontrolled settings, yet they introduce web-sourced data bias, ambiguity, and class imbalance. Aff-Wild2 moves the task towards video and affect under naturalistic conditions. It follows that no single dataset can by itself establish generalisation (S. Li et al., 2017; Z. Lian et al., 2024; Lucey et al., 2010; Lyons et al., 1998).

Table 6
Relevant FER datasets and threats to external validity

Dataset	Type of evidence	Strength	Principal risk
CK+	Controlled sequences and action units.	Clean labels and prototypical expressions.	Limited environmental and demographic diversity.
JAFPE	Controlled static images.	Useful for historical comparison.	Small size and restricted population.
FER2013	Greyscale web images.	Widely used and readily reproducible benchmark.	Low resolution and label noise.
FER+	Relabelling through label distributions.	Represents annotation ambiguity.	Inherits the visual limitations of FER2013.
RAF-DB	Naturalistic images with basic and compound expressions.	Closer to uncontrolled settings.	Requires more complete demographic reporting.
AffectNet	Large-scale categories and affective dimensions.	Substantial visual variability.	Noise, web-source bias, and imbalance.
Aff-Wild2	Video, valence/arousal, and expressions.	Tests temporal information in naturalistic settings.	Costly protocols and sensitivity to preprocessing.

Demographic bias in FER may arise from unequal representation of age, sex, skin tone, ethnicity, or cultural context. It may also arise from differences in how annotators interpret expressions. A simple range statistic can describe group disparity:

$$\Delta_{bias} = \max_g M_g - \min_g M_g \quad (7)$$

Where M_g is the performance measure for group (g). A large Δ_{bias} shows that aggregated accuracy conceals unequal performance. The statistic does not resolve the ethical problem or establish a complete fairness criterion. Its more modest purpose is to require studies to disclose whether performance is comparable across relevant, ethically documented groups (Dominguez-Catena et al., 2022, 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022).

Robustness concerns stability under realistic changes: occlusion, blur, poor illumination, compression, noise, pose, face coverings, inaccurate crops, and temporal variation. One simple robustness ratio compares clean and perturbed performance:

$$R_{rob} = 1 - \frac{M_{clean} - M_{perturbed}}{M_{clean} + \epsilon} \quad (8)$$

Where ϵ prevents division by zero. Values close to one indicate little degradation. The ratio must nevertheless be reported with the perturbation type and severity; otherwise, apparently similar robustness scores may describe tests of very different practical significance (Gaya-Morey et al., 2026; Georgiou et al., 2026; Z. Lian et al., 2024; Shirsath et al., 2022).

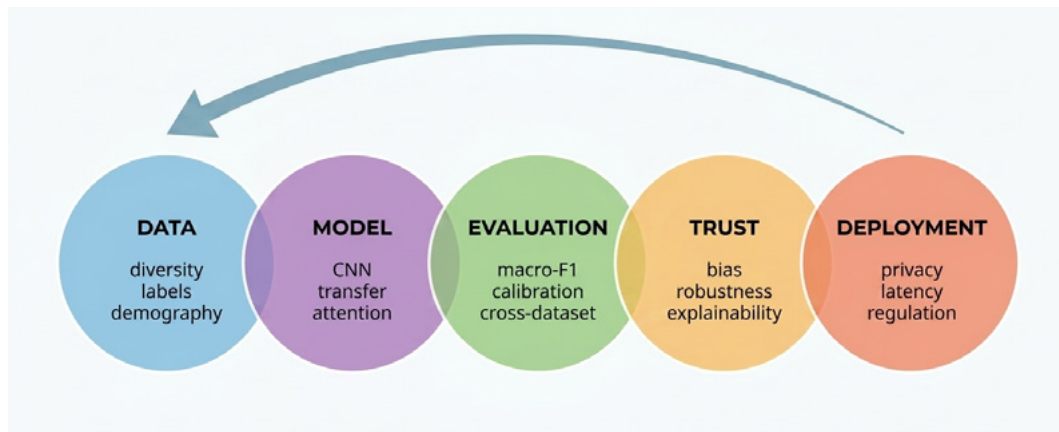
Explainability provides a diagnostic means of determining whether a model attends to plausible facial regions or to artefacts. Gradient-weighted class activation mapping (Grad-CAM), for example, estimates class-specific saliency from gradients with respect to activation maps. Such visualisations do not prove that a model ‘understands’ emotion. They can, however, reveal shortcuts: a happiness prediction driven by the background or a watermark would plainly undermine the model’s validity. Explanations should therefore accompany quantitative evaluation rather than replace it (Kim et al., 2022; Mao et al., 2025; Selvaraju et al., 2017; Zheng et al., 2023).

Explainability also needs an audit trail. A defensible FER audit should record dataset provenance and consent constraints; subgroup error and calibration; robustness tests and perturbation severity; explanation method and known failure modes; model or version changes; and deployment monitoring or incident feedback. Visual saliency is therefore one audit artefact rather than the audit itself. End-to-end algorithmic auditing frameworks reinforce the need to connect design choices, documentation, testing, governance, and post-deployment review (Raji et al., 2020).

Figure 3 synthesises the proposed trustworthiness framework. Its conceptual contribution is to connect data, model design, evaluation, trustworthiness, and deployment in a continuous chain. A failure observed after deployment should feed back into data curation, metrics, and validation protocols. Without that feedback, an FER system remains static even as its environment changes (European Parliament and Council of the European Union, 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022; Wang et al., 2022).

Figure 3

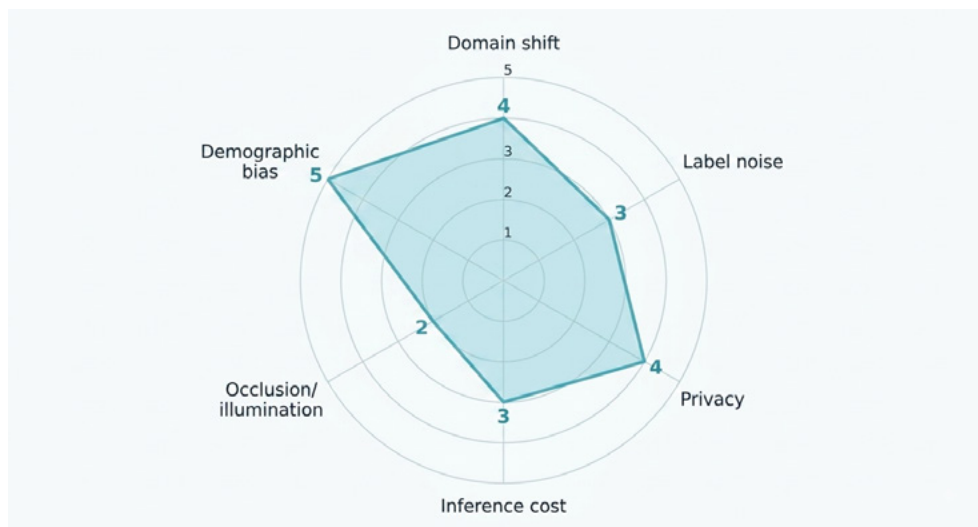
Proposed framework for evaluating FER beyond accuracy.



Note. The framework is merely conceptual: Table 3 applies its evidence domains source by source, Equation (1) provides a reproducible coverage measure, and Table 5 aggregates the resulting reporting patterns by model family.

Figure 4

Risks that reduce the external validity of FER models.



Note. The 1–5 values are an ordinal risk synthesis from the coded evidence and review discussion (1 = limited or less recurrent concern; 5 = recurrent or high-consequence concern). They are not probabilities or pooled effect sizes. The highest level assigned to demographic bias reflects its recurrence across dataset, fairness, and governance evidence; domain shift and privacy are also treated as high-impact external-validity risks.

Table 7
Minimum reporting matrix for trustworthy FER research.

Dimension	Required information	Control question
Data	Source, size, classes, distribution, preprocessing, and split.	Could another researcher reproduce the model input?
Model	Architecture, parameters, backbone, training, and data augmentations.	Is it clear what was trained and at what cost?
Metrics	Accuracy, macro-F1, confusion matrix, and class-wise measures.	Does overall performance conceal weak classes?
Generalisation	Cross-validation and cross-dataset validation were feasible.	Does the model work beyond its original domain?
Bias	Group-disaggregated performance where appropriately collected demographic data are available.	Is performance comparable across relevant groups?
Robustness	Perturbations, occlusion, illumination, compression, and pose.	Are predictions stable under realistic conditions?
Explainability	Grad-CAM, attention, or landmark analysis.	Does the model attend to plausible facial regions?
Deployment	Latency, memory, privacy, consent, and regulation.	Is the proposed use technically and ethically defensible?

Lightweight models are necessary for mobile applications, low-cost robotics, and privacy-sensitive local inference. MobileNet, quantisation, pruning, and distillation can reduce parameter counts and latency. Efficiency, however, is not an independent virtue. A lightweight model that disproportionately reduces performance for minority classes or underrepresented groups may be less trustworthy than a larger but more balanced model. Computational cost, fairness, and robustness must therefore be reported together (European Parliament and Council of the European Union, 2024; T. Lian & Celiktutan, 2025; Mao et al., 2025; Sandler et al., 2018).

Discussion

The strongest FER studies are no longer those that report accuracy alone. A marginal improvement on a benchmark may be scientifically less informative than carefully documented cross-dataset validation, group-level bias analysis, or a realistic perturbation test. CNNs and transfer learning remain strong baselines, whilst attention, transformers, and multimodality provide useful means of addressing ambiguity and context. Their value depends, however, on evaluation across diverse settings rather than internal data splits alone (Gaya-Morey et al., 2026; Georgiou et al., 2026; S. Li & Deng, 2022; Wang et al., 2022).

The quantitative pattern in Table 5 clarifies the imbalance. Performance evidence is nearly universal in the selected architecture papers, whereas explicit fairness reporting is absent from the subset architecture studies and is instead concentrated in dedicated bias studies. Attention and transformer papers in the matrix show broader coverage of generalisation, robustness, explainability, and deployment evidence, but the family contains only three selected sources; the finding should therefore be read as a reporting trend, not a claim that transformers are intrinsically more trustworthy. This is precisely why TEC is retained as an evidence-coverage descriptor rather than a league-table score.

An objection must nevertheless be considered. Broader evaluation is costly: external datasets may use incompatible labels, demographic attributes may be unavailable or inappropriate to infer, and robustness or latency tests may demand resources that smaller teams do not possess. It would therefore be misleading to treat every missing measure as evidence of a defective

model. Yet this practical constraint does not justify silent omission. When an evaluation cannot be conducted, the limitation should be stated, and claims should be narrowed accordingly. Methodological restraint is itself part of trustworthy reporting.

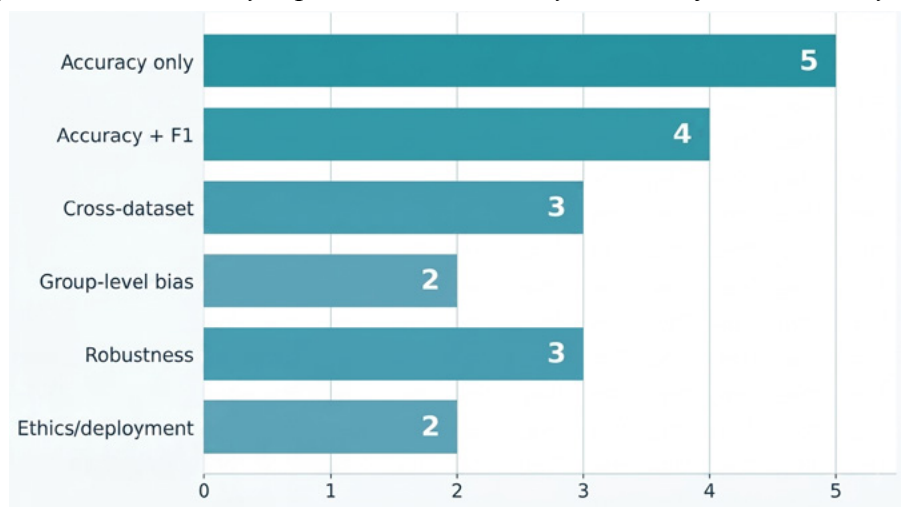
The synthesis that follows is thus not a demand for a single universal score but for evidence proportionate to the claim. The guiding question should change from “Which model has the highest accuracy?” to “What evidence warrants confidence in this model when the data, population, and context change?” This shift turns a descriptive comparison into a critical evaluation. The proposed framework does not claim a new state of the art; it specifies the conditions under which comparisons between FER systems become scientifically defensible (Barrett et al., 2019; Dominguez-Catena et al., 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022).

This restraint is particularly important because the validity of the target construct is contested. Barrett et al. (2019) show that the same facial movement can occur in different emotional contexts and that an emotion category can be expressed through heterogeneous facial configurations. Consequently, high classification accuracy against a dataset label establishes agreement with that annotation scheme under that protocol; it does not establish direct access to a person’s internal emotional state. Context, temporal information, culture, elicitation procedure, annotator uncertainty, and alternative non-facial signals should therefore delimit the language used in claims. In high-stakes settings, FER should not be treated as a standalone emotion detector without independent contextual validation.

The same reasoning applies to practical use. FER may support adaptive interfaces, personalised education, social robotics, digital assistance, and well-being applications when consent, transparency, and proportionality are established. The same technology can also enable emotion surveillance, workplace monitoring, or sensitive classification without adequate context. The European Union’s Artificial Intelligence Act places significant restrictions on emotion-recognition systems, particularly in education and employment, subject to limited medical or safety exceptions. Ethics and regulation must consequently form part of system design and evaluation rather than an ornamental final section (European Parliament and Council of the European Union, 2024; Z. Lian et al., 2024; Mohammad, 2022; Wang et al., 2022).

Figure 5

Gap between commonly reported and necessary evidence for trustworthy FER.



Note. The figure uses the same 1–5 ordinal synthesis logic as Figure 2: higher values denote evidence types that are more routinely reported in the reviewed literature. The source-level matrix in Table 3 supplies the audit trail; the bars are an interpretive scale of reporting frequency, not raw study counts.

Figure 5 illustrates a recurrent asymmetry: accuracy is reported far more often than robustness, fairness, or deployment evidence. This imbalance explains why some studies appear strong in their results yet weak in external validity. The minimum reporting matrix offers a remedy by aligning each type of claim with the evidence needed to support it (Dominguez-Catena et al., 2022; Gaya-Morey et al., 2026; Georgiou et al., 2026; Page et al., 2021).

Table 8
Practical implications for the evaluation and reporting of FER studies.

Aspect	Risk if omitted	Recommended practice
Citations	Claims rely on weak or untraceable evidence.	Distinguish peer-reviewed studies, foundational work, and technical preprints.
Comparability	Scores from incompatible protocols appear equivalent.	Report splits, preprocessing, label mapping, and uncertainty.
Method	The review or experiment cannot be reproduced.	State sources, time window, criteria, coding, and validation protocol.
Contribution	Architectural novelty is confused with stronger evidence.	Relate each claimed advance to generalisation, robustness, or fairness.
Ethics	Technical performance obscures potential harm.	Address privacy, consent, bias, intended use, and applicable regulation.
Resources	Efficiency claims lack an operational basis.	Report latency, memory, hardware, parameters, and energy were feasible.

This review has four principal limitations. First, the search may have missed recent unindexed studies or literature behind subscription barriers. Second, several preprints provide timely evidence but do not carry the same epistemic weight as peer-reviewed articles. Third, comparisons across datasets remain imperfect because protocols and label spaces vary. Finally, many datasets lack complete and ethically collected demographic documentation, which restricts bias analysis. These limitations do not invalidate the synthesis, but they delimit its claims and reinforce the need for transparent reporting (Dominguez-Catena et al., 2024; Gaya-Morey et al., 2025; Georgiou et al., 2026; Mohammad, 2022).

Two additional limitations follow from the analysis. First, TEC measures reporting coverage, not correctness, causal validity, or deployment safety; a paper can report all six domains and still reach an invalid conclusion. Secondly, quality and TEC coding necessarily compress heterogeneous evidence into ordinal categories. The full source-level matrix is therefore provided so that readers can inspect the coding rather than rely on the aggregate index alone. Future updates should use preregistered coding rules and, where multiple coders are available, report inter-rater agreement.

Conclusions

This review has shown that recent FER models have advanced substantially in benchmark performance, yet a consequential gap remains between accuracy and real-world trustworthiness. CNNs, transfer learning, hybrid models, attention, transformers, multimodality, and lightweight architectures each provide useful capabilities; no family, however, guarantees generalisation, fairness, or robustness. Such properties must be demonstrated through complementary metrics, cross-dataset validation, group-disaggregated analysis, perturbation tests, explanations, and explicit deployment conditions (Georgiou et al., 2026; S. Li & Deng, 2022; T. Lian & Celiktutan, 2025; Wang et al., 2022).

The proposed trustworthy-FER framework answers the research question by connecting methodological structure, evaluation measures, and reporting requirements. Its central claim is deliberately restrained: aggregated accuracy remains informative, but it cannot bear the full weight of a deployment claim. Trust depends on the relationship between data, model, evaluation, human context, and operational constraints (Dominguez-Catena et al., 2022, 2024; European Parliament and Council of the European Union, 2024; Mohammad, 2022).

The operationalisation makes this claim testable at the level of evidence reporting. TEC quantifies whether a study documents performance, generalisation, fairness, robustness, explainability, and deployment evidence, while the quality appraisal distinguishes complete reporting from weak methodological description. Neither quantity should be interpreted as a universal trustworthiness certificate. Their purpose is to make missing evidence visible and to align the strength of a claim with the evidence supplied.

Future work should therefore pursue six priorities: more diverse, documented, and ethically collected datasets; routine cross-dataset evaluation; macro-F1, calibration, and group-disaggregated reporting; robustness tests based on realistic perturbations; explainability and traceability in human-facing applications; and validation of lightweight models without sacrificing fairness. Progress on these fronts would move the field from systems that recognise expressions within benchmarks towards affective computing systems whose limitations are measurable, transparent, and responsibly governed (Gaya-Morey et al., 2025; Gaya-Morey et al., 2026; Z. Lian et al., 2024; Page et al., 2021).

Acknowledgements and Declarations

Disclosure of Interests

The authors have no competing interests to declare.

Author Contributions

The authors declare equal contribution and sharing of authorship roles for this publication.

AI Usage Statement

Generative artificial intelligence tools were used to support language editing, outline refinement, and bibliographic normalisation. The authors remain responsible for verifying all screening decisions, interpretations, claims, and final wording, and assume full responsibility for the manuscript's content.

References

- Agrawal, I., Kumar, A., Swathi, D., Yashwanthi, V., & Hegde, R. (2021). Emotion recognition from facial expression using CNN. *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, 1–6. <https://doi.org/10.1109/R10-HTC53172.2021.9641578>
- Bahmanabady, E., Imani, M., & Ghassemian, H. (2023). Fusion of multi-level CNN with LBP features for facial emotion recognition. *2023 31st International Conference on Electrical Engineering (ICEE)*, 463–467. <https://doi.org/10.1109/ICEE59167.2023.10334654>
- Banoth, S., Kukreja, V., Thapliyal, N., Aeri, M., & Sharma, R. (2024). Beyond basic emotions: High-accuracy facial expression classification using a CNN-SVM hybrid model. *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 1–6. <https://doi.org/10.1109/ICETCS61022.2024.10544213>

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68. <https://doi.org/10.1177/1529100619832930>
- Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 279–283. <https://doi.org/10.1145/2993148.2993165>
- Dominguez-Catena, I., Paternain, D., & Galar, M. (2022). *Assessing demographic bias transfer from dataset to model: A case study in facial expression recognition*. <https://doi.org/10.48550/arXiv.2205.10049>
- Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5209–5226. <https://doi.org/10.1109/TPAMI.2024.3361979>
- European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act)*. Official Journal of the European Union, L 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Gaya-Morey, F. X., Buades-Rubio, J. M., Palanque, P., Lacuesta, R., & Manresa-Yee, C. (2025). *Deep learning-based facial expression recognition for the elderly: A systematic review*. <https://doi.org/10.48550/arXiv.2502.02618>
- Gaya-Morey, F. X., Manresa-Yee, C., Martinie, C., & Buades-Rubio, J. M. (2026). Evaluating facial expression recognition datasets for deep learning: A benchmark study with novel similarity metrics. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/TAFFC.2026.3693316>
- Georgiou, S., Psiris, A., Evangelatos, S., Lagkas, T., Argyriou, V., Sarigiannidis, P., Varlamis, I., & Papadopoulos, G. Th. (2026). *Facial expression recognition in the deep learning era: A systematic multi-criteria review of methods, models, datasets, performance, challenges, and future research directions*. <https://doi.org/10.48550/arXiv.2606.08612>
- Goodfellow, I. J., Erhan, D., Carrier, P.-L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., & others. (2013). *Challenges in representation learning: A report on three machine learning contests*. arXiv. <https://doi.org/10.48550/arXiv.1307.0414>
- Gupta, V., & Sejwar, V. (2022). Facial expression recognition with combination of geometric and textural domain features extractor using CNN and machine learning. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 400–407. <https://doi.org/10.1109/ICAISS55157.2022.10010853>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Karatay, B., Bestepe, D., Sailunaz, K., Ozyer, T., & Alhajj, R. (2022). A multimodal emotion recognition system based on CNN-transformer deep learning technique. *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, 145–150. <https://doi.org/10.1109/CDMA54072.2022.00029>
- Khairuddin, Y., & Chen, Z. (2021). *Facial emotion recognition: State of the art performance on FER2013*. <https://doi.org/10.48550/arXiv.2105.03588>
- Khajuria, O., Kumar, R., & Gupta, M. (2023). Facial emotion recognition using CNN and VGG-16. *2023 International Conference on Inventive Computation Technologies (ICICT)*, 472–477. <https://doi.org/10.1109/ICICT57646.2023.10133972>
- Kim, J.-H., Kim, N., & Won, C. S. (2022). *Facial expression recognition with swin transformer*. <https://doi.org/10.48550/arXiv.2203.13472>

- Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
- Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593. <https://doi.org/10.1109/CVPR.2017.277>
- Li, T.-H. S., Kuo, P.-H., Tsai, T.-N., & Luan, P.-C. (2019). CNN and LSTM based facial expression analysis model for a humanoid robot. *IEEE Access*, 7, 93998–94011. <https://doi.org/10.1109/ACCESS.2019.2928364>
- Lian, T., & Celiktutan, O. (2025). *A feature-level bias evaluation framework for facial expression recognition models*. <https://doi.org/10.48550/arXiv.2505.20512>
- Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X., Yi, J., Liu, R., Xu, K., Liu, B., Cambria, E., Zhao, G., Schuller, B. W., & Tao, J. (2024). MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 41–48. <https://doi.org/10.1145/3689092.3689959>
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. <https://doi.org/10.1109/CVPRW.2010.5543262>
- Lyons, M. J., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 200–205. <https://doi.org/10.1109/AFGR.1998.670949>
- Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., & Huang, A. (2025). POSTER++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157, 110951. <https://doi.org/10.1016/j.patcog.2024.110951>
- Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2), 239–278. https://doi.org/10.1162/coli_a_00433
- Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18–31. <https://doi.org/10.1109/TAFFC.2017.2740923>
- Negi, A. S., Arora, A., Bisht, S., Devliyal, S., Kumar, B. V., & Kaur, G. (2024). Facial emotion detection using CNN and VGG16 model. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–5. <https://doi.org/10.1109/I2CT61223.2024.10543618>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Poulose, A., Kim, J.-H., & Han, D. S. (2021). Feature vector extraction technique for facial emotion recognition using facial landmarks. *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 1072–1076. <https://doi.org/10.1109/ICTC52510.2021.9620798>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
- Rawat, U., & Rai, C. S. (2023). Improving facial emotion recognition through transfer learning with deep convolutional neural network models. *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 1335–1339. <https://doi.org/10.1109/ICTACS59847.2023.10390186>

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- Sharafi, M., Yazdchi, M., & Rasti, J. (2023). Audio-visual emotion recognition using k-means clustering and spatio-temporal CNN. *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 1–6. <https://doi.org/10.1109/IPRIA59240.2023.10147192>
- Shirsath, S., Vikhe, V., & Vikhe, P. (2022). Xception CNN-ensemble learning based facial emotion recognition. *2022 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–4. <https://doi.org/10.1109/ICCUBEA54992.2022.10011111>
- Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83–84, 19–52. <https://doi.org/10.1016/j.inffus.2022.03.009>
- Zheng, C., Mendieta, M., & Chen, C. (2023). POSTER: A pyramid cross-fusion transformer network for facial expression recognition. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3138–3147. <https://doi.org/10.1109/ICCVW60793.2023.00339>

An academic chatbot with hybrid retrieval and local large language models

Chatbot académico con recuperación híbrida y modelos de lenguaje locales

Jhon Paúl Guadalupe Andrade¹ <https://orcid.org/0009-0009-2656-3219>, Blanca Topón Visarrea² <https://orcid.org/0000-0001-5307-5450>, Christian Iza¹ <https://orcid.org/0000-0002-0392-8766>

¹Universidad Indoamérica, Ingeniería en Tecnologías de la Información,
Quito, Ecuador
jguadalupe6@indoamerica.edu.ec, christianiza@uti.edu.ec

²Universidad Indoamérica, Ingeniería Industrial, Quito, Ecuador
blancatopon@uti.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/07/12

Accepted: 2026/08/21

Published: 2026/10/14

Abstract

Higher-education institutions devote a substantial share of their staff time to answering queries whose answers are already set out in institutional regulations. This paper presents the design and evaluation of an academic chatbot based on retrieval-augmented generation that lets users query those documents in natural language and receive grounded answers in Spanish. The system combines a structure-aware ingestion pipeline, a hybrid retriever that fuses dense semantic search and lexical search through reciprocal rank fusion, and a locally hosted large language model. A distinguishing goal is data sovereignty: every component runs on institutional hardware, so no data is sent to external services. Quality is evaluated with the RAGAS framework over five regulatory documents: the corpus-wide set attains a global score of 0.90, with faithfulness of 0.97. The acceptance criterion is not met, however, under its conjunctive reading: answer relevancy stays below its threshold across all sets, and p95 latency reaches 6.47 seconds against a 5-second target. Quality further degrades on long, heavily cross-referenced regulations. A contrast questionnaire administered to nineteen users places the perception index at 3.64 out of 5 and exposes a relevant discrepancy: only 57.9% agree that the system does not produce fabricated information, against the faithfulness the automatic judge assigns it.

Summary: Introduction, Materials and Methods, Results and Discussion, Conclusions.

How to cite: Guadalupe, J., Topón, B. & Iza, C. (2026). An academic chatbot with hybrid retrieval and local large language models. *Revista Tecnológica - Espol*, 38(E1), 50-70. <https://doi.org/10.37815/rte.v38nE1.1595>

Keywords: retrieval-augmented generation, higher education, artificial intelligence, natural language processing, information privacy, automatic evaluation.

Resumen

Las instituciones de educación superior dedican una parte considerable del tiempo de su personal a responder consultas cuya respuesta ya consta en la normativa institucional. Este trabajo presenta el diseño y la evaluación de un *chatbot* académico basado en generación aumentada por recuperación que permite consultar dichos documentos en lenguaje natural y obtener respuestas fundamentadas en español. El sistema combina un pipeline de ingesta consciente de la estructura documental, un recuperador híbrido que fusiona la búsqueda semántica densa y la léxica mediante la fusión de rango recíproco y un modelo de lenguaje alojado localmente. Un objetivo distintivo es la soberanía de datos: todos los componentes se ejecutan en el hardware institucional, de modo que ningún dato se envía a servicios externos. La calidad se evalúa con el marco RAGAS sobre cinco documentos normativos: el conjunto en el nivel del corpus alcanza una puntuación global de 0,90, con fidelidad de 0,97. No obstante, el criterio de aceptación no se satisface en su lectura conjuntiva: la relevancia de la respuesta permanece bajo su umbral en todos los conjuntos y el percentil 95 de la latencia alcanza 6,47 segundos frente al objetivo de 5. La calidad se degrada, además, en reglamentos extensos con referencias cruzadas. Un cuestionario de contraste con diecinueve usuarios sitúa el índice de percepción en 3,64 sobre 5 y expone una discrepancia relevante: solo el 57,9 % coincide en que el sistema no produce información inventada, frente a la fidelidad que le atribuye el juez automático.

Palabras clave: generación aumentada por recuperación, educación superior, inteligencia artificial, procesamiento del lenguaje natural, privacidad de la información, evaluación automática.

Introduction

Student support services at higher-education institutions absorb a recurring and largely avoidable workload: a high proportion of incoming queries concern information already codified in academic bylaws, enrollment calendars, disciplinary rules, and graduation manuals, lengthy documents whose consultation is mediated by human staff. Answers are therefore structurally constrained by office hours and by the order in which requests arrive. Measurements in comparable settings report a mean latency of roughly 8.5 hours for the human channel against a few seconds for an automated one (Odede & Frommholz, 2024). That difference, of approximately three orders of magnitude, comes from an institutional context distinct from the one studied here and is invoked solely as an indication of the magnitude of the problem that motivates automation.

Conversational interfaces built on large language models (LLMs) are natural candidates for absorbing this demand. An off-the-shelf LLM used in isolation, however, has no knowledge of a particular institution's regulations and tends to hallucinate plausible but incorrect statements. Fine-tuning on institutional data is costly and must be repeated whenever the regulations change; nor does it eliminate hallucination. retrieval-augmented generation (RAG), proposed by Lewis et al. (2020), addresses the problem by decoupling knowledge from the generative model: at query time, relevant chunks are retrieved from a verified knowledge base and injected into the prompt so that answers remain anchored to auditable institutional sources.

A second constraint is data sovereignty. Institutional documents may contain sensitive internal procedures, and Ecuador's Organic Law on Personal Data Protection (Asamblea Nacional del Ecuador, 2021) imposes obligations on the processing of personal data. Commercial LLM APIs route every query and every document chunk to third-party infrastructure, which is undesirable in this context and additionally introduces a recurring cost per token processed. This work therefore adopts a fully local deployment.

The objective of this work is to design, deploy, and evaluate a conversational assistant that answers queries about institutional academic regulations with grounded and traceable answers, under the constraint that no data leaves the institution's infrastructure. The paper presents that assistant over the document base of Universidad Indoamérica. Its contributions are fourfold: a privacy-preserving RAG architecture in which retrieval, vectorization, and generation run entirely on institutional hardware; a structure-aware ingestion pipeline that converts PDFs into structured Markdown and chunks them along the document hierarchy, preserving the natural semantic unit of regulatory text, namely the article or section; a hybrid retrieval strategy that fuses dense and lexical search through reciprocal rank fusion, recovering both conceptual matches and exact institutional terms; and an evaluation under the design science paradigm centered on automatic answer-quality metrics, together with verification of the data-sovereignty property and of functional correctness, accompanied by a public and reproducible implementation.

These contributions are organized around three research questions. RQ1 asks whether a fully local RAG pipeline—with embeddings, a vector store, and generation on the premises, with no recourse to any external API—can answer institutional regulatory questions at a quality that meets predefined acceptance thresholds. RQ2 asks whether hybrid retrieval, the strategy that fuses dense and lexical signals through reciprocal rank fusion, improves retrieval and answer quality over dense-only and lexical-only retrieval when queries are posed over institutional regulatory documents written in Spanish, that is, over bylaws, procedural instructions, and manuals such as those in the corpus described below. RQ3 examines how answer quality varies with the length and structure of the source document, from compact procedural instructions to long bylaws with abundant cross-references.

Related Work

Since its introduction by Lewis et al. (2020), RAG has become the dominant pattern for grounding LLM answers in domain corpora, with rapid uptake in education (Swacha & Gracel, 2025). Several systems address the university domain: JayBot (Odede & Frommholz, 2024), whose latency gap between the human and the automated channel frames this motivation; BARKPlug (Neupane et al., 2024); and the Unimib Assistant (Antico et al., 2024). A RAG assistant for a large Brazilian university (Kuratomi et al., 2025) reports quality figures used here as an external reference point. In the Ibero-American setting, several Spanish-language efforts apply natural language processing and LLM/RAG techniques (Manjarrés-Betancur & Echeverri-Torres, 2020; Espinosa-Luna et al., 2023; Guerra et al., 2025; Coronel Crespo & Ulloa Bernal, 2025), which confirms both the regional relevance of the problem and the comparative scarcity of fully local deployments.

As regards hybrid retrieval, dense retrieval captures semantic similarity but may miss exact lexical matches that are decisive in regulatory text, for instance, "Art. 47" or institutional acronyms. BM25 (Robertson & Zaragoza, 2009) remains the standard lexical baseline and is complementary to dense methods. Combining the rankings of heterogeneous retrievers is not trivial, since their scores inhabit different scales; reciprocal rank fusion (Cormack et al., 2009) sidesteps normalization by fusing rank positions alone and is the strategy adopted here. Hybrid

RAG has been shown to improve accuracy in university admission chatbots (Nguyen & Quan, 2025) and in industrial customer service (Xu et al., 2024). Practical reports on building RAG systems from PDF corpora (Khan et al., 2024) further stress that document parsing and chunking are decisive for final quality, which motivates the use of a layout-aware converter (Livathinos et al., 2025) and of header-based chunking rather than fixed-window splitting.

Relative to prior work, the present system is distinguished by combining three properties in a single deployable system: fully local operation, with local embeddings and a locally hosted LLM, in the service of data sovereignty; structure-aware ingestion plus hybrid retrieval tuned to Spanish regulatory text; and an evaluation centered on automatic answer-quality metrics, with verification of data sovereignty and of functional correctness, under an explicit design science protocol and with an open-source release. Table 1 positions the system against representative university RAG assistants.

Table 1
Positioning Against Representative University RAG Assistants

SYSTEM	LANGUAGE	RETRIEVAL	LOCAL (NO EXT. API)	EVALUATION
JayBot (Odede & Frommholz, 2024)	English	Dense	No (ChatGPT)	User study
BARKPlug v2 (Neupane et al., 2024)	English	Dense	No (GPT-3.5)	RAGAS, SUS
Kuratomi et al. (2025)	Portuguese	Dense	n/r	Top-k accuracy
URAG (Nguyen & Quan, 2025)	Vietnamese	Dense (semantic)	Yes (local)	Accuracy
This work	Spanish	Hybrid (RRF)	Yes (local)	RAGAS + ablation

Note. Attributes as reported by each work; n/r: not reported. This work is distinguished by combining hybrid retrieval with a fully local deployment, free of external APIs.

Materials and Methods

Research approach

The study is applied research with a mixed-methods design in which the quantitative component predominates validation rests on the four answer-quality metrics of the RAGAS framework, compared against predefined acceptance thresholds, and is complemented by qualitative inspection of representative answers and their cited sources.

The work follows the design science research (DSR) paradigm formulated by Hevner et al. (2004), aimed at the construction and evaluation of technological artifacts that solve real-world problems. In the terminology of this paradigm, "artifact" denotes the system built and evaluated; the term is used hereafter only where the statement refers to the DSR protocol. DSR is operationalized in six sequential phases with feedback between them: problem identification, definition of objectives with verifiable acceptance criteria, design and development, demonstration through deployment on institutional infrastructure with real documents, evaluation, and communication.

Document Corpus

The knowledge base gathers the institutional documents responsible for the bulk of recurring student questions —admissions and recognition of prior study, academic rules, library services, and quality assurance—, each ingested into the vector store with the 1024-dimensional BGE-M3 embedding model (Chen et al., 2024). Table 2 lists the five documents of the evaluated corpus, ranging from short procedural instructions to long bylaws with cross-references.

This heterogeneity is deliberately preserved, since it allows the behavior of retrieval to be characterized on both compact and dense regulatory text. The same pipeline accepts additional documents without modification.

Table 2
Documents in the Evaluated Institutional Corpus and Their Ingestion Footprint

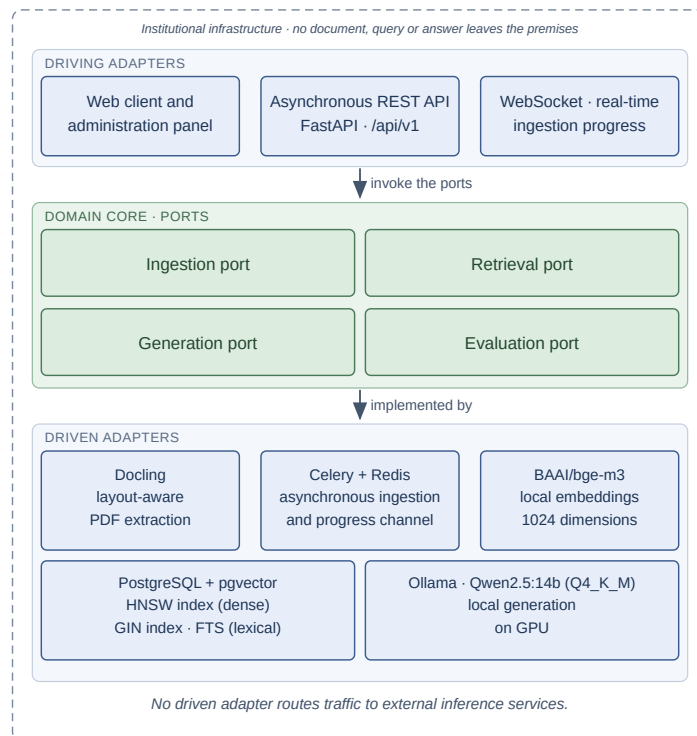
DOCUMENT	PAGES	CHUNKS
Instructions for appointing teaching and research assistants	8	9
Instructions for credit recognition through advanced study	10	19
Bylaws on credit recognition and transfer	26	42
Library collection management manual	14	19
Institutional quality assurance policy	6	17

Note. Retrievable chunks after structure-aware chunking.

System architecture

The system realizes the RAG pattern as a set of containerized microservices communicating over an internal network (Figure 1). A reverse proxy terminates TLS and applies rate limiting; an asynchronous API orchestrates retrieval and generation and exposes REST and WebSocket endpoints; a task worker runs document ingestion in the background; an in-memory broker manages the task queue and a publish/subscribe channel for real-time progress; a local LLM server performs inference on a GPU; and a PostgreSQL instance with the pgvector extension stores document metadata, chunk embeddings, and their lexical representation. Critically, both the embedding model and the LLM run on institutional hardware, so neither documents nor queries leave the premises.

Figure 1
Containerized Microservice Architecture



Note. Retrieval, vectorization, and generation run on institutional hardware; no data is sent to external APIs.

The reference deployment runs on a single server with a 16 GB data-center GPU (NVIDIA Tesla V100), a 28-core, 56-thread processor, and 125 GB of RAM under Linux. The 16 GB of graphics memory simultaneously host the quantized 14-billion-parameter LLM and the embedding model, which makes a fully local deployment feasible on conventional institutional hardware.

The data model underpinning the architecture is deliberately compact (Figure 2): a documents table records the metadata and processing state of each PDF, and a chunks table stores, for each retrievable unit, its textual content, the metadata of its source headers, its embedding as a pgvector vector (1024) column —over which the HNSW index is built— and its tsvector representation for lexical search, indexed with GIN. This one-to-many relation between document and its chunks is what enables traceability: every retrieved chunk knows the document and the section it comes from.

Figure 2
Data Model of the Document and Vector Store



Note. The embedding column uses the pgvector (1024) type with an HNSW index; the fts column uses the tsvector type with a GIN index.

Technology stack and deployment

So that the system is replicable by other institutions, Table 3 details the full technology stack with the versions used. Every component is open source or free to use, which removes licensing as a barrier to adoption. The main language is Python 3.11. The generative model is served through Ollama in Q4_K_M quantization, which reduces the graphics memory required by the 14-billion-parameter model from roughly 28 GB at 16-bit precision to about 9 GB, allowing it to coexist with the embedding model on a single 16 GB GPU.

Table 3
Technology Stack of the System

LAYER	TECHNOLOGY (VERSION)	ROLE
API	FastAPI 0.111 + Uvicorn 0.30	Asynchronous web API (REST and WebSocket)
Tasks	Celery 5.4	Asynchronous execution of ingestion
Broker	Redis 7	Task queue and progress pub/sub
Generation	Qwen2.5:14b (Q4_K_M) on Ollama	Local 14B-parameter LLM
Embeddings	BAAI/bge-m3 (1024 dim.)	Local multilingual vectorization
Database	PostgreSQL 15 + pgvector 0.7	Metadata and vector search (HNSW)
Lexical retr.	PostgreSQL FTS (Spanish dictionary) + GIN index	Lexical signal over tsvector, fused by RRF
PDF processing	Docling 2.14 + PyMuPDF 1.24	Structural conversion and extraction
Chunking	LangChain 0.2.6	Splitting by header hierarchy
Reverse proxy	Nginx 1.25	TLS, rate limiting, WebSocket
Orchestration	Docker 26 + Compose 2.27	Containers and internal network
Authentication	JWT (Supabase Auth)	Administrator access

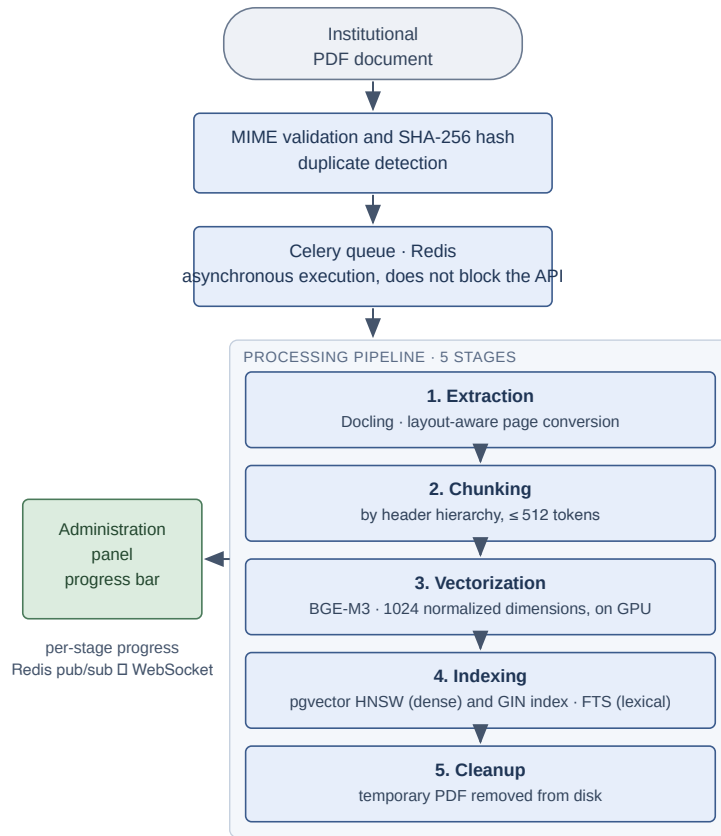
Note. Every component is open source or free to use; the main language is Python 3.11. The lexical signal is resolved with PostgreSQL's native full-text search engine and not with a BM25 implementation.

The complete deployment is defined in a single Docker Compose file: each service runs in its own container with isolated dependencies, containers communicate by service name over an internal network without exposing ports externally, persistent volumes hold the language model and the broker data, and GPU access is granted bypass-through. Replicating the system at another institution therefore reduces cloning the public repository, defining the environment variables, bringing the services up on a machine with an NVIDIA GPU, and loading the institution's own documents from the administration panel.

Ingestion pipeline

Ingestion transforms an institutional PDF into vectorized chunks stored in pgvector; it runs asynchronously so as not to block the API (Figure 3). First, structure-aware conversion: a layout-preserving document converter (Livathinos et al., 2025) detects headers, lists, and tables and exports structured Markdown, preserving the logical hierarchy of the bylaw, its chapters, and its articles. Second, header-based chunking: the Markdown is split along its header hierarchy so that each chunk constitutes a natural semantic unit inheriting the metadata of its headers for traceability. Unlike fixed-window splitting, this procedure does not cut across articles and does not lose coherence. Third, vectorization: each chunk is encoded into a 1024-dimensional vector with the multilingual BGE-M3 model (Chen et al., 2024) in batches on the GPU, with no per-call cost since it runs locally. Finally, indexing: chunks are inserted in batches, where an HNSW index (Malkov & Yashunin, 2020) enables approximate nearest-neighbor search in logarithmic time and a GIN index over the tsvector column enables lexical search. Each stage publishes its progress on a channel that a WebSocket relay to the administrator's browser.

Figure 3
Document Ingestion Pipeline



Ingestion is asynchronous and offline: it does not affect query latency.

Note. Extraction, structure-aware conversion, header-based chunking, vectorization, and indexing in pgvector.

Hybrid retrieval and generation

At query time, the system runs two retrievers in parallel and fuses their results (Figure 4). In dense retrieval, the query is vectorized with the same BGE-M3 model used at ingestion—which guarantees a shared vector space—and a pgvector call returns the top chunks by cosine similarity above a minimum threshold.

In parallel, lexical retrieval is resolved with PostgreSQL's native full-text search engine: the content of each chunk is stored as a tsvector under the Spanish dictionary—with stemming and stop-word removal—and indexed with GIN, so that the query, converted to a tsquery, recovers exact technical terms that a purely semantic match might miss. The full-text engine is used rather than a strict BM25 implementation (Robertson & Zaragoza, 2009) because it concentrates transactional, vector, and lexical storage in a single system and allows fusion to be executed entirely inside the database; this choice is revisited among the limitations.

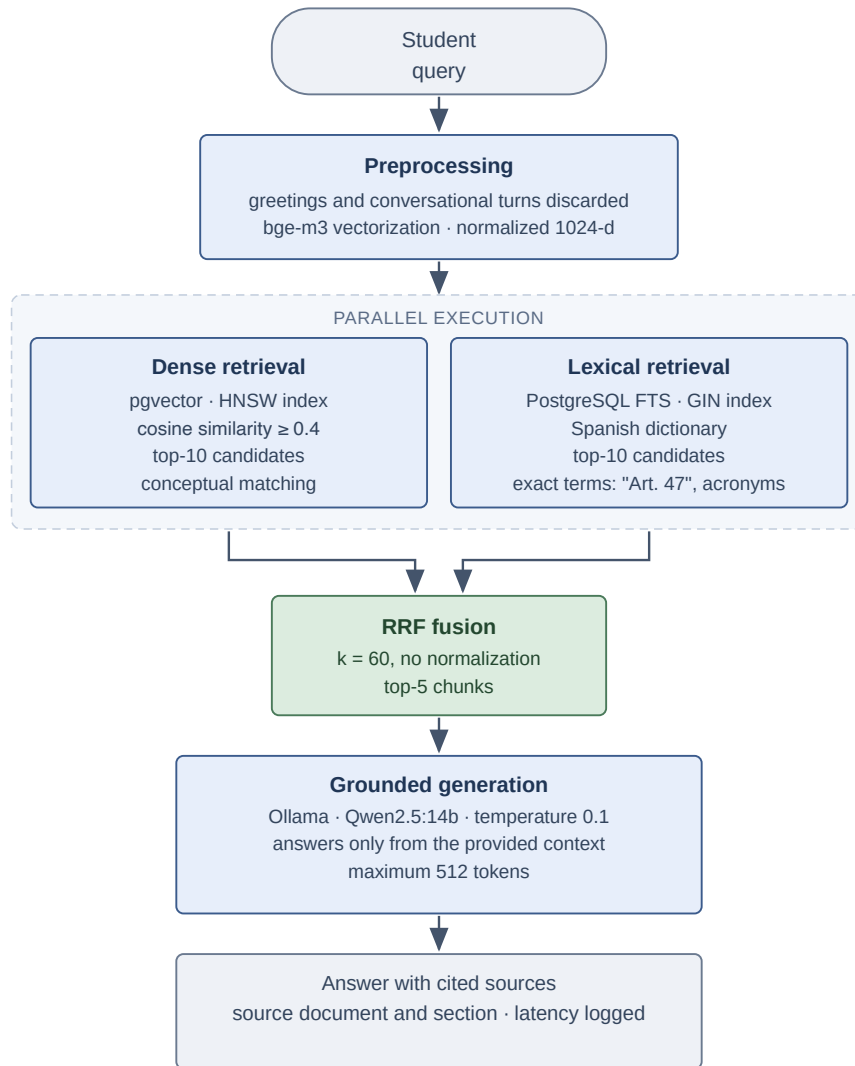
The two rankings are combined through reciprocal rank fusion (RRF) as defined by Cormack et al. (2009) in equation (1):

$$\text{scoreRRF}(c) = \sum_{r \in R} 1 / (k + \text{rank}_r(c)) \quad (1)$$

Where R is the set of retrievers, $\text{rank}_r(c)$ is the position of chunk c in retriever r , and $k = 60$ is the standard RRF constant. The fusion requires no score normalization (it operates on rank positions alone), which makes it robust and free of additional hyperparameters. The highest-ranked chunks after fusion pass to the generation stage.

For generation, an instruction-tuned, locally hosted LLM (Qwen Team, 2025) produces the answer from a prompt that injects the retrieved chunks and instructs the model to answer only from the provided context to state when information is unavailable, and to reply in Spanish. Decoding uses a low temperature to favor deterministic answers, and the answer is returned with its sources (the document and section of each chunk), which gives the student full traceability.

Figure 4
Query Pipeline



Retrieval and generation run on institutional hardware, with no calls to external APIs.

Note. Dense and lexical retrieval in parallel, reciprocal rank fusion, and grounded generation with a local LLM. Answers are returned with their cited sources.

Configuration parameters

Table 4 consolidates the operating parameters of the pipeline with their rationale, information usually omitted in literature and indispensable for reproducing the results. The values were tuned empirically during the DSR demonstration phase and remained fixed throughout the evaluation.

Table 4
Operating Parameters of the RAG Pipeline

PARAMETER	VALUE	RATIONALE
Maximum chunk size	≈ 512 tokens	Budget preserving complete semantic units
Minimum chunk length	50 characters	Discards chunks without useful content
Embedding dimension	1024	Native dimension of BGE-M3
HNSW index	m = 16, ef_construction = 64	Balance between speed and search quality
Candidates per retriever	top-10	Coverage prior to fusion
Cosine similarity threshold	0.4	Filters semantically distant chunks
RRF constant	k = 60	Standard value (Cormack et al., 2009)
Chunks injected into the prompt	top-5	Sufficient context without diluting attention
Generation temperature	0.1	Deterministic and precise answers
Maximum generated tokens	512	Concise answers
Rate limiting	60/30 requests per minute	Student chat / administrative API

Security and data sovereignty

Students query anonymously, constrained only by rate limiting at the reverse proxy, while administrators authenticate through signed JSON Web Tokens in order to upload or delete documents. Transport uses TLS with standard security headers, and data minimization is applied: the audit tables record queries, retrieved chunks, answers, and latency for evaluation purposes, but no personal identifier beyond a session token. Above all, since embeddings and generation run locally, no institutional document, query, or answer is transmitted to an external service.

Construction of the evaluation set

The reference question–answer pairs were built with the evaluation service embedded in the system itself, which generates them from the already indexed corpus chunks using the same locally hosted qwen2.5:14b model as a generator. The procedure produced twenty-two pairs distributed across three sets: eight questions spanning the five documents, four on the compact procedural instruction, and ten on the long bylaw; an additional, independent set of ten questions supports the comparative retriever experiment.

This procedure carries three methodological consequences that should be made explicit before the results are presented. First, the pairs were not validated by an independent expert in the institutional regulations, so their correctness and representativeness rest on the quality of the generator model rather than on external human judgement. Second, the same synthetic set was used both for empirical tuning of the pipeline parameters and for the final reported evaluation, with no held-out partition; this absence of separation introduces an optimistic bias of unquantified magnitude. Third, and by construction, both the reference answers and the reference contexts come from the very chunks the retriever is expected to recover, a circumstance that directly conditions the reading of the context metrics.

The thematic representativeness of the set rests on the fact that the questions were derived from the chunks of the five documents that concentrate the bulk of recurring student queries and that their distribution reproduces the heterogeneity of length and structure of the corpus. It does not rest, by contrast, on a deliberate stratification by question type or difficulty level, which was not part of the design and is recorded among the limitations.

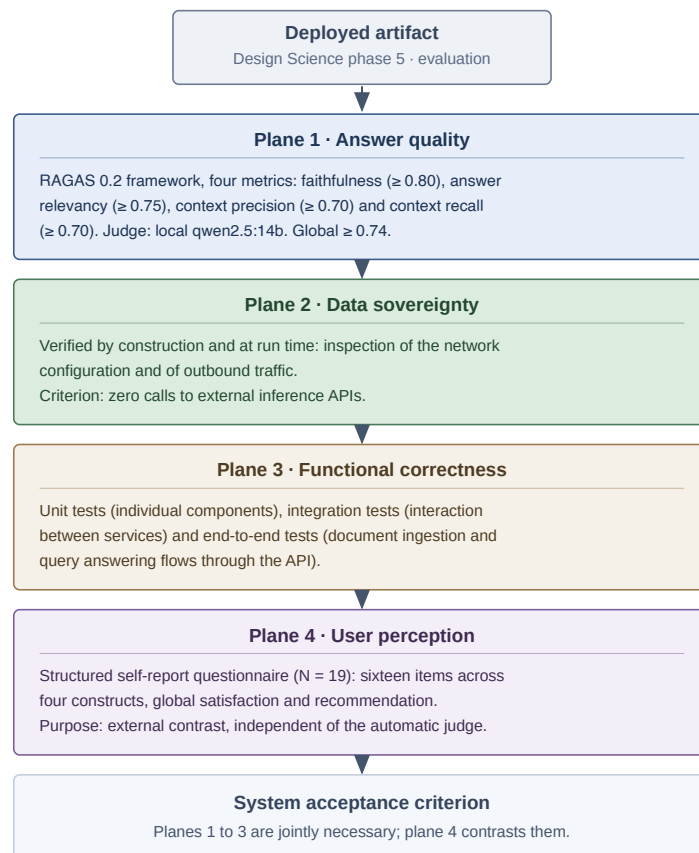
Evaluation design

In accordance with DSR phase 5, the artifact is assessed on four complementary planes: the technical quality of the generated answers, the data-sovereignty property of the deployment, the functional correctness of the software, and the perception of the consulting users (Figure 5). The first three planes are jointly necessary and the fourth operates as an external contrast to the preceding ones.

Answer quality is measured with the RAGAS framework (Es et al., 2024) through four metrics, each scored from zero to one by an LLM judge: faithfulness, that is, the extent to which the answer is grounded in the retrieved context, with a threshold of 0.80; answer relevancy, with a threshold of 0.75; and context precision and context recall, both with a threshold of 0.70. The global score is their arithmetic mean, with an acceptance criterion of 0.74, consistent with the values reported for comparable institutional RAG systems (Kuratomi et al., 2025).

To isolate the contribution of the fusion strategy, an ablation experiment is carried out, that is, a comparison of the complete system against versions of itself from which a component has been removed, so that the difference in performance can be attributed to that component and not to some other factor. Three retrieval configurations are compared over an identical question set: lexical-only, dense-only, and the hybrid retriever that fuses both rankings with RRF. All three use the same top-k, the same generation model and prompt, and the same judge, so any difference is attributable to retrieval.

Figure 5
Evaluation Pipeline



Note. The first three planes verify the system with instruments it provides itself; the fourth contributes the judgement of consulting users as an external contrast.

Beyond quality, the design adopts a target end-to-end latency at the 95th percentile of 5 seconds, consistent with the benchmark established for automated support in comparable university settings (Odede & Frommholz, 2024). Latency is measured on the reference GPU by separately timing retrieval and generation over 22 questions after model warm-up. The privacy property is verified by construction and at run time through inspection of the network configuration and of egress during operation. Functional correctness is verified with unit, integration, and end-to-end tests exercising the two complete flows through the API.

The three planes above share a limitation of principle: they measure the system with instruments the system itself provides, without eliciting the judgement of those who are to use it. The fourth plane remedies that gap through a structured self-report questionnaire, administered online after a session of free use of the prototype. The instrument gathers sixteen items on five-point agreement scales, grouped into four constructs —interface usability, conversational intelligence, document grounding and trust, and perceived performance and acceptance—, plus a global satisfaction scale, an eleven-point recommendation scale, and an optional open-ended item. The usability block includes a reverse-worded item, adapted from the System Usability Scale (Brooke, 1996), with the deliberate purpose of detecting responses issued by mere acquiescence, and item order is randomized at each administration. Internal consistency is estimated with Cronbach's (1951) alpha coefficient, and subgroup contrasts are resolved with the Mann-Whitney U test. The questionnaire's purpose is not to estimate parameters of the university community —the convenience sample precludes that— but to provide an external, independent contrast against metrics computed by a model of the same family as the generator.

Results and Discussion

The system in operation

Figure 6 presents the administration panel with the institutional corpus loaded: for each document it displays its processing state, the number of pages and chunks generated, and the management actions; during ingestion, the real-time progress bar reflects the pipeline stages. The panel itself allows evaluation runs to be launched and their reports downloaded: Figure 7 shows the RAGAS report generated by the system. This built-in self-evaluation capability turns quality measurement into a routine administrator operation rather than an external laboratory exercise.

Figure 6
Administration Panel with the Institutional Corpus Loaded

Documentos del corpus				5 documentos · 106 fragmentos		
#	DOCUMENTO	CATEGORÍA	ESTADO	PÁGS.	FRAGM.	ACTUALIZADO
01	SO-02-008-CSU-2024-firmada.pdf	■ OTROS	■ LISTO	26	42	2026-06-18
02	instructivo-de-homologacion-de-estudios-...	■ OTROS	■ LISTO	10	19	2026-06-18
03	INSTRUCTIVO-PARA-LA-VINCULACION-DE-AYUDA...	■ OTROS	■ LISTO	8	9	2026-05-28
04	manual-acervo-bibliografico.pdf	■ OTROS	■ LISTO	14	19	2026-05-28
05	POLITICAS-DE-ASEGURAMIENTO-DE-LA-CALIDAD...	■ OTROS	■ LISTO	6	17	2026-05-28

Note. Vector reproduction of the interface, which operates in Spanish; the page and chunk figures match those in Table 2.

Figure 7
RAGAS Evaluation Report Generated by the System

Reporte de Evaluación RAGAS

Generado el 28 de May de 2026 a las 18:30 · 8 muestras evaluadas

Score Global: 90.3%

El sistema muestra excelente calidad. Las respuestas son fieles al contexto y altamente relevantes.

Métricas Principales

Métrica	Score	Descripción
Faithfulness	96.88%	Fidelidad de la respuesta al contexto recuperado
Answer Relevancy	64.34%	Relevancia de la respuesta respecto a la pregunta
Context Precision	100.00%	Precisión del retrieval (chunks pertinentes)
Context Recall	100.00%	Cobertura del retrieval (chunks necesarios)

Detalle por Pregunta

#	Pregunta	Faith.	A.Rel.	C.Prec.	C.Rec.
1	¿Cuál es el objetivo principal de la política de aseguramiento de la calidad en ...	1.00	0.58	1.00	1.00
2	¿Cuál es la importancia del liderazgo en el Sistema de Gestión de la Calidad seg...	1.00	0.64	1.00	1.00
3	¿Cuál es la importancia de la investigación básica según el fragmento?	0.75	0.89	1.00	1.00
4	Quién es responsable de aplicar el Manual en las bibliotecas institu...	1.00	0.50	1.00	1.00
5	¿Cuál es la resolución estándar utilizada durante la digitalización de las obras...	1.00	0.71	1.00	1.00
6	¿Cuál es el objetivo del instructivo según se establece en el Artículo 2?	1.00	0.62	1.00	1.00
7	Cuál es una de las responsabilidades principales del ayudante de investigación s...	1.00	0.57	1.00	1.00
8	¿Cuál es el siguiente paso después de que el profesor entrega el informe de acep...	1.00	0.62	1.00	1.00

Generado automáticamente por el servicio de evaluación integrado · Métricas calculadas con RAGAS 0.2 usando qwen2.5:14b como juez.

Note. Vector reproduction of the interface, which operates in Spanish; the column means reproduce the aggregates reported in Table 5.

Answer quality

Table 5 reports the RAGAS scores obtained, computed with RAGAS 0.2 using the same locally hosted qwen2.5:14b model as the judge. To probe behavior at different granularities, three sets are used: a corpus-wide sample of eight questions and two single-document deep dives, over a compact 8-page procedural document (Doc. A) and over a dense 26-page bylaw with cross-references (Doc. B).

Table 5
RAGAS Results Across the Three Evaluation Sets

Evaluation Set	No. Quest	F	AR	CP	CR	Global
Corpus-wide sample	8	0.97	0.64	1.00	1.00	0.90
Doc. A: assistantship instructions (8 p.)	4	1.00	0.63	0.95	1.00	0.90
Doc. B: credit recognition bylaws (26 p.)	10	0.62	0.55	0.90	0.70	0.69
Acceptance threshold	—	≥ 0.80	≥ 0.75	≥ 0.70	≥ 0.70	≥ 0.74

Note. RAGAS 0.2 with qwen2.5:14b as judge. F: faithfulness; AR: answer relevancy; CP: context precision; CR: context recall.

The sense in which the acceptance criterion is declared satisfied must be made precise. The global score of 0.90 comfortably exceeds the 0.74 threshold, but that threshold is the mean of four individual thresholds, and one of them —answer relevancy— is not reached in any of the three sets. Acceptance is therefore declared over the aggregate and not over each metric separately; this distinction, usually left implicit in the literature that employs RAGAS, is made explicit here because it conditions the scope of the conclusions.

At the corpus level, the system attains a global score of 0.90, with near-perfect faithfulness (0.97) and with context precision and recall at their ceiling. The result is stable across questions: the per-question standard deviation is only 0.02, as is for the compact Doc. By contrast, Doc. B is highly dispersed (standard deviation of 0.21), so its aggregate should be read as indicative rather than precise. Doc. A reproduces the corpus profile, whereas the dense Doc. B constitutes the hard case: its global score falls below the threshold, dragged down by a faithfulness of 0.62 and a context recall of 0.70, since in a long, heavily cross-referenced bylaw the required provision is usually spread across distant chunks that the retriever recovers only partially. Across all three sets, answer relevancy is consistently the lowest metric: the answers are faithful but longer than the minimal span that RAGAS rewards. Prompt-level conciseness and recall on long documents are therefore the clearest avenues for improvement.

The values of 1.00 for context precision and recall call for a cautious reading. Both metrics compare the retrieved chunks against the reference contexts, and those contexts are, by construction, the very chunks from which each question was generated. What the value attests is therefore that the retriever consistently recovers the chunk from which the question derives, not that it generalizes to questions formulated independently of the indexed corpus. This is a ceiling effect attributable to the design of the evaluation set and not overfitting in the strict sense—the system is neither trained nor gradient-tuned on these data, but neither does it constitute a transferable estimate of performance on real queries. That the effect disappears as soon as the task hardens is shown by Doc. B itself, whose context recall falls to 0.70 even though its evaluation set was built using an identical procedure.

The four metrics are, moreover, computed with a language model acting as a judge that belongs to the same family as the generator and that is, strictly speaking, the same model. The configuration induces a risk of favorable self-assessment on two fronts: the judge shares with the generator both the stylistic regularities that the relevancy metric rewards and any biases in interpreting the regulatory text. The user contrast reported below provides evidence that this risk materializes. An evaluation with grading by institutional staff competent in the regulations, accompanied by a report of inter-rater agreement, remains a prerequisite for any adoption beyond the pilot character of the present work.

Retriever ablation

Table 6 reports the ablation experiment, run over an independent set of ten questions kept identical across the three configurations; since it is a different sample from the main one, the absolute score of the hybrid retriever differs slightly from the previous figure, but the three rows are directly comparable with one another.

Table 6
Retriever Ablation: Comparison over an Identical Question Set

Retriever	F	AR	CP	CR	Global
Lexical only (PostgreSQL FTS, conjunctive terms)	0.57	0.13	0.20	0.20	0.27
Dense only (pgvector/HNSW)	0.78	0.63	0.89	0.90	0.80
Hybrid (dense + lexical, RRF)	0.89	0.62	0.88	0.90	0.82

Note. n = 10, top-k = 5; qwen2.5:14b for generation and as RAGAS judge. Only the retrieval stage differs between rows.

Two findings stand out. First, the lexical-only configuration obtains a global score of 0.27: the full-text query, which requires conjunctive matching of every content word, returns no chunk at all for 8 of the 10 questions posed in natural language, so context precision and recall

fall to 0.20, and the model, deprived of grounding, cannot answer. The scope of this result should be delimited: it describes the behavior of the lexical signal as implemented—with conjunctive semantics and without inverse-frequency weighting—and does not warrant the conclusion that a well-specified lexical baseline, of the BM25 type, is inadequate for regulatory prose. The literature, in fact, reports in the opposite direction on university regulations in Romance languages (Kuratomu et al., 2025). What the ablation establishes robustly is more modest: the system's lexical signal, on its own, does not sustain the pipeline, and its value lies in the fusion.

Second, hybrid retrieval attains the best global score (0.82), above dense-only retrieval (0.80), and, above all, raises faithfulness from 0.78 to 0.89: fusing exact lexical matching with the dense semantic signal anchors the answer more firmly to the retrieved provisions. The difference in global score, 0.02 over ten paired questions, is too small to stand on its own; the interpretable gain from fusion is the one in faithfulness, and that is how it should be stated.

Operational and privacy results

Table 7 reports individual query latency on the reference GPU. Retrieval has a negligible time cost relative to generation, since hybrid fusion executes as a single indexed PostgreSQL query; local generation accounts for approximately 96% of end-to-end time. The median response time falls within the 5-second target, whereas the 95th percentile exceeds it by 29%. The latency objective is therefore unmet in the terms in which it was formulated, and local generation—not retrieval—is the stage responsible. Ingestion is asynchronous and offline, so it does not affect query latency; its cost is concentrated in PDF extraction and scales with document complexity, from 24 seconds for a 6-page policy to close to 12 minutes for the 26-page scanned bylaw.

Table 7
Individual Query Latency on the Reference GPU

STAGE	MEDIAN	P95	MEAN	MAX.
Hybrid retrieval	0.15 s	0.19 s	0.16 s	0.21 s
Local generation (14B)	3.40 s	6.28 s	3.69 s	7.24 s
End to end	3.57 s	6.47 s	3.85 s	7.38 s

Note. NVIDIA Tesla V100, $n = 22$ questions after warm-up. Generation accounts for practically all of the time; retrieval contributes less than 5%.

As regards privacy, the deployed system issues no call to external inference APIs: all vectorization, retrieval, and generation operations are served from institutional hardware, satisfying the data-sovereignty requirement and aligning the system with the Organic Law on Personal Data Protection (Asamblea Nacional del Ecuador, 2021).

Qualitative analysis

Table 8 shows three representative interactions from the evaluation run with hybrid retrieval, translated from Spanish into English and presented with their per-answer scores. The system extracts precise factual provisions—an approval date, a grade threshold, a procedural deadline—and appends the source marker linking each answer to its originating chunk, which allows the student to verify every answer against the institutional document.

Table 8
Representative question-answer pairs from the hybrid retrieval run

QUESTION	SYSTEM ANSWER (GROUNDED)	F / CP / CR
When was the instruction on the appointment of teaching and research assistants approved by the University Council?	Approved by the University Council at its ordinary session of 24 April 2019. [S]	1.00 / 1.00 / 1.00
What is the minimum grade required to apply as a teaching assistant?	The applicant must have passed at least 80% of the program credits and hold a grade of no less than 8.0/10 in the corresponding course. [S]	1.00 / 1.00 / 1.00
How long does the campus General Secretariat have to verify compliance once the reports and minutes are approved?	One (1) business day. [S]	1.00 / 0.95 / 1.00

Note. [S] marks the cited source chunk returned to the user. Questions and answers have been translated from the original Spanish for presentation.

Contrast with consulting users

Twenty-two people responded to the questionnaire; nineteen consented to participate and three declined. The records of the latter three were discarded without replacement. The composition of the sample conditions the reading of everything that follows: twelve participants (63.2%) are prospective students and seven (36.8%) enrolled students, and none belongs to the academic-administrative staff who are the primary users of the process being addressed. The majority group had therefore had no prior contact with the regulations the system indexes, a circumstance that strengthens the validity of judgements about interface clarity and weakens that of judgements requiring verification of regulatory content against the original document.

Table 9
User Perception by Construct and Internal Consistency of the Instrument

CONSTRUCT	MEAN	SD	95% CI	CRONBACH'S A
Interface usability	3.43	0.55	[3.17, 3.70]	0.206
Conversational intelligence	3.64	0.87	[3.23, 4.06]	0.863
Document grounding and trust	3.79	0.79	[3.41, 4.17]	0.900
Perceived performance and acceptance	3.71	0.80	[3.33, 4.09]	0.900
Global index (16 items)	3.64	0.67	[3.32, 3.97]	0.930

Note. N = 19; agreement scales from 1 to 5. The reverse-worded item was recoded before averaging.

The global index stands at 3.64, with a 95% confidence interval that excludes the neutral point of the scale: perception is favorable in a way distinguishable from indifference, though far from enthusiastic. The proportion of satisfied participants, the 73.7% located in the two favorable categories of the global satisfaction scale, exceeds the 70% threshold set in the evaluation plan.

The internal consistency of the instrument is high in the aggregate and in three of the four constructs but collapses in usability. The source of the anomaly is identifiable: that construct is the only one containing the reverse-worded item, and, once recoded, this item correlates negatively with the mean of the remaining fifteen items ($r = -0.486$; $p = 0.035$), where consistency would require the opposite. The pattern is what the literature documents as acquiescent responding, so the usability mean should be read as an average of four independent judgements rather than as the measurement of a latent construct.

With those caveats in place, the contribution of this plane lies in its convergence with the technical evidence at three specific points. First, the worst-rated item in the whole instrument is the one concerning response time ($M = 3.42$; 47.4% agreement), and that judgement coincides with the only operational requirement that the system's own measurement reported as unmet, the 95th percentile of latency. User perception thus corroborates, by a route independent of the instrumentation, that latency is the priority technical deficiency. Second, the rating of the item on the absence of fabricated information ($M = 3.68$; 57.9% agreement) is markedly more severe than the 0.97 faithfulness the judge attributes to the system over the same corpus: human users, even without verifying the original documents, are considerably less confident than the automatic metric, which empirically supports the suspicion of benevolence on the part of a same-family judge. Third, and in a favorable direction, the best-rated construct is precisely document grounding and trust, with 84.2% agreement that the system deserves trust comparable to that of the reference administrative offices; source traceability, which is the system's distinctive design property, is also the one users recognize most clearly. Willingness to adopt it is real but not unconditional: 63.2% consider that it would increase their efficiency relative to manual search, and 57.9% would prefer it as their main consultation channel.

The Mann-Whitney U test detects no significant difference in the global index between prospective and enrolled students ($U = 55.5$; $p = 0.271$), though with subsamples of this size, power is insufficient to interpret that absence as evidence of equivalence. Nor is the Net Promoter Score derived from the recommendation scale reported: the response distribution piles up at the midpoint, a pattern characteristic of an eleven-point scale read as a mark out of ten, so the resulting value is attributable to the wording of the question.

Discussion

Two patterns organize the results. First, the per-document analysis is more informative than the aggregate: quality is solid and stable on compact procedural documents but degrades on the 26-page bylaw with cross-references, where provisions spread across distant chunks are harder to retrieve in full and to use as grounding for answers; this shows that fixed chunking and top-k, adequate for short documents, ought to adapt to document length. Second, the ablation shows why hybrid retrieval matters for regulatory text: the lexical signal in isolation, in its conjunctive configuration, does not answer natural-language questions; dense search in isolation attains lower faithfulness; and RRF fusion combines the two without score normalization. The only metric that remains uniformly low is answer relevancy, which makes prompt-level conciseness an orthogonal refinement of the whole system.

Local deployment eliminates per-token cost and third-party data exposure at the price of bounded GPU capacity, the main factor that will shape latency under concurrency and the direct cause of the unmet objective at the 95th percentile. The user contrast reinforces that diagnosis by an independent route and adds a warning that automatic evaluation could not have produced: the gap between the faithfulness the judge attributes to the system and the trust users declare suggests that metrics from a judge of the same family as the generator should be treated as an upper bound rather than an unbiased estimate.

Two lessons follow for anyone wishing to replicate the system. Chunking by document structure, rather than by fixed window, is the ingestion decision with the greatest effect on retrieval quality in regulatory text, so investing in high-quality structural conversion pays off more than tuning the retriever afterwards. And evaluation should be planned from the outset and embedded in the system itself, reserving a partition of questions independent of the one used for parameter tuning and preferably using a judge model from a family different from the generator's.

Limitations

The quality evaluation rests on twenty-two question–answer pairs distributed across sets of four to ten questions. At that sample size, the reported confidence intervals, computed by normal approximation over bounded scores concentrated near the ceiling of the scale, lack sufficient asymptotic justification and should be read as an approximate indication of dispersion; a percentile bootstrap interval would have been preferable. A set of at least thirty questions, stratified by document, by type, and by difficulty level, and validated by an independent expert in the regulations, is a condition for narrowing these estimates.

The pairs were generated synthetically with the same model used for generation, and the resulting set served both for empirical tuning of the parameters and for the final evaluation, without a held-out partition; this introduces an optimistic bias of unquantified magnitude. RAGAS is moreover computed with a judge that is the same model used for generation, which may bias the scores: the user contrast provides an indication that the bias exists but does not allow it to be quantified.

The perception questionnaire suffers from four limitations that severely bound its inferential scope. Its sample is small and non-probabilistic, which prevents extrapolation and deprives subgroup contrasts of power. Its composition does not correspond to the primary target population, since no participant belongs to the academic-administrative staff. The open-ended item received not a single response, so the entire interpretation rests on closed scales. And the design lacks a control condition and a pre-intervention measure, so the ratings describe perception of the system in absolute terms and do not warrant attributing to it an improvement over the current consultation channel.

The corpus and the evaluation are confined to a single institution and to Spanish, so external validity is not established. Latency was measured only for individual, sequential queries, without characterizing behavior under sustained multi-user concurrency. The chunking parameters and the similarity and RRF constants were tuned empirically and were not the object of a systematic sweep. Finally, the lexical baseline of the ablation experiment corresponds to the full-text implementation actually deployed, with conjunctive semantics and without inverse-frequency weighting; a canonical BM25 implementation was not evaluated, so the comparison should not be read as a contrast between retrieval families but between the specific configurations that make up the system.

Conclusions

RQ1 is answered partially and conditionally. The fully local pipeline exceeds the global acceptance threshold on the corpus-wide set (0.90 against 0.74) and on the compact procedural document, with no call to external APIs; it does not, however, satisfy the criterion under its conjunctive reading. Answer relevancy remains below its individual threshold in all three sets; the long bylaw falls below both the global and the faithfulness thresholds; and the 95th percentile of latency exceeds the design target, even though the median lies within it. The evidence therefore supports a bounded statement—a fully local deployment attains acceptable quality on compact regulatory documents, with median latency within target—rather than a general claim of sufficiency. The three shortfalls are, moreover, of different natures: relevancy depends on prompt verbosity, degradation on long documents on the chunking strategy, and the 95th percentile of latency on GPU capacity; none is attributable to the locality of the deployment, which is the property RQ1 interrogates.

RQ2 is supported, though with limited scope. In the experiment where only the retrieval strategy changes, the hybrid retriever outperforms both single-strategy configurations over the same set of ten questions: it improves faithfulness by 0.11 points over dense-only retrieval and avoids the missing-context failures that the lexical signal produces when acting alone. The difference in global score, by contrast, is a mere 0.02 points, too small to support on its own a conclusion about the general superiority of the method.

RQ3 is answered through the contrast between the compact documents and the long bylaw with cross-references: document length and its cross-references explain most of the observed variation in answer quality, while the local character of the deployment shows no attributable effect. To the three technical planes is added a fourth contrast with nineteen consulting users, whose value lies not in its statistical magnitude but in its methodological independence: it corroborates that latency is the priority deficiency and qualifies, in a restrictive direction, the faithfulness the automatic judge attributes to the system.

Future work comprises the construction of an enlarged and stratified evaluation set, with a held-out partition independent of the one used for tuning, and its validation by institutional staff competent in the regulations; length-adaptive chunking and top-k, aimed at the recall gap on long bylaws; prompt refinements to raise answer relevancy; the evaluation of a canonical BM25 baseline against the current full-text signal; and a pilot under real concurrency characterizing latency and availability at scale.

Acknowledgements and Declarations

This work received no external funding and was carried out at Universidad Indoamérica, Quito, Ecuador. Following the CRediT taxonomy, the individual contributions were: J. P. Guadalupe Andrade, conceptualization, software, investigation, validation, formal analysis, and writing of the original draft; B. Topón Visarrea, conceptualization, methodology, supervision, and writing, review, and editing; and C. Iza, methodology, data curation, and writing, review, and editing. All authors read and approved the final version of the manuscript and declare no conflicts of interest.

The authors declare that, in preparing this paper, AI tools (Claude, by Anthropic, through the Claude Code tool) were used as support for style editing, text restructuring, and translation of technical content provided by the authors, as well as for minor formatting support, across all sections of the paper. Such tools were not used to generate research results, experimental data, or citations. All technical content, design decisions, measurements, and conclusions are the authors' own. The authors have reviewed and take full responsibility for the entirety of the content. Within the system described, a language model is additionally employed as a functional component and, during evaluation, as a judge model computing the RAGAS metrics.

The implementation is released as open-source software for transparency and reproducibility; the repository link is made available to the editorial office for inclusion in the published version.

References

- Antico, L., Giordano, C., Koyuturk, D. B., & Ognibene, D. (2024). *Unimib Assistant: Designing a student-friendly RAG-based chatbot for all their needs* [Preprint], arXiv. <https://doi.org/10.48550/arXiv.2411.19554>
- Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales* [Organic Law on Personal Data Protection]. Registro Oficial, Suplemento 459.

- Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis.
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2318–2335). Association for Computational Linguistics.
- Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). ACM. <https://doi.org/10.1145/1571941.1572114>
- Coronel Crespo, B. P., & Ulloa Bernal, C. F. (2025). *Desarrollo de un chatbot multiagente basado en modelos LLM y arquitectura RAG para la interacción en lenguaje natural con bases de datos organizacionales* [Undergraduate thesis, Universidad de Cuenca]. Repositorio Institucional de la Universidad de Cuenca.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), 297–334. <https://doi.org/10.1007/BF02310555>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics.
- Espinosa-Luna, G., Castillo-Oliva, R., Montañez-Díaz, M., & Mendoza-De-los-Santos, J. (2023). Implementación de un chatbot basado en modelo de lenguaje de inteligencia artificial para responder preguntas frecuentes de estudiantes universitarios. *Revista Científica de Sistemas e Informática*, 3 (2), e570. <https://doi.org/10.51252/rcsi.v3i2.570>
- Guerra, M., Monti, C., & Fernández, A. (2025). Sistema de generación aumentada por recuperación (RAG) para el acceso a información académica. In *Actas del XXXI Congreso Argentino de Ciencias de la Computación (CACIC 2025)* (pp. 1165–1169). Red UNCI.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28 (1), 75–105. <https://doi.org/10.2307/25148625>
- Khan, A. U. R., Hasan, M., Kemell, K.-K., Rasku, J., & Abrahamsson, P. (2024). *Developing retrieval-augmented generation (RAG) based LLM systems from PDFs: An experience report* [Preprint], arXiv. <https://doi.org/10.48550/arXiv.2410.15944>
- Kuratomi, G., Pirozelli, P., Cozman, F. G., & Peres, S. M. (2025). *A RAG-based institutional assistant* [Preprint], arXiv. <https://doi.org/10.48550/arXiv.2501.13880>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., & Vagenas, P. (2025). Docling: An efficient open-source toolkit for AI-driven document conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence (Demonstration Track)*. AAAI Press.
- Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42 (4), 824–836. <https://doi.org/10.1109/TPAMI.2018.2889473>
- Manjarrés-Betancur, G., & Echeverri-Torres, J. (2020). Asistente virtual académico utilizando tecnologías cognitivas de procesamiento de lenguaje natural. *Revista Politécnica*, 16 (31), 85–96. <https://doi.org/10.33571/rpolitec.v16n31a7>

- Neupane, S., Bhatt, M., Aryal, S., Lafleur, E., Poudel, S., Fathi, H., Calyam, P., & Gupta, R. (2024). *From questions to insightful answers: Building an informed chatbot for university resources* [Preprint], arXiv. <https://doi.org/10.48550/arXiv.2405.08120>
- Nguyen, M.-T., & Quan, T.-T. (2025). URAG: Implementing a unified hybrid RAG for precise answers in university admission chatbots — A case study at HCMUT. In *Proceedings of the 13th International Symposium on Information and Communication Technology (SOICT 2024)* (CCIS 2351, pp. 95–109). Springer.
- Odede, I., & Frommholz, I. (2024). JayBot — Aiding university students and admission with an LLM-based chatbot. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24)* (pp. 391–395). ACM. <https://doi.org/10.1145/3627508.3638293>
- Qwen Team. (2025). *Qwen2.5 technical report* (arXiv:2412.15115). arXiv. <https://doi.org/10.48550/arXiv.2412.15115>
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3 (4), 333–389. <https://doi.org/10.1561/15000000019>
- Swacha, J., & Gracel, J. (2025). Retrieval-augmented generation (RAG) chatbots for education: A survey of applications. *Applied Sciences*, 15 (8), 4234. <https://doi.org/10.3390/app15084234>
- Xu, Z., Cruz, M. J. N., Guevara, M., Wang, T., Deng, M., Liu, L., & Pan, S. (2024). Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)* (pp. 2905–2909). ACM. <https://doi.org/10.1145/3626772.3661370>

Numerical Simulation Engines and Deep Learning Architectures for Inverse Design of Multilayer Metamaterials

Motores de Simulación Numérica y Arquitecturas de Aprendizaje Profundo para el Diseño Inverso de Metamateriales Multicapa

Federico Rodas¹ <https://orcid.org/0000-0002-0000-0002>, Laura Malache² <https://orcid.org/0009-0001-0000-0001>, Osmani Mujica Pérez³ <https://orcid.org/0009-0008-2243-9694>

¹Universidad de las Fuerzas Armadas - ESPE, Departamento de Ciencias Exactas, Sangolquí, Ecuador
fmrodas@espe.edu.ec

²Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia - CEDIA, Cuenca, Ecuador
laura.malache@cedia.com.ec

³Universidad de Las Américas - UDLA, Quito, Ecuador
osmani.mujica@udla.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/08/25

Accepted: 2026/09/01

Published: 2026/10/14

Abstract

The inverse design of multilayer metamaterials —artificial structures with engineered electromagnetic, acoustic, thermal, or mechanical properties— represents a high-dimensional optimisation challenge that has driven the integration of deep learning into computational photonics and materials science. This paper presents a systematic literature review conducted under the PRISMA 2020 protocol, aimed at mapping the numerical simulation methods employed as forward engines or training data generators within deep learning-assisted inverse design pipelines. Starting from 392 records retrieved from Scopus and Web of Science over the 2010–2025 period, a multi-stage screening process based on a heuristic prioritisation index termed CARS (Citation-Adjusted Relevance Score) identified 44 core articles. Eleven distinct numerical methods were identified and organised into two conceptual layers: classical simulation engines (finite element method, topology optimisation, guided-mode expansion, adjoint method, and

Summary: Introduction, Materials and Methods, Results and Discussion, Conclusions.

How to cite: Rodas, F., Malache, L. & Mujica, O. (2026). Numerical Simulation Engines and Deep Learning Architectures for Inverse Design of Multilayer Metamaterials. *Revista Tecnológica - Espol*, 38(E1), 71-83.
<https://doi.org/10.37815/rte.v38nE1.1645>

genetic algorithms) and deep learning architectures (tandem CNN, transformer, GAN, PINN, and diffractive neural networks). Results show a strong dominance of the electromagnetic-photonic domain (72%), accelerated growth of deep learning-based publications from 2022 onward, and a clear trend toward hybrid physics, data and learning pipelines. The review concludes by proposing a prioritised experimental roadmap for future research.

Keywords: metamaterials, inverse design, deep learning, numerical simulation, systematic review, topology optimisation.

Resumen

El diseño inverso de metamateriales multicapa —estructuras artificiales con propiedades electromagnéticas, acústicas, térmicas o mecánicas diseñadas a medida— constituye un problema de optimización de alta dimensionalidad que ha impulsado la integración del aprendizaje profundo en la fotónica computacional y la ciencia de materiales. Este artículo presenta una revisión sistemática de la literatura realizada siguiendo el protocolo PRISMA 2020, orientada a mapear los métodos de simulación numérica empleados como motores directos o generadores de datos de entrenamiento dentro de flujos de trabajo de diseño inverso asistido por aprendizaje profundo. A partir de 392 registros recuperados de Scopus y Web of Science en el periodo 2010–2025, un proceso de cribado multietapa basado en un índice heurístico de priorización denominado CARS (Citation-Adjusted Relevance Score) permitió identificar 44 artículos centrales. Se identificaron once métodos numéricos distintos, organizados en dos capas conceptuales: motores clásicos de simulación (método de elementos finitos, optimización topológica, expansión en modos guiados, método adjunto y algoritmos genéticos) y arquitecturas de aprendizaje profundo (CNN en tándem, transformer, GAN, PINN y redes neuronales difractivas). Los resultados evidencian un predominio del dominio electromagnético-fotónico (72 %), un crecimiento acelerado de las publicaciones basadas en aprendizaje profundo a partir de 2022, y una tendencia clara hacia flujos híbridos que combinan física, datos y aprendizaje. La revisión concluye proponiendo una hoja de ruta experimental priorizada para futuras investigaciones.

Palabras clave: metamateriales, diseño inverso, aprendizaje profundo, simulación numérica, revisión sistemática, optimización topológica.

Introduction

Multilayer metamaterials are periodic or quasi-periodic artificial structures whose effective electromagnetic, acoustic, thermal, or mechanical response depends on the geometry, thickness, and material properties of each constituent layer (Lin et al., 2018; Liu et al., 2024). Because their response arises from structural rather than chemical composition, they can be engineered to exhibit properties unavailable in nature: negative refractive index, near-perfect absorption, acoustic bandgaps, or programmable heat flux.

The inverse design problem—recovering layer parameters from a desired spectral or field response—is fundamentally ill-posed, nonlinear, and one-to-many: multiple geometries may produce the same target response (Ma et al., 2020; Naseri & Hum, 2021). Classical approaches iterate between a forward solver, such as a finite element or transfer-matrix code, and an optimisation heuristic, requiring thousands of computationally expensive simulations.

Deep learning offers an alternative: trained once on a synthetic dataset of forward simulations, a deep learning model can infer designs in milliseconds at inference time (Jia et

al., 2022; Lininger et al., 2021). Despite rapid progress, however, the field lacks a comprehensive synthesis of which numerical simulation methods have been most productive as training data generators, and how these map to deep learning architectures and application domains. This gap is the motivation for the present work.

Objectives and contributions

The general objective of this research is to systematically identify and classify the numerical simulation methods used in the inverse design of multilayer metamaterials, with emphasis on their role as data generators for training deep learning architectures. The specific contributions are as follows:

1. A PRISMA 2020 systematic review of 392 records yielding a corpus of 44 core articles on numerical methods for multilayer metamaterial inverse design.
2. A quantitative analysis of method frequency, application domains, and temporal trends over the 2010–2025 period.
3. A taxonomy of eleven numerical methods and deep learning architectures organised according to their function within the design pipeline.
4. A prioritised experimental roadmap, presented explicitly as a research direction derived from the synthesised evidence rather than as a validated outcome.

Related work

Deep learning for photonic inverse design

The application of neural networks to nanophotonic and metamaterial design has grown rapidly since 2018. Work by Liu et al. (2022) demonstrated that fully connected networks could map optical spectra to nanostructure parameters for single-layer gratings. Chen et al. (2019) extended this approach to graphene-based photonic metamaterials using an adaptive neural network, while Lininger et al. (2021) showed that convolutional tandem models generalise adequately to multilayer thin-film stacks.

An important observation across this body of work is that the inverse design problem is ill-conditioned: multiple layer configurations can produce nearly identical spectra. Tandem architectures address this difficulty by pre-training a forward network that enforces physical consistency (Jia et al., 2022; Zhu et al., 2022). Approaches based on generative adversarial networks (Hodge et al., 2020; Naseri & Hum, 2021) follow a different route, learning the full distribution of valid designs rather than a point estimate, which can accommodate the one-to-many character of the inverse problem.

Optimisation Methods for Metamaterial Design

Classical optimisation methods predate deep learning in the metamaterial design literature. Topology optimisation, originally developed for structural mechanics, has been applied to multilayer photonic structures by Lin et al. (2018) for metaoptic lens design and by Fu et al. (2016) for broadband perfect absorbers. The reliance of topology optimisation on adjoint gradients (Mansouree et al., 2020) makes it suitable for problems with thousands of design variables, but its computational cost per iteration limits the number of designs that can be evaluated, preventing its practicality as a standalone dataset generator.

Genetic algorithms (Cao et al., 2024; Jiang et al., 2023) avoid gradient computation and accommodate discrete design variables, such as integer layer counts or discrete material choices, but scale poorly beyond approximately fifty variables. For this reason, in the reviewed literature they are typically employed in combination with a surrogate model that replaces expensive forward simulations.

Forward Simulation Engines

The choice of forward simulation engine strongly influences the feasibility of large-scale dataset generation. The transfer-matrix method (García-Esteban et al., 2021) and guided-mode expansion (Zanotti et al., 2024) are semi-analytical and compute the optical response of periodic multilayer stacks in milliseconds per structure. Finite-difference time-domain and finite element methods provide full-wave simulations at a cost of minutes to hours per structure (Peralta & Fachinotti, 2017), which can make them impractical for generating the 10^5 samples typically required to train a deep learning model.

Materials and Methods

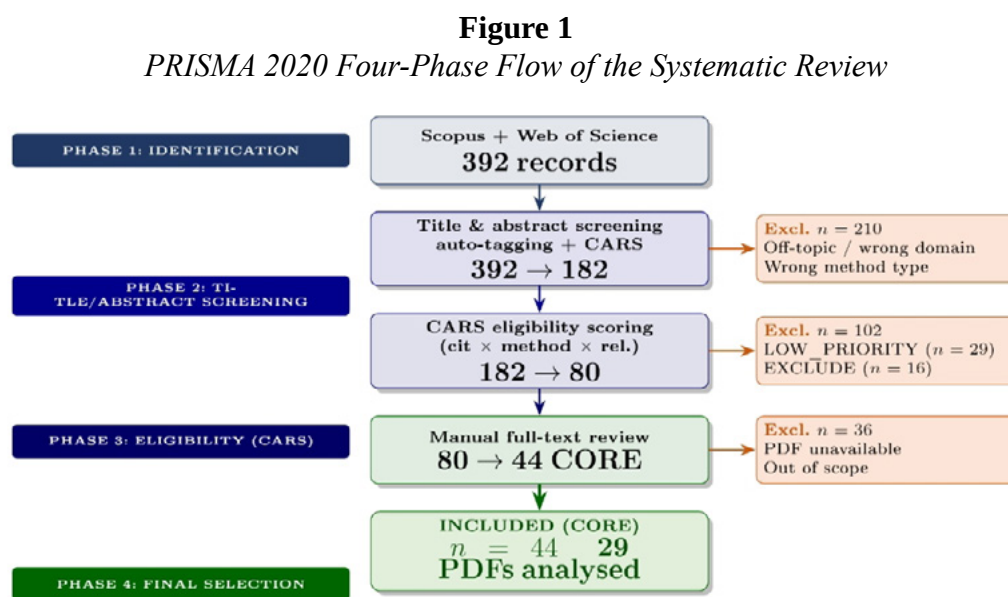
Search strategy and databases

Records were retrieved from Scopus and Web of Science in January 2026 using Boolean combinations of the terms multilayer metamaterial, inverse design, neural network, deep learning, topology optimisation, finite element method, and adjoint method, together with domain terms (electromagnetic, acoustic, thermal, mechanical). No temporal restriction was applied; the effective range of retrieved records spans 2010–2025.

The resulting corpus of 44 articles is intentionally narrow by design. The review targets a highly specific intersection: structures that are multilayer metamaterials and are addressed through inverse design using a numerical simulation or optimisation method. This triple constraint excludes the large body of work on single-interface metasurfaces, bulk metamaterial characterisation, and forward-only simulation studies. Within this niche, 44 included articles provide the evidence base used to characterise methodological trends, domain distributions, and dominant pipeline patterns (Page et al., 2021).

PRISMA 2020 Screening protocol

The review was conducted following the PRISMA 2020 guidelines (Page et al., 2021). The four-phase flow is illustrated in Figure 1. Screening files were managed through seven successive workbooks implementing automated tagging, CARS computation, and manual full-text verification.



Note. Adapted from the PRISMA 2020 protocol (Page et al., 2021). Databases consulted: Scopus and Web of Science, 2010–2025.

CARS (Citation-Adjusted Relevance Score)

Each record was assigned to a composite prioritisation index weighting three orthogonal dimensions, according to Equation 1:

$$\text{CARS} = 0.40 \cdot \text{cit_pct} + 0.35 \cdot \text{meth} + 0.25 \cdot \text{rel} \quad (1)$$

Where $\text{cit_pct} \in [0, 1]$ is the annual citation percentile of each record, normalised within its publication-year cohort (that is, a 2024 article is compared only with other 2024 articles in the set), so that recent records are not penalised relative to older ones. The meth component $\in \{0.30, 0.65, 0.85, 1.00\}$ is an ordinal score of methodological specificity: 0.30 for records with no recognised method keywords; 0.65 for a single general keyword; 0.85 for a specific deep learning keyword; and 1.00 for combinations including DEEP_LEARNING together with at least one complementary methodological term. Finally, $\text{rel} \in \{0.20, 1.00\}$ is a binary thematic relevance indicator, set to 1.00 when the title or abstract explicitly addresses the inverse design of multilayer structures, and 0.20 otherwise.

The weight assignment was determined through a heuristic calibration process: the three weights were varied over a grid and evaluated against a manually labelled gold-standard set of twenty records (ten highly relevant and ten irrelevant) until the resulting ranking correctly placed all ten relevant records within the top 20 % of scores. The final weights reflect the practical judgement that citation impact provides one indirect signal for prioritising influential studies in a field without standardised benchmarks, that methodological specificity is the most decisive criterion for inclusion, and that thematic alignment, while necessary, operates as a binary gate rather than a fine discriminant.

It must be stated explicitly that CARS constitutes an operationally designed heuristic prioritisation tool, and not a universal bibliometric index. It was conceived specifically to manage the large-volume, multi-keyword screening challenge of this review, and its weights should not be interpreted as absolute importance estimates transferable to other literature searches. Its function is analogous to that of a scoring rubric: transparent, reproducible within this study, and open to recalibration in future replications.

Table 1
CARS Classification of the 182 Records that passed title and abstract screening

Class	Cars range	N	%	Role in the review
CORE	≥ 75	44	24.2	Primary synthesis corpus
HIGH_VALUE	60–74	36	19.8	Background context (deferred)
SUPPORTING	45–59	47	25.8	Supplementary references (deferred)
LOW_PRIORITY	< 45	29	15.9	Excluded from synthesis
EXCLUDE	—	16	8.8	Off-topic after title screening
Not scored	—	10	5.5	Duplicate or missing metadata
Total Phase 2		182	100	

Note. Non-scored records were retained by Phase 2 but lacked sufficient metadata for CARS computation; all were manually reviewed and ultimately excluded at Phase 4. Prepared by the authors.

Exclusion Breakdown and Full-Text Access

Table 2 details the reasons for exclusion at each phase, accounting for the transitions $392 \rightarrow 182 \rightarrow 80 \rightarrow 44$.

Table 2
Exclusion breakdown by phase and reason

Phase	Exclusion reason	n	Retained
2 – Screening	Off-topic (no metamaterial design)	121	182
	Wrong material type (single-layer, bulk)	58	
	Language other than English	31	
3 – CARS	Low CARS: LOW_PRIORITY (< 45)	29	80
	Low CARS: EXCLUDE (off-scope)	16	
	CARS 45–74 (HIGH_VALUE/SUPPORTING, deferred)	57	
4 – Full text	HIGH_VALUE below CORE threshold (deferred)	36	44
Total excluded		348	44 CORE

Note. Prepared by the authors based on the screening workbooks available in the public repository.

Regarding full-text access, of the 44 core articles, 29 full texts were retrieved and analysed in depth; the remaining 15 could not be accessed because Universidad de las Fuerzas Armadas ESPE, as a public university, does not hold institutional subscriptions to all publisher platforms, primarily paywalled content from Elsevier and IEEE Xplore. These 15 articles were nonetheless retained in the corpus: their inclusion was supported by complete title, abstract, and citation metadata, as well as by cross-referencing their key contributions through citations contained in the 29 accessible full texts. This limitation is explicitly acknowledged as a threat to internal validity.

Results and Discussion

Temporal distribution

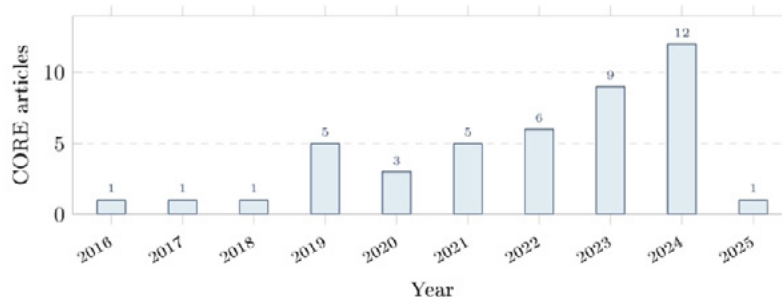
Table 3 and Figure 2 reveal a pronounced acceleration from 2022 onward, with 2024 alone accounting for 27% of the entire corpus. This behaviour reflects the convergence between mature deep learning architectures —particularly Transformer and deep convolutional networks— and the availability of large synthetic datasets generated by semi-analytical solvers (Cai et al., 2023; Teng et al., 2023).

Table 3
Annual distribution of core articles (2016–2025)

YEAR	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	TOTAL
Articles	1	1	1	5	3	5	6	9	12	1	44
% corpus	2	2	2	11	7	11	14	20	27	2	100

Note. Prepared by the authors.

Figure 2
Temporal distribution of core articles (2016–2025)



Note. The number of core articles increases markedly from 2022 onward. Prepared by the authors.

Taxonomy of computational methods and architecture

Eleven distinct methods were identified and organised into two conceptually differentiated layers of the inverse design pipeline. Layer A groups the classical simulation methods, that is, physics-based forward engines that generate the training data —spectrum, displacement field, or temperature map— from a given structural configuration. Layer B groups the deep learning architectures, namely data-driven models that learn to invert the forward map, enabling millisecond inference at deployment time.

This separation is useful for pipeline design: the choice of Layer A engine influences the volume and fidelity of the dataset, while Layer B architecture influences inverse accuracy and generalisation capability. Table 4 presents the complete taxonomy structured accordingly.

Table 4
Computational methods and architectures identified in the 44 core articles, separated by layer

Method	N	Period / domain	Role in the pipeline
Layer A — Classical Simulation (Forward Engines)			
Topology Optimisation	7	2016–24 / EM, Ac, Me	Material density distribution; adjoint gradients
Adjoint Method	2	2019–20 / EM	Efficient gradient in high-dimensional spaces
Finite Element Method	3	2017–23 / Th, Me	Full 3D PDE solver; validation engine
Genetic Algorithm	4	2013–25 / EM, Me	Derivative-free search for discrete layer stacks
Isogeometric Optimisation	2	2022–24 / Me	Spline-based CAD-exact shape optimisation
Layer B — Deep Learning Architectures			
CNN + Optimisation	12	2020–24 / EM, Th, Ac	Spectral forward surrogate; drives tandem loop
Tandem DNN	8	2021–24 / EM, Th	Bidirectional network; resolves one-to-many mapping
Transformer	5	2023–25 / EM	Multi-head attention on broadband spectra
GAN	4	2019–21 / EM	Generative one-to-many geometry synthesis
PINN	3	2021–24 / EM, Ac	Physics (PDE) constraints in the loss function
Diffractional Neural Network	2	2019–20 / EM	Optical layers as neural layers; end-to-end training

Note. EM = electromagnetic/photonic; Th = thermal; Ac = acoustic; Me = mechanical. Prepared by the authors.

Domain distribution and cross-analysis

Table 5 cross-tabulates methods with application domains. The electromagnetic-photonic domain dominates the corpus with 32 of 44 articles (72.7%), concentrating convolutional network-based approaches. The mechanical domain relies more heavily on classical simulation —finite elements, genetic algorithms, and isogeometry— while the acoustic domain exhibits the fastest recent growth, suggesting a potential research opportunity in deep learning-assisted acoustic metamaterial design.

Table 5
Method occurrences across core articles by application domain

Method	Em/phot.	Thermal	Acoustic	Mech.	Total
CNN + Optimisation	9	2	1	0	12
Tandem DNN	6	1	0	1	8
Transformer	5	0	0	0	5
GAN	4	0	0	0	4
PINN	2	0	1	0	3
Diffraction Neural Network	2	0	0	0	2
Topology Optimisation	3	1	2	1	7
Adjoint Method	2	0	0	0	2
Finite Element Method	0	1	0	2	3
Genetic Algorithm	3	0	0	1	4
Isogeometric Optimisation	0	0	0	2	2
Total	36	5	4	7	52

Note. The total sum exceeds 44 because articles employing multiple methods are counted in each of them. Prepared by the authors.

Method–architecture combinations

The dominant pattern identified in the corpus is the convolutional forward-inverse tandem loop, present in twelve articles, in which a convolutional network pre-trained on forward simulations is concatenated with a dense inverse network that maps target spectra to layer parameters (Lininger et al., 2021; Wang et al., 2023; Xiong et al., 2023). It should be noted that the individual studies differ in their datasets, target spectral ranges, and evaluation metrics; direct cross-study performance comparisons are therefore indicative rather than conclusive.

The second most prominent pattern is synthesis based on generative adversarial networks, present in four articles, where the generator directly produces stack geometries (Hodge et al., 2020; Naseri & Hum, 2021). The third corresponds to the Transformer-type spectral model, present in five articles, which in the reviewed studies were applied to long-range spectral dependencies in broadband terahertz and near-infrared problems, although comparative evaluation under equivalent conditions remains an open research task (Cai et al., 2023; Teng et al., 2023).

Physics-informed approaches incorporate Maxwell’s equations or elasticity equations into the training loss function. The reviewed works report dataset reductions of up to one order of magnitude relative to their own purely data-driven baselines (Noureen et al., 2023; Yago et al., 2024); whether this advantage generalises across different layer geometries and solvers requires further study.

Key trends

From standalone networks to hybrid pipelines

Early studies, between 2019 and 2021, deployed simple dense networks trained on small datasets (Chen et al., 2019; García-Esteban et al., 2021). From 2022 onward, the dominant pipeline combined a physics-based forward solver—capable of generating between 10^4 and 10^6 simulations—with a deep learning surrogate model trained on that data (Hu et al., 2024; Knightley et al., 2023). This shift substantially improves generalisation by separating the data generation problem, addressed through numerical simulation, from the inference problem, solved through deep learning.

Emergence of attention mechanisms

The Transformer architecture, originally proposed for natural language processing, has been adapted to spectral inverse design since 2023 (Cai et al., 2023; Teng et al., 2023). Its capacity to capture global correlations across spectral bands may be advantageous in broadband problems where resonances interact throughout the entire spectrum. The self-attention mechanism allows the model to simultaneously weight contributions from all spectral positions, unlike convolutional networks, which operate with local receptive fields that must be stacked to achieve global coverage. In the context of multilayer metamaterials this is particularly relevant, given that a resonance in one spectral band is physically coupled to adjacent bands through the interference conditions between layers.

Physics-constrained learning as a data-reduction strategy

Physics-informed networks and adjoint-integrated networks reduce the required number of forward simulations by incorporating partial differential equation constraints directly into the loss function (Noureen et al., 2023). This is particularly relevant for three-dimensional finite element forward models, where each simulation may take hours. Within their respective experimental settings, the reviewed works individually report training-data reductions relative to their own baselines, with accuracy impacts that the authors characterise as acceptable (Noureen et al., 2023; Yago et al., 2024); the magnitude of these gains is problem-specific and should not be generalised without comparable experimental conditions.

Threats to Validity

Internal validity

The CARS index introduces a citation-count bias that may underrepresent recent high-quality work published within two years of the search date. To partially mitigate this effect, the *cit_pct* component was normalised within the publication-year cohort rather than globally. The two-layer classification of method family may likewise obscure architectural variations within each family; future work should apply a finer-grained taxonomy. The manual relevance indicator, assigned by a single reviewer, introduces reviewer subjectivity; a dual-review protocol with inter-rater agreement assessed using Cohen's kappa measurement is recommended for future replications.

A further threat is partial full-text coverage: 15 of the 44 core articles, equivalent to 34%, could not be retrieved in full for the institutional access reasons already stated. These articles were retained based on abstract and metadata evidence but nuances visible only in the full text—experimental protocols, exact dataset sizes, architecture hyperparameters—may have been missed.

External validity and database bias

The review covers only Scopus and Web of Science. Both databases offer solid coverage of peer-reviewed journals and conference proceedings from Springer, IEEE, and Elsevier, but present limited indexing of: machine learning conference proceedings such as NeurIPS, ICLR, and ICML, where architecturally innovative work is often published before its journal versions; preprint repositories such as arXiv, which may contain the most recent results from the 2024–2025 period; and regional or Spanish-language journals relevant to the Ecuadorian and Latin American research context.

Language and grey literature bias

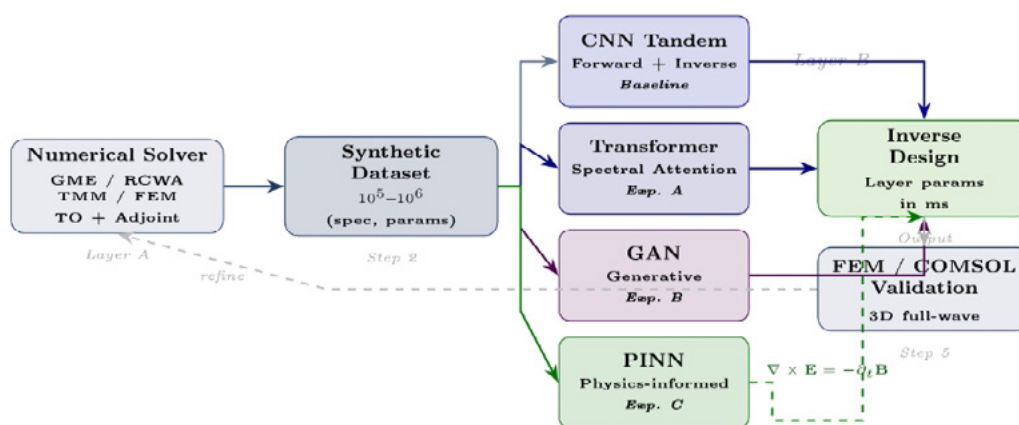
The search was restricted to English-language records. Although English is the dominant language of scientific publication in this domain, relevant contributions may also have been

published in Chinese and other languages. Likewise, technical reports, doctoral theses, patents, and software documentation were excluded. For a domain that depends on computational tools, this exclusion may underrepresent emerging engineering practices. Future reviews should consider complementary searches in ProQuest Dissertations relevant institutional repositories.

Proposed experimental pipeline

Based on the systematic analysis, the following experimental pipeline is proposed as a research direction for the development of a deep learning training framework oriented to the inverse design of multilayer metamaterials. It must be emphasised that this pipeline has not been implemented or validated; it represents the logical next step derived from the evidence synthesised in the review, and each stage corresponds directly to methods identified in the analysed corpus. Figure 3 illustrates the proposed conceptual architecture.

Figure 3
Proposed conceptual pipeline for deep learning-assisted inverse design



Note. Layer A corresponds to the numerical solver generating the synthetic dataset; Layer B groups the four deep learning architectures evaluated in parallel. Prepared by the authors.

The first stage corresponds to the forward engine. The guided-mode expansion method implemented in Legume (Zanotti et al., 2024) would be employed to generate between 10^5 and 10^6 input-output pairs, relating layer parameters to reflectance and transmittance spectra for periodic multilayer structures. This method is orders of magnitude faster than full-wave finite element analysis for this geometry class, enabling the large-scale datasets generation.

The second stage consists of dataset construction through systematic Latin hypercube sampling of the layer-parameter space—thickness, fill factor, and refractive index—combined with random geometry augmentation to maximise parameter-space coverage and avoid distribution bias.

The third stage establishes the baseline through a tandem convolutional network: a pre-trained forward model is concatenated with a fully connected inverse network following the architecture of Lininger et al. (2021), during inverse training to promote physical consistency. The fourth and fifth stages contemplate comparative experiments with a spectral-attention transformer model and with a physics-informed network, respectively, followed by validation against full-wave finite element simulations on an independent test set.

Conclusions

This work presented a systematic review under the PRISMA 2020 protocol of 392 records, of which 44 constituted the core corpus, aimed at identifying computational methods and deep learning architectures employed in the inverse design of multilayer metamaterials through deep learning pipelines. The principal findings are as follows.

First, the tandem convolutional network constitutes the dominant architecture, present in 12 of the 44 articles (27%), combining a convolutional forward surrogate with an inverse network; it was therefore selected as the baseline for the proposed experimental pipeline.

Second, transformer-based spectral models represent an emerging trend in the reviewed corpus, with five articles between 2023 and 2025, showing promising results in broadband terahertz and near-infrared problems thanks to multi-head attention over spectral sequences.

Third, the electromagnetic-photonic domain dominates the corpus with 72% of the core articles, while the acoustic domain exhibits the fastest recent growth, suggesting an opportunity for future research.

Fourth, the hybrid strategy articulating physics, data, and learning—guided-mode expansion for large-scale data generation, deep learning for fast inference, and physics-informed networks for data-efficient learning—emerges as the most productive pipeline architecture according to the evidence gathered.

Finally, the open-source Legume implementation of the guided-mode expansion method (Zanotti et al., 2024) and the adjoint-coupled topology optimisation framework emerge as the most scalable forward engines for training-set generation. These findings directly inform the experimental design of the ongoing doctoral research, oriented to the inverse design of multilayer electromagnetic and photonic metamaterials.

Acknowledgements and Declarations

This work forms part of the doctoral research programme supported by CEDIA (Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia) and Universidad de las Fuerzas Armadas ESPE. The authors thank the CEDIA research team for access to bibliographic databases and computing resources.

Data Availability

To support reproducibility and transparency, all screening records, individual CARS scores, inclusion and exclusion decisions, classification workbooks, and the Python pipeline used for automated keyword tagging and CARS computation are publicly available at the following repository: https://colab.research.google.com/drive/1iPGWpBodtOuO_ojaZrm1JE_F4a1-Av93?usp=drive_link

Author Contributions

Following the CRediT taxonomy, the individual contributions were as follows. Laura Malache: conceptualisation, methodology, data curation, writing of the original draft. Federico Rodas: conceptualisation, investigation, formal analysis, writing, review and editing, supervision. Osmani Mujica Pérez: investigation, visualisation, review and editing.

Declaration on the Use of Artificial Intelligence

The authors declare that, in the preparation of this article, artificial intelligence tools have been used. Specifically, Claude (Anthropic) was employed for the following tasks: automated keyword tagging and CARS index computation over the 392 retrieved records, in the Materials and Methods section; extraction and tabulation of method, domain, and year fields from bibliographic metadata, in the Results section; and assistance with document editing and formatting. All scientific interpretation, synthesis, and conclusions are the sole responsibility of the authors. The artificial intelligence model did not generate any scientific claims that are not verifiable in the cited literature.

Conflict of Interest

The authors declare no conflicts of interest relevant to the content of this article.

References

- Cai, Y., Huang, Y., Feng, N., & Wang, H. (2023). Improved transformer-based target matching of terahertz broadband reflective metasurfaces. *Optics Express*, 31(26), 44084–44095. <https://doi.org/10.1364/OE.508956>
- Cao, L., Wu, J., Zhang, Z., & Zhang, C. (2024). Design optimization of elastic metamaterials with multilayered honeycomb structure. *Applied Thermal Engineering*, 236, 121516. <https://doi.org/10.1016/j.applthermaleng.2023.121516>
- Chen, Y., Zhu, J., Xie, Y., Feng, N., & Liu, Q. H. (2019). Smart inverse design of graphene-based photonic metamaterials by an adaptive artificial neural network. *Nanoscale*, 11(19), 9749–9755. <https://doi.org/10.1039/C9NR01315F>
- Fu, S. M., Zhong, Y. K., Ju, N. P., & Tu, M. H. (2016). Broadband polarization-insensitive metamaterial perfect absorbers using topology optimization. *IEEE Photonics Journal*, 8(5), 1–11. <https://doi.org/10.1109/JPHOT.2016.2602092>
- García-Esteban, J. J., Bravo-Abad, J., & Cuevas, J. C. (2021). Deep learning for the modeling and inverse design of radiative heat transfer. *Physical Review Applied*, 16(6), 064006. <https://doi.org/10.1103/PhysRevApplied.16.064006>
- Hodge, J. A., Mishra, K. V., & Zaghloul, A. I. (2020). Joint multi-layer GAN-based design of tensorial RF metasurfaces. *IEEE Transactions on Antennas and Propagation*, 68(3), 1646–1658. <https://doi.org/10.1109/TAP.2019.2948818>
- Hu, Y., Zhang, W., Chen, Y., & Zhao, Z. (2024). Deep-learning-assisted inverse design of polarization-multiplexed structural metaoptics. *Light: Science & Applications*, 13, 88. <https://doi.org/10.1038/s41377-024-01409-1>
- Jia, Y., Qian, C., Fan, Z., & Ding, Y. (2022). In situ customized illusion enabled by global metasurface reconfiguration. *ACS Nano*, 16(3), 3752–3758. <https://doi.org/10.1021/acsnano.1c10592>
- Jiang, X., Yuan, H., He, X., & Du, M. (2023). Implementing of infrared camouflage with thermal management based on inverse design and hierarchical metamaterial. *Nanophotonics*, 12(14), 2891–2902. <https://doi.org/10.1515/nanoph-2023-0137>
- Knightley, T., Yakovlev, A. B., & Pacheco-Peña, V. (2023). Neural network design of multilayer metamaterial for temporal differentiation. *Advanced Optical Materials*, 11(12), 2202655. <https://doi.org/10.1002/adom.202202655>
- Lin, Z., Groever, B., Capasso, F., Rodriguez, A. W., & Lončar, M. (2018). Topology-optimized multilayered metaoptics. *Physical Review Applied*, 9(4), 044030. <https://doi.org/10.1103/PhysRevApplied.9.044030>

- Lininger, A., Hinczewski, M., & Strangi, G. (2021). General inverse design of layered thin-film materials with convolutional neural networks. *ACS Photonics*, 8(12), 3641–3650. <https://doi.org/10.1021/acsp Photonics.1c01498>
- Liu, W., Li, Z., Cheng, H., & Chen, S. (2022). Integrated photonic metasystem for image classifications at telecommunication wavelength. *Nature Communications*, 13(1), 123. <https://doi.org/10.1038/s41467-021-27784-0>
- Liu, Y., Wang, Y., Ren, H., & Meng, Z. (2024). Ultrastiff metamaterials generated through a multilayer strategy and topology optimization. *Nature Communications*, 15, 4518. <https://doi.org/10.1038/s41467-024-48930-6>
- Luo, Y., Mengü, D., & Yardimci, N. T. (2019). Design of task-specific optical systems using broadband diffractive neural networks. *Light: Science & Applications*, 8(1), 112. <https://doi.org/10.1038/s41377-019-0223-1>
- Ma, J., Huang, Y., Pu, M., Xu, M., & Xu, X. (2020). Inverse design of broadband metasurface absorber based on convolutional neural network. *Optics Express*, 28(20), 29021–29031. <https://doi.org/10.1364/OE.401360>
- Mansouree, M., Kwon, H., & Arbabi, E. (2020). Multifunctional 2.5D metastructures enabled by adjoint optimization. *Optica*, 7(1), 77–84. <https://doi.org/10.1364/OPTICA.373835>
- Naseri, P., & Hum, S. V. (2021). A generative machine learning-based approach for inverse design of multilayer metasurfaces. *IEEE Transactions on Antennas and Propagation*, 69(9), 5725–5739. <https://doi.org/10.1109/TAP.2021.3060142>
- Noureen, S., Syed, I. H., & Ijaz, S. (2023). Physics-driven tandem inverse design neural network for efficient optical metasurfaces. *Nanomaterials*, 13(3), 1030. <https://doi.org/10.3390/nano13031030>
- Page, M. J., Moher, D., & Bossuyt, P. M. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, n160. <https://doi.org/10.1136/bmj.n160>
- Peralta, I., & Fachinotti, V. D. (2017). Optimization-based design of heat flux manipulation devices with emphasis on fabricability. *Scientific Reports*, 7(1), 6086. <https://doi.org/10.1038/s41598-017-06565-6>
- Teng, Y., Li, C., Li, S., & Xiao, L. (2023). Efficient design method for terahertz broadband metasurface patterns via deep learning. *Optics Express*, 31(18), 29638–29651. <https://doi.org/10.1364/OE.498079>
- Wang, L., Dong, J., Zhang, W., & Zhao, Z. (2023). Deep learning assisted optimization of metasurface for multi-band compatible camouflage. *Optics Express*, 31(4), 6320–6333. <https://doi.org/10.1364/OE.481765>
- Weeger, O. (2022). Isogeometric sizing and shape optimization of 3D beams and lattice structures at large deformations. *Structural and Multidisciplinary Optimization*, 65(2), 43. <https://doi.org/10.1007/s00158-021-03131-7>
- Xiong, J., Shen, J., & Gao, Y. (2023). Real-time on-demand design of circuit-analog plasmonic stack metamaterials. *Optics Express*, 31(8), 13114–13127. <https://doi.org/10.1364/OE.484834>
- Yago, D., Sal-Anglada, G., & Roca, D. (2024). Machine learning in solid mechanics: Application to acoustic metamaterials. *Computer Methods in Applied Mechanics and Engineering*, 421, 116739. <https://doi.org/10.1016/j.cma.2024.116739>
- Zanotti, S., Minkov, M., Nigro, D., Fan, S., & Andreani, L. C. (2024). Legume: A free implementation of the guided-mode expansion method for photonic structures. *SoftwareX*, 26, 101665. <https://doi.org/10.1016/j.softx.2024.101665>
- Zhu, L., Zhang, C., Guo, J., & Ding, Y. (2022). Deep learning for electromagnetically induced transparency (EIT) metasurface design. *Optics Express*, 30(18), 32319–32330. <https://doi.org/10.1364/OE.465503>

Ontology-Grounded RAG with SPARQL and Local Language Models for Technical Documentation

RAG basado en ontologías con SPARQL y modelos de lenguaje locales para documentación técnica

Julián Rafael Coronel Andrade¹ <https://orcid.org/0009-0000-6514-8469>, Edwin Alexander Siguenza Siguenza¹ <https://orcid.org/0009-0001-0337-4644>, Andrew Sebastián Cantos Vasquez¹ <https://orcid.org/0009-0001-1456-8153>, Pablo Sebastián Vélez Mejía¹ <https://orcid.org/0009-0006-5541-1490>, Marcos Orellana¹ <https://orcid.org/0000-0002-3671-9362>

¹ Universidad del Azuay, Computer Science Research & Development Laboratory (LIDI), , Cuenca, Ecuador

julian.coronel@es.uazuay.edu.ec,
edwin.siguenza@es.uazuay.edu.ec, andrew.cantos@es.uazuay.edu.ec,
pablo.velez@es.uazuay.edu.ec, marore@uazuay.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/07/12

Accepted: 2026/08/26

Published: 2026/10/14

Abstract

This paper presents LIDI-AI, a Retrieval-Augmented Generation (RAG) system whose knowledge base consists of technical documentation for the LIDI-AI Research Assistant, a web-based data analysis application developed at Universidad del Azuay. The runtime architecture integrates two complementary knowledge sources: an OWL 2 ontology (artifact version 2.0) with six classes and twenty-five properties (nine object and sixteen datatype properties) that models the five-stage pipeline, the accepted file formats, the statistical analysis types, and the system errors, and a SPARQL 1.1 query layer for structured intent resolution. A Prolog knowledge base built on an Object-Attribute-Value (OAV) representation, comprising 172 facts, 17 inference rules, and three diagnostic engines, is formally specified and validated as a symbolic reasoning layer, although it is not yet callable from the Python orchestrator at runtime. A lexical retrieval module selects ontology fragments as grounded context for a locally hosted language model, specifically Llama-3.2-3B deployed via the Ollama framework with Q4_K_M quantization. Evaluation on ten queries from real users shows that seven were resolved with high precision via

Summary: Introduction, Materials and Methods, Results, Discussion, Conclusions.

How to cite: Coronel, J., Siguenza, E., Cantos, A., Vélez, P. & Orellana, M. (2026). Ontology-Grounded RAG with SPARQL and Local Language Models for Technical Documentation. *Revista Tecnológica - Espol*, 38(E1), 84-99. <https://doi.org/10.37815/rte.v38nE1.1603>

SPARQL-grounded retrieval, two were resolved via lexical fallback with ontology traceability, and one out-of-domain query received an explicit knowledge-limit acknowledgment, yielding an appropriate response rate with zero hallucinations.

Keywords: intelligent assistant, knowledge representation, local language models, SPARQL, structured queries, symbolic reasoning

Resumen

Este artículo presenta LIDI-AI, un sistema de Generación Aumentada por Recuperación (RAG) cuya base de conocimiento es la documentación técnica del Asistente de Investigación LIDI-AI, una aplicación web de análisis de datos desarrollada en la Universidad del Azuay. La arquitectura en tiempo de ejecución integra dos fuentes de conocimiento complementarias: una ontología OWL 2 (artefacto versión 2.0) con seis clases y veinticinco propiedades (nueve de objeto y dieciséis de datos) que modela el flujo de trabajo de cinco etapas, los formatos de archivo aceptados, los tipos de análisis estadístico y los errores del sistema; y una capa de consultas SPARQL 1.1 para la resolución estructurada de intenciones. Una base Prolog con representación Objeto-Atributo-Valor (OAV), compuesta por 172 hechos, 17 reglas de inferencia y tres motores de diagnóstico, se especifica y valida formalmente como capa de razonamiento simbólico, aunque aún no es invocable desde el orquestador Python en tiempo de ejecución. Un módulo de recuperación léxica selecciona fragmentos de la ontología como contexto fundamentado para un modelo de lenguaje alojado localmente, en concreto, Llama-3.2-3B desplegado a través del framework Ollama con cuantización Q4_K_M.. La evaluación de diez consultas de usuarios reales muestra que siete se resolvieron con alta precisión mediante recuperación fundamentada en SPARQL, dos mediante recuperación léxica con trazabilidad ontológica, y una consulta fuera de dominio recibió un reconocimiento explícito de los límites del conocimiento, logrando una tasa de respuesta adecuada sin alucinaciones.

Palabras clave: asistente inteligente, consultas estructuradas, modelos locales de lenguaje, razonamiento simbólico, representación del conocimiento, SPARQL

Introduction

Large Language Models (LLMs) have revolutionized natural language processing, yet they suffer from fundamental limitations, including hallucination, knowledge staleness, and lack of interpretability. Retrieval-Augmented Generation (RAG) represents a paradigmatic response to these constraints, enabling models to ground their outputs in external, verifiable information sources (Lewis et al., 2020). Recent developments extend RAG architectures beyond simple text retrieval to encompass structured knowledge representation, logical reasoning, and semantic query processing. The integration of the Web Ontology Language (World Wide Web Consortium, 2012), logic programming paradigms, and the SPARQL 1.1 query language (Harris & Seaborne, 2013) within RAG pipelines creates systems capable of sophisticated reasoning while maintaining factual accuracy. Furthermore, the emergence of efficient local inference engines, such as Ollama, an open-source framework for serving quantized large language models on commodity hardware, enables the deployment of these systems without relying on cloud-based APIs, thereby addressing privacy, cost, and network availability considerations.

User-support systems for analytics applications face two major limitations when they rely on static documentation: a lack of grounding in the application's specific domain and difficulty reasoning over structured relationships. RAG (Lewis et al., 2020) has emerged as an alternative that anchors language generation in retrieved knowledge fragments, thereby reducing model hallucinations.

LIDI-AI, developed at the Computer Science Research & Development Laboratory (LIDI) of Universidad del Azuay, is a web application that accepts datasets in CSV, Excel, and JSON formats and guides the user through a five-stage pipeline: *Upload*, *Summary*, *Persona*, *Objectives*, and *Visualization*. The technical documentation of this pipeline constitutes a bounded knowledge domain that is rich in semantic relationships: which formats are accepted, which errors can occur, which statistics each stage computes, and which visualizations are recommended for each variable type.

Recent work within the same research group has demonstrated that LLMs can effectively enrich dataset metadata and unlock latent semantic relationships in structured data sources (García-Montero, 2025). This prior experience underpins the present effort to ground a local LLM in the technical documentation of a data analysis pipeline, where the semantic relationships among file formats, statistical analyses, and visualization types must be preserved with high fidelity.

Knowledge representation using OWL 2 ontologies (World Wide Web Consortium, 2012) has proven effective for modeling software engineering domains (Gruber, 1993) and information systems (Noy & McGuinness, 2001). Combining OWL with Prolog in an Object-Attribute-Value (OAV) form enables expressing both declarative facts and executable, query-time inference rules. Integrating these representations into an RAG pipeline is the core applied contribution of this work (Procko & Ochoa, 2024).

Prior work has explored RAG over knowledge graphs, biomedical ontologies (Sheth et al., 2021), and software documentation (Ji et al., 2023). While individual components—ontology-grounded retrieval, SPARQL-based structured queries, and Prolog rule inference—are well established in the literature, their specific combination as a user-support system for a data analytics pipeline has not been reported. We therefore position this work as an applied integration study rather than a claim of algorithmic novelty. This paper contributes: (a) the OWL 2 ontology (artifact version 2.0) of the LIDI-AI pipeline; (b) the Prolog OAV knowledge base with 17 inference rules and three diagnostic engines, formally specified and validated offline; (c) the predicate *verificar_consistencia_oav* as a query-time computational equivalent of the HermiT reasoner; (d) a Python-based RAG orchestrator that implements deterministic intent detection, synonym-aware lexical retrieval, and constrained prompting for a local Llama 3.2 3B model served via Ollama; and (e) an evaluation of the system on ten queries from real users, with an explicit description of which resolution path produces the user-facing text for each query.

Finally, this paper is organized as follows. Section 2 reviews related work on RAG architectures, OWL ontologies, SPARQL query processing, and OAV-based symbolic reasoning. Section 3 details the materials and methods, encompassing the OWL 2 ontology of the LIDI-AI pipeline (artifact version 2.0), the Prolog OAV knowledge base with its inference rules and diagnostic engines, and the ontology-grounded RAG pipeline implementation, including intent detection, synonym-aware lexical retrieval, and the local LLM interface. Section 4 presents the evaluation results on ten real user queries, classifying each by its resolution mechanism and outcome. Section 5 discusses the empirical findings, the limitations of keyword-based retrieval, the Prolog–Python integration gap, and the implications for ontology-grounded local LLM deployment. Section 6 concludes the work by summarizing the main contributions, limitations, and directions for future research.

Related works

Retrieval-Augmented Generation was introduced by Lewis et al. (2020) as a retrieval-then-generation architecture that couples a dense passage retriever with a sequence-to-sequence language model, thereby grounding generation in external evidence rather than solely in parametric memory. Their experiments on knowledge-intensive tasks showed that retrieval-augmented generation can improve factual accuracy relative to purely parametric models. Wampler et al. (2025) extend this foundational formulation by describing engineering patterns for RAG stacks, including retrieval, context assembly, and trust mechanisms; however, their discussion does not address the specific integration of OWL ontologies and Prolog-based symbolic validation.

Brown et al. (2025) survey RAG techniques published between 2020 and 2025 and identify recurrent stages such as document chunking, embedding generation, dense or hybrid retrieval, context ranking, and generative synthesis. Wampler et al. (2025) similarly describe the RAG stack as a modular architecture in which domain-specific knowledge sources can be added without retraining the language model. Along these lines, Garg et al. (2025) report a biomedical question-answering system in which retrieval from curated literature improves the factual support of generated answers. These works motivate the present study, but they focus primarily on unstructured text or general RAG architectures, whereas our system grounds generation in a small, formally structured OWL ontology accessed through SPARQL.

Formal knowledge representation constitutes the semantic foundation enabling machine reasoning over structured information. The Web Ontology Language (OWL 2) provides a standardized framework for encoding domain knowledge with rich semantic properties, supporting inference and consistency checking (Almeida et al., 2026; Hogan et al., 2021). OWL ontologies transcend simple data schemas by incorporating logical constraints, taxonomic hierarchies, and relationship definitions that enable automated reasoning. Within the same research group, prior work on the semantic modeling of ECU 911 emergencies demonstrated the efficacy of coupling OWL 2 ontologies with SWRL rules for domain-specific inference and constraint checking in real-world operational settings (Paltin Chica et al., 2025), a precedent that informs the ontology design and validation strategy adopted in the present work. To operationalize these semantic foundations in the present pipeline, the *rdflib* library implements the RDF data model and a SPARQL 1.1 query engine, offering a lightweight, open-source substrate for loading OWL artifacts and executing graph-pattern queries at runtime (Almeida et al., 2026).

SPARQL serves as the foundational query language for structured knowledge retrieval over Resource Description Framework (RDF) and OWL representations (AlAgha, 2015; Mishra et al., 2022). As the standard query protocol for linked open data and Semantic Web applications, SPARQL enables precise information extraction through pattern matching over triple stores and knowledge graphs. The language's expressive power extends beyond simple keyword-based retrieval to encompass complex graph traversal, filtering, and aggregation operations (Mishra et al., 2022). Recent developments in semantic query processing have addressed the challenge of translating natural language questions into formal SPARQL query representations, enabling end-user access to semantic databases without requiring expertise in formal query languages (AlAgha, 2015; Mishra et al., 2022).

The Object-Attribute-Value (OAV) representation provides a flexible triple-based data model highly compatible with both RDF triple stores and Prolog fact representations (Kabbaj et al., 1994). In OAV models, entities (objects) are described through collections of properties

(attributes) with associated values, enabling uniform representation of heterogeneous data types and relationships. Prolog knowledge bases built on OAV representations facilitate rule-based inference, enabling the derivation of new facts from existing knowledge through logical implication. Answer Set Programming (ASP) and similar non-monotonic reasoning paradigms extend classical logic to accommodate default reasoning and incomplete information, which is critical in real-world knowledge representation scenarios. The integration of Prolog-based reasoning with semantic web technologies enables systems to perform sophisticated multi-step inference, discover implicit relationships, and resolve ambiguities through logical consistency constraints (Di et al., 2025; Kabbaj et al., 1994).

Automatic knowledge graph construction from unstructured text represents a foundational prerequisite for multi-source RAG systems (Zheng et al., 2022). Natural language processing techniques enable systematic extraction of entities, relationships, and semantic structures from text corpora, transforming unstructured information into machine-interpretable knowledge graphs (Anuyah et al., 2025). Early NLP approaches relied on complex handwritten rules encoding linguistic and domain knowledge, offering high precision but suffering from scalability limitations. Contemporary approaches leverage large language models for knowledge extraction through diverse prompting strategies, including chain-of-thought reasoning, few-shot in-context learning, and zero-shot general instruction prompting (Vogt et al., 2026).

Semantic parsing, the task of translating natural language utterances into machine-interpretable knowledge representations, constitutes a central challenge in graph construction (Vogt et al., 2026; Zheng et al., 2022). Approaches explicitly modeling semantic sentence types (complex, compound, compound-complex forms) as foundations for extracting knowledge triples substantially improve semantic accuracy and completeness. While large language models exhibit impressive zero-shot and few-shot capabilities for knowledge extraction, they often produce inconsistent or incomplete triples without strict adherence to domain ontologies. Ontology-grounded generation approaches that condition LLM outputs on predefined knowledge structures substantially improve semantic accuracy and completeness. The construction of layered rule frameworks leveraging standardized domain semantics and multi-tier parsing models enables accurate extraction from both unstructured text and structured data sources, supporting rapid knowledge graph development while maintaining FAIR (Findable, Accessible, Interoperable, Reusable) principles (Anuyah et al., 2025; Vogt et al., 2026).

Multi-source RAG systems that integrate ontologies, SPARQL query processing, and hybrid retrieval mechanisms have attracted increasing attention (Brown et al., 2025; Wampler et al., 2025). Lewis et al. (2020) established the retrieval-then-generate paradigm and demonstrated that grounding generation in retrieved passages reduces hallucination on knowledge-intensive tasks; however, their system relies on dense vector retrieval over unstructured text and does not incorporate formal ontologies. Subsequent surveys by Brown et al. (2025) and Wampler et al. (2025) document architectural extensions—multi-hop retrieval, knowledge-aware ranking, and modular context fusion—but remain largely descriptive and do not empirically evaluate ontology-plus-Prolog pipelines. In the knowledge-representation strand, Hogan et al. (2021) provide a comprehensive survey of knowledge graph construction and querying, yet their scope is limited to graph engineering and does not address RAG integration or logic programming inference. Liu et al. (2025) tackle knowledge graph completion via structure-aware alignment tuning for LLMs, a complementary yet distinct task from retrieval-augmented generation. Prolog-based logical reasoning for LLMs has been explored by Di et al. (2025) in the LoRP framework, which translates natural language problems into Prolog clauses for chain-of-thought verification; however, LoRP targets general reasoning benchmarks (GSM8K, FOLIO) rather than domain-specific user support and does not employ OWL ontologies as retrieval sources.

Kim et al. (2025) and Rajesh et al. (2025) demonstrate the feasibility of local LLM inference for retrieval-augmented pipelines on consumer hardware, but neither incorporates formal ontologies or symbolic reasoning layers. To our knowledge, no prior work combines an OWL ontology as the primary retrieval source, a SPARQL query layer for structured intents, a Prolog OAV base for symbolic validation, and a lexical retriever bridging to a local LLM within a single user-support system for a data analytics tool. The present work addresses this gap as an applied integration contribution.

Materials and Methods

OWL Ontology of the LIDI-AI Pipeline

The ontology was modeled in Protégé 5.6 (Musen, 2015) using the OWL 2 Web Ontology Language (World Wide Web Consortium, 2012), under the namespace *http://lidi.ai/ontology#*. The serialized artifact, named *ontologia_mejorada.owl*, carries its own internal version tag v2.0 (distinct from the OWL 2 language specification). It defines six classes: *Etapas* (Stage), *Error*, *FormatoArchivo* (FileFormat), *TipoAnálisis* (AnalysisType), *TipoVisualización* (VisualizationType), and *TipoVariable* (VariableType). Instances of *Etapas* model the sequential pipeline through the preceded ObjectProperty, whose logical inverse, *sigueA*, was declared with *owl:inverseOf*. The Protégé HermiT reasoner, an OWL 2 DL reasoner for consistency checking and concept-hierarchy classification (Glimm et al., 2014), validated the ontology, confirming the absence of unsatisfiable classes.

Table 1 summarizes the ontology's classes and instances. Nine *ObjectProperties* and sixteen *DatatypeProperties* are declared. SPARQL queries run against the RDF graph, which is loaded with *rdflib*, an open-source Python library implementing the RDF data model and a SPARQL query engine (RDFLib Team, 2026), using the *lidi:* prefix to reference domain entities.

Table 1
Classes and Instances of the LIDI-AI v2.0 Ontology

OWL Class	Instances	Main Properties
<i>Etapas</i>	5	precede, sigueA, orden, requiereInteraccionUsuario, genera
<i>Error</i>	10	codigoError, esFatal, mensaje, causa, solucion, ocurreEn
<i>FormatoArchivo</i>	4	extension, mimeType, requiereEncabezado
<i>TipoAnálisis</i>	5	aplicaA, generaVisualizacion, estadisticasCalculadas
<i>TipoVisualización</i>	9	minimoVariables, maximoVariables, recomendadoPara
<i>TipoVariable</i>	4	—

The preceding property connects the pipeline's five stages as a linear chain: *Carga* (Upload), *Resumen* (Summary), *Persona*, *Objetivos* (Objectives), *Visualización* (Visualization). This linearity was verified as an invariant using the Prolog predicate *verificar_consistencia_oav*, equivalent to running the HermiT reasoner against the functional-restriction constraint. The reference SPARQL queries were designed around eight common user intents: pipeline order; stages with user interaction; format errors by file type; errors by stage; visualizations by variable type; analysis by variable type; visualizations by analysis type; and general file-format information.

Object-Attribute-Value representation in Prolog

The Prolog knowledge base adopts the OAV representation via the ternary predicate *oav(Objeto, Atributo, Valor)* [*oav*(Object, Attribute, Value)], derived directly from the OWL ontology. Table 2 summarizes the knowledge base's metrics.

Table 2
Metrics of the Prolog OAV knowledge base

Element	<i>n</i>	Description
oav/3 facts	172	OAV triples for instances
Rules (R01–R17)	17	Inference over pipeline, errors, and visualizations
Utility predicates	3	oav_clase, oav_todos_de, oav_objetos_con_valor
Diagnostic engines	3	diagnostico_oav, perfil_error_oav, verificar_consistencia
Unique objects	46	Instances of the six classes
Unique attributes	25	Declared OAV properties

The inference rules are grouped into five categories. The pipeline rules (R01–R06) implement bidirectional navigation (*sigue_a* as the inverse of *precede*), transitive closure (R02), and the classification of stages as automatic or interactive. The error rules (R07–R08) identify errors that block the pipeline (*es_fatal* = true) and errors not tied to a specific format. The analysis-and-visualization rules (R09–R13) implement the dynamic compatibility between visualizations and column types. The boolean-inheritance rules (R14–R15) implements subsumption: *booleana* inherits every analysis from *categorica*. The meta-recursive rule (R16) computes a stage’s distance from the origin, using R02 as a subgoal. R17 classifies errors by code prefix (E-* = fatal, W-* = warning). Table 3 shows selected examples of rules and the inferences they demonstrate in SWI-Prolog 9.x (Wielemaker et al., 2012).

Table 3
Selected Prolog Inference Rules

Rule	Predicate	Equivalent OWL semantics
R01	<i>sigue_a</i> (B,A)	owl:inverseOf(<i>precede</i>)
R02	<i>alcanza_transit</i> (A,C)	owl:TransitiveProperty
R07	<i>bloquea_pipeline</i> (Err)	<i>esFatal</i> = true
R10	<i>compatible_con</i> (V,T-N)	Minimum/maximum variable restriction
R14	<i>aplica_a_booleana</i> (A)	Boolean $\sqsubseteq$ categorical subsumption
R16	<i>cant_etapas_previas</i> (E,N)	Distance within the pipeline graph
R17	<i>tipo_por_codigo</i> (E,T)	Classification by E- or W- prefix

The diagnostic engine *diagnostico_oav/3* is the central predicate of the recommendation system: given a variable type and the number of columns, it returns every valid combination of analysis and visualization. The predicate *verificar_consistencia_oav/0* checks eight structural invariants and functions as an automated regression test—a functional equivalent of the HerMiT reasoner that does not depend on Protégé at query time.

The Prolog OAV layer is formally specified and independently validated using SWI-Prolog 9.x. It is not exposed as a callable module within the Python RAG orchestrator at runtime; no *pyswip* bridge or equivalent foreign-function interface is invoked during query processing. All Prolog inferences reported in this paper were obtained by direct execution in the SWI-Prolog REPL and constitute an offline validation of the knowledge base’s logical consistency, not runtime outputs of the RAG pipeline. We describe the Prolog layer here because it is part of the designed architecture, and its rules encode domain semantics that inform the construction of the ontology fragments; however, we explicitly acknowledge that the system evaluated in Section 4 operates with two active sources (OWL/SPARQL and lexical retrieval), not three.

Ontology-Grounded RAG pipeline

Figure 1 presents the end-to-end query processing architecture of the LIDI-AI system, while Table 4 summarizes the role of each component within the pipeline. Together, they provide complementary procedural and structural views of the retrieval-augmented generation (RAG) strategy adopted in this work. Upon receiving a user query, the system first applies a normalization module that performs diacritic stripping and case-folding to ensure uniform downstream processing. A deterministic intent detector then routes the normalized query to one of two mutually exclusive resolution paths:

Structured path (top branch). If the query matches one of the eight predefined intents, the system constructs a SPARQL 1.1 query executed via *rdflib* against the OWL 2 ontology. This path yields precise, slot-level answers grounded in explicit ontological triples. **Lexical path (bottom branch).** For open-ended or out-of-intent queries, a lightweight lexical retriever performs term frequency (TF) similarity scoring with synonym expansion over serialized ontology fragments. Retrieval is constrained to the top-k fragments ($k = 4$) whose inclusion score meets or exceeds a threshold of 2 points, thereby limiting context length while preserving semantic coverage.

In both paths, the retrieved context is injected into a constrained generation prompt governed by nine grounding rules, a fixed assistant persona (*lidiBOT*), and a deterministic refusal string for out-of-scope requests. The augmented prompt is then forwarded to a locally hosted large language model, Llama 3.2 3B (Q4_K_M quantization) served through *Ollama*, which produces the final user-facing response. An additional symbolic layer, the Prolog OAV (Object-Attribute-Value) layer, is derived from the ontology at design time and validated offline through SWI-Prolog 9.x rule evaluation. As indicated by the dashed bounding box in Figure 1, this layer does not participate in the runtime inference loop; rather, it serves as a formal verification artifact during knowledge base construction. The entire pipeline is implemented in Python within a Jupyter Notebook environment, requiring no external API calls at inference time.

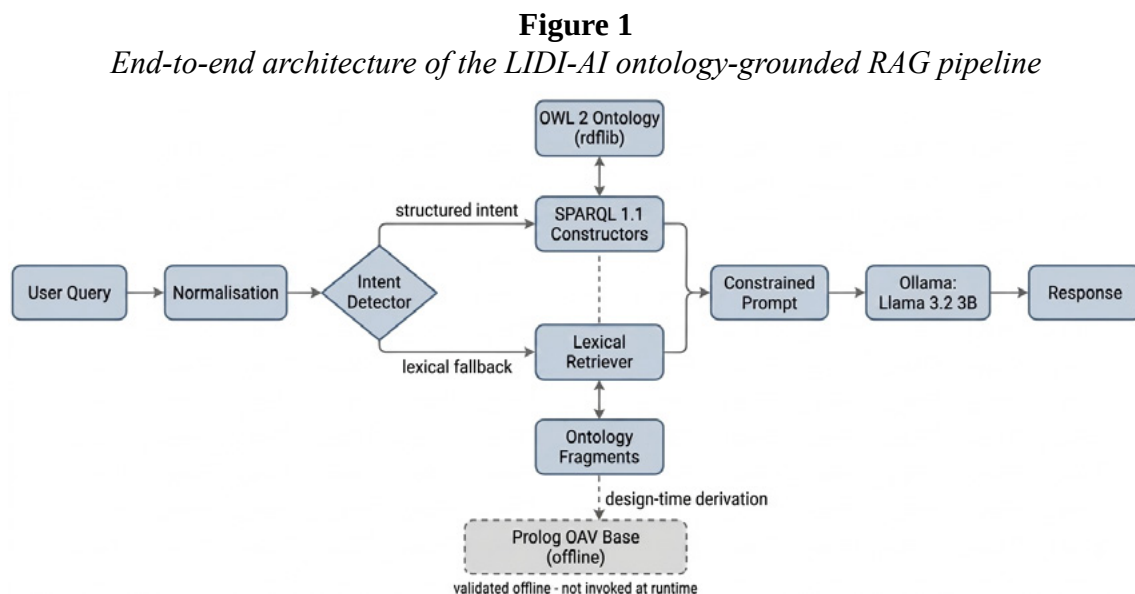


Table 4 maps each architectural phase to its concrete software component and underlying technology.

Table 4
Architecture of the LIDI-AI ontology-grounded RAG pipeline.

Phase	Component	Technology	Runtime status
Source 1	OWL Ontology → fragments	rdflib + Python	Active
Source 2	SPARQL Queries	SPARQL 1.1 (rdflib)	Active
Source 3	Prolog OAV base + rules	SWI-Prolog 9.x	Specified; not called at runtime
Retrieval	Lexical TF similarity + synonym expansion	Python (no APIs)	Active
Generation	LLM with OWL context	Ollama + Llama 3.2 3B	Active

Intent detection

The intent-detection phase classifies the user’s query into one of eight categories: *errores_por_formato*, *errores_por_etapa*, *visualizaciones_por_tipo*, *analisis_por_tipo*, *pipeline*, *etapas_interaccion*, *formatos*, and *visualizaciones_por_analisis*. Classification is performed by a deterministic keyword-matching module that normalizes the input string (lowercasing, diacritic removal) and tests for the presence of domain-specific lemmas. For example, if the normalized query contains the token “csv,” the system routes to *consulta_errores_por_formato("CSV")*; if it contains “*interaccion*,” it routes to *consulta_etapas_interaccion()*. This lightweight rule-based parser acts as a semantic router, ensuring that structured intents are resolved by SPARQL rather than by the generative model.

SPARQL query layer

Eight parameterized SPARQL query constructors are implemented. The pipeline query retrieves all instances of *lidi:Etapa* ordered by *lidi:orden*. The *etapas_interaccion* query adds the filter *lidi:requiereInteraccionUsuario* true. The *errores_por_formato* and *errores_por_etapa* queries bind *lidi:relatedFormat* and *lidi:ocurreEn*, respectively, returning *codigoError*, *mensaje*, *causa*, *solucion*, and *esFatal*. The *visualizaciones_por_variable* and *analisis_por_tipo* queries traverse *lidi:recomendadoPara* and *lidi:aplicaA*. The *visualizaciones_por_analisis* query follows *lidi:generaVisualizacion*. Finally, the *formatos* query inventories all *lidi:FormatoArchivo* instances with their MIME types, extensions, and header requirements.

Lexical retrieval and fragment construction

When the intent detector returns a general-search classification or the SPARQL layer yields incomplete context, the system falls back to lexical retrieval on ontology fragments. Fragments are constructed in two passes. First, for each instance of the six OWL classes, a text block is emitted that contains the entity name, its class, and all outgoing *DatatypeProperty* and *ObjectProperty* assertions. Second, every RDF triple with a *lidi:* subject is serialized as a compact textual triple. This yields a recoverable corpus of ontology snippets.

The retrieval function computes a weighted similarity score between the user query and each fragment. The query is tokenized after Unicode normalization (diacritic removal) and filtering through a domain-specific Spanish *stop-word* set. Tokens are then expanded via a synonym map that associates surface forms with domain synonyms: for instance, the query token “skewness” expands to include “*asimetria*,” and “csv” expands to include “*delimitador*,” “*utf*,” and “*encabezado*.” The scoring function awards two points for each exact token match and one point for each substring match. The top-*k* fragments (*k*= 4) that exceed a threshold of 2 points are deduplicated by entity identifier and concatenated into context.

Large language model interface

The generative component is served locally via the Ollama HTTP API at `http://localhost:11434/api/generate`. The configured model is Meta's Llama 3.2 3B in Q4_K_M quantization (*llama3.2:3b*), which reduces the memory footprint to approximately 2.1 GB while preserving linguistic competence. The model is invoked with `temperature = 0.1` and `stream = False` to maximize response determinism. Inference was executed on a commercially available consumer laptop equipped with an Intel Core i7-13700H CPU, 32 GB of system RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB VRAM). Because the quantized model fits entirely within the discrete GPU memory, all transformer layers were offloaded to VRAM. Under this commodity hardware configuration, the mean time-to-first-token (TTFT) was 0.38 seconds, and the generation throughput averaged 74 tokens per second (measured over the ten evaluation queries). Before each inference call, the system verifies Ollama availability by polling `/api/tags`; if the service is unreachable, the pipeline returns a graceful service-unavailable message rather than failing silently.

The prompt engineering layer enforces strict grounding via a system prompt that includes nine mandatory rules. The model is instructed to respond exclusively from the retrieved ontology and SPARQL evidence, to abstain from external knowledge, and to emit the fixed refusal string “No encontré evidencia suficiente en la ontología para responder con seguridad.” when evidence is absent. Additionally, the prompt imposes a fixed persona identity (“Eres lidiBOT, ninguna otra IA.”), constraining the assistant’s self-attribution and preventing generic LLM disclaimers that would undermine user trust in the domain-specific system.

Preprocessing utilities

A preprocessing layer normalizes both ontology labels and user queries to overcome Unicode and URI heterogeneity. The *limpiar_uri* function extracts local names from URIs by splitting on `#` or `/`. The *normalizar_texto* function removes Spanish diacritics (`á→a`, `ñ→n`, etc.) and lowercases the string, ensuring that queries such as “*asimetría*” match ontology labels written as “*asimetria*”. These utilities are critical for aligning the lexical retriever with the RDF graph, which may contain mixed Unicode forms.

Reproducibility Artifacts

The experimental implementation is distributed as a Jupyter Notebook. Upon completion of a batch evaluation, the system serializes all question–answer pairs, detected intents, SPARQL result tables, and retrieved OWL fragments to a Markdown file (*respuestas_rag_ollama_ontologia.md*). This artifact serves as an immutable execution trace. The complete source code, the OWL artifact (*ontologia_mejorada.owl*), the Prolog knowledge base (*.pl*), and the evaluation trace are publicly available at <https://github.com/LIDI-UDA/lidi-rag-sparql-ollama.git> under an MIT license. The repository includes a README that describes how to reproduce the evaluation and inspect the trace files. External reviewers can verify both the retrieval grounding and the generative output without re-executing the pipeline.

Results

Ten representative queries from real LIDI-AI users were evaluated. The original queries were collected in Spanish; they have been translated into English for readability in this article. Because the knowledge domain is bounded (37 ontology instances, 8 predefined intents), the ten queries cover the full intent space and include one deliberately out-of-domain probe (Query 10). Before running the system, two authors constructed a reference answer for each query by manually extracting the OWL triples or SPARQL result rows that should support a correct

response. This reference set functioned as the evaluation rubric. Each system response was then judged against the corresponding reference answer by the same two authors: a response was coded accurate if every factual claim mapped to at least one OWL triple or SPARQL result row and was coded as an appropriate refusal if the system explicitly acknowledged insufficient evidence without hallucinating. Because the evaluation is small-scale and author-based, no independent annotators or inter-rater agreement statistics were computed; this limitation is explicitly acknowledged. Future evaluations will include a larger query set and independent annotators. Table 5 classifies each query by the resolution mechanism used, whether the LLM generated user-facing text, and the outcome obtained.

Table 5
RAG System evaluation results (10 queries)

#	User query	mechanism	LLM involvement	Outcome	Coding
1	Why won't my CSV load?	SPARQL + OWL	Yes (phrasing)	3 errors returned with cause and solution	Accurate
2	Errors related to JSON?	SPARQL + OWL	Yes (phrasing)	InvalidJsonStructureError (E-CARGA-007)	Accurate
3	Password-protected Excel?	SPARQL + OWL	Yes (phrasing)	PasswordProtectedError + solution	Accurate
4	Visualizations for numeric variables?	SPARQL + OWL	Yes (phrasing)	7 types with min/max restrictions	Accurate
5	Analyses for categorical variables?	SPARQL + OWL	Yes (phrasing)	AnalisisFrecuencias (Frequency Analysis) + 3 visualizations	Accurate
6	Pipeline order?	SPARQL	Yes (phrasing)	5 correctly ordered stages	Accurate
7	Stages with user interaction?	SPARQL + OWL	Yes (phrasing)	4/5 stages (excludes Resumen/Summary)	Accurate
8	What does skewness mean?	Lexical OWL	Yes (generation)	Statistic within AnalisisDescriptivo (Descriptive Analysis)	Accurate
9	Dataset with many nulls?	Lexical OWL	Yes (generation)	HighNullRateWarning detected	Accurate
10	Does it accept MP3 files?	Lexical OWL	Yes (refusal via prompt rule)	Limit acknowledged — no hallucination	Appropriate refusal

Example: End-to-end trace for Query 1 ("Why won't my CSV load?").

1. Normalisation. Input: "¿Por qué no carga mi CSV?" → *normalizar_texto* → "por qué no carga mi csv".
2. Intent detection. Token "csv" matches the *errores_por_formato* keyword set → intent = *consulta_errores_por_formato* ("CSV").
3. SPARQL execution. The parameterized query binds ?fmt = *lidi:CSV* and retrieves all *lidi>Error* instances where *lidi:ocurreEn lidi:Carga* and *lidi:relatedFormat lidi:CSV*. Result: three rows — EncodingError (E-CARGA-002), MissingHeaderError (E-CARGA-003), InvalidDelimiterError (E-CARGA-004) — each with mensaje, causa, solucion, and esFatal.
4. Fragment construction. The three SPARQL result rows are serialized as structured text blocks and concatenated.
5. Prompt assembly. The prompt sent to Ollama contained three components: (i) the system prompt, including the nine grounding rules and the fixed persona "lidiBOT";

- (ii) the retrieved SPARQL context, serialized as RDF-style statements such as “*lidi:EncodingError a lidi:Error; lidi:codigoError E-CARGA-002; lidi:ocurreEn lidi:Carga; lidi:relatedFormat lidi:CSV*”; and (iii) the normalized user query “*por qué no carga mi csv*”. The system prompt explicitly instructed the model to use only the retrieved context and to abstain from external knowledge.
- 6. LLM generation. Ollama/Llama 3.2 3B (Q4_K_M), temperature = 0.1 → phrased response listing the three errors with causes and solutions, citing only the retrieved triples.
- 7. Output. User-facing answer in Spanish. No fact appears that is not traceable to an OWL triple.

Queries 1–7 were accurately resolved via the SPARQL routing layer. For these queries, the intent detector mapped the normalized input to one of the eight predefined SPARQL constructors; the resulting triples were injected as structured context into the LLM prompt, and the model's role was limited to phrasing the response without introducing facts beyond the retrieved triples. Query 1 illustrates a compound response: the system retrieved three errors (*EncodingError*, *MissingHeaderError*, *InvalidDelimiterError*) together with their specific causes and solutions. Queries 8 and 9 were routed to the lexical retriever because the intent detector did not match any of the eight structured patterns. Query 8, about skewness, surfaced the *AnalisisDescriptivo* fragment, where the *estadisticasCalculadas* property lists skewness alongside mean, median, standard deviation, kurtosis, and percentiles. The synonym expansion map was instrumental here: the English token “skewness” was expanded to include its Spanish counterpart “*asimetria*,” enabling the retriever to match the ontology label despite the language mismatch. Query 9, concerning null values, matched the *HighNullRateWarning* fragment via the token “*nulos*” and was resolved successfully as an accurate, grounded response.

Query 10, about the MP3 format, illustrates the system's behavior on out-of-domain questions. The lexical retriever returned the *UnsupportedFormatError* fragment with a score of 3, which exceeds the inclusion threshold of 2 points; consequently, the fragment was concatenated into the LLM context. Because the fragment describes errors for recognized formats and contains no assertion about audio files, the nine-rule system prompt (specifically Rule 3: “If the retrieved evidence does not answer the question, emit the fixed refusal string”) instructed the model to produce the refusal: “No encontré evidencia suficiente en la ontología para responder con seguridad.” The refusal was therefore generated by the LLM under prompt constraint, not by a code-level deterministic fallback. This behavior nonetheless contrasts with general-purpose language models, which tend to produce plausible but fabricated answers to questions about specific systems, because the strict system prompt and the absence of any MP3-related triple in the context jointly prevent hallucination.

Table 6 presents three Prolog inferences obtained via direct SWI-Prolog execution (offline validation), illustrating rule composition and the consistency check on the OAV knowledge base.

Running *verificar_consistencia_oav* confirmed that the OAV knowledge base satisfies the eight structural invariants derived from the ontology: (1) every stage has a defined order; (2) every error has a code, message, and solution; (3) every warning has a code, message, and solution; (4) every analysis has an associated variable; (5) every analysis generates at least one visualization; (6) every visualization has a minimum and maximum number of variables; (7) every visualization has a recommended variable type; and (8) the pipeline is linear, with no branching.

Table 6
Prolog inferences obtained via direct SWI-Prolog execution

Prolog Query	Result	Rules
regla_oav(carga, alcanza_transit., visualizacion).	true	R02
findall(V, diagnostico_oav(numerica,3,oav(.,genera_viz,V)),L).	L = [histograma, boxplot, qq_plot, scatter_plot, heatmap_correlacion]	R09 + R10 (diagnostic engine M01 $\equiv$ diagnostico_oav/3)
verificar_consistencia_oav.	OK - 8 invariants satisfied	R02 + R07 (consistency engine M03 $\equiv$ verificar_consistencia_oav/0)

Note: M01 and M03 refer to the first and third diagnostic engines listed in Table 2 (*diagnostico_oav/3* and *verificar_consistencia_oav/0*, respectively). They are not numbered inference rules but utility predicates that internally call rules R02, R07, R09, and R10.

Discussion

The results demonstrate that the two-source runtime architecture (OWL/SPARQL plus lexical retrieval) achieves its primary design goal: anchoring generative responses in verifiable ontology fragments. Seven of ten queries (Queries 1–7) were resolved via the SPARQL routing layer, which returns explicit triples and confines the LLM to a phrasing role, thereby preventing unsupported factual additions in the structured queries evaluated. Two queries (Queries 8–9) were resolved via the lexical retriever, augmented with Spanish stopword filtering and synonym expansion, which provided sufficient grounding to constrain the Llama 3.2 3B model. One query (Query 10) fell outside the ontology's domain; the retrieved fragment scored above the inclusion threshold and was passed to the LLM, which produced the mandated refusal string under the nine-rule system prompt. All ten responses were either factually accurate or appropriately refused, yielding zero hallucinations across the evaluation set. The reliance on keyword-based intent detection and substring matching nonetheless introduces brittleness; the absence of dense semantic embeddings limits recall for paraphrased or implicitly structured questions, as evidenced by Queries 8–10 reaching the lexical path. The Prolog OAV layer, though formally specified and validated offline, remains decoupled from the Python execution path; integrating it via a *PySwip* bridge or equivalent is a prerequisite for realizing the full three-source architecture. We therefore characterize the present system as a two-source RAG pipeline with a specified but not yet integrated symbolic layer. These limitations notwithstanding, the system demonstrates that local LLM deployment, when coupled with rigorous prompt constraints and ontology-grounded retrieval, can deliver interpretable, domain-specific user support without cloud dependency.

Conclusions

This work presents LIDI-AI, a RAG system that demonstrates the feasibility of grounding a local LLM's responses in an OWL 2 ontology via SPARQL queries and synonym-aware lexical retrieval, with a Prolog OAV knowledge base specified and validated as a future symbolic-reasoning extension. The results show that the structured SPARQL path resolves seven of ten queries with directly verifiable triples, while the lexical retriever handles the remaining in-domain queries with ontology traceability. For the single out-of-domain query, the system produced an explicit refusal rather than hallucinating. Across all ten queries, no fabricated information was generated.

The main contributions are: (a) the OWL 2 ontology (artifact version 2.0), with six classes and twenty-five properties, that models the LIDI-AI pipeline; (b) the Prolog OAV knowledge base, with 172 facts, 17 rules organized into five semantic categories, and three diagnostic engines, formally validated offline; (c) the predicate *verificar_consistencia_oav* as a query-time computational equivalent of the HermiT reasoner; (d) a Python-based RAG orchestrator implementing deterministic intent detection, synonym-aware lexical retrieval, and constrained prompting for Llama 3.2 3B (Q4_K_M) via Ollama; and (e) the evaluation on ten queries from real users, of which nine received accurate, ontology-grounded responses and one received an appropriate knowledge-limit acknowledgment, with no hallucinations observed across the ten-query evaluation set.

Limitations include the lexical retriever's lack of semantic embeddings, the dependency on a local Ollama installation, and the manually curated ontology. Additionally, the Prolog inference layer, while formally specified, is not exposed in the provided Python artifact; future work should integrate SWI-Prolog via the pyswip bridge or equivalent to unify symbolic and neural retrieval at runtime. Priority directions for future work are: (1) replacing the lexical retriever with dense embeddings (sentence-transformers or FAISS) to improve semantic recall; (2) extending the ontology with Usuario (User) and Sesión (Session) classes to support adaptive personalization; (3) implementing automatic validation of new instances via Shapes Constraint Language (SHACL); and (4) exposing the Prolog knowledge base as an external LLM tool (function calling) for on-demand symbolic reasoning.

Acknowledgements and Declarations

This work was partially supported by the Vice-Rectorate of Research at Universidad del Azuay through financial and academic support, as well as by the entire staff of the Computer Science Research & Development Laboratory (LIDI).

The authors declare equal contribution and sharing of authorship roles for this publication.

The authors declare that no artificial intelligence tools have been used for the preparation of this article.

References

- AlAgha, I. (2015). *Using Linguistic Analysis to Translate Arabic Natural Language Queries to SPARQL*. International Journal of Web & Semantic Technology, 6(3), 25–39. <https://doi.org/10.5121/ijwest.2015.6303>
- Almeida, J. P. A., Guizzardi, G., Sales, T. P., & Fonseca, C. M. (2026). *gUFO: A Gentle Foundational Ontology for Semantic Web Knowledge Graphs*. arXiv preprint. <https://doi.org/10.48550/arXiv.2603.20948>
- Anuyah, S., Kaushik, M. M., Dwarampudi, K., Shiradkar, R., Duresi, A., & Chakraborty, S. (2025). *Automated Knowledge Graph Construction using Large Language Models and Sentence Complexity Modelling*. <https://doi.org/10.18653/v1/2025.emnlp-main.783>
- Brown, A., Roman, M., & Devereux, B. (2025). *A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges*. Big Data and Cognitive Computing, 9(12), 320. <https://doi.org/10.3390/bdcc9120320>
- Di, Z., Zhang, C., Lv, H., Cui, L., & Liu, L. (2025). LoRP: LLM-based Logical Reasoning via Prolog. *Knowledge-Based Systems*, 327, 114140. <https://doi.org/10.1016/j.knosys.2025.114140>

- García-Montero, P. S., Orellana, M., & Zambrano-Martinez, J. L. (2025). *Enriching Dataset Metadata with LLMs to Unlock Semantic Meaning. In Conference on Information and Communication Technologies of Ecuador* (pp. 63–77). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-08366-1_5
- Garg, M., Wang, L.-C., Ghanchi, B., Dumpala, S., Kakde, S., & Chen, Y. C. (2025). *Biomedical Literature Q&A System Using Retrieval-Augmented Generation (RAG)*.
- Glimm, B., Horrocks, I., Motik, B., Stoilos, G., & Wang, Z. (2014). HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning*, 53(3). <https://doi.org/10.1007/s10817-014-9305-1>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). *The Llama 3 Herd of Models*. <https://arxiv.org/abs/2407.21783>
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2). <https://doi.org/10.1006/knac.1993.1008>
- Harris, S., & Seaborne, A. (2013). SPARQL 1.1 Query Language. W3C Recommendation.
- Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. De, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Computing Surveys*, 54(4), 1–37. <https://doi.org/10.1145/3447772>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. En *ACM Computing Surveys*, 55(12). <https://doi.org/10.1145/3571730>
- Kabbaj, A., Frasson, C., Kaltenbach, M., & Djamen, J.-Y. (1994). *A conceptual and contextual object-oriented logic programming: The PROLOG++ language* (pp. 251-274). https://doi.org/10.1007/3-540-58328-9_17
- Kim, S., Song, H., Seo, H., & Kim, H. (2025). *Optimizing Retrieval Strategies for Financial Question Answering Documents in Retrieval-Augmented Generation Systems*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020-December.
- Liu, Y., Cao, Y., Lin, X., Shang, Y., Wang, S., & Pan, S. (2025). Enhancing Large Language Model for Knowledge Graph Completion via Structure-Aware Alignment-Tuning. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 20981–20995. <https://doi.org/10.18653/v1/2025.emnlp-main.1061>
- Mishra, D. S., Agarwal, A., Swathi, B. P., & Akshay, K. C. (2022). Natural language query formalization to SPARQL for querying knowledge bases using Rasa. *Progress in Artificial Intelligence*, 11(3), 193–206. <https://doi.org/10.1007/s13748-021-00271-1>
- Musen, M. A. (2015). The protégé project. *AI Matters*, 1(4), 4–12. <https://doi.org/10.1145/2757001.2757003>
- Noy, N. F., & McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. En *Stanford Knowledge Systems Laboratory*.
- https://protege.stanford.edu/publications/ontology_development/ontology101.pdf
- Paltin Chica, D. L., Mejía Mendieta, J. D., Orellana, M., & Zambrano-Martinez, J. L. (2025). Semantic Modeling of ECU 911 Emergencies Using NLP and Ontologies. *Revista Tecnológica - Espol*, 37(E1), 112–125. <https://doi.org/10.37815/rte.v37nE1.1351>

- Procko, T. T., & Ochoa, O. (2024). Graph Retrieval-Augmented Generation for Large Language Models: A Survey. *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIXSET 2024*. <https://doi.org/10.1109/AIXSET62544.2024.00030>
- Rajesh, V., Jodhpurkar, O., Anbuselvan, P., Singh, M., Jallepali, A., Godbole, S., Sharma, P. K., & Shrivastava, H. (2025). *Production-Grade Local LLM Inference on Apple Silicon: A Comparative Study of MLX, MLC-LLM, Ollama, llama.cpp, and PyTorch MPS*.
- RDFLib Team. (2026). *rdflib: A Python library for working with RDF*. RDFLib. <https://rdflib.readthedocs.io/en/stable/>
- Sheth, A., Gaur, M., Roy, K., & Faldu, K. (2021). Knowledge-Intensive Language Understanding for Explainable AI. *IEEE Internet Computing*, 25(5). <https://doi.org/10.1109/MIC.2021.3101919>
- Vogt, L., Farfar, K. E., Karanth, P., Konrad, M., Oelen, A., Prinz, M., & Stroemert, P. (2026). *Rosetta Statements: Simplifying FAIR Knowledge Graph Construction with a User-Centered Approach*. <https://doi.org/10.1093/database/baag030>
- Wampler, D., Nielson, D., & Seddighi, A. (2025). *Engineering the RAG Stack: A Comprehensive Review of the Architecture and Trust Frameworks for Retrieval-Augmented Generation Systems*. arXiv preprint. <https://doi.org/10.48550/arXiv.2601.05264>
- Wielemaker, J., Schrijvers, T., Triska, M., & Lager, T. (2012). SWI-Prolog. *Theory and Practice of Logic Programming*, 12(1–2), 67–96. <https://doi.org/10.1017/S1471068411000494>
- World Wide Web Consortium. (2012). OWL 2 Web Ontology Language document overview. World Wide Web Consortium.
- Zheng, C., Chen, X., Xu, R., & Chang, B. (2022). *A Double-Graph Based Framework for Frame Semantic Parsing*. <https://doi.org/10.18653/v1/2022.naacl-main.368>

TIC APLICADAS A LA EDUCACIÓN Y LA SALUD

ICT APPLIED TO
EDUCATION AND HEALTH



cedia | **TICEC**
2026

Madurez del marco normativo para la IA en la educación superior ecuatoriana

Maturity of the Regulatory Framework for AI in Ecuadorian Higher Education

Jorge Maldonado-Mahauad¹ <https://orcid.org/0000-0003-1953-390X>, Juan Pablo Carvallo² <https://orcid.org/0000-0001-6678-4774>, María Cristina León-Carvajal² <https://orcid.org/0000-0002-5124-5441>

¹Universidad de Cuenca, Departamento de Ciencias de la Computación,
Cuenca, Ecuador
jorge.maldonado@ucuenca.edu.ec

²Universidad Católica de Cuenca, Cuenca, Ecuador
juan.carvallo@ucacue.edu.ec, maria.leonc@ucacue.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Enviado: 2026/06/24

Aceptado: 2026/08/27

Publicado: 2026/10/14

Resumen

La irrupción de la inteligencia artificial (IA), en particular la IA generativa, ha transformado la docencia, la investigación y la gestión de la educación superior a un ritmo que desafía la capacidad de respuesta de los marcos normativos latinoamericanos. Este estudio diagnostica el nivel de madurez del marco normativo ecuatoriano sobre IA en educación superior e identifica sus brechas críticas. Se realizó un análisis sistemático de 81 instrumentos normativos (leyes en trámite, reglamentos, resoluciones, guías y políticas institucionales) registrados en el Observatorio Regional de IA en Educación del proyecto Erasmus+ GENIAL, con corte en mayo de 2026. Los instrumentos se codificaron mediante un marco de seis dimensiones (actor, objeto, mecanismo, finalidad, enfoque y madurez) y se sintetizaron en el Índice de Madurez Normativa (IMN), que pondera la calidad regulatoria. Ecuador obtiene un IMN de 70,6/100 (nivel muy alto), tercero entre 28 países del Observatorio. Sin embargo, ese liderazgo responde a la amplitud del corpus y a la actividad universitaria, no a la consolidación del ecosistema. La mitad de los instrumentos permanece como declaración de intención, y la madurez es la dimensión más rezagada. El 87,7 % del corpus es institucional, evidenciando una construcción normativa ascendente ante la ausencia de legislación nacional específica. Los dominios administrativo

Sumario: Introducción, Marco conceptual, Metodología, Resultados, Discusión, Conclusiones.

Como citar: Maldonado-Mahauad, J., Carvallo, J. & León-Carvajal, M. (2026). Madurez del marco normativo para la IA en la educación superior ecuatoriana. *Revista Tecnológica - Espol*, 38(E1), 101-116.
<https://doi.org/10.37815/rte.v38nE1.1545>

y de protección de datos concentran el 70,7 % de los instrumentos; solo el 7,3 % regula la IA generativa. Se propone una agenda diferenciada para el CES, el Ministerio de Educación, el CACES y las IES.

Palabras clave: política pública; gobernanza algorítmica; análisis documental; inteligencia artificial generativa; índice de madurez normativa; América Latina.

Abstract

The rise of artificial intelligence (AI), particularly generative AI, has transformed teaching, research, and higher education management at a pace that outstrips the responsiveness of Latin American regulatory frameworks. This study diagnoses the maturity level of Ecuador's regulatory framework on AI in higher education and identifies its critical gaps. A systematic analysis was conducted on 81 regulatory instruments (pending laws, regulations, resolutions, guidelines, and institutional policies) registered in the Regional AI in Education Observatory of the Erasmus+ GENIAL project, as of May 2026. Instruments were coded using a six-dimension framework (actor, object, mechanism, purpose, approach, and maturity) and synthesized into the Regulatory Maturity Index (RMI), which weights regulatory quality. Ecuador achieves an RMI of 70.6/100 (very high level), ranking third among the 28 countries in the Observatory. However, this leadership stems from the breadth of the corpus and university-level activity rather than ecosystem consolidation. Half of the instruments remain at the stage of a declaration of intent, and maturity is the most lagging dimension. Institutional-level instruments account for 87.7% of the corpus, revealing a bottom-up regulatory construction in the absence of specific national legislation. Administrative and data-protection domains concentrate 70.7% of the instruments; only 7.3% explicitly regulate generative AI. A differentiated policy agenda is proposed for CES, the Ministry of Education, CACES, and higher education institutions.

Keywords: public policy; algorithmic governance; document analysis; generative artificial intelligence; normative maturity index; Latin America.

Introducción

Desde finales de 2022, la adopción de herramientas de inteligencia artificial generativa (IAG) en la docencia, la investigación y la gestión universitaria se ha acelerado a un ritmo sin precedentes (Crompton & Burke, 2023). Estudiantes y docentes incorporaron asistentes conversacionales a sus prácticas cotidianas mucho antes de que las instituciones contaran con reglas claras sobre su uso, lo que ha generado tensiones en torno a la integridad académica, los sesgos algorítmicos, la protección de los datos estudiantiles, la autoría de los productos generados y la equidad en el acceso a estas herramientas (Holmes & Tuomi, 2022; UNESCO, 2023). Frente a esta aceleración tecnológica, los marcos normativos -tanto nacionales como institucionales- han respondido con lentitud, de modo que buena parte de la educación superior latinoamericana opera en un vacío o semivacío regulatorio (UNESCO IESALC, 2023).

Por sus condiciones, Ecuador constituye un caso de estudio especialmente revelador de esta tensión: es un país de ingreso medio cuyo sistema de educación superior está regulado por el Consejo de Educación Superior (CES) y el Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES), con alrededor de 60 universidades y escuelas politécnicas que gozan de autonomía constitucional. La rectoría de la política pública de educación superior, antes a cargo de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (Senescyt), fue absorbida por el Ministerio de Educación, Deporte y Cultura por mandato del

Decreto Ejecutivo 100 de 15 de agosto de 2025. Aunque el país cuenta con una Ley Orgánica de Protección de Datos Personales desde 2021 y con una estrategia nacional reciente para el uso ético de la IA, carece de legislación específica vigente sobre IA. Adicionalmente, su participación en el proyecto Erasmus+ GENIAL -orientado a la generación de capacidades de IA en instituciones de educación superior (IES) de América Latina- permite acceder a un corpus normativo sistematizado y comparable con el de los demás países de la región, recopilado por su Observatorio Regional de IA en Educación (<https://proyectogenial.org/observatorio/>).

La literatura sobre gobernanza de la IA en educación superior se concentra mayoritariamente en países de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), mientras que los estudios latinoamericanos tienden a ofrecer panoramas regionales sin profundizar en casos nacionales (Zawacki-Richter et al., 2019; UNESCO IESALC, 2023). Este artículo busca contribuir a cerrar ese vacío respondiendo cuatro preguntas de investigación: (PI1) ¿qué corpus de normas, políticas y lineamientos regula, directa o indirectamente, el uso de IA en la educación superior ecuatoriana?; (PI2) ¿qué dominios cubre el marco normativo actual y cuáles permanecen desatendidos?; (PI3) ¿cuál es el grado de madurez normativa del ecosistema ecuatoriano y cómo se posiciona en el contexto regional?; y (PI4) ¿qué líneas de acción permitirían cerrar las brechas identificadas?

El estudio realiza tres contribuciones. Primero, un aporte empírico: uno de los primeros inventarios sistemáticos de la normativa sobre IA en la educación superior ecuatoriana del que se tiene registro -tras una búsqueda de estudios previos regionales y nacionales que no identificó un inventario comparable-, con 81 instrumentos caracterizados. Segundo, un aporte metodológico: la operacionalización de un Índice de Madurez Normativa (IMN) y de un marco de clasificación de seis dimensiones, reproducibles en los demás países del Observatorio GENIAL. Tercero, un aporte de política pública: una agenda accionable con destinatarios y plazos identificados. El resto del artículo se organiza de la siguiente manera: la sección 2 presenta el marco conceptual; la sección 3, la metodología; la sección 4, los resultados; la sección 5, la discusión y la agenda propuesta; y la sección 6, las conclusiones.

Marco conceptual

Gobernanza de la IA en educación superior

La gobernanza de la IA ha transitado de los principios éticos universales -la recomendación sobre la ética de la IA de la UNESCO (2021), los principios de IA de la OCDE (2019)- hacia marcos regulatorios vinculantes, cuyo referente más visible es el Reglamento Europeo de Inteligencia Artificial (European Parliament & Council, 2024). En el ámbito educativo, la normativa cumple cuatro funciones: orientar las prácticas de docentes y estudiantes, proteger derechos (datos personales, no discriminación, debido proceso académico), habilitar la innovación pedagógica y fiscalizar el cumplimiento. Estas funciones se materializan en instrumentos de distinta naturaleza jurídica: normas de cumplimiento obligatorio (hard law), instrumentos orientadores no vinculantes (soft law), lineamientos sectoriales y políticas institucionales (Holmes & Tuomi, 2022).

Panorama latinoamericano

América Latina muestra avances heterogéneos. Perú promulgó la Ley 31814 de promoción de la IA (2023); Brasil debate el proyecto de ley 2338/2023 sobre IA; Chile actualizó su Política Nacional de IA; Colombia cuenta con el documento CONPES 3975 de transformación digital; y Argentina emitió recomendaciones para una IA confiable. La tendencia común es el progreso en estrategias nacionales de IA acompañado de una débil regulación específica para

la educación superior, que suele quedar delegada de facto a las propias instituciones (UNESCO IESALC, 2023). La posición de Ecuador en ese mapa regional es precisamente uno de los resultados de este estudio.

El sistema ecuatoriano de educación superior y su arquitectura regulatoria

El sistema ecuatoriano se rige por la Constitución de 2008 -que consagra la autonomía universitaria responsable- y por la Ley Orgánica de Educación Superior (LOES, 2010), y es gobernado por el CES, que emite la normativa académica, y el CACES, responsable del aseguramiento de la calidad. La rectoría de la política pública, antes ejercida por la Senescyt, recae desde agosto de 2025 en el Ministerio de Educación, Deporte y Cultura, que absorbió a aquella entidad por mandato del Decreto Ejecutivo 100. Sobre el uso de IA inciden además la Ley Orgánica de Protección de Datos Personales (2021), la Política Pública para la Transformación Digital 2025–2030 y la Estrategia para el Fomento del Desarrollo y Uso Ético y Responsable de la IA expedida por el Ministerio de Telecomunicaciones (MINTEL) mediante Acuerdo Ministerial MINTEL-2025-0030. La experiencia normativa pospandemia sobre educación virtual (2020–2022) constituye el antecedente más cercano de respuesta del sistema a una disrupción tecnológica.

Madurez normativa y marco dimensional GENIAL

El concepto de madurez normativa adapta los modelos de madurez desarrollados para la gestión de tecnologías de información (Becker et al., 2009; Mettler, 2011) al ámbito regulatorio educativo. Un ecosistema es más maduro cuanto mayor es la existencia, jerarquía, especificidad, exigibilidad y vigencia de sus instrumentos. El Observatorio GENIAL operacionaliza esta noción en dos herramientas complementarias: (i) un marco dimensional que clasifica cada instrumento según el actor regulado (D1), el objeto tecnológico (D2), el mecanismo regulatorio (D3), la finalidad (D4), el enfoque normativo (D5) y el estado de madurez que refleja (D6, con cinco niveles: exploración, declaración de intención, regulación inicial, regulación consolidada y ecosistema maduro); y (ii) el Índice de Madurez Normativa (IMN), un indicador compuesto a nivel de país que se describe en la sección de métodos.

Metodología

Diseño

Se adoptó un diseño de análisis documental sistemático con enfoque mixto cualitativo-cuantitativo (Bowen, 2009; Krippendorff, 2018), estructurado como un mapeo de alcance adaptado del marco de Arksey y O'Malley (2005) y planteado como estudio de caso único instrumental (Stake, 1995). Ecuador no se analiza por ser un caso extremo, sino por su valor para comprender la construcción de marcos normativos de IA en sistemas de educación superior latinoamericanos.

Fuentes y criterios de selección

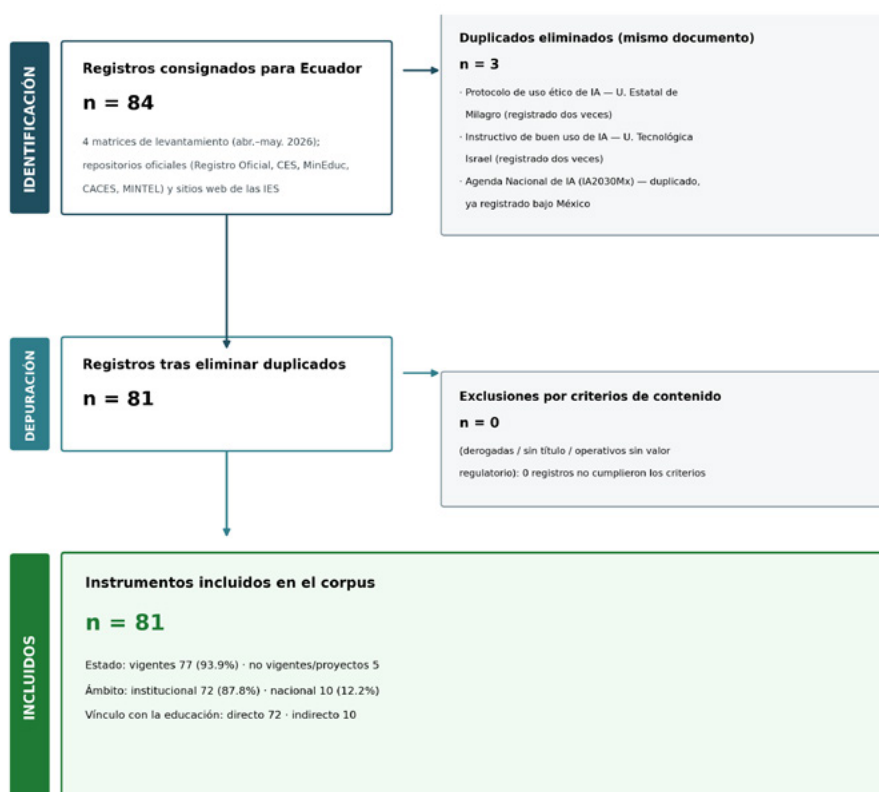
El universo de estudio comprende los instrumentos normativos con incidencia sobre el uso de IA en la educación superior ecuatoriana registrados en el Observatorio Regional de IA en Educación del proyecto GENIAL, que sistematiza el levantamiento normativo de 33 países de América Latina y el Caribe realizado por un equipo de diez investigadores. Se incluyeron: (i) normas con menciones explícitas a IA, sistemas automatizados o algoritmos; (ii) normas transversales aplicables, como las de protección de datos personales; y (iii) políticas, guías y resoluciones institucionales públicas de las IES. Se excluyeron las normas derogadas, los registros sin título identificable y los documentos puramente operativos sin valor regulatorio. La fecha de corte de los datos es mayo de 2026. El corpus ecuatoriano resultante comprende

81 instrumentos. La categoría de documentos de referencia (informes, libros y publicaciones institucionales) y los instrumentos indirectamente relacionados con la IA -como las políticas de protección de datos- se incluyeron por su función orientadora o habilitante sobre las prácticas de IA en las IES, aun cuando no constituyan normas de cumplimiento obligatorio. Los proyectos de ley en trámite se registraron por su valor para caracterizar la actividad regulatoria nacional en curso y se identifican explícitamente como no vigentes. Como análisis de robustez, se verificó el comportamiento de los principales hallazgos al restringir el corpus a los instrumentos efectivamente vigentes y con vínculo directo con la IA.

Procedimiento de recolección y registro

Figura 1

Diagrama de flujo PRISMA-ScR del proceso de levantamiento, depuración y selección del corpus normativo ecuatoriano (corte: mayo de 2026). Fuente: elaboración propia a partir de las matrices del Observatorio GENIAL



La búsqueda se realizó en repositorios oficiales (Registro Oficial, portales del CES, el Ministerio de Educación, Deporte y Cultura - que absorbió las funciones de la Senescyt-, el CACES y el MINTEL) y en los sitios web institucionales de las IES, con triangulación en repositorios académicos. Cada instrumento se registró con trece metadatos: título, tipo de norma, estado (vigente o en proyecto), organismo emisor, dominio temático (administrativo, protección de datos, pedagógico, ético o técnico), vínculo con la educación (directo o indirecto), dedicación del texto a la IA (articulado completo, sección/capítulo o mención breve), fecha de publicación, URL oficial, ámbito (nacional o institucional), observaciones, país e investigador responsable. El proceso de levantamiento, depuración y selección se documenta mediante un diagrama de flujo tipo PRISMA-ScR (Figura 1): se consignaron 84 registros para Ecuador a partir de repositorios oficiales y sitios web institucionales; se eliminaron 3 duplicados (documentos idénticos registrados dos veces) y el corpus final quedó conformado por 81 instrumentos. No se registraron exclusiones adicionales por criterios de contenido -normas derogadas, registros sin

título identificable o documentos puramente operativos sin valor regulatorio-, dado que en las matrices solo se consignaron instrumentos que cumplieran los criterios de inclusión.

Codificación con el marco dimensional

Cada instrumento se clasificó en las seis dimensiones del marco GENIAL -adaptado del modelo de referencia para la analítica del aprendizaje de Chatti et al. (2012)- mediante un procedimiento de codificación asistida basado en reglas explícitas (definidas en un codebook disponible en el repositorio de replicación del estudio <https://drive.google.com/file/d/1IzwUR-vs3iiXMm8eOQb5dQlR1fUnxMFM/view?usp=sharing>) con supervisión humana. El sistema asigna, además, un nivel de confianza a cada clasificación en función de la riqueza de la evidencia disponible (alto, medio o bajo): en el corpus ecuatoriano, 56 instrumentos (69,1 %) obtuvieron confianza alta, 4 (4,9 %) media y 21 (25,9 %) baja; estos últimos fueron verificados manualmente contra el documento original antes del análisis. Cuando la información disponible resultó insuficiente para una dimensión, se codificó como “no determinado” en lugar de forzar una asignación.

Cálculo del Índice de Madurez Normativa (IMN)

El IMN es un indicador compuesto intensivo que, en lugar de sumar el volumen de normas, promedia la calidad de la regulación existente. Se compone de seis subíndices normalizados de 0 a 100, uno por dimensión del marco, organizados en dos bloques:

1. El de calidad: mecanismo (S3), enfoque (S5) y madurez (S6), calculados como la media por norma de una escala de intensidad.
2. El de amplitud: actor (S1), objeto (S2) y finalidad (S4), calculados como la proporción de categorías cubiertas.

El índice es el promedio ponderado de los seis subíndices:

$$\text{IMN} = w1.S1 + w2.S2 + w3.S3 + w4.S4 + w5.S5 + w6.S6$$

Donde la suma de los pesos es igual a 1.

El bloque de calidad mide con qué fuerza regula cada país, el de amplitud mide cuántos frentes distintos toca su regulación. Cada subíndice se expresa como una proporción (entre 0 y 1) que se multiplica por 100 para expresarlo en una escala de 0 a 100, igual que los demás. Se adopta como referencia un escenario que prioriza la calidad, donde los pesos para las diferentes dimensiones son D1 (10 %), D2 (10 %), D3 (20 %), D4 (10 %), D5 (20 %), D6 (30 %) y se evalúan cinco escenarios de ponderación como análisis de sensibilidad, siguiendo las recomendaciones para la construcción de indicadores compuestos (Nardo et al., 2008). La definición operativa de cada subíndice, las escalas de intensidad, el procedimiento de normalización a 0–100 y la formulación completa se detallan en el material complementario del proyecto GENIAL (http://proyctogenial.org/wp-content/uploads/2026/07/2_Metodologia_IMN_GENIAL_vfinal_Formato.pdf)

Para evitar que los países con pocas normas obtengan posiciones infladas, los subíndices de calidad se ajustan por credibilidad de Bühlmann (1967) mediante un factor:

$$Z = n / (n + k)$$

Donde n es el número de normas del país y k una constante de credibilidad. Cuando n es grande, Z se acerca a 1 y prevalece el dato propio; cuando n es pequeño, el puntaje se apoya en la media regional. La constante k se estimó por credibilidad de Bühlmann

$$k = EPV/VHM$$

Donde EPV es la varianza de las normas dentro de los países (varianza de proceso) y VHM es la varianza entre las medias verdaderas de los países. Los valores obtenidos por dimensión se pueden consultar en el material complementario. De este modo, los países con menos evidencia se encogen hacia la media regional. Cada país recibe, además, una etiqueta de confianza (alta- $n \geq 40$, media- $15 \leq n < 40$, o provisional- $n < 15$) según su número de normas. El índice se interpreta en seis niveles: muy alto (70–100), alto (60–69), medio (50–59), medio-bajo (40–49), bajo (25–39) y muy bajo (0–24). Conviene precisar que el IMN mide la madurez normativa como un constructo compuesto que integra amplitud (actores, objetos y finalidades cubiertos) y calidad (mecanismo, enfoque y estado de madurez del ecosistema); por ello, un país puede alcanzar un nivel agregado alto por su amplitud aun cuando el subíndice específico de madurez del ecosistema (S6) permanezca rezagado, como ocurre en el caso ecuatoriano. Los cinco escenarios de ponderación evaluados, con sus pesos y los puntajes resultantes, se detallan en la Tabla 1. Como análisis de robustez, la posición de Ecuador se mantiene estable en el tramo superior del *ranking* a través de los cinco escenarios.

Tabla 1
IMN ajustado por credibilidad de una selección de países bajo los cinco escenarios

País	n	E1	E2	E3	E4	E5
Brasil	81	77,6	72,2	72,5	79,4	70,1
Chile	105	75,2	72,0	74,4	75,8	69,6
Ecuador	81	76,4	70,6	70,4	78,5	68,6
Perú	29	74,0	70,4	71,2	74,9	68,2
Paraguay	18	68,1	67,6	70,6	67,4	65,2
Guatemala	15	70,8	67,6	69,3	71,5	65,4
Colombia	36	70,6	67,4	68,4	71,4	65,1
Nicaragua	6	68,2	66,7	67,6	68,0	64,4
Venezuela	10	68,1	66,0	66,4	68,2	63,7
México	179	70,3	65,5	65,5	71,9	63,7

Nota: Elaboración propia a partir del Observatorio GENIAL (corte: mayo de 2026).

Aseguramiento de la calidad y consideraciones éticas

El estudio analiza exclusivamente documentos normativos de acceso público, por lo que no involucra sujetos humanos ni requiere aprobación de un comité de ética. La trazabilidad se garantiza mediante el registro de cada decisión de limpieza y codificación en una bitácora de ejecuciones y los materiales de replicación (*codebook*, registro consolidado y fórmula del índice) están disponibles a solicitud al autor de correspondencia.

Resultados

PI1. ¿Qué corpus de normas, políticas y lineamientos regula directa o indirectamente el uso de IA en la educación superior ecuatoriana? Panorama general del corpus normativo

El corpus ecuatoriano comprende 81 instrumentos, de los cuales 76 (93,8 %) están vigentes y 5 (6,2 %) se encuentran en proyecto o en fase de borrador. Su rasgo estructural más notable es la distribución por ámbito: 71 instrumentos (87,7 %) son de carácter institucional -emitidos por 56 IES distintas- y solo 10 (12,3 %) son de alcance nacional. En el nivel nacional destacan la Estrategia para el Fomento del Desarrollo y Uso Ético y Responsable de la IA (Acuerdo Ministerial MINTEL-2025-0030, enero de 2026), la Política Pública para la Transformación Digital 2025–2030, las orientaciones del Ministerio de Educación para el uso pedagógico de herramientas de IA (2024) y tres registros legislativos correspondientes a dos proyectos de ley sobre IA en trámite en la Asamblea Nacional (expedientes 450889 y 453516). Ningún instrumento con rango de ley específicamente dedicado a la IA se encuentra vigente. La Tabla 2 presenta la composición del corpus por tipo de instrumento y ámbito.

Tabla 2

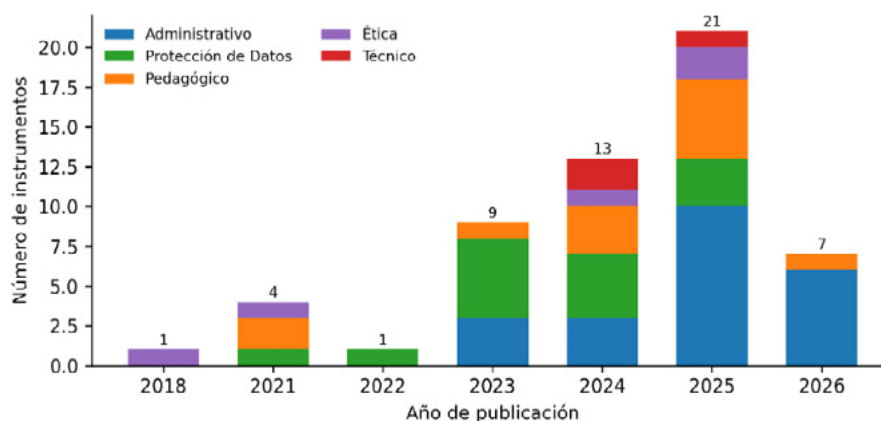
Composición del corpus normativo ecuatoriano por tipo de instrumento y ámbito

Tipo de instrumento	Institucional	Nacional	Total	%
Reglamento	16	0	16	19,8
Otro (informes, libros, publicaciones)	14	2	16	19,8
Guía ética	13	0	13	16,0
Lineamiento académico	12	0	12	14,8
Resolución rectoral	9	0	9	11,1
Guía pedagógica	6	1	7	8,6
Ley (proyectos en trámite)	0	3	3	3,7
Resolución (acuerdos ministeriales)	0	3	3	3,7
Estrategia nacional	0	1	1	1,2
Libro blanco	1	0	1	1,2
Total	71	10	81	100

Nota: Elaboración propia a partir del Observatorio GENIAL (corte: mayo de 2026).

Figura 2

Distribución temporal del corpus normativo por dominio



Nota: Se presentan los 56 instrumentos con fecha identificable; 25 instrumentos del corpus carecen de fecha de publicación verificable.

La distribución temporal confirma el efecto disparador de la IA generativa: de los 56 instrumentos con fecha de publicación identificable, 50 (89,3 %) se publicaron a partir de 2023, con un pico de 21 instrumentos en 2025 y 7 adicionales en lo que va de 2026 (Figura 2). Los 25 instrumentos restantes carecen de fecha verificable, en su mayoría políticas institucionales publicadas sin datación explícita.

PI2. ¿Qué dominios cubre el marco normativo actual y cuáles permanecen desatendidos? Cobertura por dominio y especificidad

Dos dominios concentran el 70,7 % del corpus: el administrativo (34 instrumentos; 42,0 %) y el de protección de datos (24 instrumentos; 29,6 %), este último impulsado por la implementación de la Ley Orgánica de Protección de Datos Personales de 2021, que llevó a las IES a publicar políticas de privacidad y tratamiento de datos. El dominio pedagógico reúne 14 instrumentos (17,3 %), mientras que los dominios ético (5 instrumentos; 6,2 %) y técnico (3 instrumentos; 3,7 %) resultan marginales. La especificidad del tratamiento es, sin embargo, dispar, pues como muestra la Figura 3, el dominio pedagógico presenta la mayor proporción de instrumentos con articulado completo dedicado a la IA (13 de 14; 92,9 %), mientras que en el dominio administrativo casi un tercio de los instrumentos (10 de 34) aborda la IA solo mediante menciones breves.

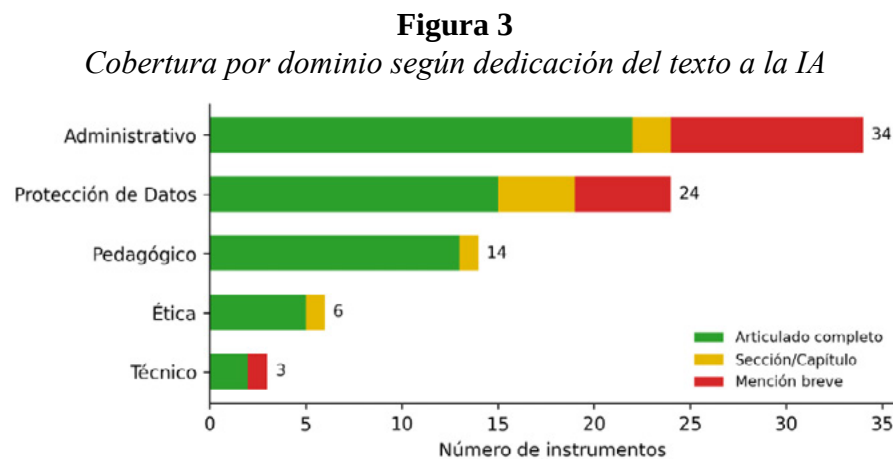


Tabla 3
Clasificación del corpus según el framework dimensional GENIAL (categorías principales)

Dimensión	Categoría más frecuente	n	Segunda categoría	n	Categorías ausentes o marginales
D1 Actor	Instituciones IES	72	Gobierno/sector público	20	Estudiantes (11); Docentes (10)
D2 Objeto	IA en general	38	Datos personales educativos	34	Analíticas de aprendizaje (0); biometría (0)
D3 Mecanismo	Lineamiento orientador	45	Obligación	39	Sanción y control (3); prohibición (1)
D4 Finalidad	Gobernanza y gestión	42	Protección de derechos	38	Monitoreo estudiantil (0); personalización (1)
D5 Enfoque	Reflexivo	45	Estratégico	26	Restringido (1)
D6 Madurez	Declaración de intención	40	Regulación inicial	34	Ecosistema maduro (0); exploración (0)

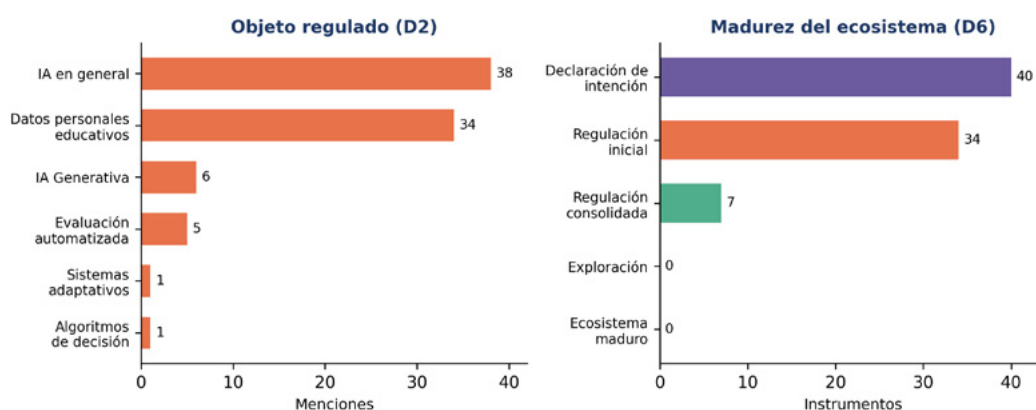
Nota: Las dimensiones D1–D5 admiten clasificación múltiple, por lo que las frecuencias corresponden a menciones; D6 es de asignación única. En D3, 8 instrumentos no permitieron determinar el mecanismo regulatorio.

La Tabla 3 sintetiza la clasificación del corpus en las seis dimensiones del marco. En la dimensión de actores (D1), las IES aparecen como sujeto regulado o protegido en 72 instrumentos (88,9 %), mientras que los estudiantes son nombrados explícitamente en 11 (13,6 %) y los docentes en 10 (12,3 %): la normativa ecuatoriana se ocupa de las instituciones antes que de las personas. En cuanto al objeto regulado (D2), predominan la “IA en general” (38 menciones) y los datos personales educativos (34); la IA generativa solo es regulada de manera explícita por 6 instrumentos (7,3 % del corpus) y la evaluación automatizada por 5, mientras que las analíticas de aprendizaje y el reconocimiento biométrico no son objeto de ningún instrumento.

El mecanismo regulatorio (D3) confirma el predominio del derecho blando, aunque con una presencia creciente de las obligaciones: los lineamientos orientadores (45 menciones) encabezan, seguidos de las obligaciones (39) y los principios éticos (24), mientras que los mecanismos de exigibilidad fuerte siguen siendo excepcionales (3 instrumentos con sanción y control; 1 con prohibición). En la finalidad (D4) dominan la gobernanza y gestión (42) y la protección de derechos (38) sobre las finalidades propiamente educativas: la integridad académica y la innovación pedagógica registran 15 menciones cada una y la evaluación con retroalimentación solo 3. El enfoque normativo (D5) es mayoritariamente reflexivo (45) o estratégico (26), con presencia mínima de enfoques restrictivos. Finalmente, en la dimensión de madurez (D6) -que sintetiza el estado de consolidación del ecosistema regulatorio (existencia, jerarquía, especificidad, exigibilidad y vigencia de los instrumentos) e incorpora la fuerza del mecanismo (D3) como uno de sus componentes y no como su único determinante-, la mitad del corpus (40 instrumentos; 49,4 %) se ubica en el estado de declaración de intención y el 42,0 % (34) en regulación inicial, mientras que solo el 8,6 % (7) alcanza la regulación consolidada y ningún instrumento refleja un ecosistema maduro (Figura 4).

Figura 4

Objeto regulado (D2) y madurez del ecosistema (D6) en el corpus ecuatoriano



PI3. ¿Cuál es el grado de madurez normativa del ecosistema ecuatoriano y cómo se posiciona en el contexto regional? Índice de Madurez Normativa y posición regional

Ecuador obtiene un IMN de 70,6/100, que lo sitúa en el nivel muy alto de madurez y en la tercera posición entre los 28 países con datos del Observatorio, por detrás de Brasil (72,2) y Chile (72,0), y por delante de Perú (70,4), Colombia (67,4) y, notablemente, México (65,5; décimo) y Argentina (61,7) (Figura 5). La Tabla 4 descompone el índice por subíndice, donde la fortaleza de Ecuador reside en la amplitud de su regulación -cubre la totalidad de las categorías de actor ($S1 = 100$) y casi todas las de finalidad ($S4 = 87,5$)- y en un mecanismo comparativamente fuerte ($S3 = 71,7$), impulsado por la presencia de obligaciones. En contraste, su subíndice más bajo es la madurez del ecosistema ($S6 = 51,9$), ya que pese a la densidad y

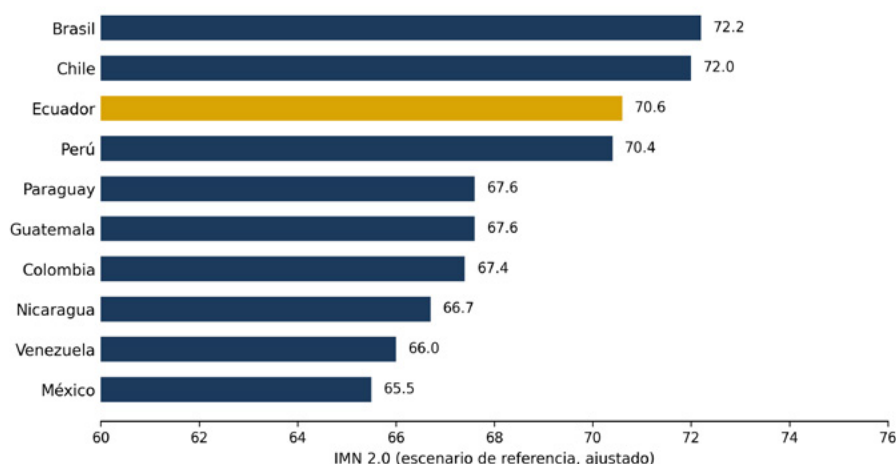
amplitud del corpus, la mitad de los instrumentos no supera la declaración de intención, lo que confirma que el alto puntaje agregado de Ecuador proviene de su actividad regulatoria y su cobertura, no de la consolidación del marco.

Tabla 4
Descomposición del puntaje bruto del IMN para Ecuador

Subíndice	Bloque	Valor (0–100)	Interpretación
S1 Actor	Amplitud	100,0	Cubre todas las categorías de actor
S4 Finalidad	Amplitud	87,5	Amplia cobertura de finalidades
S2 Objeto	Amplitud	75,0	Buena cobertura tecnológica
S5 Enfoque	Calidad	72,3	Predominio reflexivo–estratégico
S3 Mecanismo	Calidad	71,7	Presencia relevante de obligaciones
S6 Madurez	Calidad	51,9	Mitad del corpus en declaración de intención (dimensión más débil)
IMN	-	70,6	Nivel muy alto · 3.º de 28 · confianza alta

Nota: Subíndices del IMN en escala 0–100. Bloque de calidad: S3, S5 y S6 (media por norma de una escala de intensidad); bloque de amplitud: S1, S2 y S4 (proporción de categorías cubiertas). El IMN de referencia (ajustado por credibilidad) de Ecuador es 70,6.

Figura 5
IMN normalizado de los diez primeros países de la región (Ecuador resaltado)



PI4. ¿Qué líneas de acción permitirían cerrar las brechas identificadas? Mapa de brechas críticas

La integración de los hallazgos anteriores permite identificar cinco brechas críticas, sintetizadas en la Tabla 5 según una tipología de brechas de existencia, especificidad, jerarquía, exigibilidad y actualización.

Tabla 5
Mapa de brechas críticas del marco normativo ecuatoriano

Tipo de brecha	Descripción	Evidencia
Existencia	Ausencia de ley nacional vigente sobre IA y de regulación sobre analíticas de aprendizaje y biometría en educación	6 de 81 instrumentos (7,4 %)

Tipo de brecha	Descripción	Evidencia
Especificidad	La IA generativa, disparadora de la ola normativa, es regulada explícitamente por una fracción mínima del corpus	87,7 % institucional; 12,3 % nacional
Jerarquía	La regulación se concentra en el nivel institucional sin un marco nacional común que la armonice	45 lineamientos orientadores vs. 3 con sanción y control
Exigibilidad	Predominio de instrumentos orientadores sin mecanismos de cumplimiento	25 instrumentos sin fecha; ética 6,2 % y técnico 3,7 % del corpus
Actualización	Un tercio del corpus carece de fecha verificable y los dominios ético y técnico permanecen marginales	26 instrumentos sin fecha; ética 7,3 % y técnico 3,7 % del corpus

Nota: Elaboración propia.

Discusión

El hallazgo central de este estudio es la paradoja que define al caso ecuatoriano: una posición regional muy alta (tercer lugar, nivel muy alto en el IMN) que, sin embargo, no se sostiene en la consolidación del marco sino en la amplitud y la actividad regulatoria de las propias universidades. El índice intensivo lo hace evidente: Ecuador maximiza los subíndices de amplitud -cobertura total de actores y casi total de finalidades- y muestra un mecanismo relativamente fuerte, pero su subíndice de madurez del ecosistema es el más bajo y la mitad del corpus no supera la declaración de intención. Esto evidencia que la cantidad de normas y la madurez regulatoria son cosas distintas. En Ecuador, 56 IES produjeron el 87,7 % del corpus, lo que configura una construcción normativa de abajo hacia arriba (*bottom-up*). Este patrón tiene una lectura positiva: el sistema universitario no esperó al legislador y respondió con agilidad, sobre todo a partir de 2023; y otra problemática: la regulación resultante es heterogénea, mayoritariamente orientadora, de exigibilidad débil y sin garantías de coherencia entre instituciones.

Tensiones estructurales del caso ecuatoriano

Los resultados ponen de relieve tres tensiones. La primera es la relación entre autonomía universitaria y marcos comunes. La autonomía consagrada constitucionalmente explica la vitalidad normativa institucional, pero también el riesgo de fragmentación; el desafío para el CES es emitir lineamientos marco habilitantes sin invadir competencias institucionales. La segunda es la asimetría entre velocidad tecnológica y capacidad regulatoria: los dos proyectos de ley sobre IA siguen en trámite mientras el 89,3 % del corpus fechado se produjo en apenas tres años, lo que sugiere que los tiempos legislativos no están sincronizados con el ciclo de la tecnología. La tercera se relaciona con la equidad institucional: si la regulación depende de la capacidad técnica de cada IES, las instituciones con menos recursos podrían quedar rezagadas también en gobernanza de la IA, lo que profundizaría brechas preexistentes -una hipótesis plausible que este diseño no evalúa directamente-; que 56 de las cerca de 60 IES del sistema registren al menos un instrumento es alentador, pero la profundidad de esos instrumentos varía considerablemente.

Diálogo con la literatura

Estos hallazgos son coherentes con el diagnóstico regional de UNESCO IESALC (2023), que advierte que la respuesta normativa latinoamericana a la IA generativa ha sido principalmente institucional y orientadora, y con la caracterización de Holmes y Tuomi (2022) sobre la brecha entre la adopción de la IA educativa y su gobernanza. El caso ecuatoriano añade evidencia empírica cuantificada a esa conversación, ya que muestra que un país puede alcanzar una posición agregada alta en el *ranking* regional sin ley nacional, impulsado por la amplitud y

la actividad institucional, pero también que ese tipo de madurez es cualitativamente distinto -más de cobertura que de consolidación- del de los marcos que combinan amplitud con exigibilidad. Frente al referente del Reglamento Europeo de IA (European Parliament & Council, 2024), basado en niveles de riesgo y obligaciones exigibles, el corpus ecuatoriano se asemeja más al momento preregulatorio europeo, denso en principios y orientaciones, débil en exigibilidad. Metodológicamente, el estudio sugiere que el IMN funciona como herramienta diagnóstica reproducible incluso en contextos de baja institucionalización regulatoria, aunque su validez externa debe seguir evaluándose mediante replicaciones (Becker et al., 2009; Mettler, 2011).

Agenda para la política pública y la acción institucional derivada de la PI4

De las brechas identificadas se deriva una agenda con destinatarios y plazos diferenciados, sintetizada en la Tabla 6. En el corto plazo (0–12 meses), el CES debería emitir un lineamiento marco sobre el uso de IA en docencia y evaluación -cerrando la brecha de jerarquía sin esperar a la ley-, y el MINEDUC debería articular la dimensión de educación superior dentro de la estrategia nacional de IA vigente e impulsar un programa nacional de alfabetización en IA para el profesorado universitario. En el mediano plazo (1–3 años), corresponde reformar el Reglamento de Régimen Académico, para incorporar criterios de integridad académica ante la IA generativa, e integrar indicadores de gobernanza algorítmica en los modelos de evaluación del CACES. En el largo plazo (3–5 años), la aprobación de una ley de IA con disposiciones específicas para educación superior consolidaría el marco. Las IES, por su parte, deberían adoptar políticas mínimas de uso aceptable, transparencia y protección de datos estudiantiles, así como comités de ética con capacidad real de revisión de sistemas de IA, prestando especial atención a los objetos hoy desregulados, como las analíticas de aprendizaje, la evaluación automatizada y los sistemas biométricos.

Tabla 6
Agenda propuesta de política pública por actor y plazo

Actor	Acción prioritaria	Plazo	Brecha que atiende
CES	Lineamiento marco sobre uso de IA en docencia y evaluación	Corto	Jerarquía
CES	Reforma del Reglamento de Régimen Académico (integridad académica e IAG)	Mediano	Especificidad
Min. de Educación (antes Senescyt)	Articulación de la educación superior en la estrategia nacional de IA	Corto	Jerarquía
Min. de Educación (antes Senescyt)	Programa nacional de alfabetización en IA para docentes universitarios	Corto	Existencia
CACES	Indicadores de gobernanza algorítmica en modelos de evaluación institucional	Mediano	Exigibilidad
Asamblea Nacional	Ley de IA con disposiciones específicas para educación superior	Largo	Existencia
IES	Políticas mínimas de uso aceptable, transparencia y datos estudiantiles	Corto	Especificidad
IES	Comités de ética con capacidad de revisión de sistemas de IA	Mediano	Exigibilidad

Nota: Plazos: corto = 0–12 meses; mediano = 1–3 años; largo = 3–5 años.

Limitaciones

Cuatro limitaciones acotan el alcance de los resultados. Primera, el IMN mide la presencia, la calidad formal y la madurez de la norma, no su aplicación efectiva ni su impacto en las prácticas docentes. Segunda, la cobertura de políticas institucionales depende de su publicidad, lo que introduce un posible sesgo hacia las IES más transparentes. Tercera, el

25,9 % de las clasificaciones dimensionales obtuvo inicialmente confianza baja y, aunque estas clasificaciones fueron verificadas manualmente, la codificación asistida basada en reglas con supervisión humana podría diferir del juicio de un codificador independiente. Para acotar esta limitación se realizó una validación intercodificador sobre una submuestra estratificada de 142 instrumentos del corpus regional (15,0 % del total, 27 países, 12 de ellos ecuatorianos), con acuerdo sustancial en cinco de las seis dimensiones (κ de Cohen entre 0,607 y 0,793) y moderado en madurez (D6; $\kappa = 0,576$), y con robustez del *ranking* de países confirmada mediante correlación de Spearman ($\rho = 0,790$, $p < 0,001$) entre codificadores en el bloque de calidad; la submuestra ecuatoriana, aunque reducida (12 instrumentos), no mostró discrepancias sistemáticas frente al patrón regional. Cuarta, la velocidad del ciclo regulatorio posIA generativa -evidente en el pico de 2025- implica que este diagnóstico, con corte a mayo de 2026, puede desactualizarse con rapidez; un tercio del corpus sin fecha verificable refuerza la necesidad de mejorar la trazabilidad documental del Observatorio.

Conclusiones

Este estudio presentó uno de los primeros inventarios sistemáticos de la normativa sobre IA en la educación superior ecuatoriana del que se tiene registro: 81 instrumentos codificados con un marco de seis dimensiones y sintetizados en el Índice de Madurez Normativa (IMN). Tres conclusiones se desprenden del análisis. Primera, Ecuador presenta una madurez normativa agregada alta en el contexto regional (IMN = 70,6; tercero de 28 países, nivel muy alto), pero esa posición se explica por la amplitud y la actividad regulatoria de las universidades más que por la consolidación del marco: la mitad del corpus permanece en declaración de intención y la madurez del ecosistema es su dimensión más débil. El problema, por tanto, no es la ausencia total de norma, sino su baja consolidación, especificidad y exigibilidad. Segunda, la construcción normativa ha sido de abajo hacia arriba: las universidades -56 IES con al menos un instrumento- habrían asumido un rol protagónico ante la ausencia de legislación nacional vigente. Este patrón sugiere, como interpretación derivada de los hallazgos y no evaluada directamente por el diseño documental, que la autonomía universitaria podría operar como palanca y no como obstáculo, siempre que se acompañe de marcos comunes habilitantes. Tercera, las brechas críticas son menos técnicas que éticas y pedagógicas, la IA generativa, las analíticas de aprendizaje y la evaluación automatizada -los objetos que más afectan a estudiantes y docentes- son precisamente las menos reguladas. Además, la norma ecuatoriana se dirige a las instituciones antes que a las personas: las IES son nombradas en 72 instrumentos (88,9 %), mientras que los estudiantes aparecen explícitamente en 11 (13,6 %) y los docentes en 10 (12,3 %); en conjunto, solo 20 instrumentos (24,7 %) nombran a algún actor humano del sistema.

Como trabajo futuro se plantea: (i) replicar el análisis en los demás países del Observatorio GENIAL para consolidar un diagnóstico comparativo permanente; (ii) complementar el análisis documental con trabajo de campo sobre la implementación efectiva de las políticas institucionales; (iii) realizar un seguimiento longitudinal anual del IMN ecuatoriano; e (iv) incorporar doble codificación independiente con un informe de grado de acuerdo intercodificador para robustecer el marco dimensional.

Reconocimientos y Declaraciones

Este trabajo ha sido financiado por el Vicerrectorado de Investigación de la Universidad de Cuenca y el proyecto GENIAL (Erasmus+, Unión Europea), orientado a la generación de capacidades de inteligencia artificial en instituciones de educación superior de América Latina. Los autores agradecen al equipo de investigadores del Observatorio Regional de IA en Educación

por el levantamiento normativo que hizo posible este estudio. El contenido de este artículo es responsabilidad exclusiva de los autores y no refleja necesariamente la opinión de la Unión Europea.

Contribuciones de los autores: Autor 1: conceptualización, metodología, análisis formal, curación de datos, redacción del borrador original, supervisión y administración del proyecto, levantamiento de información, codificación, verificación y validación de la información. Coautor 2: levantamiento de información, codificación, verificación y validación de la información y revisión del manuscrito. Coautor 3: levantamiento de información, codificación, verificación y validación de la información y revisión del manuscrito.

Declaración sobre el uso de IA: los autores declaran que, en la elaboración del presente artículo, se utilizaron herramientas de inteligencia artificial (Claude, de Anthropic; modelo Claude Sonnet 4.6) en tres tareas específicas: (i) el apoyo a la limpieza y consolidación del registro documental; (ii) la codificación asistida basada en reglas explícitas de un codebook, siempre bajo supervisión humana; y (iii) el desarrollo de los scripts de cálculo del IMN y de los estadísticos de acuerdo. La herramienta no sustituyó el juicio humano en la clasificación reportada: todas las asignaciones dimensionales fueron revisadas y validadas por los autores, y las clasificaciones de baja confianza se verificaron manualmente contra el documento original. Cuando la IA se empleó como codificador de referencia, se hizo de forma transparentemente etiquetada y separada de la codificación humana, únicamente con fines de comparación. Los autores asumen plena responsabilidad sobre todo el contenido.

Disponibilidad de datos: el registro consolidado de instrumentos normativos, el codebook del marco dimensional y los scripts de cálculo del IMN están disponibles en el repositorio abierto del estudio http://proyectogenial.org/wp-content/uploads/2026/07/2_Metodologia_IMN_GENIAL_vfinal_Formato.pdf y, en su defecto, a solicitud al autor de correspondencia.

Referencias

- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Asamblea Nacional del Ecuador. (2010). Ley Orgánica de Educación Superior. Registro Oficial Suplemento 298, 12 de octubre de 2010. URL: <https://www.ces.gob.ec/documentos/Normativa/LOES.pdf>
- Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales*. Registro Oficial Suplemento 459, 26 de mayo de 2021. URL: https://www.telecomunicaciones.gob.ec/wp-content/uploads/2021/06/Ley-Organica-de-Datos-Personales.pdf&ved=2ahUKEwjw_PqT1PGWAxWBTDABHaQCPIAQy_kOegYIAAgKEAM&opi=89978449&cd&psig=AOvVaw0XCi_nU_4iBL_Pljwff2tbc&ust=1789598524455000
- Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management: A procedure model and its application. *Business & Information Systems Engineering*, 1(3), 213–222. <https://doi.org/10.1007/s12599-009-0044-5>
- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27–40. <https://doi.org/10.3316/QRJ0902027>
- Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin*, 4(3), 199–207.
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5/6), 318–331.

- Congreso de la República del Perú. (2023). *Ley N.º 31814, Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*. Diario Oficial El Peruano, 5 de julio de 2023.
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20, 22. <https://doi.org/10.1186/s41239-023-00392-8>
- European Parliament & Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542–570. <https://doi.org/10.1111/ejed.12533>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.
- Mettler, T. (2011). Maturity assessment models: A design science research approach. *International Journal of Society Systems Science*, 3(1/2), 81–98. <https://doi.org/10.1504/IJSSS.2011.038934>
- Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing.
- Ministerio de Educación del Ecuador. (2024). *Orientaciones para el uso pedagógico de herramientas de inteligencia artificial en el proceso de enseñanza-aprendizaje*. Ministerio de Educación.
- Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2025). *Política pública para la transformación digital del Ecuador 2025–2030*. MINTEL. URL: https://www.gobiernoelectronico.gob.ec/wp-content/uploads/2025/03/INSTRUMENTO-Politica-Publica-para-la-Transformacion-Digital-Ecuador-2025-2030-MINTEL-signed_f.pdf&ved=2ahUKEwjMhKSI2PGWAxWKSzABHaEANQUQy_kOegYIAAgREAM&opi=89978449&cd&psig=AOvVaw0EVo-7pv8utAOuWgeQFqzJ&ust=1789599573695000
- Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2026). Acuerdo Ministerial MINTEL-MINTEL-2025-0030: Estrategia para el fomento del desarrollo y uso ético y responsable de la inteligencia artificial en el Ecuador. MINTEL. URL: <https://www.registroficial.gob.ec/suplemento-no-206/>
- OECD. (2019). *Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449)*. OECD Legal Instruments. URL: <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>
- Stake, R. E. (1995). *The art of case study research*. SAGE Publications.
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa
- UNESCO. (2023). *Guidance for generative AI in education and research*. <https://doi.org/10.54675/EWZM9535>
- UNESCO IESALC. (2023). *ChatGPT e inteligencia artificial en la educación superior: Guía de inicio rápido*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>

<https://doi.org/10.37815/rte.v38nE1.1548>

Artículos originales

“Estación Infantil”: Plataforma virtual gamificada para la educación inicial

“Estación Infantil”: A Gamified Virtual Platform for Early Childhood Education

Miguel Cobos¹ <https://orcid.org/0000-0002-6809-9639>, Sebastián Báez¹ <https://orcid.org/0009-0007-5355-0996>

¹Universidad Tecnológica Indoamérica, Facultad de Ciencias de la Educación
Quito, Ecuador

miguelcobos@indoamerica.edu.ec, bryan_sebas98@hotmail.com



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Enviado: 2026/06/26

Aceptado: 2026/07/28

Publicado: 2026/10/14

Resumen

La transición hacia entornos de educación no presencial iniciada con la pandemia de COVID-19 evidenció la escasez de recursos digitales especializados en todos los niveles, y también en el subnivel de educación inicial. Este artículo presenta el diseño, desarrollo y validación tecnopedagógica de “Estación Infantil”, una plataforma web híbrida desarrollada en Unity y Google Sites, fundamentada en la gamificación, la narrativa interactiva y la teoría del flujo. El sistema traduce experiencias de aprendizaje del currículo ecuatoriano en cuatro entornos interactivos desarrollados con WebGL orientados al trabajo mediado con niños de 3 a 5 años. La validación se realizó mediante un estudio cuantitativo en el que se aplicó estadística descriptiva e inferencial (prueba de Kruskal–Wallis) a una muestra intencional de 41 participantes vinculados a la formación y al ejercicio docente en educación inicial. Se aplicó un instrumento tipo Likert compuesto por siete dimensiones para evaluar los recursos didácticos, la claridad, la pertinencia, el diseño, la retroalimentación, la legibilidad y la usabilidad; la escala presentó una consistencia interna adecuada ($\alpha \approx 0,71$). Los resultados muestran valoraciones altas en todas las dimensiones, con medias globales entre 4,44 y 4,66. El análisis inferencial confirmó una valoración homogénea en las dimensiones pedagógicas, con diferencias significativas entre perfiles solo en legibilidad y usabilidad. Se concluye que “Estación Infantil” constituye un recurso digital viable para el refuerzo de aprendizajes en educación inicial, al integrar alineación curricular, accesibilidad tecnológica y diseño lúdico orientado a la mediación pedagógica.

Sumario: Introducción, Materiales y Métodos, Resultados y Discusión, Conclusiones.

Como citar: Cobos, M. & Báez, S. (2026). “Estación Infantil”: Plataforma virtual gamificada para la educación inicial. *Revista Tecnológica - Espol*, 38(E1), 117-135. <https://doi.org/10.37815/rte.v38nE1.1548>

Palabras clave: gamificación educativa, educación inicial, Unity WebGL, tecnopedagogía, currículo de educación inicial, validación tecnopedagógica.

Abstract

La transición hacia entornos de educación no presencial iniciada con la pandemia de COVID-19 evidenció la escasez de recursos digitales especializados en todos los niveles, y también en el subnivel de educación inicial. Este artículo presenta el diseño, desarrollo y validación tecnopedagógica de “Estación Infantil”, una plataforma web híbrida desarrollada en Unity y Google Sites, fundamentada en la gamificación, la narrativa interactiva y la teoría del flujo. El sistema traduce experiencias de aprendizaje del currículo ecuatoriano en cuatro entornos interactivos desarrollados con WebGL orientados al trabajo mediado con niños de 3 a 5 años. La validación se realizó mediante un estudio cuantitativo en el que se aplicó estadística descriptiva e inferencial (prueba de Kruskal–Wallis) a una muestra intencional de 41 participantes vinculados a la formación y al ejercicio docente en educación inicial. Se aplicó un instrumento tipo Likert compuesto por siete dimensiones para evaluar los recursos didácticos, la claridad, la pertinencia, el diseño, la retroalimentación, la legibilidad y la usabilidad; la escala presentó una consistencia interna adecuada ($\alpha \approx 0,71$). Los resultados muestran valoraciones altas en todas las dimensiones, con medias globales entre 4,44 y 4,66. El análisis inferencial confirmó una valoración homogénea en las dimensiones pedagógicas, con diferencias significativas entre perfiles solo en legibilidad y usabilidad. Se concluye que “Estación Infantil” constituye un recurso digital viable para el refuerzo de aprendizajes en educación inicial, al integrar alineación curricular, accesibilidad tecnológica y diseño lúdico orientado a la mediación pedagógica.

Keywords: educational gamification, early childhood education, Unity WebGL, techno-pedagogy, early childhood curriculum, techno-pedagogical validation.

Introducción

La transición acelerada hacia entornos de educación no presencial durante la pandemia de COVID-19 evidenció una escasez crítica de recursos digitales diseñados específicamente para la educación inicial (García Gutiérrez et al., 2023), tanto en términos de pertinencia curricular como de usabilidad para niños de 3 a 5 años. En Ecuador, aunque el Ministerio de Educación generó lineamientos y materiales como el Currículo de Educación Inicial (2014) y la guía “Aprendamos Juntos en Casa” (2022; obra original publicada en 2020), la mayor parte de las herramientas tecnológicas disponibles correspondían a plataformas genéricas, poco alineadas con destrezas y experiencias de aprendizaje concretas del nivel.

En respuesta a este vacío, se diseñó la plataforma gamificada “Estación Infantil”, cuyo objetivo es traducir en experiencias interactivas los ámbitos y destrezas del currículo nacional, con énfasis en hábitos de higiene, autonomía, organización del entorno y expresión artística en la primera infancia. Los cuatro juegos que integran la plataforma se basan en las experiencias de aprendizaje descritas en el plan “Aprendamos Juntos en Casa”. Los juegos son: “Me gusta estar limpio”, “A colgar y organizar la ropa”, “Organizo mis juguetes” y “Soy artista, me gusta la pintura”; además, se articulan con los ejes de desarrollo y los ámbitos de identidad y autonomía, convivencia, expresión artística y relaciones lógico-matemáticas definidos en el Currículo de Educación Inicial.

La actualización del Currículo Priorizado de Educación Inicial (2025) mantiene la estructura del currículo de 2014, pero incorpora un énfasis explícito en habilidades

comunicacionales, lógico-matemáticas, digitales y socioemocionales para niños de 3 a 5 años, incluyendo recomendaciones sobre el uso guiado de recursos tecnológicos. Estas orientaciones son coherentes con la literatura contemporánea que advierte sobre los riesgos del uso intensivo y no supervisado de pantallas en la infancia, tanto en términos de desarrollo cognitivo como socioemocional (Desmurget & Cortés Fernández, 2020; Haidt, 2024). En este marco, la plataforma "Estación Infantil" se concibe bajo un principio de acompañamiento mediado, es decir, como un recurso para el trabajo guiado por el docente y no como un entorno de uso autónomo de los niños.

Durante la fase de desarrollo de la plataforma se realizó una validación inicial exploratoria con un pequeño grupo de docentes de educación inicial, a partir de sesiones sincrónicas en línea en las que observaron la interacción de niños con los juegos y emitieron juicios cualitativos sobre su pertinencia y calidad. Esa fase preliminar permitió ajustar aspectos de diseño y consolidar las dimensiones centrales de evaluación tecnopedagógica, pero no aportó aún evidencia cuantitativa suficiente ni una muestra diversificada de participantes.

A partir de esta experiencia, se diseñó una segunda fase de validación ampliada, orientada a obtener evidencia cuantitativa robusta mediante un instrumento tipo Likert que evalúa siete dimensiones de la percepción tecnopedagógica de la plataforma. En esta fase participaron 41 personas con distintos niveles de formación y experiencia en educación inicial: docentes en ejercicio, estudiantes de nivel avanzado y docentes expertos con formación de posgrado, lo que permitió analizar la plataforma desde perspectivas complementarias dentro del campo educativo.

En este contexto, el objetivo del presente estudio es analizar la percepción tecnopedagógica de la plataforma gamificada "Estación Infantil" como recurso de refuerzo para la educación inicial alineado con las orientaciones del Currículo Priorizado de Educación Inicial 2025. Específicamente, se busca: (a) describir los niveles de valoración de la plataforma en siete dimensiones clave (recurso didáctico, claridad, pertinencia, diseño, retroalimentación, legibilidad y usabilidad), (b) comparar estas percepciones según el perfil formativo de los participantes y (c) estimar la consistencia interna del instrumento de evaluación utilizado.

Materiales y Métodos

Diseño y desarrollo de la plataforma "Estación Infantil"

Fundamentos teóricos

El diseño de "Estación Infantil" se sustenta en tres pilares teóricos complementarios que orientaron tanto las decisiones de desarrollo técnico como las de diseño pedagógico. El primero es la gamificación, entendida como la aplicación de mecánicas y elementos propios del diseño de videojuegos —como objetivos claros, retroalimentación inmediata, sistemas de recompensa y progresión por niveles— (Capatina et al., 2024), aplicados en contextos no lúdicos para incrementar la motivación y el compromiso del usuario. Investigaciones internacionales recientes muestran un fuerte consenso sobre los beneficios de la gamificación y del carácter lúdico para fomentar el compromiso y los resultados de aprendizaje en la primera infancia, siempre que se prioricen la narrativa, la retroalimentación positiva y la mediación docente (Fabre-Mitjans et al., 2025; Kareem y Khalid, 2024; Pedraza Vázquez et al., 2022).

El segundo es la teoría del flujo (Csíkszentmihályi, 2014), que define un estado de concentración óptima en el que el nivel de desafío de la actividad se equilibra con las capacidades del usuario, generando una experiencia de inmersión que favorece la retención y el procesamiento profundo de la información. El diseño de cada juego buscó deliberadamente provocar este estado

en niños de 3 a 5 años mediante instrucciones simples, progresión gradual del reto y refuerzo positivo constante. El tercer pilar es el uso de la narrativa en el diseño de videojuegos (Cobos & Salvador, 2022, 2023), que estructura la experiencia a través de personajes, contextos cotidianos y secuencias de acción significativas para el usuario; funciona como hilo conductor que da coherencia a las actividades y reduce la desorientación de los niños dentro del entorno digital.

De manera transversal a estos tres pilares, la plataforma fue concebida bajo el principio de acompañamiento mediado: no está diseñada para el uso autónomo de los niños frente a la pantalla, sino como un recurso de interacción guiado por el docente o el adulto responsable. Esta decisión responde a las recomendaciones de la literatura sobre los límites del tiempo de pantalla en la primera infancia y es plenamente coherente con las orientaciones del Currículo Priorizado de Educación Inicial del Ministerio de Educación (2025), que reconoce el uso de juegos digitales apropiados para la edad, siempre con el acompañamiento del docente o de la familia, como una estrategia pertinente para el desarrollo de habilidades digitales desde la primera infancia.

Arquitectura técnica

La plataforma adopta una arquitectura de aplicación web híbrida, orientada a maximizar la accesibilidad desde cualquier dispositivo con navegador web y a minimizar las barreras de instalación para los docentes. Sus componentes principales son:

Plataforma contenedora (Google Sites): funciona como interfaz principal de acceso a la plataforma. Se seleccionó por su acceso sin costo, su integración nativa con el ecosistema educativo de Google Workspace y su capacidad de incrustar contenido externo mediante enlaces.

Motor de desarrollo (Unity Engine): los cuatro juegos fueron programados en C# dentro del entorno Unity y compilados en formato WebGL, lo que permite su ejecución directa en el navegador web sin necesidad de instalar plugins adicionales. Esta decisión técnica fue determinante para garantizar que los juegos sean accesibles desde los dispositivos que habitualmente usan los docentes en el aula.

Diseño gráfico (Adobe Illustrator): todos los elementos visuales fueron creados como gráficos vectoriales con estética flat design, lo que garantiza tiempos de carga rápidos, escalabilidad a distintos tamaños de pantalla y una paleta de colores vivos y contrastados, apropiada para captar la atención de niños de 3 a 5 años.

Retroalimentación auditiva (Adobe Audition): se grabaron y procesaron voces en off con instrucciones y refuerzos positivos en español neutro, normalizando el volumen y eliminando ruido de fondo. Este canal de audio es el mecanismo principal de guía para los niños, dado que no se requiere lectura autónoma para usar la plataforma.

Flujo de diseño de escenas y mecánicas

El desarrollo de cada juego siguió un flujo de trabajo iterativo en Unity que integró recursos gráficos, componentes de física 2D, scripts personalizados y elementos de audio, con el objetivo de garantizar tanto la coherencia visual como la claridad de las mecánicas de interacción. La Figura 1 ilustra este flujo en el caso del juego “Me gusta estar limpio”, mostrando la configuración interna de una escena típica en el entorno de desarrollo.

En el panel de “jerarquía” se organizan los elementos principales de la escena (fondo, objetos interactivos, botones y sistema de eventos), lo que permite estructurar el espacio del baño en capas lógicas y facilitar la reutilización de componentes en diferentes escenas. La “vista de

escena” presenta el entorno bidimensional con una cuadrícula superpuesta, utilizada durante el diseño para alinear los objetos y definir las zonas activas de interacción en función de la altura y la posición de los niños en relación con la pantalla.

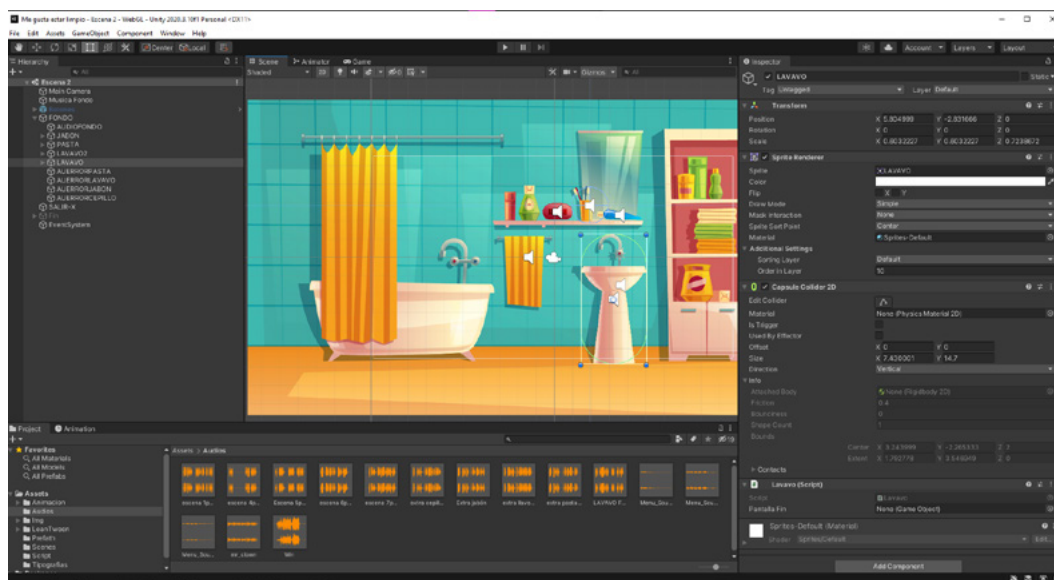
En el “inspector” se observa la configuración del objeto “lavabo”, que combina un componente SpriteRenderer para su representación visual, un CapsuleCollider2D y un RigidBody2D para gestionar las colisiones y la detección de eventos de clic, así como un script específico asociado al comportamiento del objeto. Esta combinación de componentes permite que la interacción de los niños, como, por ejemplo, hacer clic en la llave del lavabo, el jabón, el cepillo de dientes y la pasta dental, se traduzca en eventos programados que actualicen el estado del juego, desencadenen animaciones y activen mensajes de retroalimentación auditiva.

En el panel de “assets” (recursos) se aprecia la biblioteca de clips de audio asociados a la escena, que incluye instrucciones, refuerzos positivos y sonidos de logro. Estos recursos se vinculan a los eventos definidos en los scripts de cada objeto, de modo que la experiencia de juego integra de forma coherente la dimensión visual, la interacción y la guía auditiva, en consonancia con el principio de acompañamiento mediado que orienta el diseño de la plataforma.

Es importante señalar que la Figura 1 no pretende mostrar la totalidad del código ni todos los eventos programados, sino ejemplificar el tipo de arquitectura interna utilizada; un proceso de diseño similar, combinando jerarquías de objetos, componentes de física 2D, scripts específicos y recursos de audio, se aplica de manera sistemática en la construcción de los cuatro juegos que conforman “Estación Infantil”.

Figura 1

Entorno de desarrollo en Unity: arquitectura técnica y configuración de componentes para el juego “Me gusta estar limpio”



Implementación de mecánicas

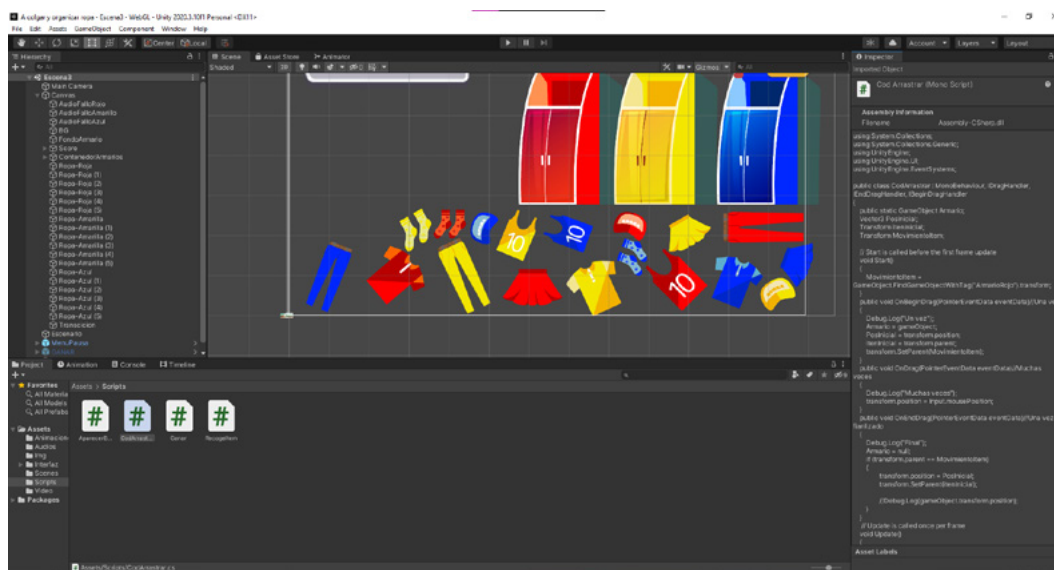
Además de la organización de escenas y recursos, el diseño de “Estación Infantil” requirió la programación de mecánicas de interacción específicas que operan como puente entre los objetivos pedagógicos y la experiencia de juego. La Figura 2 muestra un ejemplo de esta implementación en el juego “A colgar y organizar la ropa”, donde se utiliza un script en C# (CodArrastrar) para gestionar la mecánica de arrastrar y soltar prendas hacia el armario del color correspondiente.

En la escena se dispone un conjunto de objetos de ropa y tres armarios codificados por color (rojo, amarillo y azul), cada uno con sus respectivos colliders y etiquetas. El script *CodArrastrar*, asociado a las prendas mediante el “inspector”, implementa las interfaces *IBeginDragHandler*, *IDragHandler* y *IEndDragHandler* del sistema de eventos de Unity, lo que permite capturar el inicio del arrastre, el movimiento del objeto siguiendo la posición del puntero y la liberación final sobre un área de destino.

De forma simplificada, la lógica del script puede describirse en tres pasos: (a) al iniciar el arrastre, la prenda se desacopla de su posición original y se asocia a un objeto “contenedor” que gestiona el movimiento; (b) durante el arrastre, la posición del sprite se actualiza continuamente según las coordenadas del puntero; y (c) al finalizar el arrastre, el script verifica si la prenda ha sido soltada sobre el armario correcto, en cuyo caso mantiene la nueva posición y activa un refuerzo positivo, de lo contrario, la devuelve a su posición inicial y emite un audio de intento fallido. Este patrón de interacción traduce la destreza curricular de clasificación por color en una mecánica concreta de manipulación y retroalimentación inmediata, alineada con los principios de gamificación y con la teoría del flujo aplicada a tareas de dificultad moderada para niños de 3 a 5 años. En la Figura 2 puede observarse el entorno de Unity con el script.

Figura 2

Implementación de la mecánica de arrastrar y soltar en “A colgar y organizar la ropa”



Alineación con el currículo nacional

Una de las contribuciones centrales de “Estación Infantil” es su alineación explícita y verificable con el currículo ecuatoriano vigente. A diferencia de plataformas genéricas como Kahoot, Educaplay o Cokitos, que ofrecen recursos de uso general desvinculados de marcos curriculares específicos, “Estación Infantil” fue diseñada tomando como referencia directa el Currículo de Educación Inicial 2014 del Ministerio de Educación del Ecuador (2014) y las fichas de experiencias de aprendizaje del plan “Aprendamos Juntos en Casa” (2020/2022). La reciente publicación del Currículo Priorizado de Educación Inicial (2025) confirma esta alineación: el propio documento ministerial declara que mantiene las características y estructura del currículo de 2014 y, al mismo tiempo, introduce un énfasis explícito en habilidades digitales que la plataforma ya abordaba desde su diseño original.

La Tabla 1 presenta el mapeo entre cada juego de la plataforma, el ámbito curricular correspondiente, la destreza específica del Currículo Priorizado 2025 y la experiencia de aprendizaje del plan ministerial de la que se desprende.

Tabla 1
Alineación curricular de los juegos de "Estación Infantil"

Juego	Ámbito curricular (2025)	Destreza específica vinculada	Experiencia "Aprendamos Juntos en Casa"
Me gusta estar limpio	Identidad y autonomía	Practicar hábitos de higiene personal (lavado de manos, cara) de forma progresivamente autónoma.	Ficha n. °5: "Me gusta estar limpio"
A colgar y organizar la ropa	Identidad y autonomía	Practicar hábitos de orden ubicando objetos en el lugar correspondiente; reconocimiento de colores.	Ficha n.º 7: "Organizar la ropa"
Soy artista, me gusta la pintura	Expresión artística	Realizar actividades creativas con técnicas grafo-plásticas; reconocer colores secundarios en objetos e imágenes.	Ficha n.º 24: "Me gusta la pintura"
Organizo mis juguetes	Expresión corporal y motricidad / identidad y autonomía	Realizar actividades de coordinación visomotriz con niveles crecientes de dificultad; practicar hábitos de orden y concentración.	Ficha n.º 10: "Organizo mis juguetes"

Descripción de la plataforma y sus actividades

La interfaz principal de "Estación Infantil", mostrada en la Figura 3(a), presenta un diseño visual atractivo y centralizado, que ofrece acceso directo a cada uno de los cuatro módulos de juego. La navegación es intuitiva y está mediada por instrucciones de audio, de manera que el docente puede guiar la experiencia sin que el niño necesite leer. La Figura 3(b) muestra las instrucciones y el enlace de entrada para algunos juegos, en particular, los del juego "Soy artista, me gusta la pintura".

Figura 3
Interfaz principal de "Estación Infantil"



Cada uno de los cuatro juegos se describe a continuación:

Me gusta estar limpio

Este juego desarrolla hábitos de higiene personal mediante la simulación de rutinas de aseo en un entorno de baño interactivo. La niña o el niño hace clic en una secuencia de objetos —grifo, jabón, cepillo y pasta de dientes— para completar la rutina de lavado de manos y cepillado de dientes. El refuerzo positivo se emite mediante animaciones y voces en off al

completar correctamente cada paso. La mecánica de interacción es secuencial y se basa en clic directo, apropiada para niños de 3 a 5 años, siempre mediada por un docente. La Figura 4 muestra el escenario de este juego.

Figura 4

Escenario del juego “Me gusta estar limpio” en modo de ejecución

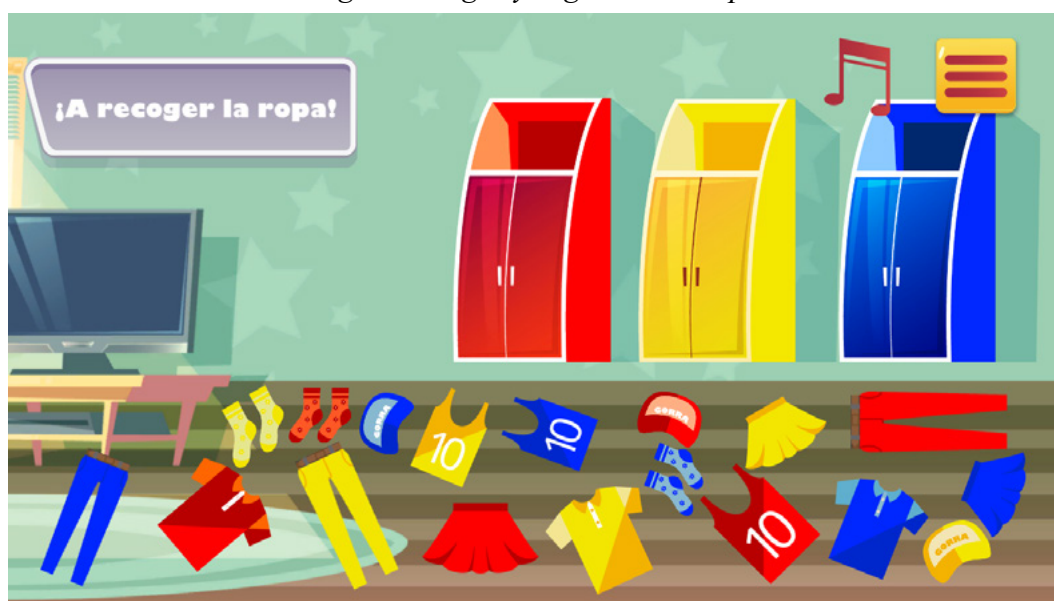


A colgar y organizar la ropa

Este juego está diseñado para desarrollar la clasificación por colores y los hábitos de orden. En él se solicita a los niños arrastrar prendas de vestir hacia el armario del color correspondiente mediante la mecánica de arrastrar y soltar. La progresión visual y auditiva confirma el acierto o sugiere intentarlo de nuevo, sin penalización negativa. La Figura 5 muestra parte del escenario del juego.

Figura 5

Juego “A colgar y organizar la ropa”

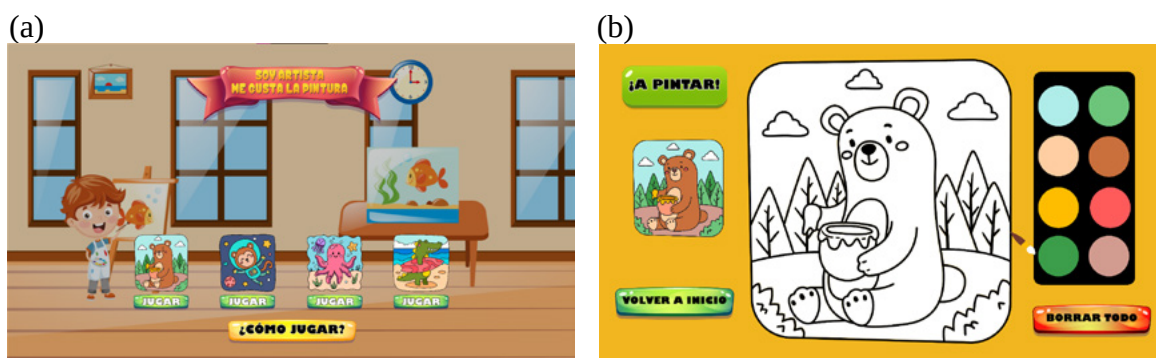


Soy artista, me gusta la pintura

Esta actividad promueve la expresión artística y el reconocimiento de colores. El niño o la niña selecciona colores de una paleta y los aplica sobre secciones de un dibujo en blanco, con una imagen de referencia visible como guía, tal como se observa en la Figura 6(b). El juego estimula la atención, la coordinación fina y la asociación entre nombre y percepción del color, en línea con las destrezas de expresión artística del currículo. En la Figura 6(a) se observa la interfaz de entrada del juego, en la que se da a elegir el objeto para pintar.

Figura 6

Juego "Soy artista, me gusta la pintura"

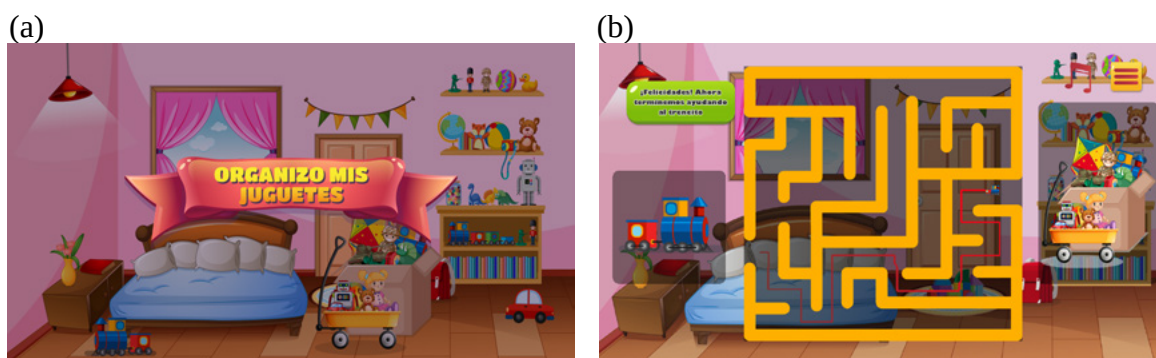


Organizo mis juguetes

Este juego refuerza la concentración, la orientación espacial y la coordinación visomotriz mediante un recorrido tipo laberinto en el que la niña o el niño guía un juguete por un camino predefinido hasta un punto de llegada. La interacción se realiza a través de las teclas de dirección del teclado, lo que permite trabajar el control del movimiento en dos ejes y la anticipación de trayectorias. El diseño del laberinto fue calibrado para niños de 4 a 5 años, incrementando ligeramente la complejidad respecto a los otros juegos: se requiere planificar rutas, evitar obstáculos y mantener la atención sostenida para completar el recorrido. Al finalizar el recorrido correctamente, el sistema activa una animación de celebración y un refuerzo sonoro, cerrando el ciclo de experiencia en coherencia con el sistema de retroalimentación positiva de la plataforma. En la Figura 7(a) puede observarse la interfaz de entrada del juego antes de acceder al menú de juego; y en la Figura 7(b) se muestra uno de los retos para organizar los juguetes, que consiste en llevar el tren hacia el contenedor destinado a guardar los juguetes.

Figura 7

Juego "Organizo mis juguetes"



Los cuatro juegos están integrados en la plataforma Google Sites, que actúa como interfaz web de acceso y los incrusta directamente mediante enlaces a juegos independientes,

garantizando una experiencia de acceso fluida y sin necesidad de descargas adicionales. Al finalizar cada módulo, el sistema emite una insignia visual y un refuerzo sonoro que actúa como indicador de logro para el estudiante y como señal de finalización verificable por el docente.

Participantes

La validación tecnopedagógica de la plataforma “Estación Infantil” se realizó en 2026 con una muestra intencional de 41 participantes vinculados a programas de formación docente en educación inicial de una universidad ecuatoriana. Se consideraron cuatro perfiles diferenciados para captar diversas perspectivas de experticia pedagógica:

Estudiantes de 6.º nivel, modalidad semipresencial

Este grupo estuvo conformado por ocho participantes, todos ellos docentes titulados en Educación Inicial que cursan una segunda titulación universitaria en Educación Básica. Este grupo aporta el criterio de profesionales en ejercicio que, además, se encuentran en un proceso de actualización académica.

Estudiantes de 6.º nivel, modalidad presencial

Compuesto por veinte participantes en etapa avanzada de formación inicial como docentes de Educación Inicial. Su participación permite valorar la plataforma desde la perspectiva de futuros docentes que se encuentran en contacto directo con el currículo priorizado vigente.

Estudiantes de 8.º nivel en proceso de titulación, modalidad presencial

Este grupo cuenta con siete participantes, quienes han culminado prácticamente su trayectoria de formación y se hallan elaborando su trabajo de titulación. Este perfil aporta una mirada más crítica y madura sobre la pertinencia curricular y el diseño de recursos didácticos.

Docentes expertos en educación con formación de maestría o doctorado

En este grupo, los seis participantes cumplen el rol de jueces expertos para la validación de contenido tecnopedagógico de la plataforma. Su perfil profesional se caracteriza por una sólida y avanzada formación académica. Todos poseen títulos de licenciatura y maestría en áreas afines: ciencias de la educación, educación inicial/parvularia, psicología educativa, neuropsicología. Actualmente, cinco de ellos se encuentran en fase final de doctorado en educación, mientras que uno cursa activamente dicho programa. Asimismo, cuentan con una amplia trayectoria que oscila entre 11 y 40 años de experiencia combinada, desempeñando funciones de docencia universitaria, dirección de centros educativos, práctica directa en aulas de educación inicial y funciones de evaluación en el Ministerio de Educación. Esta alta cualificación permite contar con un criterio riguroso y fundamentado para evaluar la pertinencia curricular, la calidad del diseño instruccional y la viabilidad del recurso digital en contextos reales de la primera infancia.

Se empleó un muestreo intencional, dado que se buscó incluir participantes con experiencia o formación específica en el ámbito de la educación inicial y la tecnología educativa. Todos los participantes fueron informados sobre el propósito académico de la investigación y respondieron de forma voluntaria al instrumento de evaluación.

Instrumento

Para la validación se diseñó un instrumento ad hoc de percepción tecnopedagógica de la plataforma “Estación Infantil”, implementado como formulario digital en Google Forms. El formulario inicia con una breve explicación del objetivo del estudio y una nota metodológica que enfatiza que la plataforma está concebida para un uso bajo acompañamiento mediado, es decir, en interacción guiada por el docente y no como recurso autónomo de los niños.

El instrumento está compuesto por siete ítems obligatorios, cada uno asociado a una dimensión específica:

- D1. Recurso didáctico: eficacia de la plataforma para alcanzar objetivos de aprendizaje en la infancia.
- D2. Claridad del material multimedia e instrucciones.
- D3. Pertinencia de las actividades para subniveles de educación inicial (3 a 5 años).
- D4. Diseño de módulos, colores y dinámicas para despertar interés intrínseco y atención.
- D5. Calidad y oportunidad del sistema de retroalimentación (voces en off y efectos visuales).
- D6. Tipo de letra y legibilidad de textos en contexto de supervisión docente en el aula.
- D7. Usabilidad, controles (ratón/teclado) y jugabilidad general del entorno interactivo.

Cada ítem se responde mediante una escala de Likert de 5 puntos con anclas verbales claramente definidas: 1 = "Totalmente en desacuerdo (no cumple en absoluto)", 2 = "En desacuerdo (presenta falencias importantes)", 3 = "Ni de acuerdo ni en desacuerdo (cumplimiento medio)", 4 = "De acuerdo (cumple satisfactoriamente)" y 5 = "Totalmente de acuerdo (cumple de forma excelente)". Adicionalmente, se registra el perfil del participante evaluador, lo que permite el análisis comparativo por grupos.

La consistencia interna del instrumento se evaluó mediante el coeficiente alfa de Cronbach (Cronbach, 1951), calculado con el software estadístico jamovi (2025) a partir de las 41 respuestas válidas, con lo que se obtuvo un valor de $\alpha \approx 0,71$ para los siete ítems. Este coeficiente es adecuado para escalas de Likert orientadas a estimar la coherencia interna de un conjunto de ítems relacionados y se interpreta como un indicador de fiabilidad aceptable para un instrumento multidimensional en fase de validación inicial.

Es importante precisar que, aunque el instrumento contempla siete dimensiones específicas, el alfa de Cronbach se calculó sobre la escala total con el fin de estimar la fiabilidad del constructo global de percepción tecnopedagógica, asumiendo que las facetas evaluadas convergen en una valoración integral y coherente del recurso educativo.

Procedimiento

La validación se desarrolló en sesiones presenciales supervisadas por el investigador principal, siguiendo el modelo de uso previsto de la plataforma. En cada sesión, los participantes:

- Accedieron individualmente al sitio web de la plataforma "Estación Infantil".
- Interactuaron con los cuatro juegos ("Me gusta estar limpio", "A colgar y organizar la ropa", "Soy artista, me gusta la pintura" y "Organizo mis juguetes"), completando cada uno de los módulos en secuencia.
- La finalización de cada juego fue verificada mediante el sistema interno de retroalimentación de la plataforma, que emite una insignia visual y un refuerzo sonoro al concluir exitosamente la actividad, lo que funcionó como marcador observable de cumplimiento.
- Una vez completados los cuatro juegos, los participantes respondieron el formulario digital de validación tecnopedagógica desde sus propios dispositivos, bajo la supervisión del investigador, quien se aseguró de que se resolvieran las dudas de comprensión del instrumento, sin influir en las respuestas.

Este procedimiento garantizó que las valoraciones se basaran en experiencias de uso directo y completo de la plataforma, en un contexto de acompañamiento docente coherente con las orientaciones del Currículo Priorizado de Educación Inicial 2025 respecto al uso guiado de recursos digitales en la primera infancia.

Análisis de datos

Los datos se exportaron desde Google Forms en formato CSV y fueron cargados en jamovi para el análisis estadístico. Con la ayuda del software, estos datos se procesaron mediante estadística descriptiva (medias, desviaciones estándar, medianas, valores mínimos y máximos) para cada una de las siete dimensiones, tanto para el total de la muestra como segmentados por perfil de participante.

Además de los descriptivos globales, se generaron tablas de medias por grupo de participantes con el fin de comparar la percepción tecnopedagógica según el perfil formativo, es decir, docentes en ejercicio, estudiantes de niveles avanzados y docentes expertos.

Resultados y Discusión

Descriptivos globales

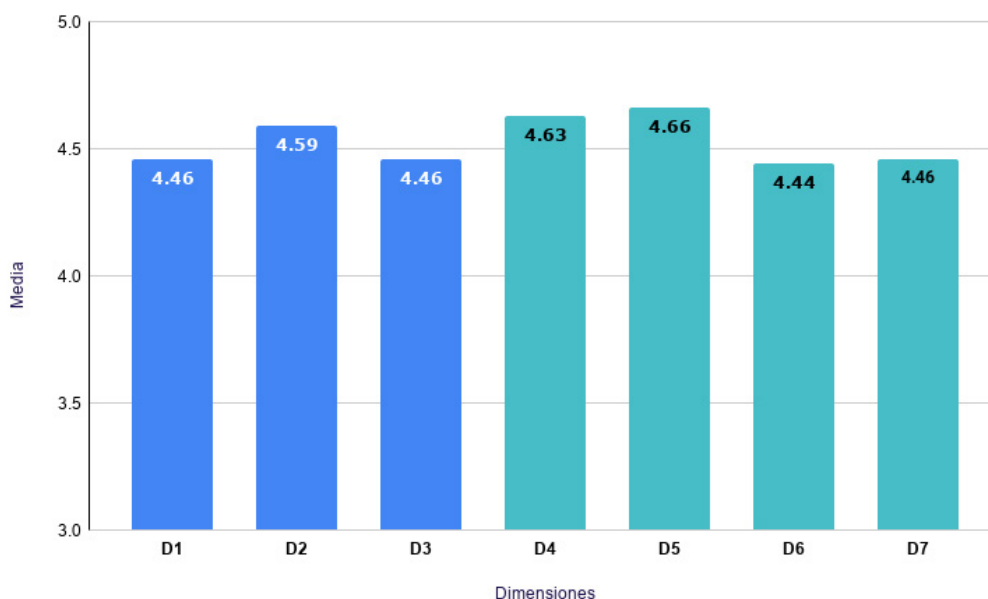
La Tabla 2 resume los estadísticos descriptivos de las siete dimensiones de la escala (1 = totalmente en desacuerdo, 5 = totalmente de acuerdo) para el total de 41 participantes.

Tabla 2
Estadísticos descriptivos globales de la percepción tecnopedagógica (n = 41)

Dimensión	Enunciado abreviado	Media	Desv. est.	Mediana	Mín-Máx
D1	Plataforma como recurso didáctico	4,46	0,67	5	3–5
D2	Claridad del material e instrucciones	4,59	0,50	5	4–5
D3	Pertinencia para niños/as de 3 a 5 años	4,46	0,74	5	2–5
D4	Diseño, colores e interés intrínseco	4,63	0,58	5	3–5
D5	Retroalimentación audiovisual	4,66	0,53	5	3–5
D6	Legibilidad de textos para supervisión docente	4,44	0,74	5	2–5
D7	Usabilidad y jugabilidad	4,46	0,60	5	3–5

La Figura 8 resume visualmente las medias globales por dimensión e ilustra la diferencia entre los componentes pedagógicos de la escala (D1–D3) mostrados en color azul y los componentes tecnoexperienciales (D4–D7) mostrados en color verde. Se observa que las valoraciones más altas se concentran en el diseño motivacional (D4) y la retroalimentación audiovisual (D5), mientras que la legibilidad de textos para docentes (D6) presenta la media relativamente más baja, aunque se mantiene por encima de 4,4 puntos.

En términos globales, todas las dimensiones se sitúan entre 4,44 y 4,66 puntos, lo que indica niveles de acuerdo elevados respecto a la eficacia didáctica, pertinencia curricular y calidad tecnológica de la plataforma. La mediana es 5 en las siete dimensiones, lo que muestra que la mayoría de los participantes tiende a valorar la experiencia en el extremo superior de la escala.

Figura 8*Medias globales por dimensión de percepción tecnopedagógica (n = 41)***Resultados por tipo de participante**

Se distinguieron cuatro perfiles de participantes: estudiantes de 6.º nivel, modalidad semipresencial (segunda carrera, docentes ya titulados), estudiantes de 6.º nivel, modalidad presencial, estudiantes de 8.º nivel en proceso de titulación, y docentes expertos con formación de maestría o doctorado en educación. En la Tabla 3 puede observarse la media de cada una de las dimensiones medidas correspondientes a los distintos grupos de participantes.

Tabla 3*Medias por dimensión según perfil de participante*

Grupo	D1	D2	D3	D4	D5	D6	D7
6.º semipresencial (2.ª carrera, n = 8)	4,62	4,75	4,50	4,75	4,62	4,88	4,00
6.º presencial (n = 20)	4,55	4,45	4,35	4,70	4,80	4,70	4,55
8.º titulación (n = 7)	4,43	4,57	4,57	4,29	4,14	4,14	4,29
Docentes expertos (n = 6)	4,00	4,83	4,67	4,67	4,83	3,33	5,00

Los docentes titulados de 6.º semipresencial muestran un patrón homogéneo y alto, con medias entre 4,00 y 4,88, lo que respalda la pertinencia pedagógica percibida por profesionales en ejercicio. Los estudiantes de 6.º presencial valoran especialmente la calidad de la retroalimentación (D5 = 4,80) y el diseño visual (D4 = 4,70), sugiriendo una experiencia de usuario positiva en términos de motivación e interés.

El grupo de 8.º nivel en titulación mantiene medias iguales o superiores a 4 en todas las dimensiones, aunque es algo más crítico en diseño y legibilidad (D4–D6 ≈ 4,14–4,29), lo que puede asociarse a una mayor experiencia acumulada en prácticas educativas. Los docentes expertos presentan un patrón interesante: otorgan las medias más altas a claridad (D2), pertinencia (D3), diseño (D4) y retroalimentación (D5), y evalúan la usabilidad/jugabilidad con la máxima puntuación posible (D7 = 5,00), mientras que son más exigentes en la legibilidad de textos (D6 = 3,33).

Consistencia interna del instrumento

La consistencia interna de la escala de siete ítems se evaluó mediante el coeficiente alfa de Cronbach, obteniéndose un valor de $\alpha \approx 0,71$ sobre las 41 respuestas. Es importante precisar que, si bien el instrumento aborda múltiples facetas, el cálculo del alfa se realizó sobre la escala total para estimar la fiabilidad del constructo global de “percepción tecnopedagógica”. Este resultado indica una fiabilidad aceptable para la escala total como medida integral, lo que permite interpretar los resultados con confianza razonable, ya que refleja la coherencia general de los ítems al converger en una valoración global que integra tanto componentes pedagógicos (D1–D3) como tecnoexperienciales (D4–D7) en esta fase de validación inicial.

Análisis inferencial entre perfiles de participantes

Con el fin de comparar estadísticamente las percepciones según el perfil de los participantes, se aplicó la prueba no paramétrica de Kruskal–Wallis, cuyos resultados se presentan en la Tabla 4. No se identificaron diferencias estadísticamente significativas entre los cuatro perfiles en las dimensiones D1 a D5 ($p > .05$). En cambio, se observaron diferencias globales significativas en D6, legibilidad de textos para supervisión docente, $\chi^2(3) = 17.714$, $p < .001$, y en D7, usabilidad y jugabilidad general, $\chi^2(3) = 10.962$, $p = .012$. Estos resultados indican que la valoración de estas dos dimensiones varió según el perfil de los participantes; no obstante, por el carácter exploratorio del estudio y el tamaño reducido de algunos grupos, las diferencias deben interpretarse con cautela.

Tabla 4

Prueba de Kruskal–Wallis para comparar las valoraciones según perfil de participante

Dimensión	χ^2	gl	p	Interpretación
D1. Recurso didáctico	4.587	3	.205	No significativa
D2. Claridad	3.833	3	.280	No significativa
D3. Pertinencia	0.492	3	.921	No significativa
D4. Diseño motivacional	2.863	3	.413	No significativa
D5. Retroalimentación	7.690	3	.053	No significativa
D6. Legibilidad	17.714	3	< .001	Significativa
D7. Usabilidad y jugabilidad	10.962	3	.012	Significativa

Discusión de los resultados

Los resultados obtenidos confirman que “Estación Infantil” presenta una valoración tecnopedagógica alta y consistente en los distintos perfiles de participantes evaluados, lo que respalda su viabilidad como recurso digital de refuerzo para el subnivel de educación inicial. Las medias globales entre 4,44 y 4,66 en las siete dimensiones, junto con medianas de 5 en todos los casos, sugieren que la plataforma logra articular de manera satisfactoria aspectos pedagógicos, estéticos y funcionales en un mismo entorno interactivo.

Un primer hallazgo relevante es la coherencia entre la validación empírica y la alineación curricular documentada en el diseño de la plataforma. “Estación Infantil” no se construyó como un videojuego genérico adaptado posteriormente al aula, sino como un recurso concebido desde su origen a partir de experiencias de aprendizaje y destrezas específicas del Currículo de Educación Inicial y del plan “Aprendamos Juntos en Casa”. En ese sentido, los puntajes altos en D1 (plataforma como recurso didáctico) y D3 (pertinencia para niños de 3 a 5 años) indican que los participantes reconocen esta correspondencia entre mecánica lúdica y finalidad educativa, lo cual fortalece la validez de contenido de la propuesta.

Un segundo aspecto para destacar es la fortaleza de las dimensiones vinculadas con la experiencia de los niños y el diseño motivacional. Las puntuaciones elevadas en D4 (diseño, colores e interés intrínseco) y D5 (retroalimentación audiovisual) muestran que la plataforma logra captar la atención y sostener la participación de los niños mediante elementos propios de la gamificación, como el refuerzo inmediato, la claridad de objetivos y la progresión de tareas. Esto es congruente con los fundamentos teóricos asumidos en el desarrollo de cada juego, especialmente con la gamificación educativa y la teoría del flujo, ya que un entorno con instrucciones claras, desafíos ajustados y retroalimentación oportuna favorece experiencias de inmersión y aprendizaje significativo en la primera infancia.

Estas diferencias en las valoraciones entre dimensiones reflejan tanto las prioridades de diseño de la plataforma como las expectativas de los evaluadores. Las puntuaciones más altas en D4 (diseño motivacional) y D5 (retroalimentación audiovisual) responden a la priorización deliberada de estímulos multisensoriales y una estética visual atractiva, elementos que activan de forma directa la motivación intrínseca en la primera infancia. Por el contrario, la puntuación relativamente más baja en D6 (legibilidad de textos) puede atribuirse a un sesgo en las expectativas: aunque la plataforma está diseñada para un uso mediado por el docente, algunos evaluadores proyectaron criterios de lectura autónoma infantil, lo que elevó las exigencias sobre el tamaño y diseño tipográfico. Esta brecha no debilita la propuesta, sino que señala un aspecto puntual y técnicamente abordable en futuras iteraciones.

En este sentido, la valoración positiva de la retroalimentación audiovisual (D5) es coherente con estudios actuales que demuestran que la retroalimentación inmediata y multisensorial es uno de los predictores más fuertes del aprendizaje en entornos gamificados (Fabre-Mitjans et al., 2025). En el contexto de la primera infancia, Trawick-Smith (2025) subraya que el rol del docente como mediador es determinante para que el juego digital se traduzca en aprendizajes significativos, lo cual valida el principio de acompañamiento mediado que orienta el diseño de "Estación Infantil". Finalmente, la evaluación rigurosa de estos recursos digitales requiere analizar no solo su funcionalidad técnica, sino también sus características pedagógicas y de contenido, tal como lo destacan los modelos actuales de evaluación de recursos educativos (Rodríguez et al., 2025).

La valoración otorgada por los distintos grupos permite además observar matices importantes. Los estudiantes de 6.º nivel en modalidad presencial otorgaron puntuaciones elevadas a la retroalimentación y el diseño visual, mientras que los docentes titulados de modalidad semipresencial ofrecieron un patrón de respuestas más homogéneo y alto en casi todas las dimensiones. Este comportamiento sugiere que quienes ya poseen experiencia docente en educación inicial perciben la plataforma no solo como visualmente atractiva, sino también como funcionalmente útil dentro de una lógica de mediación pedagógica.

Por su parte, el grupo de docentes expertos aporta una lectura especialmente valiosa para la interpretación de los hallazgos. Aunque este grupo fue más exigente en la dimensión de legibilidad (D6 = 3,33), evaluó con puntuaciones muy altas la claridad de instrucciones, la pertinencia curricular, la retroalimentación y, de manera sobresaliente, la usabilidad y jugabilidad general del entorno (D7 = 5,00). En las sesiones de aplicación del instrumento se observó que, pese a que se enfatizó el uso mediado de la plataforma, algunos participantes valoraron los textos como si estuvieran dirigidos a la lectura autónoma de niños, lo que pudo elevar las expectativas respecto de la competencia lectora y del tamaño y diseño tipográfico y pudo contribuir a la menor puntuación en D6. Este patrón sugiere que la crítica de los expertos no debilita la propuesta, sino que se concentra en un aspecto puntual y técnicamente abordable: explicitar con mayor

claridad el rol del adulto mediador en la interfaz y optimizar los textos y elementos visuales de apoyo al docente para evitar ambigüedades en su interpretación.

El comportamiento del grupo de 8.º nivel en proceso de titulación también merece atención. Aunque sus medias se mantuvieron por encima de 4 en todas las dimensiones, este grupo mostró valoraciones relativamente más cautelosas en D4, D5 y D6. Una posible interpretación es que estos participantes, al encontrarse en una etapa avanzada de formación y reflexión profesional, aplican criterios más analíticos al evaluar recursos educativos digitales, especialmente en aspectos de interfaz, claridad visual y adecuación para el uso en aula. Lejos de debilitar el estudio, esta variabilidad enriquece la interpretación de los resultados y muestra que la plataforma resiste evaluaciones realizadas desde distintos niveles de experiencia pedagógica.

Desde una perspectiva metodológica, la consistencia interna del instrumento ($\alpha \approx 0,71$) respalda su uso como herramienta de evaluación inicial de calidad tecnopedagógica. Aunque no se trata de una escala destinada a medir un constructo estrictamente unidimensional, el valor obtenido indica una coherencia adecuada entre los siete ítems y permite interpretar los resultados con confianza razonable. Esto es particularmente relevante porque el instrumento integra dimensiones pedagógicas (eficacia, pertinencia, claridad) con dimensiones tecnoexperienciales (diseño, retroalimentación, legibilidad y usabilidad), lo que exige un equilibrio entre diversidad conceptual y consistencia interna.

En conjunto, los hallazgos permiten sostener que “Estación Infantil” responde a una necesidad concreta del campo de la educación inicial: disponer de recursos digitales contextualizados, curriculares y técnicamente viables para apoyar procesos de reforzamiento. A diferencia de plataformas genéricas, la propuesta articula destrezas del currículo ecuatoriano con mecánicas de interacción diseñadas específicamente para niñas y niños de 3 a 5 años, y esta especificidad es reconocida positivamente por los participantes evaluadores.

No obstante, el estudio presenta limitaciones que deben considerarse para la interpretación de los hallazgos y para futuras investigaciones. En primer lugar, la muestra fue de tipo intencional y se concentró en participantes vinculados a la formación docente en una misma institución, lo que puede introducir un sesgo institucional y limita la generalización estadística de los resultados. En segundo lugar, se reconoce una limitación metodológica en el instrumento de validación: al contemplar un único ítem por dimensión, se restringe la posibilidad de calcular la confiabilidad específica por subescala o realizar un análisis factorial. Sin embargo, este diseño se adoptó como una validación exploratoria inicial del constructo global de percepción tecnopedagógica, priorizando la brevedad para los evaluadores, y el alfa de Cronbach global ($\alpha \approx 0,71$) ofrece una consistencia interna aceptable para esta fase. Futuras investigaciones deberán expandir el instrumento a entre tres y cuatro ítems por dimensión. Asimismo, cabe precisar que la validación cuantitativa es de naturaleza perceptiva; futuros estudios deberán incorporar observación directa de niños en interacción mediada, análisis de desempeño por actividad o diseños cuasiexperimentales que permitan estimar efectos sobre aprendizajes específicos.

Respecto a las mecánicas de interacción, es importante aclarar que las acciones de clic o arrastrar no están diseñadas para ser ejecutadas de forma autónoma con un ratón por el niño, sino que operan bajo el principio de acompañamiento mediado. Además, desde una perspectiva técnica, el evento de clic en la plataforma WebGL es compatible con pantallas táctiles, como computadoras portátiles 2 en 1 o pizarras digitales interactivas, permitiendo que la interacción se realice directamente con el dedo y mitigando las dificultades de motricidad fina asociadas al

ratón físico. No obstante, se reconoce como una limitación técnica que la mecánica del juego de laberinto "Organizo mis juguetes" dependa de las teclas de dirección del teclado, lo cual podría dificultar la experiencia en dispositivos exclusivamente táctiles. Este aspecto, junto con la implementación de controles virtuales en pantalla para dispositivos móviles, se priorizará para futuras iteraciones de la plataforma.

En consecuencia, la principal contribución de este trabajo no radica únicamente en presentar un software funcional, sino en mostrar un modelo de desarrollo de recurso educativo digital que combina diseño técnico, alineación curricular y validación tecnopedagógica con múltiples perfiles de evaluadores. Esta integración fortalece la relevancia de "Estación Infantil" como propuesta transferible y abre una línea de trabajo para futuras iteraciones orientadas a mejorar la legibilidad, ampliar la variedad de juegos y profundizar en la evaluación de impacto en contextos escolares reales.

Conclusiones

El estudio evidenció que la plataforma gamificada "Estación Infantil" alcanza niveles elevados de aceptación tecnopedagógica entre docentes en formación, docentes titulados y docentes expertos, con medias globales superiores a 4,4 en las siete dimensiones evaluadas y una consistencia interna adecuada del instrumento ($\alpha \approx 0,71$). Estos resultados respaldan la viabilidad del sistema como recurso didáctico de refuerzo para el subnivel de educación inicial en contextos de acompañamiento mediado.

La articulación explícita entre los juegos de la plataforma, las experiencias de aprendizaje del plan "Aprendamos Juntos en Casa" y las destrezas del Currículo Priorizado de Educación Inicial 2025 constituye una contribución relevante, al ofrecer un ejemplo concreto de cómo traducir directrices curriculares en entornos interactivos digitales. Las altas puntuaciones en pertinencia (D3), diseño motivacional (D4) y retroalimentación (D5) sugieren que esta alineación curricular es percibida positivamente por quienes se están formando o ya ejercen como docentes.

Desde la perspectiva técnica, la arquitectura híbrida basada en Unity y Google Sites, junto con la implementación de mecánicas específicas como el arrastrar y soltar y los sistemas de retroalimentación audio-visual, muestra que es posible desarrollar recursos WebGL accesibles y funcionales para el aula de educación inicial sin requerir infraestructuras complejas. El reconocimiento de la usabilidad y jugabilidad por parte de los docentes expertos refuerza la pertinencia de estas decisiones de diseño para contextos educativos reales.

No obstante, las valoraciones relativamente menores en la dimensión de legibilidad de textos, especialmente entre docentes expertos y estudiantes de 8.º nivel, señalan la necesidad de continuar optimizando los elementos gráficos y tipográficos dirigidos al docente. Futuras iteraciones del proyecto deberían incorporar ajustes en este aspecto, así como explorar nuevas actividades que amplíen la cobertura de destrezas socioemocionales y digitales en coherencia con las prioridades del currículo vigente.

Finalmente, aunque la muestra principal de percepción tecnopedagógica es intencional y se circunscribe a una sola institución de formación docente, el diseño metodológico se realizó en dos fases: se hizo una fase exploratoria con niños en un contexto escolar real bajo mediación docente y una fase de validación cuantitativa con 41 participantes vinculados a la educación inicial. Los resultados obtenidos sugieren que "Estación Infantil" puede servir como modelo replicable para el desarrollo de otras plataformas gamificadas en este nivel, combinando alineación curricular, soporte tecnológico accesible y validación tecnopedagógica con múltiples

perfiles de evaluadores. La continuidad de esta línea de trabajo podría contribuir a consolidar un ecosistema de recursos digitales contextualizados que acompañen la implementación del Currículo Priorizado de Educación Inicial 2025 en diferentes territorios y realidades escolares.

Implicaciones educativas y líneas futuras

Los hallazgos tienen implicaciones prácticas directas para el aula: la plataforma puede integrarse como recurso de refuerzo en sesiones guiadas de 10 a 15 minutos, en las que la alineación curricular explícita facilita su adopción en la planificación docente sin requerir infraestructura tecnológica compleja. Como líneas de investigación futura se proponen: (a) la validación con una muestra probabilística multicéntrica que incluya a docentes de diversas instituciones para mitigar sesgos; (b) la expansión del instrumento de evaluación con múltiples ítems por dimensión para permitir un análisis factorial; (c) la implementación de un diseño cuasiexperimental con evaluación previa y evaluación posterior para medir el impacto real en el aprendizaje de destrezas específicas; y (d) la adaptación de las mecánicas de interacción a pantallas táctiles para optimizar la experiencia en la primera infancia.

Reconocimientos y Declaraciones

Los autores declaran que, en la elaboración del presente artículo, no se han utilizado herramientas de inteligencia artificial.

Este trabajo recibió financiamiento institucional para el desarrollo de la investigación y los costos de publicación. Los autores declaran que la entidad financiadora no tuvo injerencia en el diseño del estudio, la recolección de datos, el análisis ni la redacción del manuscrito.

Los autores desean expresar su agradecimiento a los 41 participantes (docentes en formación, docentes titulados y expertos) que contribuyeron con su tiempo y experiencia en la validación tecnopedagógica de la plataforma “Estación Infantil”, así como a las instituciones de educación inicial que facilitaron el acceso para la fase exploratoria del estudio.

Declaración de conflicto de intereses:

Los autores declaran no tener ningún conflicto de interés, financiero o de otro tipo, que pueda haber influido en el trabajo reportado en este manuscrito.

Contribuciones de los autores:

Sebastián Báez: Conceptualización, software, validación inicial, investigación, diseño de recursos, escritura del borrador original, administración del proyecto.

Miguel Cobos: Metodología, validación final, análisis formal, investigación, escritura, revisión y edición.

Referencias

- Capatina, A., Juarez-Varon, D., Micu, A., & Micu, A. E. (2024). Leveling up in corporate training: Unveiling the power of gamification to enhance knowledge retention, knowledge sharing, and job performance. *Journal of Innovation & Knowledge*, 9(3), 100530. <https://doi.org/10.1016/j.jik.2024.100530>
- Cobos, M., & Salvador, P. (2022). Narrative in videogame design. A starting guide. *2nd IEEE International Conference on Advanced Learning Technologies on Education and Research*. <https://doi.org/10.1109/ICALTER57193.2022.9964824>

- Cobos, M., & Salvador, P. (2023). Proposal for the Inclusion of narrative in the video game design process. En A. Mesquita, A. Abreu, J. V. Carvalho, C. Santana, & C. H. P. De Mello (Eds.), *Perspectives and Trends in Education and Technology* (Vol. 366, pp. 457-469). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-5414-8_42
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/BF02310555> (Obra original publicada en 1951)
- Csikszentmihályi, M. (2014). *Applications of flow in human development and education*. Springer.
- Desmurget, M., & Cortés Fernández, L. (2020). *La fábrica de cretinos digitales: Los peligros de las pantallas para nuestros hijos*. Ediciones Península.
- Fabre-Mitjans, N., Jiménez-Valverde, G., Guimerà-Ballesta, G., & Calafell-Subirà, G. (2025). Digital gamification to foster attitudes toward science in early childhood teacher education. *Applied Sciences*, 15(11), 5961. <https://doi.org/10.3390/app15115961>
- García Gutiérrez, Z. D. P., Durán Pérez, Á. P., & Sánchez Quijano, N. J. (2023). Prácticas de enseñanza con primera infancia durante el COVID-19: Relato de la emergencia de los ambientes virtuales en educación inicial. *Praxis*, 19(2), 205-220. <https://doi.org/10.21676/23897856.4612>
- Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Penguin Press.
- jamovi project. (2025). *Jamovi* (Versión 2.7) [Software]. <https://www.jamovi.org>
- Kareem, Y., & Khalid, O. Z. (2024). Gamification in early childhood education: boosting engagement and learning outcomes. *Proceeding International Conference of Innovation Science, Technology, Education, Children and Health*, 316-320.
- Ministerio de Educación del Ecuador. (2014). *Currículo Educación Inicial 2014*.
- Ministerio de Educación del Ecuador. (2022). *Plan educativo: Aprendamos juntos en casa* (1.ª edición). (Obra original publicada en 2020)
- Ministerio de Educación del Ecuador. (2025). *Currículo priorizado de atención y educación de la primera infancia con énfasis en habilidades comunicacionales, lógico matemáticas, digitales y socioemocionales para el nivel de educación inicial 3-5 años*.
- Pedraza Vázquez, M. G., Ley Martínez, J. N., & Ortiz Cisneros, C. E. (2022). El uso de la gamificación como estrategia didáctica en el nivel preescolar. *VECTORES educativos*, 1(1). <https://doi.org/10.56375/ve1.1-12>
- Rodríguez, J., Marín, D., & Sanabria, A. (2025). *EVALUATION OF DIGITAL MATERIALS AND RESOURCES IN EDUCATION ANALYSIS AND PROPOSALS* [Libro]. McGraw Hill.
- Trawick-Smith, J., & Lee, K. (2025). Ethical dimensions of teacher-child play interactions in preschool. En N. Pramling, A. Farrell, C. Wallerstedt, & M. Nilsen (Eds.), *Adults in Play with Children* (Vol. 44, pp. 23-43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99629-0_2

Pulneo: Prototipo móvil gamificado con detección acústica de actividad respiratoria para fisioterapia domiciliaria

Pulneo: Gamified Mobile Prototype with Acoustic Respiratory Activity Detection for Home-Based Physiotherapy

María Gabriela Viñansaca Cabrera¹ <https://orcid.org/0009-0007-9490-4180>, David Fernando Valladarez Muñoz² <https://orcid.org/0009-0002-3160-6243>

¹Universidad Politécnica Salesiana, Cuenca, Ecuador
mvinansacac@ups.edu.ec

²Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA), Cuenca, Ecuador
david.valladarez@cedia.org.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Enviado: 2026/07/12

Aceptado: 2026/08/27

Publicado: 2026/10/14

Resumen

La fisioterapia respiratoria domiciliaria requiere repetición y retroalimentación, pero muchas sesiones se realizan sin supervisión directa. Este artículo presenta Pulneo, un prototipo móvil gamificado que detecta localmente la actividad respiratoria mediante señales acústicas y la traduce en biorretroalimentación visual durante ejercicios guiados. El modelo se entrenó con 69 634 ventanas balanceadas: 34 817 de actividad respiratoria de 1061 sujetos sanos de Coswara y 34 817 negativas de 5 608 archivos de UrbanSound8K, con una partición 70/15/15 por grupos, sin solapamiento de sujetos ni de archivos. En 10 347 ventanas de prueba, el modelo TensorFlow Lite INT8 alcanzó 97,64 % de *accuracy*, 98,09 % de precisión, 97,18 % de *recall*, F1 de 97,63 % y ROC-AUC de 99,65 %. La cuantización redujo el modelo de 88 488 a 31 600 bytes (64,29 %) y mantuvo 99,55 % de concordancia con TFLite FP32. La aplicación interpreta la salida como P(respiración) y aplica un umbral calibrado en validación ($P \geq 0,29$). Pulneo no distingue inspiración de exhalación ni ha sido validado clínicamente; los resultados respaldan su funcionamiento técnico, no su eficacia terapéutica. Un autor fisioterapeuta con maestría en terapia respiratoria realizó una revisión conceptual interna para delimitar el uso previsto, la interpretación de la biorretroalimentación y los requisitos de seguridad para futuras evaluaciones.

Sumario: Introducción, Materiales y Métodos, Resultados, Discusión, Alcance y limitaciones, Conclusiones.

Como citar: Viñansaca, M. & Valladarez, D. (2026). Pulneo: Prototipo móvil gamificado con detección acústica de actividad respiratoria para fisioterapia domiciliaria. *Revista Tecnológica - Espol*, 38(E1), 136-152.
<https://doi.org/10.37815/rte.v38nE1.1602>

Palabras clave: aprendizaje automático, biorretroalimentación, computación en el borde, cuantización, inferencia local, redes neuronales convolucionales.

Abstract

Home-based respiratory physiotherapy requires repetition and feedback, yet many sessions are performed without direct supervision. This paper presents Pulneo, a gamified mobile prototype that detects respiratory activity on-device from acoustic signals and maps it to visual biofeedback during guided exercises. The model was trained on 69,634 balanced windows: 34,817 respiratory-activity windows from 1,061 healthy Coswara subjects and 34,817 negative windows from 5,608 UrbanSound8K files, using a group-wise 70/15/15 split with no subject or source-file overlap. On 10,347 test windows, the TensorFlow Lite INT8 model achieved 97.64% accuracy, 98.09% precision, 97.18% recall, 97.63% F1-score, and 99.65% ROC-AUC. Quantization reduced the model from 88,488 to 31,600 bytes (64.29%) while preserving 99.55% decision agreement with TFLite FP32. The application interprets the output as P(breathing) and applies a threshold calibrated on validation data ($P \geq 0.29$). Pulneo does not distinguish inhalation from exhalation and has not been clinically validated; the findings support technical performance, not therapeutic efficacy. An author who is a physiotherapist with a master's degree in respiratory therapy conducted an internal conceptual review to delimit intended use, biofeedback interpretation, and safety requirements for future evaluations.

Keywords: biofeedback, convolutional neural networks, edge computing, local inference, machine learning, quantization.

Introducción

La rehabilitación pulmonar combina evaluación, entrenamiento, educación y cambio de comportamiento para mejorar la condición física y psicológica de personas con enfermedad respiratoria crónica (Spruit *et al.*, 2013). Sin embargo, la práctica domiciliaria sigue planteando problemas de adherencia, continuidad y ejecución técnica cuando el profesional no está presente. Los programas domiciliarios y de telerehabilitación pueden reducir las barreras de acceso, pero requieren mecanismos comprensibles de seguimiento y de retroalimentación que no sustituyan el juicio clínico (Cox *et al.*, 2021; Holland *et al.*, 2017; McCarthy *et al.*, 2015; Watson *et al.*, 2023).

La gamificación puede aportar metas visibles y refuerzo inmediato en tareas repetitivas y su uso en intervenciones dirigidas a personas con enfermedades respiratorias crónicas se ha estudiado como estrategia para apoyar el compromiso y el manejo de la actividad (Chen *et al.*, 2025). Los teléfonos inteligentes añaden micrófono, capacidad de cómputo local y una interfaz familiar, lo que ha motivado trabajos de análisis respiratorio acústico y biorretroalimentación móvil (Brown *et al.*, 2020; Larson *et al.*, 2012; Shih *et al.*, 2019). Breeze constituye una referencia relevante de interacción porque detecta fases respiratorias en tiempo real; Pulneo no reproduce ese objetivo. El presente trabajo utiliza un detector binario de actividad respiratoria y, por tanto, Breeze se considera un antecedente de biorretroalimentación acústica, mas no un clasificador directamente equivalente.

El procesamiento local resulta pertinente para señales acústicas potencialmente sensibles porque reduce la necesidad de transmitir audio crudo y elimina la dependencia de conectividad para la inferencia. Modelos compactos basados en MFCC y redes convolucionales pueden ejecutarse en dispositivos con recursos limitados, aunque su utilidad depende de conservar el mismo preprocesamiento entre entrenamiento y despliegue (Abadade *et al.*, 2024; Rivas-Navarrete *et al.*, 2025; Warden & Situnayake, 2019).

Pulneo integra un *pipeline* de entrenamiento, un contrato de preprocesamiento versionado y una aplicación Flutter que captura audio, extrae MFCC, ejecuta un modelo TFLite INT8 y actualiza una interfaz gamificada. El objetivo es estudiar la viabilidad técnica del uso de la actividad respiratoria acústica como indicador indirecto de continuidad durante ejercicios guiados; no diagnosticar, estimar función pulmonar ni demostrar la eficacia terapéutica.

Objetivo general

Diseñar e implementar un prototipo móvil gamificado para apoyar ejercicios respiratorios domiciliarios mediante detección acústica de actividad respiratoria ejecutada localmente con TensorFlow Lite, integrando biorretroalimentación visual, conteo de repeticiones y registro de sesión.

Objetivos específicos

- Construir un *pipeline* auditable de entrenamiento y evaluación para detección de actividad respiratoria usando respiración profunda como clase positiva y sonidos ambientales como clase negativa.
- Cuantificar el desempeño del clasificador y comparar TFLite FP32 frente a TFLite INT8 en las mismas muestras de prueba.
- Definir un contrato único de preprocesamiento, normalización, semántica de clases y umbral para el despliegue móvil.
- Implementar en Flutter captura de audio, extracción MFCC, inferencia en isolate y una máquina de estados de biorretroalimentación.
- Incorporar una revisión fisioterapéutica conceptual del protocolo para analizar la relación entre la señal acústica, la tarea respiratoria guiada, la dosificación, la seguridad y la interpretación de los indicadores de proceso.
- Delimitar explícitamente los alcances técnicos y clínicos antes de una evaluación con participantes.

Materiales y métodos

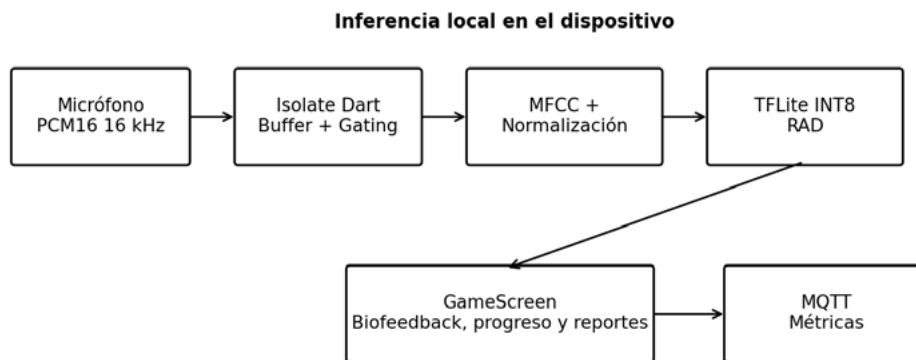
Diseño del sistema

Este trabajo se planteó como el desarrollo y la evaluación técnica de un prototipo, por lo que no corresponde a un ensayo clínico. La solución se organizó en dos componentes principales. El primero es un *pipeline* desarrollado en Python, utilizado para preparar los datos, entrenar y calibrar el modelo, evaluar su desempeño y generar los archivos necesarios para el despliegue. El segundo corresponde a una aplicación móvil desarrollada en Flutter, encargada de capturar el audio, extraer las características acústicas y ejecutar la inferencia directamente en el dispositivo.

Para mantener la consistencia entre el entrenamiento y la aplicación móvil, la configuración de despliegue se definió en el archivo `normalizer_mfcc.json`. Este archivo reúne los parámetros necesarios para reproducir el procesamiento de entrada: dimensiones esperadas, configuración de los MFCC, estadísticas utilizadas para la normalización, correspondencia entre clases, interpretación de la salida y umbral de decisión del modelo cuantizado en INT8.

Se utilizó la convención 0 = ruido/silencio y 1 = actividad respiratoria. La salida sigmoide del modelo se interpreta como la probabilidad estimada de actividad respiratoria, es decir, $P(\text{respiración})$. En la aplicación móvil, una ventana se clasifica como respiración cuando esta probabilidad es mayor o igual a 0,29. La arquitectura general del sistema se presenta en la Figura 1.

Figura 1
Arquitectura general de Pulneo



La Figura 1 resume el flujo de ejecución. El audio PCM16 mono, con una frecuencia de muestreo de 16 kHz, se procesa en un isolate de Dart, donde se aplica el contrato MFCC y se ejecuta el TFLite INT8. La salida alimenta la pantalla de juego (*gamescreen*) mediante códigos de estado semánticos. La telemetría MQTT es experimental y no transmite audio crudo.

Fundamento fisioterapéutico de la biorretroalimentación

Pulneo utiliza la presencia sostenida de actividad respiratoria acústica como señal de proceso durante un ejercicio guiado de exhalación. No mide flujo, volumen pulmonar, saturación de oxígeno, esfuerzo muscular ni fase respiratoria; por ello, el avance de la barra indica continuidad acústica detectada dentro de una instrucción externa, no una espiración terapéuticamente correcta. La calidad del patrón ventilatorio, el esfuerzo, la coordinación y la tolerancia requieren valoración profesional. La biorretroalimentación estructura la práctica y aporta una referencia temporal, sin sustituir el razonamiento clínico (Frownfelter & Dean, 2012; Pryor & Prasad, 2008).

La aplicación también distingue *clipping* o saturación del micrófono. Este evento se reporta como saturación acústica y no como “mala técnica”, porque puede depender de ganancia, distancia al micrófono, hardware o intensidad del sonido. La interpretación clínica permanece a cargo del profesional.

Conceptualización clínica y revisión conceptual fisioterapéutica del protocolo

La conceptualización clínica distinguió el objetivo terapéutico, la variable registrada por el teléfono y la respuesta de la interfaz. El ejercicio es una tarea respiratoria guiada; el teléfono solo detecta actividad acústica compatible con respiración, que la aplicación representa como continuidad. Por ello, un estado *breathing* no permite determinar volumen, flujo espiratorio, patrón diafragmático, compensaciones ni tolerancia clínica.

Desde la perspectiva fisioterapéutica, se revisó si la señal podía apoyar de forma coherente la práctica domiciliaria. El diseño priorizó continuidad y dosificación, no intensidad: la retroalimentación no debe incentivar una señal más fuerte y la saturación del micrófono se trata como un problema de adquisición, no como de desempeño respiratorio. Así se evita convertir la gamificación en una competencia por respirar con mayor intensidad (Bott *et al.*, 2009; Frownfelter & Dean, 2012; Pryor & Prasad, 2008).

La revisión fue realizada por un autor fisioterapeuta con maestría en terapia respiratoria como evaluación experta interna del prototipo. Se examinó la relación entre tarea, la señal acústica, la biorretroalimentación y las conclusiones permitidas, además de la detección positiva, el tiempo objetivo, la dosificación, el conteo de repeticiones, la saturación, los reportes y el papel del profesional.

No fue una validación formal: no se aplicó un instrumento estandarizado, ni participaron evaluadores independientes, ni se realizaron rondas de consenso. Los resultados son criterios de una revisión experta interna. Se estableció que la duración fuera configurable, que una detección continua contara como una repetición, que el nuevo conteo requiriera transición a *noBreathing* y que las alertas de adquisición no se presentaran como valoración automática de la técnica.

La pertinencia terapéutica se definió por el uso previsto, sin asumir eficacia clínica. Pulneo complementa una tarea indicada y explicada por un profesional; no selecciona tratamientos, no ajusta dosis, no decide la finalización clínica ni sustituye la valoración de signos, síntomas o compensaciones. Duración, repeticiones, frecuencia y criterios de interrupción deben individualizarse según valoración y protocolo clínico (Spruit *et al.*, 2013).

Los datos de sesión (repeticiones, tiempo de práctica, interrupciones y clipping) describen la interacción con el prototipo, pero no la función pulmonar, la respuesta fisiológica, la calidad técnica ni el beneficio terapéutico. En telemonitoreo pueden aportar información al profesional, siempre que esa información se interprete junto con la evaluación clínica y el reporte del paciente.

La evaluación se planteó por etapas: validar técnicamente el detector en condiciones reales y distintos dispositivos; comprobar la concordancia con anotaciones profesionales y usabilidad supervisada; y, solo después, estudiar la adherencia o los resultados clínicos. Esta secuencia evita equiparar el desempeño en conjuntos de datos públicos con validación fisioterapéutica.

Conjuntos de datos y partición

La clase positiva se construyó con Coswara (Sharma *et al.*, 2020): en `combined_data.csv` se seleccionaron solo sujetos con `covid_status = healthy` y el archivo `breathing-deep.wav` de cada sujeto. Cada audio se cargó en mono a 16 kHz, se normalizó por su máximo absoluto y se segmentó en ventanas de 1 s con salto de 0,5 s; los audios menores a 1 s se completaron con ceros. Se omitieron archivos no legibles y ventanas con $RMS < 0,0001$, obteniéndose 34 817 ventanas de 1061 sujetos. Para la clase negativa, los WAV de UrbanSound8K (Salamon *et al.*, 2014) se barajaron con semilla 42 y recibieron el mismo procesamiento y filtro RMS. Se procesaron archivos hasta acumular 34 817 ventanas de 5608 archivos, igualando la clase positiva; no se usaron negativos sintéticos.

La partición fue aleatoria, estratificada por clase y realizada a nivel de grupo con semilla 42. Las proporciones objetivo fueron 70 % entrenamiento, 15 % validación y 15 % prueba. Todas las ventanas de un mismo sujeto Coswara o archivo negativo permanecieron en un único subconjunto. La auditoría de la partición verificó cero solapamientos de grupos entre entrenamiento, validación y prueba. Esta decisión reduce el riesgo de que ventanas altamente relacionadas del mismo sujeto o archivo aparezcan en más de un subconjunto y permite una estimación más conservadora que una partición aleatoria a nivel de ventana.

La estrategia de partición controla el solapamiento directo entre sujetos y archivos, pero el diseño del conjunto mantiene una consideración adicional de validez externa: existe una asociación entre la clase y la procedencia, ya que las ventanas positivas provienen de Coswara y las negativas de UrbanSound8K. En consecuencia, no puede descartarse que el clasificador aproveche, además de rasgos propios de la actividad respiratoria, características del dominio de adquisición como dispositivo, ganancia, protocolo de grabación o ruido de fondo. Este aspecto no invalida el experimento ni su utilidad como evaluación técnica inicial, pero delimita su interpretación: el rendimiento reportado cuantifica discriminación dentro de los dominios públicos utilizados y constituye una referencia de ingeniería para el prototipo, no una estimación directa de generalización a grabaciones respiratorias domiciliarias. La siguiente fase de validación contempla datos recogidos con la propia aplicación y negativos del entorno objetivo para medir esa generalización de forma específica.

Preprocesamiento acústico

Los audios se convirtieron a mono y se remuestrearon a 16 kHz. Se utilizaron ventanas de 1 s (16 000 muestras) con salto de 0,5 s (8000 muestras). De cada ventana se extrajeron 40 MFCC con 128 bandas Mel, FFT de 512 puntos, ventana Hann de 400 muestras, salto STFT de 160 muestras y banda de 50 a 8000 Hz. Los MFCC constituyen una representación clásica de la envolvente espectral utilizada ampliamente en reconocimiento y análisis acústico (Davis & Mermelstein, 1980; Rabiner & Schafer, 2011), mientras que la parametrización temporal y espectral sigue principios estándar de procesamiento digital de señales (Oppenheim & Schafer, 2010).

El contrato de características especifica `center=false`, escala Mel Slaney, normalización Slaney y conversión `power_to_db` con `top_db=80`. La implementación móvil utiliza estos mismos parámetros y no aplica preénfasis. La forma de entrada resultante es $97 \times 40 \times 1$. El empleo de representaciones cepstrales y características espectrales en análisis respiratorio cuenta con antecedentes tanto en clasificación de sonidos respiratorios como en detección de fases respiratorias (Rocha *et al.*, 2019; Zhantleuova *et al.*, 2025). Los parámetros del contrato de inferencia se resumen en la Tabla 1.

Tabla 1
Parámetros de preprocesamiento acústico del contrato de inferencia.

PARÁMETRO	VALOR	USO
Frecuencia de muestreo	16000 Hz, mono	Entrada común de entrenamiento y app
Ventana / salto de audio	1 s / 0,5 s	Inferencia solapada al 50 %
MFCC / bandas Mel	40 / 128	Representación acústica compacta
FFT / ventana / salto STFT	512 / 400 / 160	Contrato de espectrograma
Mel	Slaney; norm=Slaney	Alineación Python-Dart
<code>power_to_db</code>	<code>top_db=80</code>	Escala logarítmica definida
Banda	50–8000 Hz	Rango procesado
Forma	$97 \times 40 \times 1$	Entrada fija de la CNN

Normalización y prevención de fuga de datos

La media y la desviación estándar de cada coeficiente MFCC se estimaron exclusivamente con el conjunto de entrenamiento y luego se reutilizaron en validación, prueba y aplicación móvil. La operación fue $z = (x - \mu_{train}) / (\sigma_{train} + \varepsilon)$. El contrato JSON conserva los 40 valores de media y desviación estándar, de modo que la inferencia móvil no recalcula estadísticas con datos de usuario.

Modelo de detección de actividad respiratoria

Se implementó una CNN 2D con tres convoluciones 3×3 de 16, 32 y 48 filtros, normalización por lotes, dos capas MaxPooling 2×2 , Global Average Pooling, una capa densa de 32 unidades con ReLU, Dropout de 0,25 y una salida sigmoide. El modelo contiene 20 657 parámetros totales, de los cuales 20 465 son entrenables y 192 no entrenables. Se optimizó con Adam, *binary crossentropy*, *batch size* 32, *learning rate inicial* 0,001, *early stopping* y reducción adaptativa de la tasa de aprendizaje. La selección de una arquitectura convolucional compacta se apoya en componentes estándar de aprendizaje profundo y en su aplicación previa en informática de la salud (Goodfellow *et al.*, 2016; Ravi *et al.*, 2017).

La clase positiva es actividad respiratoria (1) y la negativa ruido/silencio (0). El umbral de cada artefacto se calibró de manera independiente y exclusivamente sobre el conjunto de validación, seleccionando el valor que maximizó F1; una vez seleccionado, se mantuvo fijo durante la evaluación en test. Keras FP32 y TFLite FP32 seleccionaron 0,53, mientras que TFLite INT8 seleccionó 0,29. El umbral de despliegue de la aplicación es, por tanto, $P(\text{respiración}) \geq 0,29$. La arquitectura del detector se presenta en la Figura 2.

Figura 2
Arquitectura de la CNN 2D utilizada para el detector RAD



Cuantización y protocolo de evaluación

A partir del mismo modelo Keras se exportaron dos artefactos TFLite: TFLite FP32, utilizado como modelo de referencia para el despliegue, y TFLite INT8, utilizado como modelo cuantizado final. Keras FP32 se conserva como referencia del entrenamiento; la comparación destinada a aislar el efecto de cuantización se realiza entre los dos artefactos TFLite para mantener constante el entorno de ejecución. La cuantización INT8 utilizó 100 muestras representativas del conjunto de entrenamiento. TFLite FP32 y TFLite INT8 se ejecutaron con `tf.lite.Interpreter` sobre exactamente las mismas 10 347 ventanas de test, después de 20 muestras de calentamiento.

Se calcularon *accuracy*, precisión, *recall*/sensibilidad, especificidad, F1, ROC-AUC, PR-AUC, *balanced accuracy*, MCC y matriz de confusión. ROC-AUC y PR-AUC se calcularon sobre probabilidades continuas, antes del umbral. También se midió la concordancia entre las probabilidades y entre las decisiones, tamaño de archivo y latencia media por muestra. Las cifras de latencia corresponden al entorno de evaluación TFLite del *pipeline* y no constituyen un *benchmark* de un teléfono específico.

Entorno de software y reproducibilidad

La ejecución se realizó con Python 3.12.10 y semilla aleatoria 42. Los artefactos generados identifican TensorFlow 2.21.0 y Keras 3.15.1. El proyecto Flutter fija Dart SDK $\geq 3.9.2$ y Flutter $\geq 3.35.6$; las dependencias de inferencia y audio incluyen `tflite_flutter` 0.12.1, `record` 6.2.0 y `fft` 0.5.1. Esta configuración documenta el entorno principal utilizado para el entrenamiento, la evaluación y la integración móvil.

Implementación móvil

La aplicación Flutter captura PCM16 mono a 16 kHz y envía los buffers a un isolate. El isolate completa la inicialización del intérprete antes de abrir el flujo de audio, aplica *gating* y detección de clipping, calcula MFCC según el contrato, normaliza las características y ejecuta el modelo TFLite INT8. La salida cuantizada se descuantiza y se interpreta como P(respiración).

El protocolo de inferencia utiliza el tipo `AudioInferenceStatus` con cuatro estados: *fatalError* (−1), *noBreathing* (0), *breathing* (1) y *clipping* (2). Estos valores representan estados de la aplicación, mientras que el modelo subyacente efectúa una clasificación de dos clases. La pantalla de juego recibe estados semánticos tipados para actualizar la biorretroalimentación y las alertas.

La duración de la sesión se mide con un reloj de ejecución. El conteo de repeticiones requiere una transición a *noBreathing* para rearmar una nueva repetición, de modo que una detección respiratoria continua solo puede completar una repetición. La telemetría MQTT se mantiene como función experimental y configurable; no transmite audio crudo. El contrato de inferencia y sus códigos de estado se muestran en la Figura 3. Los componentes principales del sistema y su alcance actual se resumen en la Tabla 2.

Figura 3
Contrato de inferencia y códigos de estado utilizados por la aplicación

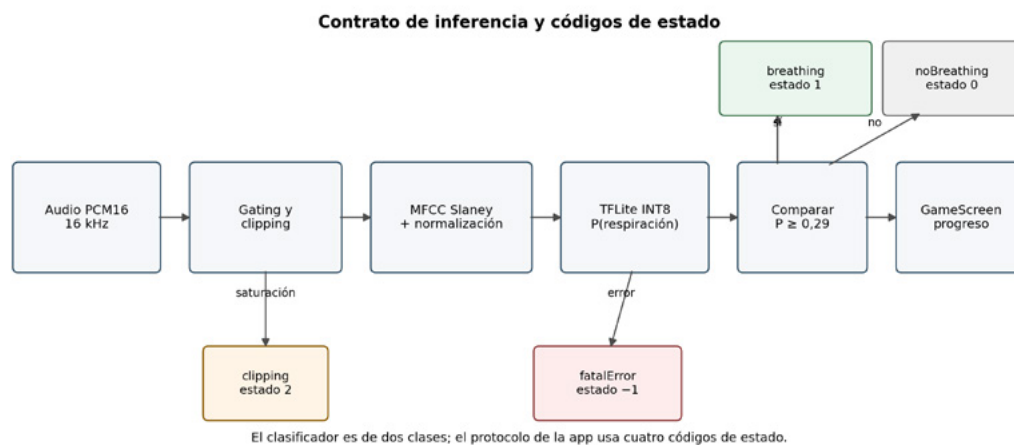


Tabla 2
Componentes principales del sistema y alcance actual

COMPONENTE	FUNCIÓN	ALCANCE ACTUAL
Pipeline Python	Datos, MFCC, entrenamiento, auditoría y exportación FP32/INT8	Evaluación técnica; no validación clínica
Modelo RAD	Dos clases: ruido/silencio frente a actividad respiratoria	No distingue inspiración/exhalación
Contrato JSON v2	Preprocesamiento, normalización, clases y <i>threshold</i> INT8	$P(\text{respiración}) \geq 0,29$
Isolate Dart	MFCC e inferencia fuera del hilo UI	Evita bloquear la interfaz
Pantalla de juego	Estado, progreso, repeticiones y <i>clipping</i>	Biorretroalimentación de proceso, no medición fisiológica
MQTT	Telemetría resumida	Experimental; sin audio crudo

Heurísticas previas a la inferencia

Las heurísticas de amplitud, *clipping* y factor de cresta pertenecen solo a la aplicación móvil: no se aplicaron a las 10 347 ventanas de test ni intervienen en las métricas reportadas.

En el entorno de ejecución, antes de MFCC/TFLite, descartan amplitud insuficiente, separan *clipping* y filtran impulsos. Controlan calidad de la entrada, pero no la técnica respiratoria ni las variables clínicas.

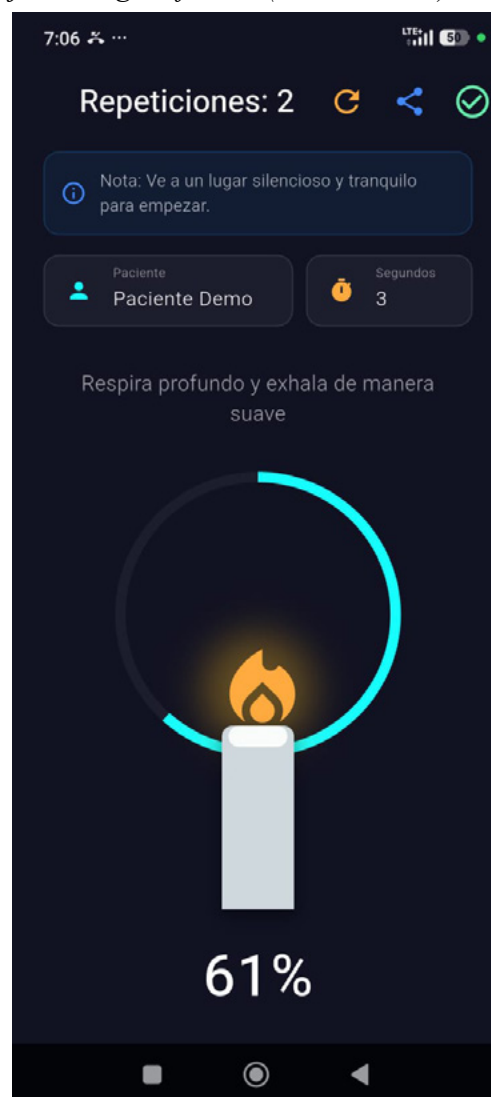
Gamificación y lógica de biorretroalimentación

La barra de progreso aumenta únicamente cuando el estado recibido es *breathing* y disminuye cuando la actividad se interrumpe. El *clipping* activa una alerta y no se contabiliza como respiración válida. Al completar el intervalo configurado, se registra una repetición y el sistema queda bloqueado hasta observar *noBreathing*, lo que impide múltiples repeticiones durante una detección continua.

Dado que el modelo no identifica la fase respiratoria, el estado *breathing* se usa como indicador acústico dentro de una instrucción guiada de exhalación. La interfaz se concibe como biorretroalimentación de proceso y no como sensor de flujo ni como clasificador específico de fase respiratoria. Esta elección mantiene sencilla la tarea de inferencia en la etapa actual y delimita con claridad qué debe comprobarse en la validación externa con datos del ejercicio objetivo. La interfaz implementada durante el ejercicio se muestra en la Figura 4.

Figura 4

Pantalla principal de ejercicio gamificado (GameScreen) durante una sesión guiada



En esta vista, el porcentaje y el anillo circular representan el progreso temporal del ejercicio configurado; el conteo superior registra repeticiones completadas y el tiempo objetivo se muestra en segundos. Estos elementos constituyen biorretroalimentación de proceso y no una medición de flujo, volumen pulmonar u otra variable fisiológica.

Plan de evaluación y criterios de seguridad

No se reclutaron pacientes ni se realizó intervención clínica. La evaluación futura se plantea en tres etapas: primero, comprobar el detector con audios del escenario objetivo anotados por profesionales, incluidos exhalaciones, inhalaciones, habla, tos, pausas, ruido doméstico y distintas distancias; segundo, con aprobación ética y consentimiento informado, evaluar comprensión, aceptabilidad, carga de uso y concordancia con la observación del fisioterapeuta; y tercero, estudiar adherencia y resultados terapéuticos. Cada etapa deberá definir los criterios de inclusión y exclusión, las reglas de detención, la supervisión y protección de datos, sin delegar decisiones clínicas a la aplicación.

La caracterización computacional de esta etapa se concentró en variables reproducibles del *pipeline*: tamaño de modelo y latencia del intérprete TFLite bajo el mismo entorno experimental. Las mediciones de consumo energético, memoria pico, batería, temperatura y latencia en modelos concretos de teléfonos inteligentes se reservan para la fase de *benchmark* móvil, donde podrán compararse bajo hardware y condiciones de uso definidos.

Resultados

Composición del conjunto de datos y partición

El conjunto de datos final incluyó 69 634 ventanas perfectamente balanceadas: 34 817 por clase. Los 1061 sujetos Coswara y 5608 archivos negativos se distribuyeron por grupos, sin solapamiento entre subconjuntos. La fracción real de ventanas fue 70,10 % en entrenamiento, 15,05 % en validación y 14,86 % en prueba, próximas a las proporciones objetivo de 70/15/15. La composición de cada subconjunto se presenta en la Tabla 3.

Tabla 3
Tamaño y composición real de train, validation y test

SPLIT	VENT.	RESP.	RUIDO	SUJETOS	NEG.
Train	48 810	24 413	24 397	743	3926
Val.	10 477	5229	5248	159	841
Test	10 347	5175	5172	159	841
Total	69 634	34 817	34 817	1061	5608

Rendimiento del clasificador

En test ($n = 10\,347$), Keras FP32 y TFLite FP32 coincidieron en las métricas principales y en la matriz de confusión: 97,59 % de *accuracy*, 98,28 % de precisión, 96,89 % de *recall*, 98,30 % de especificidad, F1 de 97,58 % y ROC-AUC de 99,71 %, con umbral 0,53 calibrado en validación. Para aislar el efecto de la cuantización, la comparación reportada en la Tabla 4 utiliza TFLite FP32 como modelo de referencia de despliegue y TFLite INT8 como modelo cuantizado, ambos ejecutados con `tf.lite.Interpreter` sobre las mismas 10 347 ventanas de prueba. TFLite INT8 alcanzó 97,64 % de *accuracy*, 98,09 % de precisión, 97,18 % de *recall*, 98,11 % de especificidad, F1 de 97,63 % y ROC-AUC de 99,65 %. Su umbral de 0,29 se fijó previamente en validación ($F1 = 96,79\%$) y no se reajustó en test.

Tabla 4
Rendimiento en test de TFLite FP32 y TFLite INT8

TIPO	TH.	ACC.	PREC.	REC.	F1	AUC
TFLite FP32	0,53	97,59	98,28	96,89	97,58	99,71
TFLite INT8	0,29	97,64	98,09	97,18	97,63	99,65

En la Tabla 4, las métricas se expresan en porcentaje, excepto el umbral. La fila TFLite FP32 corresponde al artefacto float exportado a TensorFlow Lite, no al modelo Keras; ambos coincidieron en las métricas principales de test, pero se utiliza TFLite FP32 para mantener el mismo entorno de ejecución que INT8. PR-AUC fue 99,71 % para FP32 y 99,67 % para INT8; balanced *accuracy*, 97,59 % y 97,64 %; y MCC, 0,9520 y 0,9529, respectivamente. La matriz de confusión del TFLite INT8 se presenta en la Tabla 5.

Tabla 5
Matriz de confusión del TFLite INT8 en test

CLASE REAL	PRED. RUIDO/SILENCIO	PRED. RESPIRACIÓN
Ruido/silencio	5074 (TN)	98 (FP)
Respiración	146 (FN)	5029 (TP)

Comparación FP32 frente a INT8 y huella del modelo

La comparación pareada entre TFLite FP32 e INT8 se realizó sobre las mismas 10 347 muestras y con el mismo entorno de ejecución. El F1 aumentó en 0,00054 en INT8, mientras que ROC-AUC disminuyó en 0,00058. La correlación entre probabilidades fue 0,99699, el MAE fue 0,01164 y el acuerdo de decisiones usando los umbrales de despliegue fue 99,55 %. En conjunto, estas cifras muestran que la capacidad discriminativa y el orden relativo de las puntuaciones se conservaron en gran medida tras la cuantización.

Interpretación de los umbrales de decisión. TFLite FP32 e INT8 mostraron puntuaciones altamente correlacionadas, aunque sus umbrales óptimos de validación fueron 0,53 y 0,29. Esto es compatible con salidas globalmente similares, porque el umbral que maximiza F1 depende de la distribución cercana a la frontera de decisión y de la calibración de cada artefacto. Ambos umbrales se optimizaron por separado en *validation* y se aplicaron sin reajuste en test; por tanto, la diferencia se interpreta como calibración del punto de decisión y no como evidencia de pérdida discriminativa ni se atribuye causalmente a la cuantización.

El tamaño del modelo TFLite disminuyó de 88 488 a 31 600 bytes, una reducción de 64,29 % (2,80×). El archivo Keras ocupa 310 023 bytes. La cuantización se utilizó principalmente para reducir la huella de almacenamiento; en el entorno experimental, la latencia media fue de 0,401 ms para INT8 y 0,215 ms para FP32 después del calentamiento. Por ello, el beneficio observado en esta ejecución fue la compacidad del artefacto, mientras que la eventual ventaja de velocidad, energía o memoria debe determinarse mediante *benchmarks* específicos por dispositivo. El efecto de la cuantización sobre tamaño, métricas y latencia se resume en la Tabla 6.

Tabla 6
Efecto de cuantización TFLite FP32 frente a TFLite INT8

INDICADOR	TFLITE FP32	TFLITE INT8	CAMBIO
Tamaño	88 488 bytes	31 600 bytes	-64,29 %
F1	97,577 %	97,632 %	+0,054 pp
ROC-AUC	99,711 %	99,653 %	-0,058 pp
Latencia media*	0,215 ms	0,401 ms	+87,00 %
Acuerdo de decisiones	—	99,55 % vs. FP32	0,45 % desacuerdo

* Medición con tf.lite.Interpreter, una muestra por invocación, después de 20 muestras de calentamiento; no corresponde a un benchmark de teléfonos inteligentes.

Integración técnica del prototipo

La integración móvil utiliza la misma semántica definida por el entrenamiento: P(respiración) se evalúa directamente con el *threshold* INT8 de 0,29. El protocolo de estados distingue *noBreathing*, *breathing*, *clipping* y *fatalError*. La barra de progreso avanza durante el estado *breathing*, se interrumpe fuera de ese estado y aplica una alerta independiente ante *clipping*. El reporte de sesión utiliza la duración medida y el contador de repeticiones exige una transición a *noBreathing* entre repeticiones.

Estos resultados respaldan la coherencia interna del *pipeline* y del contrato de despliegue. La evaluación cuantitativa presentada corresponde al clasificador sobre los conjuntos de datos descritos; la correspondencia entre el estado acústico y una exhalación guiada en condiciones reales constituye el objetivo específico de la siguiente fase de validación con datos del escenario de uso.

Resultado de la revisión conceptual fisioterapéutica

La revisión conceptual fisioterapéutica definió la biorretroalimentación como indicador de continuidad acústica dentro de una tarea guiada y fijó cuatro decisiones: no asociar mayor amplitud con mejor ejecución; separar *clipping* del juicio de técnica; contar una sola repetición por episodio continuo; y mantener la dosificación temporal bajo control del protocolo o del profesional.

Duración, repeticiones, interrupciones y saturación se consideran resultados de proceso; no se presentan como medidas de ventilación, función pulmonar, disnea ni respuesta terapéutica.

La evaluación futura seguirá una secuencia: concordancia técnica en el dominio de uso, usabilidad y concordancia observacional bajo supervisión, y después estudios de adherencia o efecto terapéutico. La revisión actual es experta e interna; una validación formal de contenido requerirá varios evaluadores y un método definido, y la validación clínica, estudios con participantes.

Discusión

El INT8 alcanzó F1 de 97,63 % y ROC-AUC de 99,65 % en un test separado por grupos. La ausencia de solapamiento de sujetos y archivos reduce la fuga de datos y permite usar estas cifras para comparar versiones del *pipeline* y la cuantización. Sin embargo, describen el experimento Coswara/UrbanSound8K y no demuestran desempeño clínico ni generalización domiciliaria.

La cuantización redujo el tamaño del modelo en un 64,29 % y mantuvo 99,55 % de acuerdo con TFLite FP32. El cambio de umbral de 0,53 a 0,29, calibrado independientemente en validation, no implica por sí mismo pérdida discriminativa. En el entorno medido, INT8 fue más compacto pero más lento; la estabilidad del umbral, la latencia, la energía y la memoria deberán evaluarse en teléfonos inteligentes concretos.

La principal consideración para la validez externa es la diferencia entre el dominio de entrenamiento y el escenario final de uso. La clase positiva procede de breathing-deep.wav de sujetos sanos de Coswara, con ciclos de respiración profunda registrados bajo el protocolo de ese conjunto de datos, mientras que la aplicación guía una exhalación prolongada y controlada frente al micrófono de un teléfono. Esta elección permitió construir y depurar un clasificador experimental de actividad respiratoria con datos públicos y reproducibles antes de realizar recolección propia. El rendimiento obtenido caracteriza la discriminación dentro de esos dominios públicos; la generalización al ejercicio domiciliario se trata como una pregunta experimental independiente. Fase respiratoria, distancia, orientación, micrófono, ganancia, duración, ruido y características del usuario pueden modificar la distribución acústica, por lo que la evaluación futura debe medir explícitamente el desempeño en esas condiciones.

Esta delimitación también permite diferenciar Pulneo de Breeze debido a sus distintos objetivos. Breeze fue diseñado para detectar fases respiratorias, mientras que Pulneo clasifica actividad respiratoria frente a ruido y no separa inspiración de exhalación. En Pulneo, la actividad acústica se utiliza como una señal de proceso dentro de una instrucción guiada, una simplificación deliberada que facilita la inferencia local y la integración con la interfaz gamificada en esta etapa. La evolución prevista incorpora un conjunto específico del escenario objetivo, con exhalaciones etiquetadas, múltiples teléfonos, distancias y ruidos domésticos, además de negativos difíciles como voz, tos, manipulación del equipo e inhalaciones, con el fin de evaluar si el detector mantiene su utilidad fuera de los dominios públicos originales.

Desde el punto de vista clínico, Pulneo se diseñó como una herramienta de apoyo y no como un instrumento de medición fisiológica: el micrófono aporta información acústica, mientras que variables como flujo, volumen, esfuerzo y respuesta terapéutica requieren sensores o evaluaciones diferentes. Esta delimitación permite mantener una arquitectura móvil sencilla y no invasiva para explorar la biorretroalimentación y el registro del proceso sin atribuir a la señal acústica capacidades que no posee. En esta etapa tampoco se evaluaron adherencia, calidad de vida o concordancia con fisioterapeutas; esas variables pertenecen a las fases posteriores de evaluación clínica y de usabilidad. La contribución actual es de ingeniería y diseño clínico-conceptual: un flujo auditable de datos, un clasificador compacto, un contrato explícito entre Python y Dart y una aplicación que realiza inferencia local sin transmitir audio crudo.

La conceptualización clínica aporta además un criterio para interpretar la gamificación. En este contexto, la barra de progreso no representa una magnitud fisiológica continua; representa el cumplimiento temporal de una condición acústica durante una tarea previamente explicada. Esa distinción cambia el papel de la biorretroalimentación: su finalidad inmediata es orientar atención, ritmo y persistencia, no cuantificar capacidad pulmonar. El valor potencial de la interfaz depende de que el usuario comprenda esa relación y de que el profesional pueda establecer una tarea compatible con las capacidades y objetivos individuales. Esta visión es coherente con la rehabilitación pulmonar como intervención individualizada que combina entrenamiento, educación y cambio de comportamiento, más que como ejecución aislada de una maniobra (Spruit *et al.*, 2013).

Desde el análisis fisioterapéutico, la principal tensión del prototipo se encuentra entre la simplicidad de la interacción y la complejidad del fenómeno respiratorio. Una exhalación terapéutica puede requerir control de ritmo, coordinación toracoabdominal, manejo de secreciones, tolerancia al esfuerzo y adaptación a síntomas, dimensiones que no pueden inferirse únicamente a partir de los MFCC (Bott *et al.*, 2009; Osadnik *et al.*, 2012). Por ello, incrementar la precisión del clasificador no elimina la necesidad de una capa clínica de interpretación. Incluso un detector acústico perfecto de respiración seguiría siendo insuficiente para certificar una técnica correcta si no se observan flujo, movimiento, esfuerzo y respuesta del paciente. El diseño debe mantener esta asimetría visible para evitar una falsa sensación de medición fisiológica.

La supervisión de pertinencia terapéutica también orienta el desarrollo futuro del sistema. Antes de ampliar la mecánica con puntuaciones, niveles o recompensas, cada elemento gamificado debería evaluarse por su efecto sobre la conducta que se desea facilitar. Recompensar la intensidad podría inducir esfuerzo innecesario; recompensar la velocidad podría acortar una fase que se pretende sostener; y penalizaciones excesivas podrían aumentar la frustración o la ansiedad. En cambio, metas temporales comprensibles, retroalimentación inmediata y registro de cumplimiento pueden ser compatibles con un enfoque educativo si se mantienen subordinados a la prescripción y a la tolerancia individual. La evaluación con usuarios debe estudiar estas respuestas y no asumir que una mayor interacción equivale automáticamente a mejor adherencia o mejor técnica.

Alcance y limitaciones

El conjunto positivo se restringió a sujetos sanos de Coswara para disponer de una referencia pública, homogénea y reproducible durante la construcción inicial del detector. UrbanSound8K se utilizó para aportar una clase negativa amplia de sonidos ambientales reales. Estas decisiones son apropiadas para una primera evaluación técnica controlada, aunque no pretenden representar toda la población clínica ni todos los eventos acústicos de un hogar. Por ello, la siguiente etapa incorpora como objetivo un conjunto externo recolectado con la propia aplicación, incluyendo población y condiciones de uso más próximas al escenario terapéutico.

El modelo identifica actividad respiratoria y no fue entrenado para separar fases respiratorias. En el protocolo actual, esta salida se utiliza como indicador de continuidad durante una exhalación guiada, lo que permite estudiar la interacción y el procesamiento local sin presentar el sistema como detector validado de exhalación. Una fase posterior incorporará etiquetas específicas de inhalación/exhalación y anotación profesional para determinar si una tarea más especializada mejora la correspondencia entre la señal acústica y el ejercicio terapéutico.

La evaluación de rendimiento computacional presentada se realizó con el entorno de ejecución TFLite del entorno experimental, lo que permite una comparación controlada entre FP32 e INT8 bajo las mismas condiciones. Para caracterizar el comportamiento del producto en hardware móvil, la fase de evaluación por dispositivo debe añadir *benchmarks* en modelos concretos de teléfonos inteligentes y mediciones de memoria, energía, batería y temperatura. De forma análoga, la validación con participantes se reserva para una etapa posterior con protocolo ético y anotación profesional.

La trazabilidad del experimento registra Python 3.12.10, la semilla aleatoria 42, las versiones de TensorFlow/Keras, las restricciones de Dart/Flutter y las dependencias móviles. Estos datos, junto con los artefactos de entrenamiento, normalización, partición del conjunto de datos y evaluación, permiten reconstruir las principales condiciones computacionales del experimento.

Conclusiones

Pulneo integra un *pipeline* de entrenamiento y evaluación con una aplicación Flutter para inferencia acústica local. La aplicación interpreta la salida del modelo como P(respiración), utiliza el umbral INT8 calibrado en validación y gestiona los resultados mediante códigos de estado explícitos.

En el conjunto de prueba, el TFLite INT8 alcanzó 97,64 % de *accuracy*, F1 de 97,63 % y ROC-AUC de 99,65 %. La cuantización redujo el tamaño del modelo en un 64,29 % y mantuvo un 99,55 % de acuerdo de decisiones respecto del TFLite FP32, aunque no produjo una mejora de latencia en el entorno evaluado.

Estos resultados respaldan la viabilidad técnica del detector dentro del dominio evaluado y justifican avanzar a una validación externa más próxima al uso previsto. La siguiente etapa está orientada a cuantificar la generalización con datos propios del ejercicio, *benchmarks* en teléfonos concretos y evaluación con anotación profesional y un protocolo ético. En paralelo, la conceptualización fisioterapéutica define un marco de uso en el que la señal acústica se interpreta como indicador de proceso, la dosificación permanece bajo supervisión profesional y la validación progresa desde concordancia técnica hasta estudios clínicos posteriores. Este planteamiento permite presentar el prototipo de acuerdo con la evidencia disponible: como una arquitectura técnica reproducible y clínicamente delimitada, con una ruta explícita para comprobar su desempeño en el escenario domiciliario antes de atribuirle efectos terapéuticos.

Reconocimientos y declaraciones

Los autores agradecen a las comunidades científicas que mantienen bases de datos abiertas de audio respiratorio y ambiental y a las comunidades de software libre utilizadas durante el desarrollo experimental.

Declaración sobre uso de herramientas de IA generativa

El manuscrito fue elaborado íntegramente por los autores, quienes asumieron la responsabilidad total del trabajo sin recurrir a herramientas de inteligencia artificial generativa para su preparación o redacción. Para garantizar la integridad intelectual y el rigor metodológico, la selección bibliográfica, la revisión crítica y la verificación de contenidos se realizaron de forma manual y autónoma, manteniendo un control absoluto sobre el producto final.

Contribución de autores

Los autores concibieron y desarrollaron el trabajo, revisaron el manuscrito y asumen la responsabilidad íntegra de su contenido. La metodología, los resultados, las referencias, la interpretación y las conclusiones fueron revisados y aprobados por los autores.

Consideraciones éticas

El trabajo presentado corresponde al diseño y evaluación técnica de un prototipo entrenado con bases de datos públicas. No se realizó reclutamiento de pacientes, intervención clínica ni recolección de datos personales de salud. Cualquier evaluación futura con participantes humanos requerirá aprobación ética, consentimiento informado, criterios de inclusión y exclusión, protocolo de seguridad, manejo de eventos adversos y política de protección de datos.

Conflicto de intereses

Los autores declaran no tener conflictos de intereses relacionados con el presente trabajo.

Disponibilidad de datos y código

Los conjuntos de datos públicos utilizados corresponden a Coswara y UrbanSound8K, citados en las referencias. El *pipeline* y la aplicación se encuentran en los repositorios públicos Pulneo RAD y pulneo_app (Valladarez Muñoz, 2026a, 2026b). La ejecución genera artefactos de auditoría del conjunto de datos y de la partición, métricas Keras/TFLite y una comparación FP32 e INT8, modelos TFLite y el contrato normalizer_mfcc.json.

Referencias

- Abadade, Y., Benamar, N., Bagaa, M., & Chaoui, H. (2024). Empowering healthcare: TinyML for precise lung disease classification. *Future Internet*, 16(11), Article 391. <https://doi.org/10.3390/fi16110391>
- Bott, J., Blumenthal, S., Buxton, M., Ellum, S., Falconer, C., Garrod, R., Harvey, A., Hughes, T., Lincoln, M., Mikelsons, C., Potter, C., Pryor, J., Rimington, L., Sinfield, F., Thompson, C., Vaughn, P., & White, J. (2009). Guidelines for the physiotherapy management of the adult, medical, spontaneously breathing patient. *Thorax*, 64(Suppl 1), i1–i51. <https://doi.org/10.1136/thx.2008.110726>
- Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., & Mascolo, C. (2020). Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3474–3484). Association for Computing Machinery. <https://doi.org/10.1145/3394486.3412865>
- Chen, J., Yang, T., He, Q., Pang, M., Cao, Y., Liu, Z., Li, L., Liu, H.-I., & Bullen, C. (2025). The impact of gamified interventions on the management of chronic obstructive pulmonary disease: Systematic literature review. *JMIR Serious Games*, 13, Article e69510. <https://doi.org/10.2196/69510>
- Cox, N. S., Dal Corso, S., Hansen, H., McDonald, C. F., Hill, C. J., Zanaboni, P., Alison, J. A., O'Halloran, P., Macdonald, H., & Holland, A. E. (2021). Telerehabilitation for chronic respiratory disease. *Cochrane Database of Systematic Reviews*, 2021(1), Article CD013040. <https://doi.org/10.1002/14651858.CD013040.pub2>
- Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>
- Frownfelter, D., & Dean, E. (2012). *Cardiovascular and pulmonary physical therapy: Evidence to practice* (5th ed.). Elsevier.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Holland, A. E., Mahal, A., Hill, C. J., Lee, A. L., Burge, A. T., Cox, N. S., Moore, R., Nicolson, C., O'Halloran, P., Lahham, A., Gillies, R., & McDonald, C. F. (2017). Home-based rehabilitation for COPD using minimal resources: A randomised, controlled equivalence trial. *Thorax*, 72(1), 57–65. <https://doi.org/10.1136/thoraxjnl-2016-208514>
- Larson, E. C., Goel, M., Boriello, G., Heltshe, S., Rosenfeld, M., & Patel, S. N. (2012). SpiroSmart: Using a microphone to measure lung function on a mobile phone. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 280–289). Association for Computing Machinery. <https://doi.org/10.1145/2370216.2370261>
- McCarthy, B., Casey, D., Devane, D., Murphy, K., Murphy, E., & Lacasse, Y. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, 2015(2), Article CD003793. <https://doi.org/10.1002/14651858.CD003793.pub3>
- Osadnik, C. R., McDonald, C. F., Jones, A. P., & Holland, A. E. (2012). Airway clearance techniques for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, 2012(3), Article CD008328. <https://doi.org/10.1002/14651858.CD008328.pub2>
- Oppenheim, A. V., & Schaffer, R. W. (2010). *Discrete-time signal processing* (3rd ed.). Pearson.

- Pryor, J. A., & Prasad, S. A. (2008). *Physiotherapy for respiratory and cardiac problems: Adults and paediatrics* (4th ed.). Churchill Livingstone.
- Rabiner, L. R., & Schafer, R. W. (2011). *Theory and applications of digital speech processing*. Pearson.
- Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1), 4–21. <https://doi.org/10.1109/JBHI.2016.2636665>
- Rivas-Navarrete, J. A., Pérez-Espinosa, H., Padilla-Ortiz, A. L., Rodríguez-González, A. Y., & García-Camero, D. C. (2025). Edge computing system for automatic detection of chronic respiratory diseases using audio analysis. *Journal of Medical Systems*, 49, Article 33. <https://doi.org/10.1007/s10916-025-02154-7>
- Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., Jakovljevic, N., Turukalo, T. L., Vogiatzis, I. M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Maglaveras, N., Paiva, R. P., Chouvarda, I., & de Carvalho, P. (2019). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3), Article 035001. <https://doi.org/10.1088/1361-6579/ab03ea>
- Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 1041–1044). Association for Computing Machinery. <https://doi.org/10.1145/2647868.2655045>
- Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Nirmala, R., Ghosh, P. K., & Ganapathy, S. (2020). Coswara—A database of breathing, cough, and voice sounds for COVID-19 diagnosis. In *Interspeech 2020* (pp. 4811–4815). International Speech Communication Association. <https://doi.org/10.21437/Interspeech.2020-2768>
- Shih, C.-H., Tomita, N., Lukic, Y. X., Hernández Reguera, Á., Fleisch, E., & Kowatsch, T. (2019). Breeze: Smartphone-based acoustic real-time detection of breathing phases for a gamified biofeedback breathing training. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4), Article 152. <https://doi.org/10.1145/3369835>
- Spruit, M. A., Singh, S. J., Garvey, C., ZuWallack, R., Nici, L., Rochester, C., Hill, K., Holland, A. E., Lareau, S. C., Man, W. D. C., Pitta, F., Sewell, L., Raskin, J., Bourbeau, J., Crouch, R., Franssen, F. M. E., Casaburi, R., Vercoulen, J. H., Vogiatzis, I., ... Wouters, E. F. M. (2013). An official American Thoracic Society/European Respiratory Society statement: Key concepts and advances in pulmonary rehabilitation. *American Journal of Respiratory and Critical Care Medicine*, 188(8), e13–e64. <https://doi.org/10.1164/rccm.201309-1634ST>
- Valladarez Muñoz, D. F. (2026a). *Pulneo: Pipeline RAD para respiración* [Código fuente]. GitHub. <https://github.com/DavidValladarez95/pulneo>
- Valladarez Muñoz, D. F. (2026b). *pulneo_app: Aplicación móvil gamificada para fisioterapia respiratoria domiciliaria* [Código fuente]. GitHub. https://github.com/DavidValladarez95/pulneo_app
- Warden, P., & Situnayake, D. (2019). *TinyML: Machine learning with TensorFlow Lite on Arduino and ultra-low-power microcontrollers*. O'Reilly Media.
- Watson, J. S., Jordan, R. E., Gardiner, L., Adab, P., & Jolly, K. (2023). A systematic review of the effectiveness of interventions to promote referral, adherence, and uptake of pulmonary rehabilitation for patients with chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 18, 1637–1654. <https://doi.org/10.2147/COPD.S396317>
- Zhantleuova, A. K., Makashev, Y. K., & Duzbayev, N. T. (2025). Optimizing MFCC parameters for breathing phase detection. *Sensors*, 25(16), Article 5002. <https://doi.org/10.3390/s25165002>

TRANSFORMACIÓN DIGITAL Y MADUREZ TECNOLÓGICA

DIGITAL TRANSFORMATION
AND TECHNOLOGICAL MATURITY



cedia | **TICEC**
2026

Ingeniería de una plataforma web para la evaluación de la madurez digital de microemprendimientos ecuatorianos

Engineering a Web Platform for Assessing the Digital Maturity of Ecuadorian Micro-Entrepreneurs

Ivanna Sánchez-Montero¹ <https://orcid.org/0000-0002-0029-6265>, Joel Morales Játiva¹ <https://orcid.org/0009-0009-9661-8529>, Viviana Elizabeth Cajas-Cajas² <https://orcid.org/0000-0002-7688-6966>

¹Universidad Tecnológica Indoamérica, Ambato, Ecuador
ivannasanchez@uti.edu.ec, jmorales44@indoamerica.edu.ec

²Universidad Internacional SEK, Quito, Ecuador
viviana.cajasc@uisek.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Enviado: 2026/07/11

Aceptado: 2026/08/27

Publicado: 2026/10/14

Resumen

La transformación digital constituye un desafío para los microemprendimientos ecuatorianos, debido a la limitada disponibilidad de herramientas accesibles y metodológicamente fundamentadas que permitan diagnosticar su nivel de madurez digital. Este artículo presenta el diseño y la ingeniería de una plataforma web gratuita, desarrollada en español, para la evaluación de la madurez digital de los microemprendimientos, con énfasis en la construcción del instrumento de medición y la arquitectura tecnológica que soporta su funcionamiento. El instrumento fue elaborado mediante un proceso sistemático de revisión, comparación y adaptación de modelos de madurez digital reconocidos, entre ellos, modelos para pequeñas y medianas empresas, el Digital Maturity Assessment de la Comisión Europea y el instrumento Chequeo Digital desarrollado por el Banco Interamericano de Desarrollo (BID) e implementado en Ecuador por la Escuela Superior Politécnica del Litoral (ESPOL). A partir de este proceso se definieron seis dimensiones y veintiún ítems de evaluación adaptados al contexto de los microemprendimientos ecuatorianos. El nivel de madurez digital fue calculado mediante un índice compuesto basado en la metodología de la OCDE y el Joint Research Centre (JRC), aplicando normalización mín-máx, agregación lineal con ponderación equitativa y generación automatizada de recomendaciones. La plataforma desarrollada representa una herramienta

Sumario: Introducción, Materiales y Métodos, Resultados y Discusión, Conclusiones.

Como citar: Sánchez-Montero, I., Morales, J. & Cajas-Cajas, V. (2026). Ingeniería de una plataforma web para la evaluación de la madurez digital de microemprendimientos ecuatorianos. *Revista Tecnológica - Espol*, 38(E1), 154-171. <https://doi.org/10.37815/rte.v38nE1.1591>

tecnológica orientada a facilitar diagnósticos digitales y constituye la base para estudios empíricos posteriores con microemprendimientos ecuatorianos, contribuyendo al diseño de estrategias de transformación digital contextualizadas.

Palabras clave: autodiagnóstico, emprendimientos, madurez digital, transformación digital, plataforma web.

Abstract

Digital transformation represents a challenge for Ecuadorian microenterprises due to the limited availability of accessible and methodologically grounded tools to assess their digital maturity level. This article presents the design and engineering of a free web-based platform, developed in Spanish, for evaluating the digital maturity of microenterprises, focusing on the construction of the assessment instrument and the technological architecture supporting its operation. The instrument was developed through a systematic process of review, comparison, and adaptation of recognized digital maturity models, including references from models designed for small and medium-sized enterprises, the European Commission's Digital Maturity Assessment, and the Digital Checkup instrument developed by the Inter-American Development Bank (IDB) and implemented in Ecuador by the Escuela Superior Politécnica del Litoral (ESPOL). Based on this process, six dimensions and twenty-one assessment items were defined and adapted to the Ecuadorian microenterprise context. The digital maturity level was calculated through a composite index based on the methodology proposed by the OECD and the Joint Research Centre (JRC), applying min-max normalization, linear aggregation with equal weighting, and automated recommendation generation. The developed platform provides a technological tool to facilitate digital maturity diagnostics and establishes the foundation for future empirical studies with Ecuadorian microenterprises, contributing to the design of contextualized digital transformation strategies.

Keywords: self-assessment, entrepreneurship, digital maturity, digital transformation, web platform.

Introducción

La transformación digital se ha consolidado como uno de los principales impulsores del desarrollo económico del siglo XXI, redefiniendo la competitividad, la productividad y la capacidad de innovación de las organizaciones (Vial, 2019; Verhoef et al., 2021). La crisis provocada por el COVID-19 aceleró drásticamente este proceso y evidenció que la adopción de tecnologías digitales no constituye únicamente una ventaja competitiva, sino también un factor crítico de resiliencia y, en numerosos casos, de supervivencia empresarial (Kraus et al., 2020). En economías en desarrollo, la digitalización adquiere una relevancia aún mayor, ya que permite ampliar el acceso a mercados, optimizar procesos y generar crecimiento inclusivo aun en contextos de recursos limitados (World Bank, 2016).

En América Latina, los emprendimientos y las micro, pequeñas y medianas empresas (mipymes) representan una parte fundamental del tejido productivo y son actores clave para la generación de empleo y la dinamización económica. Sin embargo, su nivel de digitalización continúa siendo limitado. La Comisión Económica para América Latina y el Caribe (CEPAL) estima que cerca del 70 % de las mipymes de la región no cuenta con presencia digital y que gran parte de estas utiliza internet de manera pasiva, sin aprovechar plenamente sus capacidades transaccionales o estratégicas (Dini et al., 2021). En Ecuador, esta situación es consistente con la

realidad regional: un autodiagnóstico aplicado a 617 empresas reveló que la mayoría se encuentra en niveles iniciales o novatos de madurez digital, reflejando una adopción principalmente reactiva, fragmentada y de corto plazo (Granda & Campoverde, 2022).

Más allá de la brecha de acceso tecnológico, persiste un problema profundo relacionado con la ausencia de mecanismos de diagnóstico claros, sistemáticos y contextualizados que permitan a los microemprendimientos conocer objetivamente su situación frente a la transformación digital. En muchos casos, los emprendedores desconocen su nivel real de desarrollo digital, carecen de criterios técnicos para identificar fortalezas y debilidades, priorizar áreas críticas de mejora o establecer una secuencia lógica para la incorporación progresiva de tecnologías. Esta limitación dificulta la formulación de estrategias de digitalización alineadas con las necesidades específicas del negocio y con su capacidad de implementación.

Como consecuencia, las decisiones de inversión en tecnologías digitales suelen sustentarse en percepciones subjetivas, recomendaciones informales o tendencias coyunturales del mercado, en lugar de basarse en un diagnóstico estructurado de las capacidades organizacionales. Ello puede generar asignaciones ineficientes de recursos, adopciones tecnológicas poco sostenibles e incluso el abandono de iniciativas digitales que no producen los resultados esperados. Esta situación resulta especialmente crítica en los microemprendimientos, donde las restricciones financieras, la limitada disponibilidad de talento especializado y la escasa planificación estratégica reducen el margen para corregir decisiones equivocadas. Paralelamente, numerosos programas públicos y privados, orientados al fortalecimiento empresarial, ofrecen soluciones estandarizadas que no consideran el nivel de desarrollo digital particular de cada organización, lo que limita la efectividad de las intervenciones y dificulta la construcción de procesos de mejora continua adaptados a las necesidades reales de cada emprendimiento (Dini et al., 2021).

En este contexto, el concepto de madurez digital constituye un marco de referencia que permite comprender la transformación digital como un proceso evolutivo y medible, superando la visión reduccionista centrada únicamente en la adquisición de tecnologías. La madurez digital puede entenderse como el grado en que una organización ha adoptado, integrado y sistematizado tecnologías digitales en sus procesos, cultura organizacional, estrategia y modelo de negocio, generando capacidades que favorecen la innovación, la eficiencia operativa y la creación de valor sostenible (Vial, 2019). Desde esta perspectiva, la digitalización representa únicamente una fase inicial enfocada en la conversión de procesos analógicos en digitales, mientras que la madurez digital implica una transformación organizacional integral en la que la tecnología se convierte en un habilitador estratégico para mejorar el desempeño empresarial y fortalecer la competitividad.

Diversos estudios coinciden en que alcanzar mayores niveles de madurez digital requiere mucho más que incorporar herramientas tecnológicas. La evidencia demuestra que el éxito de la transformación digital depende de la articulación entre factores tecnológicos y organizacionales, entre los que destacan la definición de una estrategia digital, el compromiso del liderazgo, la existencia de una cultura orientada a la innovación, el desarrollo de competencias digitales en las personas y la capacidad de gestionar el cambio organizacional (Kane et al., 2015). En consecuencia, evaluar la madurez digital implica analizar de forma integrada múltiples dimensiones que reflejan tanto las capacidades tecnológicas como las capacidades de gestión necesarias para sostener procesos de transformación a largo plazo.

Bajo este enfoque, los modelos de madurez digital se han consolidado como herramientas metodológicas que permiten diagnosticar el estado actual de una organización, identificar brechas

respecto a niveles superiores de desarrollo y orientar la definición de planes de mejora. Un modelo de madurez describe la evolución de una entidad a lo largo de una secuencia de etapas, cada una caracterizada por capacidades y prácticas organizacionales específicas, permitiendo posicionarla dentro de una trayectoria progresiva de desarrollo (de Bruin et al., 2005; Becker et al., 2009; Mettler, 2011). Además de facilitar la evaluación comparativa, estos modelos proporcionan una base objetiva para la toma de decisiones estratégicas, el seguimiento del avance de la transformación digital y la priorización de acciones que incrementen la capacidad competitiva de las organizaciones en entornos cada vez más digitalizados.

Para evaluar este nivel de madurez, diversos autores han propuesto modelos que priorizan distintas dimensiones. El modelo de Schumacher et al. (2016) fue desarrollado para contextos de industria 4.0 y enfatiza dimensiones tecnológicas y organizacionales avanzadas, mientras que Berghaus y Back (2016) identificaron cinco etapas evolutivas de transformación digital en más de 400 organizaciones. Las revisiones sistemáticas muestran que estas propuestas varían considerablemente en estructura y alcance y que muchas resultan conceptualmente ambiguas o poco adecuadas para pequeñas empresas (Teichert, 2019; Mittal et al., 2018; Thordsen et al., 2020). A pesar de esta diversidad, existe convergencia en un conjunto de dimensiones fundamentales que aparecen de manera recurrente: estrategia, tecnología, personas y cultura, procesos, datos y clientes.

Estas dimensiones constituyen la base de instrumentos consolidados orientados al segmento pyme. Entre ellos, North et al. (2020) propusieron un marco de crecimiento habilitado digitalmente, mientras que Williams et al. (2024) desarrollaron y validaron, mediante un enfoque mixto, un modelo específico para pymes basado en seis dimensiones: estrategia, productos y servicios, tecnología, personas y cultura, gestión y procesos. Asimismo, el Digital Maturity Assessment (DMA), desarrollado por la Comisión Europea a través del Joint Research Centre (JRC) y la red de European Digital Innovation Hubs, evalúa seis dimensiones y genera una puntuación compuesta en escala de 0 a 100 (European Commission, 2024; European Commission & Joint Research Centre, 2023). El diseño de esta herramienta se fundamenta en principios de construcción de indicadores compuestos y ha sido utilizado en investigaciones revisadas por pares, como estudios aplicados a pymes manufactureras en Luxemburgo (Orošnjak et al., 2025). En el contexto emprendedor, Tiggemann Prando et al. (2025) desarrollaron una escala de potencial emprendedor para la transformación digital basada en teoría de respuesta al ítem, identificando tres niveles de desarrollo: bajo, intermedio y alto.

Aunque existen herramientas regionales aplicadas a empresas ecuatorianas (Granda & Campoverde, 2022; Inter-American Development Bank, 2022) y utilizadas como referencia en estudios nacionales sobre mipymes (Armijos De La Cruz & Palacios Meléndez, 2024), persiste una brecha importante en la literatura. Los trabajos que documenten rigurosamente la ingeniería de un instrumento específicamente diseñado para microemprendimientos, que sea breve, accesible, gratuito, en español, metodológicamente transparente y capaz de generar recomendaciones accionables son escasos. Los microemprendimientos presentan características organizativas, operativas y de gestión particulares que limitan la aplicabilidad directa de modelos de madurez digital diseñados para organizaciones de mayor tamaño.

En este contexto, existe una necesidad de desarrollar enfoques que reconozcan sus recursos limitados, sus estructuras más flexibles y sus capacidades tecnológicas diferenciadas, permitiendo una evaluación de la madurez digital acorde con su realidad y sus necesidades específicas. En consecuencia, el objetivo de este trabajo es diseñar y desplegar una herramienta de diagnóstico de la madurez digital para microemprendimientos ecuatorianos, cuyo instrumento

esté sustentado en investigaciones validadas y ampliamente citadas y cuyo cálculo sea transparente y reproducible, entregado a través de una plataforma web abierta. La validación empírica de las propiedades psicométricas del instrumento con emprendedores queda fuera del alcance de este estudio y constituye una línea futura de investigación.

La relevancia y originalidad de esta propuesta se centran en tres aspectos. Primero, se enfoca en los microemprendimientos, un sector poco considerado en los modelos de madurez digital existentes. Segundo, el instrumento se construye adaptando y combinando elementos de modelos previamente desarrollados, de acuerdo con la realidad de este tipo de negocios. Tercero, permite no solo conocer el nivel de madurez digital, sino también identificar los aspectos que requieren mayor atención. En este sentido, el aporte principal del trabajo está en ofrecer una herramienta adaptada a las necesidades reales de los microemprendimientos.

Materiales y Métodos

La presente investigación es de carácter aplicado y adopta un enfoque no experimental, orientado al diseño e implementación de una plataforma web para la evaluación de la madurez digital de microemprendimientos. El estudio busca desarrollar una herramienta tecnológica fundamentada en la literatura científica que permita diagnosticar el nivel de madurez digital de los emprendimientos y proporcionar información útil para orientar procesos de mejora continua.

Metodológicamente, la investigación se enmarca en la tradición del diseño de modelos de madurez, la cual propone un proceso estructurado e iterativo para garantizar el rigor en la construcción, adaptación y operacionalización de los instrumentos de evaluación. Este enfoque establece que un modelo de madurez debe sustentarse en bases conceptuales sólidas, definir dimensiones de análisis claramente delimitadas y ofrecer criterios objetivos para valorar el nivel de desarrollo de una organización (de Bruin et al., 2005; Mettler, 2011).

El desarrollo de la investigación se estructuró en cinco fases secuenciales: (1) revisión documental de modelos de madurez digital de alcance internacional y regional, con el propósito de identificar las dimensiones e indicadores más relevantes; (2) selección, comparación y adaptación de las dimensiones e ítems del instrumento de evaluación al contexto de los microemprendimientos ecuatorianos; (3) definición del modelo de puntuación y del Índice de Madurez Digital; (4) diseño e implementación de la arquitectura y de la plataforma web para automatizar el proceso de evaluación; y (5) fundamentación metodológica del instrumento mediante la trazabilidad entre cada dimensión, sus ítems y la literatura científica que respalda su inclusión.

Método de selección del instrumento

El primer compromiso metodológico consistió en construir el instrumento mediante un proceso de triangulación de modelos y herramientas ampliamente utilizadas para la evaluación de la madurez digital, en lugar de adoptar íntegramente un único instrumento. Este enfoque responde a la recomendación de integrar elementos provenientes de diferentes modelos validados para obtener evaluaciones más completas y robustas (Williams et al., 2024). Como se muestra en la Figura 1, el procedimiento se estructuró en cuatro capas.

La capa de fundamento estableció la base conceptual de los modelos de madurez digital, identificando las dimensiones recurrentes reportadas en la literatura especializada sobre diseño de modelos de madurez (de Bruin et al., 2005; Becker et al., 2009; Mettler, 2011) y en modelos ampliamente reconocidos (Schumacher et al., 2016; Berghaus & Back, 2016; Teichert, 2019; Thordsen et al., 2020).

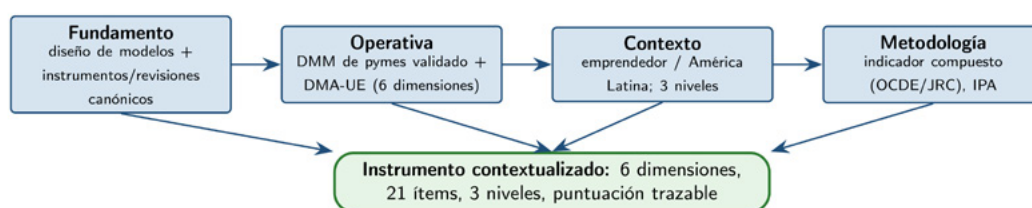
La capa operativa definió la estructura del instrumento mediante la integración de un modelo de madurez digital validado para pequeñas y medianas empresas (Williams et al., 2024), el Digital Maturity Assessment de la Comisión Europea (European Commission, 2024) y el instrumento Chequeo Digital, desarrollado por el Banco Interamericano de Desarrollo (BID) e implementado en Ecuador por la Escuela Superior Politécnica del Litoral (ESPOL) (Inter-American Development Bank, 2022; Granda & Campoverde, 2022). Estas fuentes proporcionaron dimensiones, criterios de evaluación y elementos de operacionalización que posteriormente fueron analizados, comparados y adaptados.

La capa de contextualización adecuó las dimensiones, los indicadores y los ítems a las características de los microemprendimientos ecuatorianos, incorporando además la interpretación de los niveles de madurez propuesta para emprendimientos latinoamericanos (Tiggemann Prando et al., 2025; Dini et al., 2021).

Finalmente, el componente metodológico definió el procedimiento para la construcción del Índice de Madurez Digital mediante la metodología para el desarrollo de indicadores compuestos propuesta por la OCDE y el Joint Research Centre (OECD/JRC, 2008), así como la priorización de recomendaciones utilizando el análisis importancia-desempeño (Martilla & James, 1977).

Figura 1

Método de selección del instrumento por capas



El instrumento contextualizado

El instrumento comprende seis dimensiones y veintiún ítems. Las seis dimensiones son recurrentes en los instrumentos validados previamente, restringidas a aquellas que un microemprendedor puede autoevaluar de forma fiable. Las dimensiones difíciles de autodiagnosticar en un negocio unipersonal se excluyeron, decisión de alcance consistente con las preocupaciones de aplicabilidad a las pymes señaladas en las revisiones (Thordsen et al., 2020; Mittal et al., 2018). Los veintiún ítems fueron seleccionados mediante un proceso de comparación y trazabilidad bibliográfica entre los modelos de referencia, conservando únicamente aquellos indicadores recurrentes, metodológicamente sustentados y pertinentes para caracterizar la madurez digital de los microemprendimientos ecuatorianos, además de ser susceptibles de autoevaluación por parte del emprendedor. La Tabla 1 mapea cada dimensión a su fundamento en las fuentes validadas y muestra la convergencia con el referente regional.

Cada ítem se evalúa mediante una escala ordinal de tres niveles con puntuaciones de 0 (bajo), 1 (medio) y 2 (alto), siguiendo una escala tipo Likert. Esta estructura fue seleccionada por su simplicidad de interpretación y facilidad de aplicación en procesos de autoevaluación, lo que permite que los emprendedores comprendan los resultados obtenidos. La clasificación en tres niveles coincide con enfoques validados para la evaluación de capacidades y madurez en emprendimientos (Tiggemann Prando et al., 2025).

Tabla 1

Dimensiones, ítems del instrumento y su fundamento en fuentes validadas

Dimensión	Ítem	Aspecto evaluado	Fundamento en fuentes validadas
D1. Tecnologías y habilidades digitales	I1	Equipamiento informático dedicado al negocio	Eje tecnológico recurrente (Schumacher <i>et al.</i> , 2016; Teichert, 2019); “preparación digital” del DMA-UE (European Commission, 2024); “tecnología” del modelo para pymes (Williams <i>et al.</i> , 2024). Referente regional: dimensión homónima (Granda & Campoverde, 2022).
	I2	Tipo de conexión a internet	
	I3	Programas y aplicaciones de uso general (ofimática y nube)	
D2. Comunicaciones y canales de venta	I4	Nivel transaccional de los canales digitales (información, pedido y pago en línea)	Canales digitales y comercio electrónico, desafío central de las mipymes de la región (Dini <i>et al.</i> , 2021); conectividad en la escala emprendedora (Tiggemann Prando <i>et al.</i> , 2025).
	I5	Medios de pago disponibles para los clientes	
	I6	Atención al cliente digital fuera del horario laboral	
D3. Organización y personas	I7	Aprendizaje de nuevas herramientas digitales en el último año	“Digitalización centrada en las personas” del DMA-UE (European Commission, 2024); “personas y cultura” y “gestión” del modelo para pymes (Williams <i>et al.</i> , 2024); capital humano (North <i>et al.</i> , 2020).
	I8	Plan de desarrollo de habilidades digitales del equipo	
	I9	Software para la coordinación del trabajo interno	
	I10	Gestión del soporte técnico y del desarrollo tecnológico	
	I11	Plan para incorporar o mejorar el uso de tecnología	
D4. Estrategia y transformación digital	I12	Liderazgo y gobernanza de las decisiones tecnológicas	“Estrategia de negocio digital” del DMA-UE (European Commission, 2024); “estrategia” del modelo para pymes (Williams <i>et al.</i> , 2024); primacía de la estrategia (Kane <i>et al.</i> , 2015; Vial, 2019). El ítem I15 responde a la Ley Orgánica de Protección de Datos Personales del Ecuador.
	I13	Uso de inteligencia artificial en actividades productivas	
	I14	Medidas de ciberseguridad (contraseñas, respaldos y actualizaciones)	
	I15	Protección de los datos personales de los clientes	
D5. Datos y analítica	I16	Gestión de la base de datos de clientes	“Gobernanza y conectividad de datos” del DMA-UE (European Commission, 2024); etapa de “empresa orientada por datos” (Berghaus & Back, 2016); “cultura orientada por datos” (Tiggemann Prando <i>et al.</i> , 2025).
	I17	Tipos de datos que se recopilan	
	I18	Uso de los datos para la toma de decisiones	
D6. Procesos	I19	Control de inventario o agenda de servicios	“Automatización e inteligencia artificial” del DMA-UE (European Commission, 2024); “procesos” del modelo para pymes (Williams <i>et al.</i> , 2024); digitalización de procesos (Vial, 2019; Verhoef <i>et al.</i> , 2021).
	I20	Software de gestión financiera y contable	
	I21	Automatización de flujos de trabajo	

Nota. El instrumento evalúa 21 ítems distribuidos en seis dimensiones. Cada ítem se responde en una escala ordinal de tres niveles: 0 = bajo, 1 = medio y 2 = alto. Las dimensiones se fundamentan en Chequeo Digital (Inter-American Development Bank, 2022) y Granda y Campoverde (2022), mientras que los ítems fueron elaborados a partir de Williams *et al.* (2024), European Commission (2024) y Tiggemann Prando *et al.* (2025).

La contextualización del instrumento al entorno de los microemprendimientos ecuatorianos se realizó mediante tres adaptaciones. En primer lugar, el lenguaje de los ítems y

de las opciones de respuesta se formuló utilizando acciones concretas y observables, evitando la terminología técnica para favorecer una autoevaluación clara y consistente. En segundo lugar, se incorporó un ítem relacionado con la protección de datos personales, alineado con la Ley Orgánica de Protección de Datos Personales del Ecuador, con el propósito de reflejar un requisito regulatorio relevante para el contexto nacional que no se encuentra explícitamente considerado en el instrumento europeo. Finalmente, el cuestionario se diseñó con una extensión de veintiún ítems, seleccionados por su relevancia y respaldo en la literatura, con el fin de disminuir el tiempo de respuesta y la carga cognitiva del emprendedor sin comprometer la cobertura de las dimensiones esenciales de la madurez digital.

El modelo de puntuación

El modelo de puntuación se construyó como un indicador compuesto, siguiendo el procedimiento propuesto en el Handbook on Constructing Composite Indicators de la OCDE y el Joint Research Centre (JRC), que comprende las etapas de selección de indicadores, normalización, ponderación y agregación (OECD/JRC, 2008). El diseño del modelo privilegia la simplicidad y la transparencia, de modo que el cálculo pueda ser comprendido y reproducido fácilmente por el usuario final.

Sea (i) el índice de las seis dimensiones y (j) el índice de los ítems pertenecientes a cada dimensión. Cada ítem (P_{ij}) adopta uno de los valores ordinales del conjunto $\{0,1,2\}$, mientras que (Q_i) representa el número de ítems asociados a la dimensión (i). El puntaje bruto de cada dimensión se obtiene mediante la suma de los puntajes de sus respectivos ítems:

$$(S_i = \sum_j P_{ij}) \text{ (Ec. 1)}$$

Dado que las dimensiones contienen entre tres y cinco ítems, los puntajes brutos no son directamente comparables. Para homogeneizar la escala de medición, cada puntaje se normaliza mediante el método mín-máx, considerando 0 como valor mínimo y el puntaje máximo posible de la dimensión como valor máximo ($(Q_i \times 2)$):

$$(N_i = \frac{S_i}{Q_i \times 2} \times 100) \text{ (Ec. 2)}$$

La ponderación de las dimensiones se estableció de forma equitativa ($(W_i = 1/6)$), siguiendo la recomendación de la OCDE/JRC de utilizar pesos iguales cuando no existe evidencia estadística o fundamento teórico suficiente para asignar ponderaciones diferenciadas (OECD/JRC, 2008). Este criterio también es consistente con el enfoque del Digital Maturity Assessment de la Comisión Europea, en el que todas las dimensiones contribuyen de manera equivalente al resultado global.

El Índice de Madurez Digital (IMD) se obtiene mediante agregación lineal de las seis dimensiones normalizadas:

$$(IMD = \sum_i W_i N_i = \frac{1}{6} \sum_i N_i) \text{ (Ec. 3)}$$

El índice resultante toma valores entre 0 y 100 y, debido a la igualdad de ponderaciones, equivale al promedio de las seis dimensiones normalizadas, lo que facilita su interpretación y reproducibilidad.

Tanto el IMD como los valores normalizados de cada dimensión ((N_i)) se clasifican en tres niveles de madurez: incipiente (0–33,33), en transición (33,34–66,66) y consolidado

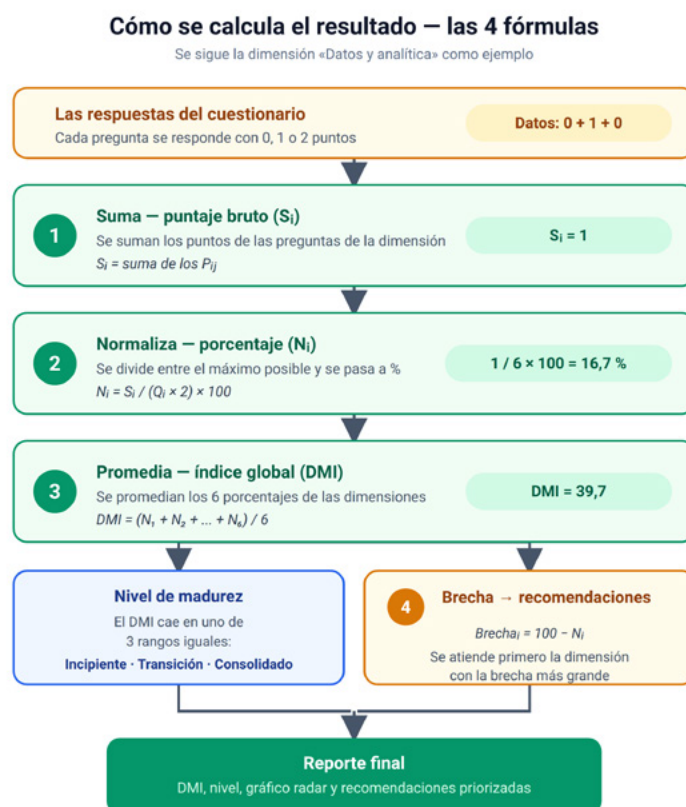
(66,67–100). Esta clasificación mantiene la estructura progresiva característica de los modelos de madurez (de Bruin et al., 2005; Becker et al., 2009) y resulta consistente con los esquemas de evaluación simplificados, utilizados en instrumentos dirigidos a emprendimientos (Tiggemann Prando et al., 2025).

Finalmente, la priorización de las recomendaciones se realizó mediante un análisis importancia-desempeño (Importance–Performance Analysis, IPA) (Martilla & James, 1977). Considerando que todas las dimensiones poseen la misma importancia relativa, la brecha de desempeño de cada dimensión se calcula como:

$$(\text{Brecha}_i = 100 - N_i) \text{ (Ec. 4)}$$

Las dimensiones se ordenan de forma descendente según su brecha, priorizando aquellas con menor nivel de desempeño para orientar las acciones de mejora. La Figura 2 resume el procedimiento completo mediante un ejemplo numérico.

Figura 2
El modelo de puntuación como cuatro operaciones encadenadas

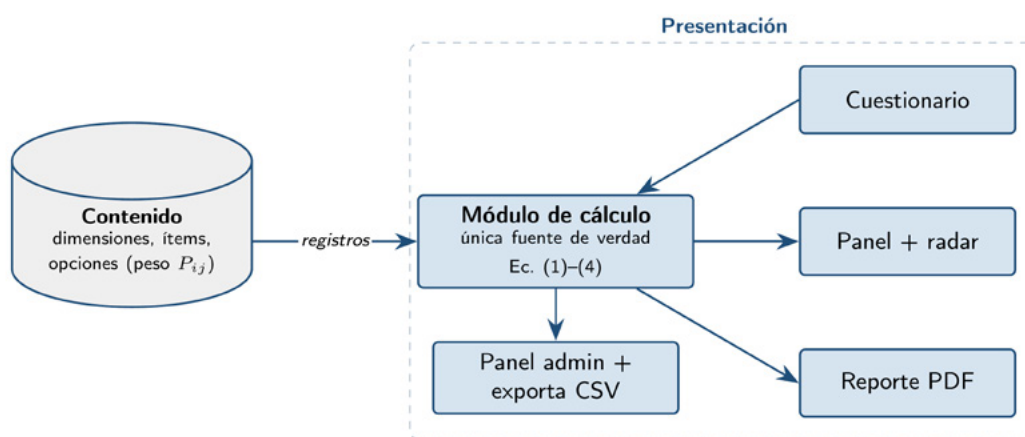


Ingeniería de la plataforma

El instrumento y el modelo se operacionalizan como una aplicación web gratuita, en español, respaldada por una base de datos relacional. La contribución de la ingeniería no radica en la elección de marcos de desarrollo, sino en un conjunto de decisiones que garantizan que el cálculo definido se ejecute de forma fiel, única y reproducible. Como se observa en la Figura 3, la plataforma separa tres responsabilidades: el contenido del instrumento (dimensiones, ítems y opciones con su puntaje 0/1/2), almacenado como datos en lugar de codificado en el programa; la lógica de cálculo, concentrada en un único módulo; y la presentación (cuestionario interactivo, panel de resultados y reporte). Dos decisiones son centrales; primera, el cuestionario

se almacena como datos editables: cada opción es un registro cuyo peso es el valor P_{ij} que alimenta la ecuación 1, de modo que el instrumento puede revisarse sin reprogramar el cálculo y la correspondencia entre respuesta y puntaje queda explícita y auditable. Segunda, toda la aritmética vive en un único módulo de cálculo invocado por igual por el cuestionario, el panel de resultados, el panel de administración y el reporte PDF. De este modo, un mismo conjunto de respuestas produce siempre el mismo diagnóstico. Este principio de única fuente de verdad evita inconsistencias entre las distintas superficies de la plataforma. La plataforma se encuentra desplegada y en operación, con acceso público y gratuito: no exige pago, licencia ni registro institucional y basta un navegador web para completar el diagnóstico y descargar el reporte. Está disponible públicamente en <https://semilla.indoamericattech.com> (Sánchez-Montero et al., 2026). Las figuras 5 y 6 documentan, respectivamente, el flujo de uso y la interfaz de resultados del sistema en funcionamiento.

Figura 3
Arquitectura de la plataforma



Resultados y Discusión

De acuerdo con el alcance definido para esta investigación, esta sección presenta los principales resultados obtenidos: el instrumento de evaluación, el modelo de puntuación, la plataforma web desarrollada y la fundamentación metodológica de su contenido. En consecuencia, no se incluyen resultados derivados de una aplicación empírica del instrumento sobre una población de microemprendimientos. El primer resultado corresponde al instrumento de evaluación, conformado por una matriz de seis dimensiones y veintiún ítems (Tabla 1), cuyos componentes fueron seleccionados mediante un proceso de trazabilidad bibliográfica y contextualizados a la realidad de los microemprendimientos ecuatorianos. Con el propósito de verificar el funcionamiento del modelo de evaluación y del algoritmo de cálculo implementado en la plataforma, se definieron tres emprendimientos como casos de prueba funcional, representativos de los niveles de madurez incipiente, en transición y consolidado. La Tabla 2 presenta el cálculo del Índice de Madurez Digital (IMD) para cada caso, mientras que la Figura 4 muestra sus respectivos perfiles radiales.

El emprendimiento A (IMD = 14,9), clasificado en el nivel incipiente, evidencia una adopción limitada de tecnologías digitales y una ausencia de capacidades en la dimensión de datos y analítica. El emprendimiento B (IMD = 39,7), correspondiente al nivel en transición, presenta fortalezas en infraestructura tecnológica, aunque mantiene brechas relevantes en las dimensiones de datos y analítica y organización. El análisis de brechas identifica como prioridades

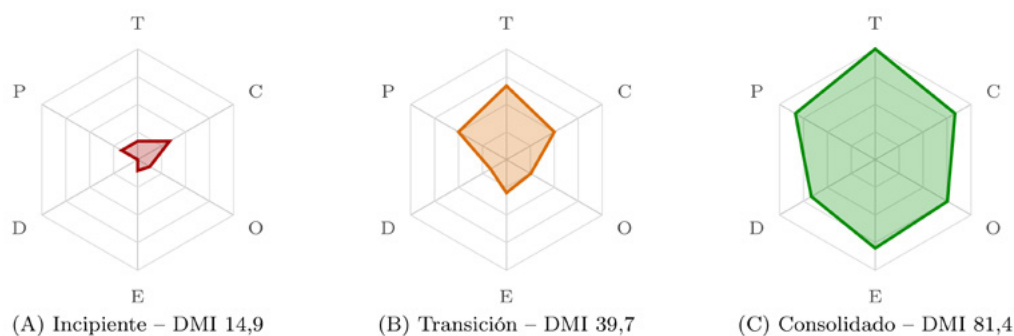
de intervención las dimensiones de datos y analítica (83,3), organización (75,0), estrategia (70,0), comunicaciones y procesos (50,0) y tecnologías (33,3). Por su parte, el emprendimiento C (IMD = 81,4), clasificado en el nivel consolidado, alcanza valores superiores al umbral de 66,7 en todas las dimensiones evaluadas. Los perfiles radiales muestran una progresión consistente entre los tres casos de prueba y facilitan la identificación del equilibrio relativo entre dimensiones. Los tres casos fueron contruidos para representar los tres niveles previstos por el modelo. En consecuencia, permiten verificar que el algoritmo produce resultados coherentes y reproducibles con las entradas diseñadas, pero no constituyen evidencia sobre el comportamiento del instrumento frente a respuestas de microemprendimientos reales.

Tabla 2
Cálculo del Índice de Madurez Digital para los tres casos de prueba funcional

Dimensión	Q_i	A. Incipiente (S_i / N_i)	B. En transición (S_i / N_i)	C. Consolidado (S_i / N_i)
Tecnologías y habilidades digitales	3	1 / 16,7	4 / 66,7	6 / 100,0
Comunicaciones y canales de venta	3	2 / 33,3	3 / 50,0	5 / 83,3
Organización y personas	4	1 / 12,5	2 / 25,0	6 / 75,0
Estrategia y transformación digital	5	1 / 10,0	3 / 30,0	8 / 80,0
Datos y analítica	3	0 / 0,0	1 / 16,7	4 / 66,7
Procesos	3	1 / 16,7	3 / 50,0	5 / 83,3
IMD (promedio de N_i)	—	14,9	39,7	81,4
Nivel de madurez	—	Incipiente	En transición	Consolidado

Nota. Q_i representa el número de ítems de la dimensión; S_i , el puntaje bruto obtenido; y N_i , el puntaje normalizado en una escala de 0 a 100. El Índice de Madurez Digital (IMD) corresponde al promedio simple de los seis valores N_i , conforme a la ponderación equitativa propuesta por OECD/JRC (2008). Los tres casos corresponden a escenarios de prueba funcional del algoritmo y no a datos empíricos.

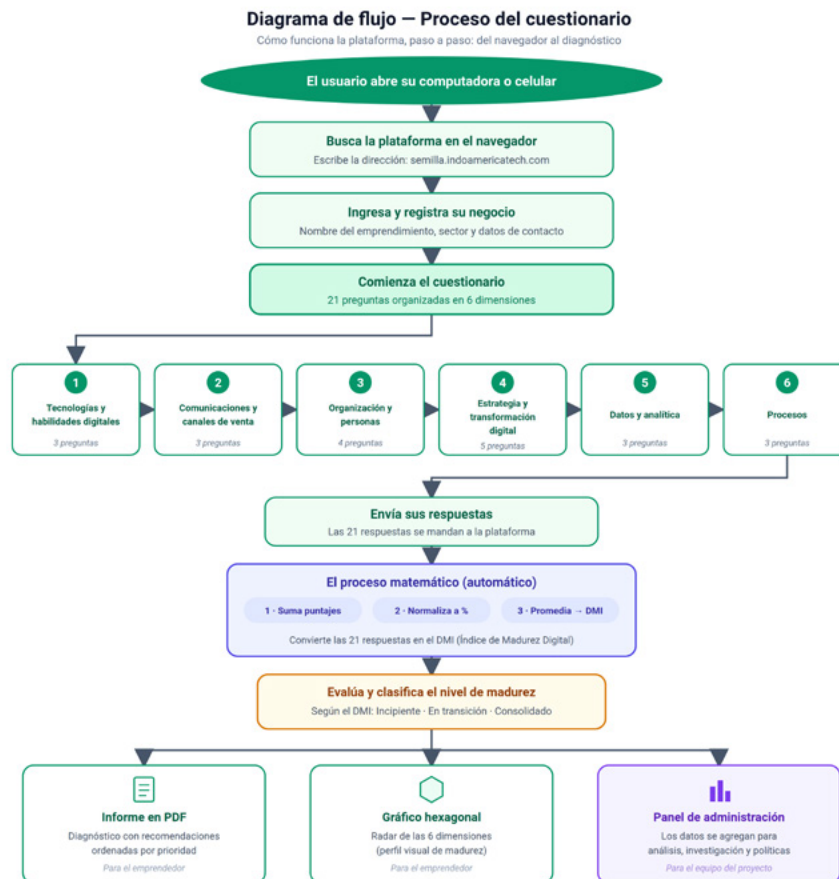
Figura 4
Perfiles radiales de los tres emprendimientos de la Tabla 2



El segundo resultado es la plataforma funcional. Como se ilustra en la Figura 5, el emprendedor sigue un flujo de cuatro pasos: registra los datos básicos de su negocio, responde los veintinueve ítems, la plataforma aplica el modelo y el emprendedor recibe un panel en pantalla (IMD, nivel y gráfico radar) y un reporte PDF descargable de dos partes: un diagnóstico y un conjunto de recomendaciones. De acuerdo con el principio de proporcionar recomendaciones y

no solo una puntuación, el reporte no comunica solo un puntaje: para cada dimensión presenta una recomendación aplicable acorde con su nivel actual, de modo que el emprendedor identifica de inmediato dónde actuar. Un panel de administración con acceso restringido agrega la muestra y la exporta como valores separados por comas para análisis posterior y constituye el puente hacia el estudio empírico previsto.

Figura 5
Flujo de extremo a extremo de la plataforma

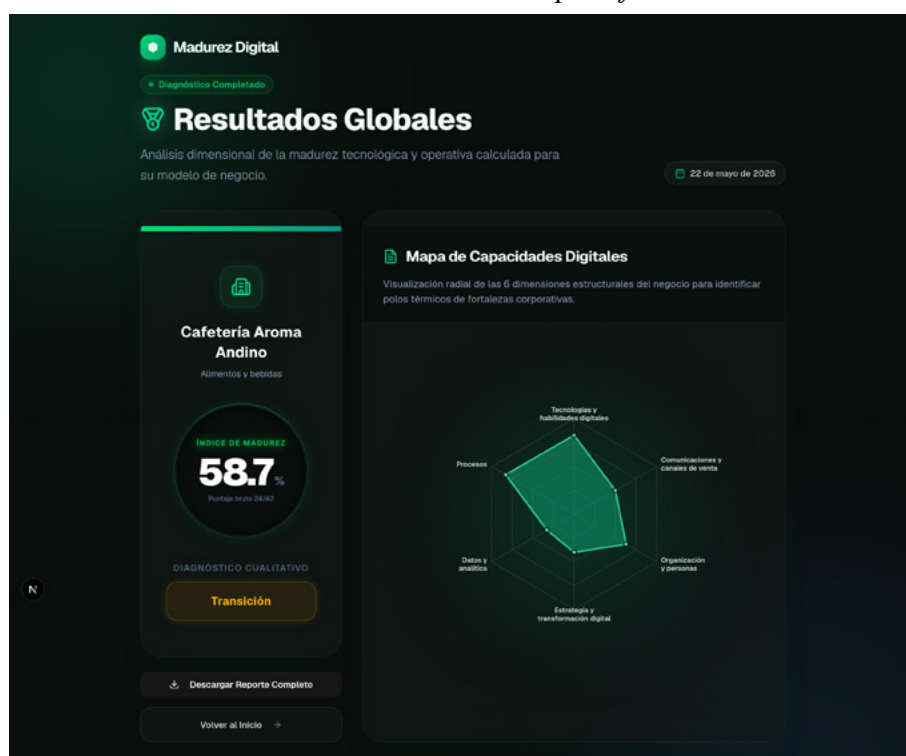


La Figura 6 presenta el panel de resultados que la plataforma entrega al emprendedor al finalizar el cuestionario. En esta interfaz se sintetiza de manera visual e inmediata la información obtenida durante la evaluación, mostrando el Índice de Madurez Digital (IMD) calculado, el nivel de madurez alcanzado de acuerdo con la escala definida por el modelo y un perfil hexagonal que representa el desempeño obtenido en cada una de las seis dimensiones evaluadas. Esta visualización permite identificar de forma rápida las fortalezas y las áreas que requieren mayor atención, facilitando la interpretación de los resultados y proporcionando una base objetiva para la priorización de acciones orientadas a fortalecer el proceso de transformación digital del microemprendimiento.

El tercer resultado corresponde a la fundamentación del contenido del instrumento. En lugar de sustentarse en un panel de expertos individuales, la validez de contenido se apoya en la trazabilidad de cada dimensión e ítem hacia fuentes metodológicas validadas y ampliamente citadas (Tabla 1). En particular, los instrumentos operativos de referencia corresponden a un modelo de madurez digital para pymes validado mediante un enfoque de métodos mixtos (Williams et al., 2024), al Digital Maturity Assessment de la Comisión Europea (European Commission, 2024) y al instrumento Chequeo Digital, desarrollado por el Banco Interamericano

de Desarrollo (BID) e implementado en Ecuador por la Escuela Superior Politécnica del Litoral (ESPOL) (Inter-American Development Bank, 2022; Granda & Campoverde, 2022).

Figura 6
Panel de resultados de la plataforma



Nota. Captura de pantalla de la plataforma desarrollada. Los valores mostrados corresponden a un caso de prueba funcional y no a datos empíricos.

Asimismo, el conjunto de dimensiones converge con las identificadas de forma recurrente en revisiones sistemáticas sobre madurez digital (Teichert, 2019; Thordsen et al., 2020). La clasificación en tres niveles adopta la interpretación validada para emprendimientos (Tiggemann Prando et al., 2025); y el modelo de puntuación reproduce el procedimiento canónico para la construcción de indicadores compuestos propuesto por la OCDE y el Joint Research Centre (OECD/JRC, 2008), complementado con un análisis de priorización importancia-desempeño (Martilla & James, 1977). Como ocurre con cualquier instrumento de medición, esta fundamentación puede complementarse mediante estudios de validación psicométrica, incorporando análisis de validez de constructo y fiabilidad mediante indicadores como el alfa de Cronbach, o modelos basados en la teoría de respuesta al ítem (Tiggemann Prando et al., 2025). Finalmente, la Tabla 3 presenta una comparación del instrumento propuesto con las dos fuentes operativas y el referente regional Chequeo Digital.

Los resultados de este trabajo evidencian que el principal desafío en el diagnóstico de madurez digital para microemprendimientos no radica únicamente en la disponibilidad de tecnología, sino en la construcción de instrumentos de evaluación que sean simultáneamente rigurosos, accesibles y contextualizados. Esta necesidad es especialmente relevante en economías emergentes, donde la baja adopción digital no siempre responde a una carencia de herramientas, sino a la ausencia de mecanismos claros para medir, interpretar y priorizar la transformación digital (Dini et al., 2021). La propuesta desarrollada responde a una limitación recurrente en la literatura: considerable parte de los modelos de madurez digital disponibles fueron diseñados para grandes organizaciones o entornos industriales, con estructuras extensas y complejas que

dificultan su aplicación en pequeñas empresas (Teichert, 2019; Mittal et al., 2018). En este sentido, el instrumento diseñado aporta una adaptación orientada al microemprendimiento ecuatoriano, manteniendo la trazabilidad hacia modelos validados y ampliamente citados, pero reduciendo la complejidad operativa para facilitar su uso en escenarios reales de autodiagnóstico.

Tabla 3
Comparación del instrumento propuesto con las fuentes operativas y el referente regional

Característica	Instrumento propuesto	DMA-UE y modelo para pymes	Chequeo Digital (referente regional)
Dimensiones	6	6	8
Ítems	21	Definidos por cada instrumento	Hasta 62
Niveles	3	Escala de 0 a 100	5
Ponderación	Equitativa	Equitativa	No especificada
Población objetivo	Microemprendimientos (Ecuador)	Pymes (Unión Europea)	Mipymes (América Latina)
Acceso	Abierto, gratuito y en español	Abierto y validado	Abierto

Nota. Adaptado de European Commission (2024), Williams et al. (2024), Chequeo Digital (Inter-American Development Bank, 2022) y Granda y Campoverde (2022).

Un hallazgo metodológicamente relevante es que la utilidad de un diagnóstico no depende únicamente de las dimensiones evaluadas, sino también de la transparencia de su sistema de puntuación. Diversos instrumentos reportan resultados agregados sin explicitar claramente su lógica de ponderación, lo que reduce su interpretabilidad y dificulta su reproducibilidad. Al reformular el modelo de puntuación como un indicador compuesto explícito (OECD/JRC, 2008), este trabajo fortalece la consistencia metodológica del diagnóstico; asimismo, la adopción de ponderación equitativa entre dimensiones reduce la subjetividad asociada al uso de pesos arbitrarios, especialmente en ausencia de evidencia empírica suficiente para justificar ponderaciones diferenciadas. Desde la perspectiva de la ingeniería de software, la arquitectura basada en una única fuente de verdad constituye un aporte para la confiabilidad del sistema, pues en plataformas de evaluación las inconsistencias entre el cuestionario, las visualizaciones y los reportes pueden comprometer la validez práctica del diagnóstico.

Las implicaciones de esta propuesta son tanto prácticas como estratégicas. En el ámbito del emprendedor, la plataforma transforma un diagnóstico abstracto en información accionable mediante una hoja de ruta priorizada, facilitando la toma de decisiones sobre inversión tecnológica. En el ámbito institucional, un instrumento estandarizado y reproducible puede fortalecer programas de apoyo, incubación o política pública, al permitir segmentar intervenciones según el nivel real de madurez digital de los emprendimientos en lugar de aplicar estrategias homogéneas (Dini et al., 2021). No obstante, los resultados deben interpretarse considerando las limitaciones del estudio. Este trabajo se centra en el diseño y la ingeniería del instrumento, por lo que no presenta aún evidencia empírica sobre validez de constructo, consistencia interna o fiabilidad psicométrica; la validez de contenido se sustenta en la trazabilidad bibliográfica de las dimensiones e ítems, que constituye una base necesaria, aunque insuficiente, para la validación completa del instrumento. Adicionalmente, por tratarse de un autodiagnóstico, la herramienta está sujeta a sesgos de autopercepción y deseabilidad social. Asimismo, el levantamiento de requisitos de la plataforma se apoyó en la literatura y en los instrumentos de referencia, y no en entrevistas a personas propietarias de microemprendimientos; la incorporación de técnicas de

elicitación con usuarios finales permitiría contrastar la formulación de los ítems y la usabilidad de la herramienta. Finalmente, aunque el instrumento conserva seis dimensiones centrales, algunas variables relevantes —como el liderazgo transformacional, la cultura organizacional o la innovación de producto— fueron simplificadas u omitidas para mantener la brevedad y la usabilidad del autodiagnóstico, lo que mejora la accesibilidad, pero reduce la profundidad analítica.

Debe señalarse que, hasta el momento, no se ha levantado información de campo, ya que los resultados presentados corresponden al diseño del instrumento, al modelo de puntuación y a la verificación funcional del algoritmo, y no a una aplicación del instrumento sobre una muestra de microemprendimientos. La recolección de datos constituye una segunda fase de la investigación, ya prevista, cuyo propósito es establecer la línea base de madurez digital del segmento estudiado, examinar las propiedades psicométricas del instrumento —validez de constructo, consistencia interna y fiabilidad— y valorar la utilidad práctica de las recomendaciones generadas. Esa línea base delimitará, a su vez, el punto de partida de los trabajos futuros derivados de esta investigación.

Conclusiones

Este artículo presenta la ingeniería de una plataforma web abierta para la evaluación de la madurez digital de microemprendimientos ecuatorianos, centrándose en el proceso de selección, adaptación y operacionalización del instrumento de evaluación y en la implementación del modelo de puntuación dentro de una arquitectura tecnológica reproducible. El aporte principal del trabajo consiste en documentar una vía reproducible para construir un instrumento de diagnóstico trazable sin recurrir a la construcción de un modelo completamente nuevo, integrando de manera sistemática evidencia proveniente de modelos de madurez digital ampliamente validados. En particular, la propuesta articula la literatura sobre modelos de madurez, un modelo de madurez digital validado para pymes, el Digital Maturity Assessment de la Comisión Europea, el instrumento Chequeo Digital desarrollado por el Banco Interamericano de Desarrollo (BID) e implementado en Ecuador por la Escuela Superior Politécnica del Litoral (ESPOL), la contextualización al ecosistema emprendedor ecuatoriano y una metodología de puntuación basada en indicadores compuestos de la OCDE/JRC.

Como resultado, se obtuvo una plataforma web gratuita, transparente y disponible en español, diseñada para generar un diagnóstico multidimensional del nivel de madurez digital, identificar las principales brechas y priorizar recomendaciones orientadas a los procesos de transformación digital de los microemprendimientos; su funcionamiento fue verificado mediante casos de prueba funcional del algoritmo de cálculo. Dado que el instrumento no ha sido aplicado todavía a microemprendimientos reales, estas capacidades deben interpretarse como potenciales y no como efectos comprobados: sostener que la herramienta mide efectivamente la madurez digital o que sus recomendaciones producen mejoras exige la validación empírica prevista para la siguiente fase del estudio. Con esa salvedad, la plataforma ofrece una herramienta estandarizada que las instituciones públicas, los gobiernos locales y las organizaciones de apoyo al emprendimiento podrían emplear para homogeneizar procesos de diagnóstico y orientar estrategias de fortalecimiento empresarial.

Como línea futura de investigación, se plantea la aplicación de la plataforma a los microemprendimientos vinculados a los programas de fortalecimiento empresarial de la prefectura de Pichincha, con el propósito de realizar la validación del instrumento, analizar sus propiedades psicométricas, evaluar su utilidad práctica y generar evidencia sobre su contribución

a los procesos de transformación digital en el contexto ecuatoriano. Esta segunda fase permitirá establecer la línea base de madurez digital del segmento y delimitará el alcance de los trabajos posteriores; hasta su culminación, los aportes aquí presentados se circunscriben al diseño, la fundamentación y la ingeniería de la herramienta.

Reconocimientos y Declaraciones

Financiamiento: La Universidad Tecnológica Indoamérica financió la presente investigación mediante la asignación de horas de investigación al equipo investigador y el aporte de recursos económicos provenientes del presupuesto institucional, destinados a la ejecución de las actividades previstas en el proyecto.

Disponibilidad de la plataforma: La plataforma descrita en este artículo es de acceso abierto y gratuito. Su dirección electrónica es: <https://semilla.indoamericattech.com>.

Contribuciones de los autores (taxonomía CRediT): Conceptualización y metodología: Ivanna Sánchez-Montero y Joel Morales Játiva; desarrollo de software y de la plataforma: Joel Morales Játiva; investigación y curación de datos: Ivanna Sánchez-Montero y Joel Morales Játiva; redacción del borrador original: Ivanna Sánchez-Montero y Joel Morales Játiva; redacción, revisión y edición: Ivanna Sánchez-Montero, Joel Morales Játiva y Viviana Elizabeth Cajas-Cajas; supervisión: Ivanna Sánchez-Montero y Viviana Elizabeth Cajas-Cajas.

Declaración sobre el uso de inteligencia artificial. Los autores declaran que, en la elaboración del presente artículo, se han utilizado herramientas de inteligencia artificial generativa (Claude, de Anthropic) como asistente para la redacción y edición del texto y del código fuente de la plataforma. Las tareas asistidas incluyeron la mejora de la claridad y la estructura de la redacción en las secciones de Introducción, Materiales y Métodos, y Resultados y Discusión, así como la depuración del código de la plataforma. Todas las decisiones técnicas, el instrumento, el modelo de puntuación, las figuras y el contenido final fueron concebidos, verificados y aprobados por los autores, quienes asumen la responsabilidad total del contenido del artículo. Ningún sistema de IA se utilizó para generar datos de investigación ni resultados experimentales.

Referencias

- Armijos De La Cruz, B. A., & Palacios Meléndez, J. G. (2024). Madurez digital del e-commerce y desarrollo empresarial de las MIPYMES, provincia de Santa Elena. *Ciencia Latina Revista Científica Multidisciplinar*, 8(5), 11275–11291. https://doi.org/10.37811/cl_rcm.v8i5.14509
- Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management. *Business & Information Systems Engineering*, 1(3), 213–222. <https://doi.org/10.1007/s12599-009-0044-5>
- Berghaus, S., & Back, A. (2016). Stages in digital business transformation: Results of an empirical maturity study. En *Proceedings of the 10th Mediterranean Conference on Information Systems (MCIS)* (pp. 1–17). AIS Electronic Library. <https://aisel.aisnet.org/mcis2016/22/>
- de Bruin, T., Freeze, R., Kaulkarni, U., & Rosemann, M. (2005). Understanding the main phases of developing a maturity assessment model. En *Proceedings of the 16th Australasian Conference on Information Systems (ACIS)*. <https://aisel.aisnet.org/acis2005/109/>
- Dini, M., Gligo, N., & Patiño, A. (2021). *Transformación digital de las mipymes: Elementos para el diseño de políticas* (Documentos de Proyectos LC/TS.2021/99). Comisión Económica para América Latina y el Caribe (CEPAL). <https://repositorio.cepal.org/handle/11362/47183>

- European Commission. (2024). *Digital maturity assessment (DMA) tool for SMEs*. European Digital Innovation Hubs Network, Joint Research Centre. <https://european-digital-innovation-hubs.ec.europa.eu/dma-tool>
- European Commission, Joint Research Centre. (2023). *Digital maturity assessment (dma) framework: Guidelines for european digital innovation hubs*. European Digital Innovation Hubs Network.
- Granda, M. L., & Campoverde, J. (2022). *¿Cuál es el nivel de digitalización de las empresas en Ecuador? Una aproximación a través de la herramienta Chequeo Digital. Reporte 2020–2021*. Escuela Superior Politécnica del Litoral (ESPOL). <https://www.espae.edu.ec/investigacion/publicaciones/cual-es-el-nivel-de-digitalizacion-de-las-empresas-en-ecuador-una-aproximacion-a-traves-de-la-herramienta-chequeo-digital-report-2020-2021/>
- Henriquez, P., & Torrico, B. (2021). *Chequeo digital: ¿Cómo acelerar la transformación digital de las mipyme en América Latina y el Caribe (ALC)?* Banco Interamericano de Desarrollo. <https://doi.org/10.18235/0003607>
- Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2015). *Strategy, not technology, drives digital transformation*. MIT Sloan Management Review & Deloitte University Press. https://shop.sloanreview.mit.edu/store/strategy-not-technology-drives-digital-transformation?utm_source=
- Kraus, S., Clauss, T., Breier, M., Gast, J., Zardini, A., & Tiberius, V. (2020). The economics of COVID-19: Initial empirical evidence on how family firms in five European countries cope with the corona crisis. *International Journal of Entrepreneurial Behavior & Research*, 26(5), 1067–1092. <https://doi.org/10.1108/IJEBR-04-2020-0214>
- Martilla, J. A., & James, J. C. (1977). Importance-performance analysis. *Journal of Marketing*, 41(1), 77–79. <https://doi.org/10.1177/002224297704100112>
- Mettler, T. (2011). Maturity assessment models: A design science research approach. *International Journal of Society Systems Science*, 3(1–2), 81–98. <https://doi.org/10.1504/IJSSS.2011.038934>
- Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing and Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). *Journal of Manufacturing Systems*, 49, 194–214. <https://doi.org/10.1016/j.jmsy.2018.10.005>
- North, K., Aramburu, N., & Lorenzo, O. J. (2020). Promoting digitally enabled growth in SMEs: A framework proposal. *Journal of Enterprise Information Management*, 33(1), 238–262. <https://doi.org/10.1108/JEIM-04-2019-0103>
- OECD/JRC. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
- Orošnjak, M., Kedziora, S., & Desloges, M. (2025). Digital maturity of SMEs in the EU: Leaders and laggards of Luxembourg's manufacturing ecosystem. *Technologies*, 13(12), 541. <https://doi.org/10.3390/technologies13120541>
- Schumacher, A., Erol, S., & Sihni, W. (2016). A maturity model for assessing Industry 4.0 readiness and maturity of manufacturing enterprises. *Procedia CIRP*, 52, 161–166. <https://doi.org/10.1016/j.procir.2016.07.040>
- Teichert, R. (2019). Digital transformation maturity: A systematic review of literature. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 67(6), 1673–1687. <https://doi.org/10.11118/actaun201967061673>
- Thorsen, T., Murawski, M., & Bick, M. (2020). How to measure digitalization? A critical evaluation of digital maturity models. En *Responsible design, implementation and use of information and communication technology* (I3E 2020) (Lecture Notes in Computer Science, Vol. 12066, pp. 358–369). Springer. https://doi.org/10.1007/978-3-030-44999-5_30
- Tiggemann Prando, L., Schmitt, J., de Souza Mendonça, A. K., Rosa, F. S., Lunkes, R. J., Bornia, A. C., & de Andrade, D. F. (2025). Entrepreneurial potential in digital transformation: Scale development using item response theory. *European Journal of Innovation Management*, 28(7), 2784–2814. <https://doi.org/10.1108/EJIM-06-2024-0727>

- Universidad Tecnológica Indoamérica. (s. f.). *Madurez digital — Diagnóstico para emprendimientos* [Plataforma web]. <https://semilla.indoamericattech.com/>
- Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, 122, 889–901. <https://doi.org/10.1016/j.jbusres.2019.09.022>
- Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>
- Williams, C., Krumay, B., Schallmo, D., & Scornavacca, E. (2024). Digital maturity model for SMEs: Validation through a mixed-method approach. *Pacific Asia Journal of the Association for Information Systems*, 16(1), Artículo 2. <https://doi.org/10.17705/1pais.16102>
- World Bank. (2016). *World development report 2016: Digital dividends*. <https://doi.org/10.1596/978-1-4648-0671-1>

Motores y brechas de la transformación digital: Un análisis del uso y adopción de tecnologías en las pequeñas y medianas empresas (pymes) de la ciudad de Guayaquil, Ecuador

Drivers and Gaps of Digital Transformation: An Analysis of Technology Use and Adoption in Small and Medium Enterprises (SMEs) of Guayaquil – Ecuador

Germania Rodríguez-Morales¹ <https://orcid.org/0000-0001-8932-9213>, María Paula Espinosa-Vélez¹ <https://orcid.org/0000-0002-9090-5721>, Carlos Granda Tandazo¹ <https://orcid.org/0000-0002-2111-4862>

¹Universidad Técnica Particular de Loja, Loja, Ecuador
grrodriguez@utpl.edu.ec, mpespinoza@utpl.edu.ec,
cwgranda@utpl.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Enviado: 2026/07/12

Aceptado: 2026/08/27

Publicado: 2026/10/14

Resumen

El presente estudio analiza la dinámica de la transformación digital en 141 pequeñas y medianas empresas (pymes) de Guayaquil, con el propósito de identificar los motores que impulsan este proceso y las brechas tecnológicas que limitan su competitividad. La investigación utiliza un enfoque híbrido fundamentado en un modelo de medición sintético que integra marcos internacionales como el Digital Maturity Model, Chequeo Digital y DIGROW. Los resultados establecen que la dimensión “uso y adopción de tecnologías” constituye el motor fundamental del ecosistema, con un puntaje de 65,20 % y una correlación de $r = 0,83$ con el éxito estratégico general. Este impulso se sustenta en la consolidación de herramientas operativas como el almacenamiento en la nube y la presencia comercial en plataformas digitales. No obstante, se evidencia una brecha crítica en la sofisticación analítica y en la seguridad; la dimensión de “datos, inteligencia artificial y ciberseguridad” presenta el desempeño más bajo, con un puntaje de 53,27 %. Esta limitación se refleja en una baja implementación de inteligencia artificial (36 %) y una mínima inversión formal en ciberseguridad (5 %). Se concluye que el ecosistema requiere superar cuellos de botella estructurales relacionados con la disponibilidad de recursos financieros y la formación técnica especializada. La evidencia sugiere que, para trascender la

Sumario: Introducción, Metodología, Resultados, Discusión, Conclusiones.

Como citar: Rodríguez-Morales, G., Espinosa-Vélez, M. & Granda, C. (2026). Motores y brechas de la transformación digital: Un análisis del uso y adopción de tecnologías en las pequeñas y medianas empresas (pymes) de la ciudad de Guayaquil, Ecuador. *Revista Tecnológica - Espol*, 38(E1), 172-187.
<https://doi.org/10.37815/rte.v38nE1.1606>

optimización operativa, es imperativo orientar las políticas públicas hacia el fortalecimiento de capacidades analíticas y de protección de activos digitales para asegurar la sostenibilidad del crecimiento empresarial en el entorno urbano ecuatoriano.

Palabras clave: adopción tecnológica, madurez digital, transformación digital, pequeñas y medianas empresas, pymes.

Abstract

This study analyzes the dynamics of digital transformation in 141 small and medium-sized enterprises (SMEs) in Guayaquil, with the aim of identifying the drivers behind this process and the technological gaps that limit their competitiveness. The research employs a hybrid approach based on a synthetic measurement model that integrates international frameworks such as the Digital Maturity Model, Chequeo Digital, and DIGROW. The results show that the “Use and Adoption of Technologies” dimension is the fundamental driver of the ecosystem, with a score of 65.20% and a correlation of $r=0.83$ with overall strategic success. This momentum is underpinned by the consolidation of operational tools such as cloud storage and a commercial presence on digital platforms. However, there is a critical gap in analytical sophistication and security; the “Data, Artificial Intelligence, and Cybersecurity” dimension shows the weakest performance at 53.27%. This limitation is reflected in low adoption of artificial intelligence (36%) and minimal formal investment in cybersecurity (5%). It is concluded that the ecosystem must overcome structural bottlenecks related to the availability of financial resources and specialized technical training. The evidence suggests that, to move beyond operational optimization, it is imperative to direct public policies toward strengthening analytical capabilities and the protection of digital assets to ensure the sustainability of business growth in Ecuador’s urban environment.

Keywords: technology adoption, digital maturity, digital transformation, small and medium-sized enterprises, SMEs.

Introducción

Las pequeñas y medianas empresas (pymes) constituyen el 99,54 % del tejido empresarial ecuatoriano y son responsables de generar el 56,18 % del empleo nacional, lo que las convierte en un pilar estratégico para la estabilidad económica y el desarrollo social del país (Instituto Nacional de Estadística y Censos, INEC, 2023). A pesar de su importancia crítica, este segmento enfrenta una brecha de productividad persistente en comparación con las grandes corporaciones, una disparidad que en el contexto de América Latina se atribuye en gran medida al acceso limitado a infraestructuras tecnológicas, a la falta de talento especializado y a las restricciones en el financiamiento para la innovación. En este escenario, la transformación digital ha dejado de ser una tendencia opcional para convertirse en un imperativo de supervivencia y escalabilidad, lo que permite a las organizaciones optimizar sus procesos internos y mejorar su propuesta de valor en mercados cada vez más competitivos y globalizados.

En el ámbito regional, estudios de la Comisión Económica para América Latina y el Caribe (CEPAL, 2021) sugieren que la adopción de tecnologías digitales avanzadas es todavía incipiente en las pequeñas organizaciones y se concentra mayoritariamente en herramientas de comunicación y marketing básico. No obstante, informes recientes, como el sondeo de la adopción digital de Movistar Empresas en 2024, indican una disposición creciente al cambio; el 91 % de las pymes ecuatorianas planeaban invertir en su digitalización durante el año 2024 (Revista Gestión, 2024). Este interés se ve respaldado por marcos normativos nacionales como

la Agenda de Transformación Digital del Ecuador 2022–2025 y la reciente Política Pública para la Transformación Digital 2025–2030, que buscan fomentar un ecosistema inclusivo y seguro. A pesar de estos esfuerzos institucionales, la realidad operativa muestra una dicotomía: mientras algunas áreas funcionales avanzan rápidamente, otras permanecen estancadas por barreras estructurales.

La investigación se enfocó específicamente en Guayaquil, un ecosistema urbano que lidera el crecimiento de usuarios de internet en el país y actúa como un centro neurálgico para el comercio y los servicios. Investigaciones previas a nivel nacional, como el reporte *Chequeo Digital 2022–2023* (Banco Interamericano de Desarrollo y Escuela Superior Politécnica del Litoral, 2024), situaban a casi la mitad de las empresas ecuatorianas en niveles iniciales de madurez digital (47,9 %). Sin embargo, la dinámica pospandemia ha acelerado la integración tecnológica, por lo que exige una actualización de los diagnósticos locales para identificar con precisión qué factores están actuando como motores del progreso y cuáles representan brechas críticas que impiden el salto hacia niveles de madurez avanzados. El propósito de este artículo es analizar el nivel de madurez digital de 141 pymes guayaquileñas, fundamentando el análisis en la relación de dependencia entre el uso efectivo de tecnologías y el éxito estratégico de la organización.

Los resultados obtenidos muestran que el avance digital en la ciudad no es uniforme y se caracteriza por la existencia de motores tecnológicos consolidados. Asimismo, la dimensión “uso y adopción tecnológica” alcanza niveles de madurez establecidos, impulsada primordialmente por el uso extendido del almacenamiento en la nube, los sistemas electrónicos de pago y una sólida presencia en redes sociales. Estas herramientas han permitido a las pymes locales mejorar su visibilidad comercial y su eficiencia operativa inmediata, actuando como la principal palanca que impulsa la madurez general del ecosistema. En este sentido, la tecnología no es solo un soporte, sino el motor que permite a las empresas adaptarse a las demandas de un consumidor urbano altamente conectado.

Por otro lado, el análisis también expone brechas tecnológicas y estratégicas que actúan como cuellos de botella para el escalamiento hacia la innovación disruptiva. La gestión estratégica de datos, la implementación de inteligencia artificial y la ciberseguridad proactiva presentan los desempeños más débiles dentro del modelo de medición utilizado. Aunque existe una percepción positiva sobre el impacto de la tecnología en la productividad, la falta de inversión formal en seguridad digital y la escasa utilización de análisis de datos para la toma de decisiones limitan la capacidad de las empresas para competir en mercados internacionales. Esta brecha analítica sugiere que las organizaciones han superado la alfabetización digital básica, pero enfrentan dificultades para transformar la información en activos estratégicos.

Finalmente, este artículo busca profundizar en la comprensión de estos fenómenos duales mediante la aplicación de un modelo sintético validado que integra marcos internacionales como el DMM de TM Forum (2017), *Chequeo Digital del BID* (2024) y el marco DIGROW de Zubillaga *et al.* (2019). A través de un enfoque híbrido, se evalúan seis dimensiones críticas no solo para determinar el estado actual del ecosistema sino también para proponer rutas de acción que permitan a las pymes de Guayaquil transitar de una digitalización de supervivencia hacia una transformación digital integral y sostenible. Los hallazgos presentados pretenden servir como un insumo estratégico para la academia, el sector público y los tomadores de decisiones empresariales, contribuyendo al fortalecimiento de la competitividad en la economía digital emergente del Ecuador.

Transformación digital y la creación de valor empresarial

La transformación digital se consolida como uno de los principales mecanismos de adaptación y generación de ventajas competitivas en entornos empresariales ante la creciente volatilidad tecnológica y la acelerada evolución de los mercados. Aunque inicialmente se entendía como la incorporación de tecnologías digitales en procesos operativos, la literatura contemporánea reconoce que implica la reconfiguración de modelos de negocio, estructuras organizacionales, procesos de gestión y relaciones con los grupos de interés. Por ello, la transformación digital no debe interpretarse exclusivamente como un proceso tecnológico, sino como una estrategia de cambio organizacional orientada a la creación y captura de valor (Vial, 2019; Verhoef *et al.*, 2021).

En este contexto, el concepto de valor se sitúa en el centro de la transformación digital. La adopción de tecnologías digitales aporta beneficios a las organizaciones siempre que contribuya a mejorar la eficiencia operativa, fortalecer la experiencia del cliente, incrementar la capacidad innovadora o desarrollar nuevas fuentes de ingresos. Bharadwaj *et al.* (2013) sostienen que las tecnologías digitales han dejado de desempeñar un papel exclusivamente funcional para convertirse en componentes esenciales de la estrategia empresarial, influyendo directamente en la forma en que las organizaciones compiten y generan ventajas sostenibles. Así, Vial (2019) argumenta que la transformación digital produce valor cuando las empresas logran combinar los recursos tecnológicos con capacidades organizacionales que permitan modificar procesos, productos y mecanismos de interacción con los clientes.

Entre los principales beneficios asociados a la transformación digital pueden citarse: primero, la mejora de la eficiencia operativa. Las tecnologías digitales facilitan la automatización de actividades, la integración de procesos y la gestión de información en tiempo real. Según la Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2025), las empresas que incorporan tecnologías digitales en sus operaciones presentan mayores oportunidades para incrementar la productividad, mejorar la coordinación interna y acceder a nuevos mercados. Segundo, el fortalecimiento de las relaciones con los clientes. Las plataformas digitales, el comercio electrónico y las tecnologías de interacción remota han transformado la manera en que las organizaciones diseñan y entregan sus propuestas de valor. Verhoef *et al.* (2021) señalan que uno de los principales resultados de la transformación digital es la posibilidad de ofrecer experiencias más personalizadas, mejorar la comunicación con los consumidores y desarrollar modelos de negocio centrados en el cliente. Tercero, la relación entre la transformación digital y la innovación empresarial, pues la digitalización proporciona nuevas oportunidades para desarrollar productos, servicios y modelos de negocio que anteriormente resultaban inviables o demasiado costosos. Kahveci (2025) destaca que la transformación digital en las pequeñas y medianas empresas debe entenderse como un proceso continuo de aprendizaje y mejora organizacional que fortalece la capacidad de adaptación, innovación y competitividad. De igual manera, Kumar *et al.* (2025) identifican la innovación organizacional como uno de los principales mecanismos mediante los cuales la transformación digital contribuye al desempeño empresarial.

Sin embargo, la evidencia empírica muestra que la incorporación de tecnología por sí sola no garantiza la generación de valor. Bongomin *et al.* (2024), en una revisión sistemática sobre el éxito de la transformación digital en las pymes, identifican que las organizaciones que obtienen mejores resultados son aquellas capaces de integrar tecnología, recursos humanos y estrategia empresarial dentro de una visión compartida de cambio organizacional. En consecuencia, el reto principal no reside únicamente en adquirir tecnología, sino en desarrollar las condiciones necesarias para aprovecharla de forma efectiva. No obstante, es necesario considerar que la generación de valor depende de factores contextuales que varían según el tamaño de la

empresa, el sector económico y las características del entorno. Omol *et al.* (2025) señalan que la madurez digital empresarial está influenciada por factores tecnológicos, organizacionales y ambientales que interactúan de manera compleja, afectando la capacidad de las empresas para avanzar hacia niveles superiores de digitalización. En los últimos años, además, la creación de valor empresarial ha comenzado a vincularse estrechamente con la capacidad para gestionar y explotar estratégicamente los datos, dada la creciente disponibilidad de información procedente de clientes, operaciones y mercados.

En este contexto, de acuerdo con estudios recientes, muchas pymes permanecen concentradas en etapas iniciales de digitalización, caracterizadas por el uso de tecnologías orientadas principalmente a la comunicación, el marketing digital y la gestión administrativa. Mick *et al.* (2024) sostienen que la ausencia de rutas claras de transformación digital y la limitada disponibilidad de recursos dificultan la transición hacia modelos basados en innovación digital, sostenibilidad y aprovechamiento de datos. Por tanto, la transformación digital puede entenderse como un proceso progresivo de construcción de capacidades orientadas a incrementar el valor generado por la organización. Analizar cómo las pequeñas y medianas empresas generan valor a partir de la digitalización resulta particularmente relevante para comprender los factores que impulsan o limitan su competitividad y para identificar los motores y brechas que condicionan su evolución en la economía digital contemporánea (Verhoef *et al.*, 2021; Kahveci, 2025; Kumar *et al.*, 2025).

Factores habilitadores de la adopción tecnológica en las pymes

La adopción tecnológica constituye uno de los pilares de la transformación digital empresarial, pero, como se ha mencionado, su desarrollo depende de una combinación de factores organizacionales, humanos y contextuales. Por tanto, las diferencias en los niveles de digitalización entre empresas no pueden explicarse únicamente por la disponibilidad de tecnología, sino por la capacidad de las empresas para disponer de recursos, desarrollar competencias y adaptarse a su entorno. En este sentido, la transformación digital de las pymes es el resultado de la interacción entre las características de la organización, las oportunidades tecnológicas y las condiciones del ecosistema en el que operan (Omol *et al.*, 2025).

Entre los factores habilitadores de la transformación digital, se distinguen los internos y los externos. En el ámbito interno, el liderazgo directivo emerge como uno de los principales habilitadores. Kahveci (2025) destaca que la conciencia digital, el compromiso de la dirección y la existencia de una hoja de ruta clara favorecen significativamente la incorporación de tecnologías y la sostenibilidad de los procesos de transformación digital. Las competencias digitales del talento humano constituyen otro elemento fundamental, pues la digitalización requiere colaboradores capaces de utilizar herramientas tecnológicas, gestionar información y adaptarse a nuevas formas de trabajo. Kumar *et al.* (2025) identifican las capacidades tecnológicas y el conocimiento digital como factores estrechamente vinculados con la innovación y el desempeño empresarial, mientras que Bongomin *et al.* (2024) señalan que el capital humano digital es uno de los factores de éxito más recurrentes en los procesos de transformación digital.

La disponibilidad de recursos financieros también desempeña un papel decisivo. Aunque muchas tecnologías digitales se han vuelto más accesibles, las inversiones relacionadas con inteligencia artificial, analítica avanzada o ciberseguridad siguen representando un desafío para numerosas organizaciones. La Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2025) indica que las restricciones presupuestarias continúan limitando la capacidad de las pymes para ejecutar proyectos de digitalización de mayor alcance.

Entre los factores externos se destacan la calidad de la infraestructura digital, el acceso a la conectividad, la disponibilidad de proveedores tecnológicos y el apoyo institucional. Diversos estudios muestran que las empresas insertas en ecosistemas digitales más desarrollados presentan mayores niveles de innovación y competitividad, lo que demuestra la importancia de las políticas públicas y de los programas de fortalecimiento empresarial para impulsar la digitalización de las pymes (OCDE, 2025; Omol *et al.*, 2025).

Finalmente, los estudios insisten en que la adopción tecnológica exitosa depende de la capacidad de las empresas para convertir la tecnología en valor organizacional. Desde esta perspectiva, no basta con incorporar herramientas digitales; es necesario desarrollar capacidades que permitan integrarlas en la estrategia, los procesos y la toma de decisiones.

Esta tendencia también se observa en el ecosistema empresarial ecuatoriano. El informe GEM Ecuador 2025 evidencia una creciente orientación hacia el uso estratégico de tecnologías emergentes, pues informa que cerca del 47 % de los emprendedores tempranos ya utiliza o prevé utilizar herramientas de inteligencia artificial en sus negocios. Este resultado sugiere una transición progresiva desde una adopción tecnológica centrada en la eficiencia operativa hacia enfoques más orientados a la innovación, la productividad y la generación de ventajas competitivas sostenibles (Amaya *et al.*, 2026).

Impulsores y brechas de la madurez digital

Tal como se ha señalado previamente, la transformación digital de las pymes no responde a una trayectoria lineal, sino a un proceso condicionado por la interacción entre factores impulsores y limitaciones estructurales.

Entre los principales impulsores identificados se destacan la adopción de tecnologías digitales, el liderazgo estratégico, la orientación a la innovación y el fortalecimiento de las competencias digitales. Las organizaciones que incorporan herramientas digitales en sus procesos, que desarrollan capacidades tecnológicas internas y que promueven una cultura de aprendizaje continuo suelen alcanzar mayores niveles de madurez y competitividad. Asimismo, la creciente disponibilidad de soluciones basadas en la nube, comercio electrónico, automatización y analítica facilita que las pymes mejoren su eficiencia operativa y amplíen sus oportunidades de crecimiento.

No obstante, el avance hacia niveles superiores de madurez digital continúa enfrentando importantes brechas. Las limitaciones financieras, la escasez de talento especializado, la insuficiente capacitación del personal y la falta de planificación estratégica son algunas de las barreras más frecuentemente reportadas por las pequeñas y medianas empresas. A ello se suman dificultades asociadas a la resistencia al cambio, la baja capacidad de absorción tecnológica y la limitada disponibilidad de recursos para implementar proyectos digitales de mayor complejidad. Estas condiciones explican por qué muchas organizaciones logran avances en herramientas de comunicación y gestión operativa, pero encuentran obstáculos para desarrollar capacidades más sofisticadas relacionadas con innovación, analítica de datos e inteligencia artificial.

La evidencia internacional también muestra que las brechas de madurez digital no se concentran únicamente en el acceso a tecnología, sino en la capacidad de transformar dicha tecnología en ventajas competitivas sostenibles. Marcos *et al.* (2024), en una revisión de modelos de madurez digital para pymes, identifican que dimensiones como estrategia, conocimiento, procesos, cultura e infraestructura tecnológica son determinantes para el avance digital empresarial.

En consecuencia, la madurez digital puede entenderse como el resultado del equilibrio entre motores que impulsan la adopción tecnológica y brechas que limitan su consolidación como capacidad estratégica. Analizar esta dualidad resulta especialmente relevante en economías emergentes, donde las empresas suelen avanzar con rapidez en la incorporación de herramientas digitales básicas, pero enfrentan mayores dificultades para desarrollar capacidades analíticas, de innovación basada en datos y mecanismos de seguridad digital.

Metodología

La investigación adoptó un enfoque híbrido que integró una fase inicial cuantitativa de carácter exploratorio para el levantamiento de datos, seguida de un proceso de análisis clasificatorio, estratificado y correlacional. El diseño se fundamentó en la necesidad de cerrar el vacío en la operacionalización de niveles de madurez digital para las pymes locales, estableciendo una base objetiva, comparable y replicable para futuras mediciones longitudinales.

Población y muestra

La población objetivo comprendió las pequeñas y medianas empresas (pymes) de la ciudad de Guayaquil, Ecuador. La estratificación se realizó siguiendo los criterios de la Decisión 702 de la Comunidad Andina, que define una estratificación de las empresas según el personal ocupado y su valor bruto de ventas anuales (Secretaría General de la Comunidad Andina, 2009).

Se empleó un muestreo no probabilístico por conveniencia, seleccionando un total de 141 pymes registradas en las bases de datos de la Cámara de Innovación y Tecnología del Ecuador (CITEC) y la Universidad Técnica Particular de Loja (UTPL). Este método se justificó por la complejidad y el alto costo de obtener un marco muestral completo en entornos urbanos emergentes, lo que permitió acceder rápidamente a tomadores de decisiones estratégicas.

Fundamento conceptual del modelo de medición

El núcleo metodológico consistió en el desarrollo y aplicación de un modelo sintético de medición de la madurez digital, construido a partir de la integración de tres marcos internacionales de referencia:

- Digital Maturity Model (DMM) de TM Forum (2017): aportó la visión holística para evaluar cómo la tecnología respalda la estrategia y la cultura organizacional.
- Chequeo Digital del BID (2024): proporcionó los parámetros de adopción tecnológica y habilidades digitales adaptados al contexto de las mipymes en América Latina.
- Marco DIGROW de Zubillaga *et al.* (2019): enfatizó la medición del crecimiento digital vinculado a la generación de valor, el aumento de ingresos y la reducción de costes.

Dimensiones y subatributos

Para garantizar una evaluación integral, el modelo se estructuró en seis dimensiones críticas, cada una con una ponderación determinada y compuesta por subatributos específicos recolectados mediante un cuestionario de 23 preguntas:

- Visión y estrategia digital (15 %): Mide la formalización de la estrategia y la sustentación del modelo de negocio en lo digital.
- Uso y adopción de tecnologías (25 %): Evalúa el uso de herramientas operativas como ERP, CRM, almacenamiento en la nube, pagos digitales y tecnologías para la exportación.

- Datos, IA y ciberseguridad (20 %): Analiza la gestión analítica de datos, el uso incipiente de inteligencia artificial y la implementación de protocolos de seguridad digital.
- Digitalización por área (15 %): Mide la integración transversal de la tecnología en producción, administración, RR. HH., logística y finanzas.
- Competencias digitales (15 %): Evalúa el nivel de manejo tecnológico de los empleados y la existencia de programas de capacitación formal.
- Impacto de la digitalización (10 %): Registra la percepción de mejora en ventas y productividad tras la implementación tecnológica.

Tratamiento de datos y operacionalización

El procesamiento de la información se dividió en cuatro fases técnicas:

1. Codificación: las respuestas cualitativas se transformaron en valores numéricos. Por ejemplo, en escalas de Likert de seis puntos, se asignaron valores desde 0 (“totalmente en desacuerdo”) hasta 100 (“totalmente de acuerdo”).
2. Normalización: los puntajes de cada pregunta se normalizaron a una escala de 0 a 100 para garantizar la comparabilidad entre dimensiones con diferente número de reactivos.
3. Cálculo dimensional: se obtuvo el puntaje promedio normalizado para cada una de las seis dimensiones.
4. Índice de madurez global: el puntaje final se calculó mediante una suma ponderada. Esta ponderación no fue arbitraria, sino que se basó en el análisis de correlación de Pearson; en esta ponderación “uso y adopción de tecnologías” recibió el mayor peso (25 %) por mostrar la correlación más alta con el éxito general del ecosistema ($r = 0,85$).

Definición de niveles de madurez

Se adoptó el método de umbrales equidistantes para clasificar a las empresas en cinco niveles, dividiendo la escala de 0 a 100 en intervalos de 20 puntos porcentuales:

- Nivel 1 (básico): 0 %–19,99 %. Uso aislado y reactivo de tecnología.
- Nivel 2 (emergente): 20 %–39,99 %. Presencia digital básica sin integración.
- Nivel 3 (intermedio): 40 %–59,99 %. Estandarización de procesos y estrategia documentada.
- Nivel 4 (establecido): 60 %–79,99 %. Alineación estratégica y uso de sistemas avanzados (ERP/CRM).
- Nivel 5 (avanzado): 80 %–100 %. Transformación digital como parte de la estructura organizacional e innovación constante.

Instrumento y recolección

El instrumento fue automatizado mediante ArcGIS Survey123¹, lo que permitió una recolección ágil y digital. El proceso de levantamiento se extendió por **24 semanas** (desde finales de 2024 hasta inicios de 2025), combinando entrevistas presenciales, encuestas telefónicas y reuniones virtuales vía Zoom. Para mitigar sesgos de percepción propios de datos autoinformados, los resultados fueron contrastados con *benchmarks* regionales del BID (2024) y la OCDE (2021).

1 <https://arcgis.com/survey123/>

Resultados

La investigación realizada sobre las pequeñas y medianas empresas de Guayaquil permite identificar una estructura de madurez digital que, aunque superior a los promedios nacionales registrados en años anteriores, presenta una marcada dicotomía entre la eficiencia operativa y la sofisticación estratégica. Los hallazgos se presentan a continuación, estructurados en torno a la caracterización del ecosistema, la distribución de la madurez y la identificación detallada de los motores y las brechas tecnológicas.

Caracterización del ecosistema

El estudio se fundamentó en una muestra de 141 pymes cuya distribución sectorial refleja la naturaleza comercial y de servicios de la economía de Guayaquil. El sector de servicios constituye el 38,3 % de la muestra, seguido por el comercial con un 36,1 % y el manufacturero con un 15,6 %. Por su parte, el sector tecnológico mantiene una representación incipiente del 4,2 % y otros sectores representan un 5,6 %, lo cual guarda relación con las tendencias nacionales de emprendimiento en las que los negocios de base tecnológica todavía se encuentran en etapas de desarrollo temprano.

Para comprender el entorno digital de las pymes encuestadas, debe considerarse que el 93 % de las empresas declara que el acceso a internet es un requisito indispensable para la ejecución de sus operaciones diarias. Esta dependencia de la conectividad se complementa con la institucionalización de la gestión tecnológica, dado que el 79 % de las organizaciones cuenta con un área o personal dedicado específicamente al manejo de tecnología, programas, plataformas y redes sociales. No obstante, la proyección hacia mercados externos sigue siendo limitada, pues únicamente el 23 % de las empresas declara realizar actividades de exportación.

Distribución por niveles de madurez digital

La aplicación del modelo de medición permitió establecer la distribución de las pymes por nivel de madurez, como se muestra en la Tabla 1.

Tabla 1
Distribución de pymes por nivel de madurez

Nivel de madurez	Número de pymes	Porcentaje de la muestra
Básico	4	2,84 %
Emergente	22	15,60 %
Intermedio	44	31,21 %
Establecido	53	37,59 %
Avanzado	18	12,77 %

Se destaca que el 81,57 % de las empresas se sitúa en niveles intermedios o superiores de madurez digital. Específicamente, el 37,59 % se clasifica en el nivel establecido, lo que indica que estas han logrado alinear su estrategia digital con los objetivos de negocio y utilizan sistemas de gestión estandarizados. El 31,21 % se ubica en el nivel intermedio, mientras que un segmento del 12,77 % ha alcanzado el nivel avanzado, en el que la transformación digital es considerada una capacidad intrínseca de la organización.

En contraste, los niveles inferiores presentan una incidencia reducida: el 15,60 % de la muestra se encuentra en el nivel emergente y solo el 2,84 % permanece en el nivel básico. Esta baja representación en el nivel más elemental es un indicador positivo, ya que contrasta con informes nacionales previos en los que casi la mitad de las empresas ecuatorianas se situaban en etapas iniciales de digitalización.

Los motores de la transformación digital en Guayaquil

Los resultados mostraron dos dimensiones que destacan por su nivel de madurez y, por tanto, se consideran motores fundamentales para la transformación digital de las pymes guayaquileñas: “uso y adopción de tecnologías”, que obtuvo el puntaje más alto del modelo con un 65,20 %, e “impacto de la digitalización”, que obtuvo un 64,66 % (ver Tabla 2).

Tabla 2
Nivel de madurez promedio por dimensión

Dimensión	Nivel promedio	Puntaje promedio	Correlación con nivel general
Visión y estrategia digital	3 (Intermedio)	53,55 %	0,71
Uso y adopción de tecnologías	4 (Establecido)	65,20 %	0,83
Datos, IA y ciberseguridad	3 (Intermedio)	53,27 %	0,85
Digitalización por área	3 (Intermedio)	56,10%	0,62
Competencias digitales	3 (Intermedio)	55,09%	0,75
Impacto de la digitalización	4 (Establecido)	64,66%	0,70

Los datos de la encuesta muestran una adopción extendida en el uso de tecnologías de operatividad básica. El almacenamiento en la nube es la herramienta con mayor penetración, utilizada por el 80 % de las organizaciones. Le siguen en nivel de adopción la disponibilidad de un sitio web propio con un 79 %, el uso de medios digitales para la promoción de productos con un 77 % y la implementación de sistemas de respaldo o recuperación de datos con un 73 %. En el ámbito comercial, el 70 % de las empresas emplea activamente las redes sociales o canales de comercio electrónico para la venta en línea, mientras que el 67 % ofrece medios de pago digitales a sus clientes.

Este despliegue tecnológico ha generado un impacto positivo percibido que se refleja en que el 67 % de los directivos encuestados afirma haber observado una mejora tangible en sus niveles de venta tras la implementación de estas herramientas. Asimismo, el 77 % reconoce que el uso de tecnología digital influye en gran medida en la productividad de su organización.

Identificación de las brechas tecnológicas y estratégicas

A pesar de la fortaleza en la adopción operativa, el estudio identifica brechas críticas en la analítica y la seguridad digital. La dimensión “datos, IA y ciberseguridad” se presenta como el cuello de botella principal del ecosistema, con el desempeño más bajo, un 53,27 %. Este hallazgo resulta particularmente relevante puesto que esta dimensión posee la correlación más alta con el éxito general del modelo ($r = 0,85$), lo que indica que es la palanca con mayor potencial para impulsar a las empresas hacia niveles avanzados, pero es precisamente donde menos progreso se registra.

Las respuestas evidencian estas brechas mediante bajos niveles de adopción de innovaciones disruptivas. Solo el 36 % de las pequeñas y medianas empresas han implementado herramientas de inteligencia artificial para mejorar la eficiencia de sus procesos. Existe también un rezago en el uso de herramientas de interacción avanzada: únicamente el 30 % emplea *chatbots* o asistentes virtuales. La gestión de la información muestra limitaciones similares, ya que solo el 37 % de las organizaciones realiza análisis de búsquedas de clientes potenciales en internet para orientar su estrategia comercial.

La brecha de seguridad digital constituye otro factor de vulnerabilidad. Aunque el 19 % de las empresas declara haber sufrido ataques o eventos de inseguridad en sus sistemas digitales durante los últimos tres años, solo el 5 % de la muestra realiza inversiones especializadas y formales en ciberseguridad proactiva. La mayoría de las organizaciones se limita a protocolos básicos de respaldo sin contar con procedimientos integrales de respuesta ante incidentes cibernéticos.

Factores determinantes y barreras estructurales

El análisis de los desafíos permite comprender las causas subyacentes de estas brechas tecnológicas. La disponibilidad de recursos económicos es señalada por el 33 % de los empresarios como el principal obstáculo para avanzar en su transformación digital. Esta limitación financiera es coherente con el hallazgo de que el 57 % de las organizaciones destina menos del 5 % de sus ingresos anuales a la digitalización, mientras que solo el 4 % realiza inversiones estratégicas superiores al 20 % de sus ingresos.

Adicionalmente, persiste una brecha de capacidades técnicas. El 22 % de las empresas identifica la falta de conocimiento técnico especializado como el segundo desafío más importante para su progreso. Aunque el 59 % de los directivos percibe que su personal tiene un nivel alto de manejo tecnológico para tareas cotidianas, esta percepción se contradice con la dificultad reportada para liderar procesos de innovación avanzada. Solo el 53 % de las empresas participó en actividades de capacitación en competencias digitales durante el último año, lo que deja a una parte considerable del talento humano sin actualización en áreas críticas como la inteligencia artificial o la ciberseguridad.

Heterogeneidad de la madurez por sector económico

El avance digital no es uniforme entre las diferentes actividades económicas de la ciudad. El sector de tecnología lidera el ecosistema con un puntaje de 70,10 %, situándose firmemente en el nivel establecido. La industria manufacturera le sigue con un 62,48 %, alcanzando también el nivel establecido, impulsada principalmente por inversiones estratégicas en sistemas administrativos complejos como ERP o CRM. No obstante, este sector enfrenta el reto de digitalizar plenamente sus áreas de producción para cerrar la brecha con sus procesos administrativos.

Los sectores de servicios y comercio, que constituyen el núcleo mayoritario de la muestra, se encuentran ubicados en el nivel intermedio con puntajes de 56,91 % y 56,21 %, respectivamente (ver Tabla 3). Estas organizaciones muestran una digitalización orientada fundamentalmente hacia el marketing digital y la interfaz con el cliente, pero carecen de la profundidad analítica y estratégica necesaria para alcanzar el nivel avanzado. Estadísticamente, las diferencias de madurez entre sectores resultaron significativas, lo que subraya la necesidad de políticas de fomento diferenciadas según la naturaleza de la actividad económica.

Tabla 3
Nivel de madurez promedio por sector económico

Sector	Número de pymes	Porcentaje de la muestra	Puntaje promedio	Nivel promedio
Servicios	54	38,3 %	56,9 %	3 (Intermedio)
Comercio	51	36,1 %	56,9 %	3 (Intermedio)
Industrias manufactureras	22	15,6 %	62,4 %	4 (Establecido)
Otros	8	5,6 %	61,3 %	4 (Establecido)
Tecnología	6	4,2 %	70,1 %	4 (Establecido)

En conclusión, los resultados muestran que el motor de la transformación digital en las empresas de Guayaquil es la adopción de herramientas operativas que mejoran la eficiencia funcional y la visibilidad comercial. Sin embargo, el ecosistema se encuentra en un nivel de madurez establecido que no logra alcanzar niveles transformadores debido a las brechas estructurales en la gestión estratégica de datos, la inversión limitada en seguridad digital y la escasez de talento especializado en tecnologías emergentes.

Discusión

Los hallazgos del presente estudio exponen una dualidad estructural en el ecosistema digital de las pequeñas y medianas empresas de Guayaquil, caracterizada por un avance operativo significativo que, no obstante, se encuentra limitado por brechas profundas en la gestión estratégica de datos y la seguridad digital. Al contrastar estos resultados con mediciones nacionales previas, como el reporte de BID & ESPOL (2024), se observa un salto evolutivo notable: mientras que en aquel periodo casi la mitad de las empresas ecuatorianas se ubicaban en un nivel inicial, la muestra actual de Guayaquil sitúa al 81,57 % de las organizaciones en niveles intermedios o superiores. Este progreso sugiere que el tejido empresarial urbano ha logrado superar la fase de supervivencia digital básica para adentrarse en una etapa de optimización funcional.

Los resultados permiten además interpretar la transformación digital como la coexistencia de motores y brechas dentro de un mismo ecosistema empresarial. Asimismo, se detecta una paradoja en las competencias digitales que merece un análisis profundo. Existe una contradicción entre la percepción de los directivos, según la cual el 59 % califica el manejo tecnológico de su personal como alto, y la realidad de los desafíos reportados, entre los que la falta de conocimiento técnico especializado se sitúa como la segunda barrera más importante para la digitalización (22 %). Esta discrepancia sugiere que las empresas cuentan con talento humano capacitado para la operatividad diaria, pero carecen de expertos capaces de liderar procesos de innovación disruptiva o analítica avanzada.

Desde una perspectiva sectorial, la heterogeneidad del progreso es evidente. Los sectores de tecnología y manufactura lideran la madurez digital con puntajes superiores al 62 %, lo cual se atribuye a una necesidad intrínseca de eficiencia industrial y procesos estandarizados. No obstante, los sectores mayoritarios de servicios y comercio permanecen anclados en una

madurez intermedia, centrada primordialmente en la interfaz con el cliente y el marketing digital. Esta fragmentación subraya la necesidad de políticas públicas diferenciadas, alineadas con la Agenda de Transformación Digital 2025–2030, que prioricen no solo el acceso a financiamiento —identificado como el obstáculo principal por el 33 % de los empresarios—, sino también la formación técnica avanzada y la cultura de seguridad digital. En definitiva, para que las empresas de Guayaquil alcancen el nivel avanzado, el ecosistema debe evolucionar de una adopción tecnológica reactiva hacia una gestión proactiva basada en datos y protección de activos digitales.

Un aspecto relevante es que los resultados reflejan una diferencia entre adoptar tecnología y desarrollar capacidades digitales estratégicas. Aunque las empresas muestran niveles elevados de utilización de herramientas digitales, el menor desempeño observado en la gestión de datos, la inteligencia artificial y la ciberseguridad indica que la digitalización se concentra principalmente en la operación diaria. Esta situación coincide con lo planteado por Bongomin *et al.* (2024), quienes sostienen que muchas pymes logran avances significativos en digitalización funcional, pero encuentran dificultades para transformar la tecnología en capacidades organizacionales capaces de generar innovación y ventajas competitivas sostenibles. Por tanto, el reto para las empresas analizadas no parece estar en el acceso a la tecnología, sino en su capacidad para integrarla dentro de procesos estratégicos basados en información, análisis y generación de conocimiento.

Los resultados también encuentran eco en los hallazgos recientes del *Global Entrepreneurship Monitor Ecuador 2025*. Dicho informe evidencia una creciente predisposición del ecosistema emprendedor hacia el uso de tecnologías avanzadas, pues cerca del 47 % de los emprendedores tempranos ya utiliza o prevé utilizar herramientas de inteligencia artificial en sus actividades empresariales (Amaya *et al.*, 2026). Sin embargo, la menor adopción observada en las pymes consolidadas de Guayaquil sugiere que existe una brecha entre la intención de innovación tecnológica y su implementación estratégica dentro de las organizaciones (Armijos *et al.*, 2024). Esta diferencia podría explicarse por las restricciones de financiamiento, talento especializado y capacidades organizacionales identificadas en el presente estudio.

Limitaciones del estudio

A pesar del valor de los hallazgos para establecer una línea base en el ecosistema local, este estudio presenta limitaciones que condicionan la interpretación de los resultados. En primer lugar, el empleo de un muestreo no probabilístico por conveniencia y el enfoque exclusivo en el área urbana de Guayaquil restringen la generalización de las conclusiones a nivel nacional o a entornos rurales del Ecuador. Asimismo, la dependencia de datos autoinformados por directivos y responsables técnicos introduce un componente de subjetividad y posibles sesgos de percepción o autoselección respecto a la madurez digital real de las organizaciones analizadas.

Por otro lado, se debe considerar que el reducido tamaño de la muestra en sectores específicos, como el de tecnología (aproximadamente un 4,32 % del total), limita la potencia estadística de los análisis comparativos intersectoriales, lo que requiere cautela al afirmar liderazgos de nicho. Finalmente, aunque el modelo sintético propuesto integra marcos de referencia internacionales, esta fase de la investigación no incluye aún pruebas de consistencia interna, como el alfa de Cronbach, o análisis factoriales confirmatorios, tareas que se reservan para fases posteriores del estudio para reforzar la validez del instrumento.

Conclusiones

La investigación permite concluir que el ecosistema de las pequeñas y medianas empresas de Guayaquil ha experimentado un avance evolutivo significativo, situando al 81,57 % de las organizaciones en los niveles intermedio, establecido y avanzado de madurez digital. Este hallazgo es fundamental, ya que indica que el tejido empresarial urbano ha logrado superar la fase de alfabetización digital inicial para enfocarse en la estandarización de procesos operativos. No obstante, la reducida representación en el nivel avanzado, que apenas alcanza el 12,77 %, subraya que la transformación digital aún no se ha consolidado como una capacidad estratégica intrínseca para la mayoría de las entidades locales.

El motor fundamental que impulsa este progreso es la dimensión “uso y adopción de tecnologías”, que alcanzó el puntaje más alto del modelo con un 65,20 % y una correlación alta con el índice general de madurez ($r = 0,83$). Las empresas han integrado con éxito herramientas de operatividad funcional y presencia en línea, destacando el almacenamiento en la nube (80 %), la disponibilidad de un sitio web propio (79 %) y el uso de medios digitales para la promoción comercial (77 %). Este despliegue tecnológico se traduce en un impacto positivo percibido en la productividad y las ventas por parte del 67 % de los directivos, lo que indica que la digitalización actual cumple un rol efectivo de optimización operativa.

En contraposición, la dimensión “datos, IA y ciberseguridad” se identifica como la principal brecha tecnológica y el punto crítico del ecosistema. A pesar de presentar la correlación más alta con el éxito del modelo general ($r = 0,85$), esta dimensión presenta el desempeño más débil con un 53,27 %. Los datos revelan rezagos estructurales en la sofisticación analítica, con una adopción de inteligencia artificial limitada al 36 % y una alarmante falta de inversión formal en ciberseguridad especializada, registrada en apenas el 5 % de la muestra. Esta asimetría indica que, aunque las organizaciones son eficientes en la adquisición de tecnología funcional, presentan deficiencias en la extracción de valor estratégico a partir de los datos.

Finalmente, el estudio concluye que persisten barreras estructurales vinculadas principalmente a la falta de recursos económicos (33 %) y a la escasez de conocimiento técnico especializado (22 %). Se detecta una paradoja en el recurso humano: mientras los directivos perciben niveles de competencia aceptables para las tareas cotidianas, la falta de talento calificado se mantiene como un obstáculo recurrente para liderar procesos de innovación avanzada. Estos hallazgos ratifican la urgencia de alinear las estrategias empresariales con la Política Pública para la Transformación Digital del Ecuador 2025–2030, priorizando el financiamiento estratégico, la formación técnica y la seguridad digital para transitar de una digitalización de supervivencia hacia una ventaja competitiva sostenible.

Reconocimientos y Declaraciones

Este trabajo ha sido financiado parcialmente por la UTPL y la Oficina de Transformación Digital de Guayaquil, proyecto “Observatorio de transformación digital y estudio de adopción de la transformación digital de las empresas de Guayaquil”, código PROY-CONS-CCE-2024-4558.

Las contribuciones de los autores del artículo se distribuyen principalmente, aunque no exclusivamente: Autor 1: administración del proyecto, adquisición de fondos, concepción, metodología, curación de datos, análisis formal, escritura y redacción. Autor 2: conceptualización, escritura y redacción. Autor 3: conceptualización, metodología y análisis formal.

Los autores declaran que, en la elaboración del presente artículo, se ha utilizado la herramienta de IA NoteBookLM, de Google, para revisar la redacción y las posibles redundancias en el contenido, así como para revisar las citas bibliográficas.

Referencias

- Amaya, A., Barriga, A., Espinoza, D., Coronel, K., Fernández, M. F., Espinosa, M. P., & Armas, R. (2026). *Global Entrepreneurship Monitor Ecuador 2025*. Ediloja. https://www.utpl.edu.ec/carreras/sites/default/files/Proyecto_GEM_abril_2026%20%281%29_compressed.pdf
- Armijos, V., Espinosa, M. P., & Rodríguez-Morales, G. (2024). *Level of digital maturity in micro, small and medium-sized companies in southern Ecuador: Strategies for strengthening*. En *Conferencia Iberoamericana de Complejidad, Informática y Cibernética, (CICIC 2024)*. <https://www.iiis.org/CDs2024/CD2024Spring/papers/CB714CK.pdf>
- Banco Interamericano de Desarrollo, & Escuela Superior Politécnica del Litoral. (2024). *Reporte Chequeo Digital 2022–2023*. Observatorio Ecuador Digital. <https://observatorioecuadordigital.mintel.gob.ec/wp-content/uploads/2024/09/CHEQUEO-DIGITAL.pdf>
- ESPOL. (2024). *Reporte Chequeo Digital 2022–2023*. Observatorio Ecuador Digital. <https://observatorioecuadordigital.mintel.gob.ec/wp-content/uploads/2024/09/CHEQUEO-DIGITAL.pdf>
- Bongomin, O., Ocen, G. G., Taremwa, G. K., Choe, J. Y., & Adewale, A. A. (2024). Toward SMEs digital transformation success: A systematic literature review. *Electronic Markets*, 34, Article 55. <https://doi.org/10.1007/s12525-024-00705-8>
- Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital business strategy: Toward a next generation of insights. *Information Systems Research*, 24(2), 471–482. <https://doi.org/10.1287/isre.1120.0344>
- Comisión Económica para América Latina y el Caribe. (2021). *Datos y hechos sobre la transformación digital*. https://www.cepal.org/sites/default/files/publication/files/46766/S2000991_es.pdf
- Instituto Nacional de Estadística y Censos. (2023). *Principales resultados del Registro Estadístico de Empresas: REEM 2023* [Reporte estadístico]. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Registro_Empresas_Establecimientos/2024/Semestre_I/Principales_Resultados_REEM_2023.pdf
- Kahveci, E. (2025). Digital transformation in SMEs: Enablers, interconnections, and a framework for sustainable competitive advantage. *Administrative Sciences*, 15(3), Article 107. <https://doi.org/10.3390/admsci15030107> [mdpi.com]
- Kumar, J., Rani, V., Rani, M., & Rani, G. (2025). Digital transformation in small and medium-sized enterprises: Key insights for entrepreneurs and policymakers to cultivate a future-ready business. *Digital Transformation and Society*, 5(2), 217–234. <https://doi.org/10.1108/DTS-10-2024-0192>
- Lyons, D. (2009, June 15). Don't 'iTune' us: It's geeks versus writers. Guess who's winning. *Newsweek*, 153(24), 27.
- Marcos, S., Reis, J. L., Pereira, R. H., & A. P., I. (2024). Digital maturity model for industries SMEs: A systematic literature review. En CAPSI 2024 Proceedings. Association for Information Systems Electronic Library (AISeL). <https://aisel.aisnet.org/capsi2024/5/>
- Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2025). *Política Pública para la Transformación Digital del Ecuador 2025–2030* [Documento de política]. https://www.gobiernoelectronico.gob.ec/wp-content/uploads/2025/03/INSTRUMENTO-Politica-Publica-para-la-Transformacion-Digital-Ecuador-2025-2030-MINTEL-signed_f.pdf
- Mick, M. M. A. P., Kovalski, J. L., & Chiroli, D. M. G. (2024). Sustainable digital transformation roadmaps for SMEs: A systematic literature review. *Sustainability*, 16(19), Article 8551. <https://doi.org/10.3390/su16198551>
- Movistar Empresas. (2024). *Sondeo de Adopción Digital 2024*. <https://www.telefonica.com.ec/el-91-de-las-pymes-ecuatorianas-invertira-en-sudigitalizacion-este-2024/>

- OECD. (2021). *The digital transformation of SMEs*. OECD Publishing. <https://doi.org/10.1787/1b9c3c3c-en>
- OECD, CAF, & SELA. (2024). *Índice de políticas para pymes: América Latina y el Caribe 2024: Hacia una recuperación inclusiva, resiliente y sostenible*. OECD Publishing. https://www.oecd.org/es/publications/2024/07/sme-policy-index-latin-america-and-the-caribbean-2024_d0ab1c40.html
- OECD. (2025). *SME digitalisation for competitiveness*. OECD Publishing. https://www.oecd.org/en/publications/sme-digitalisation-for-competitiveness_197e3077-en.html [oecd.org]
- Omol, E., Abuonji, P., & Mburu, L. (2025). SMEs' digital maturity: Analyzing influencing factors and the mediating role of environmental factors. *Journal of Innovative Digital Transformation*, 2(1), 19–36. <https://doi.org/10.1108/JIDT-01-2024-0002> [emerald.com]
- Revista Gestión (2024). El 91 % de las pymes ecuatorianas invertirá en su digitalización este 2024. <https://revistagestion.primicias.ec/cifras/el-91-de-las-pymes-ecuatorianas-invertira-en-su-digitalizacion-este-2024/>
- Secretaría General de la Comunidad Andina (2009). *Resolución No 1260*. <https://andina.vlex.com/vid/resolucion-n-1260-secretaria-1053509743>
- TM Forum. (2017). *Digital Maturity Model (DMM): A blueprint for digital transformation* [White paper]. <https://www.tmforum.org/wp-content/uploads/2017/05/DMM-WP-2017-Web.pdf>
- Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, 122, 889–901. <https://doi.org/10.1016/j.jbusres.2019.09.022>
- Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>
- Zubillaga Rego, A., Aramburu, N., Lorenzo, O., North, K., & Peletier Espiga, C. (2019). *Madurez digital de la PYME vasca*. Orkestra Instituto Vasco de Competitividad. <https://www.orkestra.deusto.es/images/investigacion/publicaciones/informes/cuadernosorkestra/madurez-digital-PYMES-vascas.pdf>

INFORMACIÓN PARA AUTORES

INFORMATION
FOR AUTHORS

cedia | **TICEC**
2026

Evaluación por pares

La **Revista Tecnológica Espol - RTE®** es una revista arbitrada que se rige por el sistema doble par anónimo. Los artículos enviados por los autores son evaluados en previamente por el Editor Asociado para comprobar si se ajustan a las normas de edición y a las políticas temáticas de la revista. Cuando el artículo pasa ese primer filtro es enviado a los evaluadores externos expertos en la temática abordada por el autor. Para cumplir y defender la ética de la investigación, estos evaluadores desconocen el nombre de el/los autores/as y la identificación de la/s institución/es a la que pertenece el artículo, encargándoseles dictaminar si responde a los intereses científicos de la revista y si procede su publicación. En la valoración final, los revisores deciden entre las siguientes opciones:

1. Aceptar este envío
2. Publicable con modificaciones
3. Reenviar para revisión
4. Reenviar a otra publicación
5. No publicable

En el caso de que haya disparidad de opinión entre revisores del Comité Científico, se someterá al juicio de un tercer experto, que dirimirá en conflicto de pareceres.

Una vez recibidas las valoraciones, la decisión final sobre la publicación de un manuscrito será tomada por el editor asignado de la **Revista Tecnológica Espol - RTE®**.

El proceso de revisión, desde el momento del depósito del manuscrito a través del OJS hasta la finalización de las revisiones ocupa un tiempo promedio de 6-8 semanas, salvo ciertas excepciones.

En el siguiente link [Guía de revisores](#), se detalla paso a paso el proceso de evaluación a seguir por los pares ciegos asignados.

Las responsabilidades específicas de los revisores se encuentran declaradas en el link [Código de ética y buenas prácticas de publicación](#).

Derechos de autor (Copyright)

Los originales publicados, en las ediciones impresa y electrónica, de la **Revista Tecnológica Espol - RTE®**, bajo derechos de primera publicación, son propiedad de la Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, República del Ecuador, siendo absolutamente necesario citar la procedencia en cualquier reproducción parcial o total de los contenidos (textos o imágenes) publicados. RTE proporciona un acceso abierto e inmediato a su contenido, pues creemos firmemente en el acceso público al conocimiento, lo cual no obsta para que la cita de la fuente sea obligatoria para todo aquél que desee reproducir contenidos de esta revista.

De igual modo, la propiedad intelectual de los artículos o textos publicados en la revista RTE pertenece al/la/los/las autor/a/es/as, quienes conservan sus derechos de autor y dominio irrestricto de su obra. Esta circunstancia ha de hacerse constar expresamente de esta forma cuando sea necesario.

Todo el contenido de la *Revista Tecnológica Espol - RTE®* mantiene una licencia de contenidos digitales otorgada por Creative Commons.



Esta obra está bajo una [Licencia Creative Commons Atribución-NoComercial 4.0 International](https://creativecommons.org/licenses/by-nc/4.0/).

Directrices para autores

Los textos postulados deben:

1. Corresponder a las categorías universalmente aceptadas como producto de investigación.
2. Ser originales e inéditos.
3. Sus contenidos responden a criterios de precisión, claridad y brevedad.

Se clasifican en:

3.1. Artículos. En esta sección se publican:

- 3.1.1. Artículos de investigación científica o tecnológica: presenta de manera detallada los resultados originales de proyectos terminados de investigación. La estructura generalmente utilizada contiene cuatro aportes importantes: introducción, metodología, resultados y conclusiones.
- 3.1.2. Artículo de revisión: resultado de una investigación terminada donde se analizan, sistematizan e integran los resultados de investigaciones, publicadas o no, ya sea en el campo científico, artístico o artístico tecnológico, con el fin de dar cuenta de los avances y las tendencias de desarrollo.

Próximos Volúmenes

PROGRAMACIÓN

Vol. 38, Núm. 2

Volumen Abierto Semestral

- i. *Convocatoria abierta:* envío de artículos hasta el 15 de septiembre de 2026
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos:* hasta el 30 de noviembre de 2026
- iii. *Publicación del volumen:* diciembre de 2026

PROGRAMACIÓN

Vol. 39, Núm. 1

Volumen Abierto Semestral

- i. *Convocatoria abierta:* envío de artículos hasta el 15 de abril de 2027
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos:* hasta el 30 de mayo de 2027
- iii. *Publicación del volumen:* junio de 2027

Próximos Volúmenes

PROGRAMACIÓN

Vol. 39, Núm. 2

Volumen Abierto Semestral

- i. *Convocatoria abierta: envío de artículos hasta el 15 de septiembre de 2027*
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos: hasta el 30 de noviembre de 2027*
- iii. *Publicación del volumen: diciembre de 2027*

PROGRAMACIÓN

Vol. 40, Núm. 1

Volumen Abierto Semestral

- i. *Convocatoria abierta: envío de artículos hasta el 15 de abril de 2028*
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos: hasta el 30 de mayo de 2028*
- iii. *Publicación del volumen: junio de 2028*

PROGRAMACIÓN

Vol. 40, Núm. 2

Volumen Abierto Semestral

- i. *Convocatoria abierta: envío de artículos hasta el 15 de septiembre de 2028*
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos: hasta el 30 de noviembre de 2028*
- iii. *Publicación del volumen: diciembre de 2028*

PROGRAMACIÓN

Vol. 41, Núm. 1

Volumen Abierto Semestral

- i. *Convocatoria abierta: envío de artículos hasta el 15 de abril de 2029*
- ii. *Notificación de artículos seleccionados después de revisión por pares ciegos: hasta el 30 de mayo de 2029*
- iii. *Publicación del volumen: junio de 2029*



rte.espol.edu.ec



@rte.espol



@rte_espol



@EspolRte



Revista RTE ESPOL

cedia



DOAJ

e-revist@s

Scopus®

publons

Google Scholar

latindex
catálogo 2020

REDIB

AmeliCA

OpenAlex

Crossref

RECIE
Red de Editoriales y Centros de Investigación Científica

MIAR

ROAD

INDEX
COPERNICUS