

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>

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

²*Universidad 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

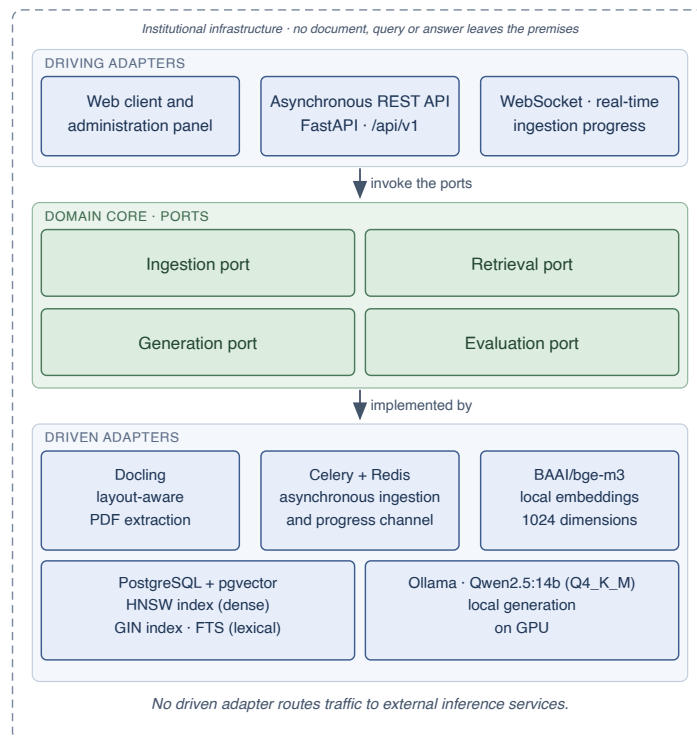
DOCUMENT	PAGES	CHUNKS
Instructions for appointing teaching and research assistants	8	9
Instructions for credit recognition through advanced study	10	19
Bylaws on credit recognition and transfer	26	42
Library collection management manual	14	19
Institutional quality assurance policy	6	17

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

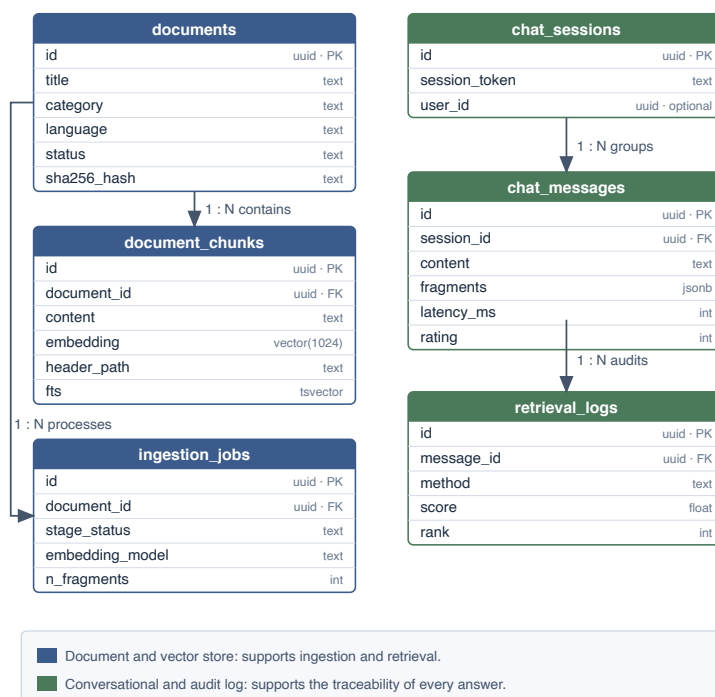


Note. Retrieval, vectorization, and generation run on institutional hardware; no data is sent to external APIs.

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

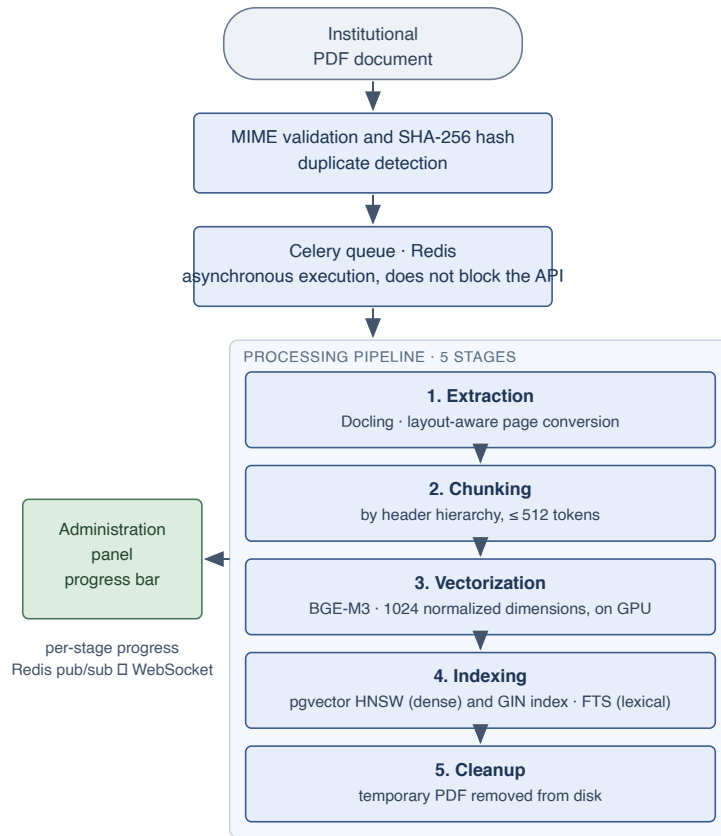
Note. Every component is open source or free to use; the main language is Python 3.11. The lexical signal is resolved with PostgreSQL's native full-text search engine and not with a BM25 implementation.

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.

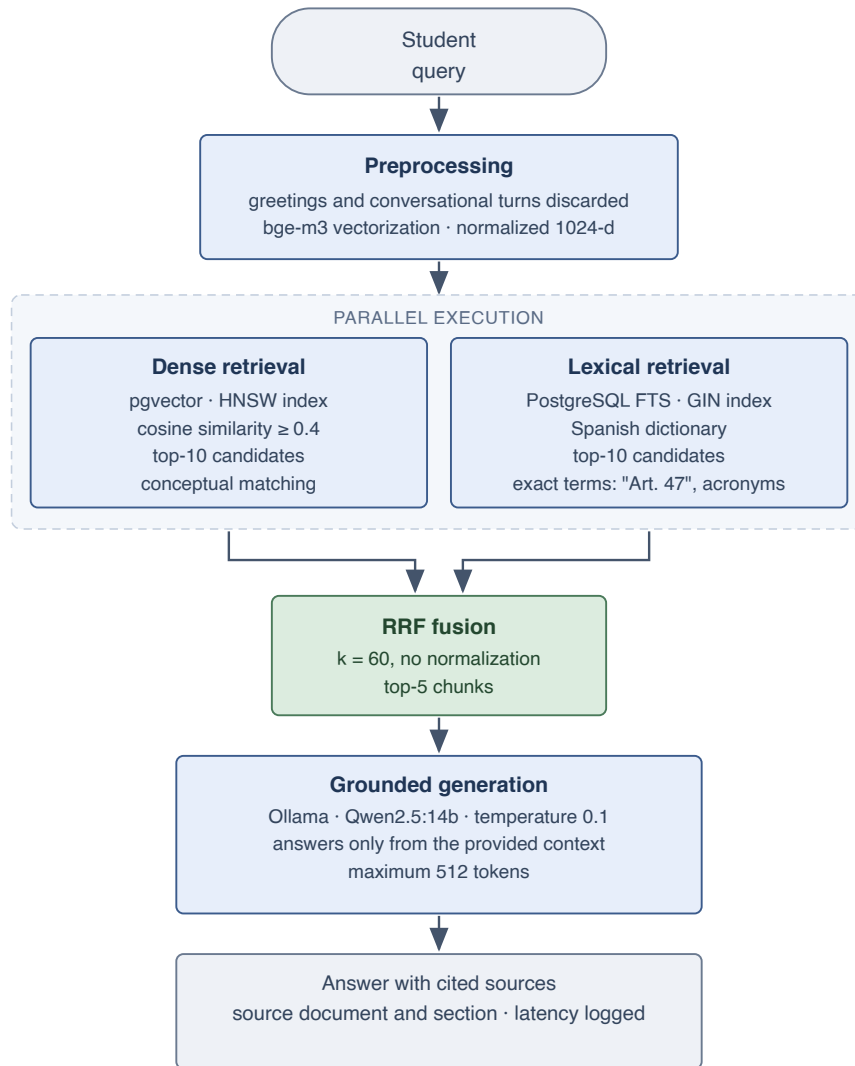
The two rankings are combined through reciprocal rank fusion (RRF) as defined by Cormack et al. (2009) in equation (1):

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

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.

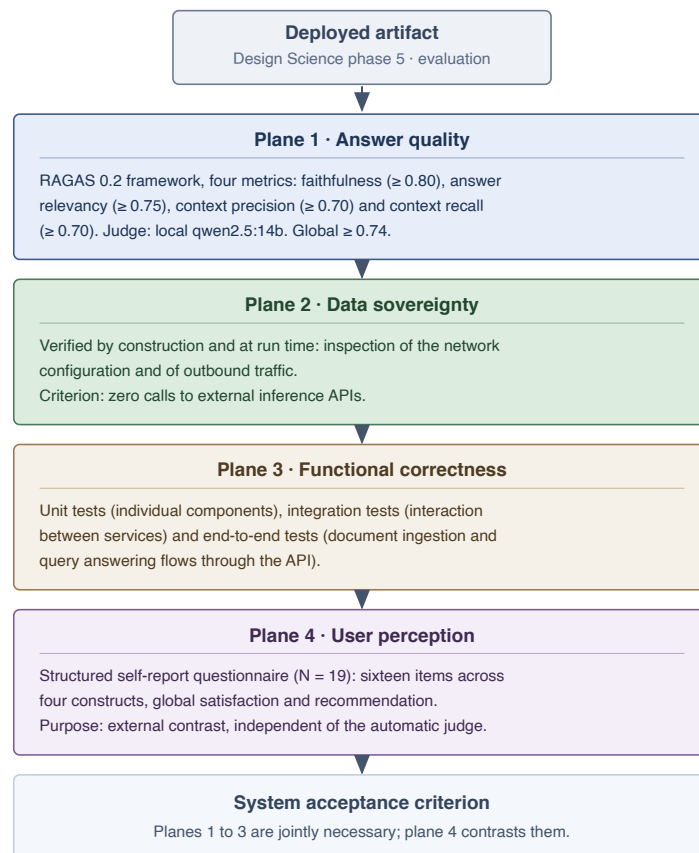
Evaluation design

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 (Kuratomi 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>