ARTICLE

An AI-enabled Digital Research Assistant for the *Legislación Mexicana* Corpus

Alberto Martínez, Rodrigo Cuellar Hidalgo, and Javier Cisneros Brito

ABSTRACT

Large-scale historical corpora present persistent challenges for research, particularly in relation to the navigation, interpretation, and extraction of meaningful information across extensive and heterogeneous collections. The Legislación Mexicana *corpus, a 42-volume compilation of legal dispositions spanning more than two centuries, exemplifies these challenges due to its size, structural complexity, and evolving terminology. Traditional approaches to working with such corpora have relied on close reading and the use of indexes. While effective within defined limits, these methods constrain the scope of inquiry and require significant time and expertise to produce meaningful results. This article presents the development of LegMexIA, an AI-enabled research assistant designed to extend interaction with the corpus beyond conventional retrieval methods. The system integrates traditional text retrieval and retrieval-augmented generation via a coordinated architecture that distinguishes between different types of user queries. Structured queries are processed using Elasticsearch and BM25 ranking, while exploratory and interpretive queries are routed through a retrieval-augmented generation workflow, where relevant text fragments are retrieved using vector similarity and assembled into contextual inputs for a language model. A key aspect of the system is the agentic decision layer that analyzes user intent and dynamically selects the appropriate processing strategy, in alignment with established reference practices in academic libraries and in a way that maintains the user's role in verification. The article also addresses key considerations related to prompt design, technological sustainability, institutional constraints, and the implications of public deployment.*

INTRODUCTION

One of the principal challenges that researchers face is working effectively with large volume collections and corpora. Extensive collections require substantial time to navigate structure, terminology, and context. This problem has been recognized for decades, from early concerns about the growing volume of recorded knowledge to later efforts in automated indexing to more recent approaches that focus on digital methods for pattern detection across large textual corpora.[1] Researchers have relied on two primary approaches to work with large volumes of text. The first involves direct engagement through reading. However, with the exponential growth of information, the growth of corpora exceeds the capacity of close reading, creating the need for alternative forms of analysis and navigation.[2] The second mode of interaction relies on indexes, which reduce the search space and provide structured points of entry into large bodies of text. Nevertheless, their usefulness is tied to the conditions under which they are created, as they reflect anticipated uses, established vocabularies, and relatively stable categories of description. In

About the Authors

Alberto Martínez (asmartinez@colmex.mx; corresponding author) is the coordinator of the Digital Scholarship Unit at Biblioteca Daniel Cosío Villegas, El Colegio de Mexico. **Rodrigo Cuellar Hidalgo** (rcuellar@colmex.mx) is a Digital Systems Specialist at Biblioteca Daniel Cosío Villegas, El Colegio de Mexico. **Javier Cisneros** (ecisneros@colmex.mx) is a Digital Systems Specialist at Biblioteca Daniel Cosío Villegas, El Colegio de Mexico.

Submitted: 6 May 2026. Accepted for Publication: 18 June 2026. Published: 21 September 2026.

https://doi.org/10.5860/ital.v45i3.17711

this manner, they represent limited use cases and do not account for the full range of possible uses over time. As research questions evolve and new contexts emerge, these structures become increasingly difficult to sustain, especially in historical collections.

The transition to digital formats expanded these possibilities. Through optical character recognition (OCR), transcription, and metadata creation, the corpus became searchable, allowing researchers to retrieve documents using keyword queries and structured filters. These developments increased precision and efficiency. However, they did not significantly alter the nature of research, as they remain primarily oriented toward locating words and phrases.[3]

BACKGROUND AND CONTEXT

The *Legislación Mexicana* corpus, compiled by José María Lozano and Manuel Dublán, is one such body of knowledge that reflects these conditions. The 42-volume collection includes 28,713 legal dispositions spanning more than two centuries. In 2012, the Daniel Cosío Villegas Library at El Colegio de México embarked on a project to prepare the corpus such that digital humanities methods could be applied to increase interaction with it. The project began with a focus on digitization, transcription, and metadata generation at the disposition level. Over time, the work led to a broader question that was grounded in the daily practices of both researchers and librarians. How can the library extend its role beyond retrieval, while maintaining the standards that govern scholarly work?

As the project progressed, the emergence of large language models (LLMs), generative artificial intelligence (AI), and retrieval-augmented generation (RAG) introduced new possibilities that extended beyond the project's original scope.[4] In light of these developments, we began exploring these technologies to determine whether interaction with the corpus could move beyond traditional search and digital humanities methods, while assessing their feasibility, capabilities, limitations, potential applications, and incorporation into the system. During these initial explorations, it quickly became evident that LLMs, in their default state, operate primarily through synthesis while obscuring the sources on which their responses are based. This became a significant consideration given that libraries strive to empower users with the means to identify, assess, and use information on their own and guide inquiry through a process of dialogue rather than provide direct answers.[5] Nevertheless, LLMs, if properly configured, could perhaps be useful to users if they're designed to align with academic library reference practices.

RELATED WORK

To identify related work, we conducted a review of the literature and other resources focused on recent applications of RAG, conversational interfaces, and AI in library, cultural heritage, and legal information systems. The search was conducted in library and information science databases, Google Scholar, arXiv, institutional repositories, and AI4LAM resources, and used citation chaining from relevant projects and benchmark papers. Search terms included "retrieval augmented generation," "RAG," "library discovery," "conversational search," "legal RAG," "AI discovery systems," "cultural heritage collections," and "historical corpora." Because the topic is still emerging, the review also included technical reports, project documentation, community repositories, conference materials, and experimental systems produced by libraries and research labs. Sources were selected when they addressed the design, evaluation, or use of AI-based retrieval systems for large textual corpora, with particular attention to retrieval accuracy, grounding in source documents, citation and metadata quality, and alignment with scholarly research practices.

Recent work in library and information systems has begun to explore the use of AI and conversational interfaces as a means of improving discovery and interaction with collections.[6] These developments reflect a broader shift toward systems that enable users to engage with large bodies of information through natural language inquiry. At the same time, they raise important questions regarding how such systems should be integrated into existing research practices, particularly in environments where the relationship among retrieval, analysis, and verification remains central.

The Open French Law RAG project by Harvard's Library Innovation Lab is an experimental system that combines multilingual RAG with LLMs to help users, including non-French speakers, search and understand a large corpus of French statutory law while grounding responses in source documents. It evaluates both the potential and limitations of this approach, showing that while it improves access and cross-language retrieval, it still faces issues with accuracy, relevance, and source reliability.[7]

During the exploratory phase, we identified LegalBench-RAG as a methodology for evaluating legal RAG systems. Its approach focuses on automating the evaluation process by using other LLMs to assess a RAG system's output. Because legal information can influence interpretation and decision-making, errors in retrieval or generated output required particular attention during the evaluation process. The benchmark therefore emphasizes the retrieval of minimal, highly relevant text segments that can support verifiable claims and enable citation.[8]

Recent projects have also explored the application of RAG to large-scale historical corpora, including historical newspapers and cultural heritage collections.[9] These works reflect ongoing challenges associated with the use of keyword-based retrieval in environments where research questions are often exploratory and interpretive in nature. As such, the integration of retrieval and generative processes has been proposed as a means of enabling new forms of interaction that allow users to move between specific documents and broader thematic relationships within a corpus. At the same time, these efforts continue to highlight issues related to metadata quality, retrieval strategy, and the need to ensure that system behavior remains aligned with the practices and expectations of research communities.

SYSTEM DESIGN AND ARCHITECTURE

The AI system for *Legislación Mexicana*, named LegMexIA, was implemented as a Django application (Python 3.11) that serves as both the public interface and the central orchestration layer. It renders the user interface, manages sessions and authentication, and hosts the administrative environment. Within this application, the conversational agent module executes directly, through which a multilayered decision process evaluates user intent, routes queries, and assembles responses without relying on external orchestration services. The agent integrates Elasticsearch for deterministic BM25-based retrieval over the legmex_nomic_4 index, while Gemini 2.5 Flash is used for the OpenRouter API for intent classification and response generation.[10] A FastAPI service exposes the RAG workflow as an independent HTTP endpoint, supporting semantic vector search using Nomic embeddings and generating analytical responses based on retrieved context.

Structured metadata, transcriptions, and collection records are stored in a structured query language (SQL) database accessed through the Django object-relational mapping (ORM). The prompt system supports remote version control through a GitHub repository, enabling updates without redeployment. Decision processes are recorded in JSON log files, supporting traceability

of system behavior. The system operates in a Linux environment on a local dedicated workstation server, where Elasticsearch, the Django application deployed through Docker, and the FastAPI service coexist as coordinated but independently deployable processes.

The system processes queries through a structured workflow. When a user asks about legal provisions related to public education in the 19th century, the Django interface receives the query and passes it to the conversational agent. The agent first transforms the question from a conversational formulation into a query based on key terms, including topic, date range, document type, and legal category. The system then uses an agentic routing layer to determine the appropriate retrieval path. Queries that depend on structured criteria, including dates, identifiers, titles, document types, precise legal terms, or the frequency of specific terms across the corpus, are processed through Elasticsearch using BM25 retrieval over the indexed corpus. This supports precise retrieval of candidate records. Exploratory queries, or queries that ask about relationships between documents, are processed through the FastAPI RAG service. In this process, Nomic embeddings support semantic retrieval of relevant text fragments, which are used by the language model to generate a response grounded in the retrieved material. These fragments, together with their metadata, are assembled as context for Gemini 2.5 Flash through the OpenRouter API. The prompt repository contains the instructions used by the agent and can be updated independently from the rest of the system.

The system combines rule-based logic with model-based classification and can revise its initial decisions when the generated response does not meet the system's criteria for relevance, completeness, or quality. Queries may shift between structured retrieval and semantic interpretation depending on their characteristics. This gives the system more flexibility than a standard RAG workflow.

These processes are coordinated through an orchestration layer that manages query interpretation, routing, execution, and recovery. The layer keeps track of the current conversation, including previous questions, returned documents, and references to earlier results. This allows the system to handle follow-up queries without treating each question as an isolated request.

EVALUATION

The performance and usability of the LegMexIA system were evaluated through a series of one-on-one sessions with four participants: three researchers in Mexican history and Latin American economic history, and one subject librarian with extensive familiarity with the *Legislación Mexicana* corpus. The participating researchers had previously worked with the corpus or with closely related primary sources, which allowed them to assess the system in relation to established research practices in Mexican legal, political, and economic history. The subject librarian had previously developed the first digital project focused on search and retrieval for the *Legislación Mexicana* corpus and was familiar with the history of the work itself, including earlier efforts to make the collection searchable and usable for research. This background allowed him to evaluate the system in relation to previous approaches to access and discovery with the corpus.

The sessions followed a guerrilla usability approach. They were structured as informal, task-based conversations with the participants, with emphasis on observing how experienced users interacted with the system, how they formulated questions, and where they identified useful or problematic results. Because the sessions were not recorded, the comments reported here are drawn from session notes.

The process began with exploratory searches designed to identify relevant materials, including the following:

1. What documents exist concerning the stamp tax and its collection between 1884 and 1900?
2. What documents list the tariff laws enacted between 1821 and 1900 and, separately, the tariff provisions issued during that period?
3. Which government decrees granted patents for inventions in Mexico between 1893 and 1901?

This was followed by more structured descriptive questions that included the following:

1. Analyze the legal provisions established for mine inspector engineers under the 1887 instructions. Identify and describe at least five specific obligations these officials were required to fulfill, explaining how each contributed to the regulation and development of mining activity in Mexico.
2. How did the methods and procedures used by spies to gather intelligence about the enemy during field operations evolve over time?
3. How did Mexico seek to promote foreign trade after independence?

The final stage involved open-ended questions that required interpretation and comparison across multiple documents. Such queries included the following:

1. How might the concept of settlement be defined through an analysis of the documents?
2. How did the concept of liberty evolve over time across the documents in the corpus?
3. Is there evidence of debates or discussions of a philosophical nature in the documents?

This progression reflects the distinction embedded in the system between retrieval-oriented queries and those that require semantic interpretation and contextual reasoning.

Representative feedback from the sessions pointed to three important considerations for future development. First, participants emphasized a preference for exhaustive retrieval over presenting only a small sample of relevant documents. However, exhaustiveness could present a challenge for RAG systems, as the current RAG system depends on selecting a limited set of passages or documents for inclusion in the model context and uses a filtering process. For this reason, the system must distinguish between tasks that call for illustrative examples and tasks that require broader or more systematic coverage of the corpus. The participants also recommended the creation of a guide to explain the system's intended uses, its limitations, the difference between retrieval and interpretation, and the need to verify generated responses against cited source documents. Participants also suggested expanding the corpus beyond Dublán and Lozano by incorporating other historical legal collections with similar relevance for the study of Mexican law and legal history.

These queries were used as part of an evaluation process, with validated responses produced and reviewed. The sessions showed that researchers could use the system to identify legislative documents they had not previously encountered, despite their familiarity with the corpus. One such example was a question regarding the nationality of investors in railway lines. The results revealed a common practice of granting citizenship to foreign investors that was already known. Nevertheless, the system provided a way to retrieve and assemble these documents for further study.

Several open-ended questions were also identified as possible starting points for further research, such as the evolution of the concept of liberty over time as well as a question regarding the concept of property. These results suggest that the LegMexIA system improves interaction with the corpus and enables new ways of exploring it, while also influencing how questions are formulated.

The evaluation process also revealed areas that require further development. Accurate interpretation of queries remains essential, and misclassification can affect results. In addition, variability in generated responses remains present and requires continued refinement. The feedback called for future versions of the system to provide clearer mechanisms for comprehensive retrieval, source verification, and user guidance, especially when researchers need exhaustive results that go beyond selective examples.

CONSIDERATIONS

Retrieval and interpretation should be treated as distinct but related functions, particularly in cases where users engage with materials that require both precise identification and broader contextual understanding. Systems benefit from recognizing query structure and applying strategies that match each type of request. Grounding and traceability remain essential, especially in research contexts where verification is central. Systems with decision-making layers and multiple retrieval strategies can respond more effectively to the variability of research behavior and should be considered when designing similar tools.

A related challenge concerns the design of semantic similarity-based retrieval strategies. In the case of LegMexIA, the documents that compose the corpus exhibit significant variation in length, ranging from very short entries to texts that exceed 150,000 words, and frequently contain multiple topics within a single document. Under these conditions, uniform chunking proves insufficient, as it does not reflect internal structure or thematic variation. Addressing this required a segmentation process capable of generating variable-length subdocuments that respond to shifts in topic. The documents were first divided into overlapping fragments of approximately 200 tokens. An embedding was then generated for each fragment using a multilingual sentence transformer model, and cosine similarity was calculated between consecutive fragments. When similarity fell below a threshold of 0.60, established through empirical testing on documents previously identified as containing multiple topics, the system marked that point as a thematic transition and began a new subdocument. Consecutive fragments that remained above the threshold were grouped into the same unit. As a result, longer documents are transformed into more coherent subdocuments, while shorter texts tend to remain intact, reducing the noise produced by long or thematically heterogeneous segments.

Prompt engineering introduced a significant shift in systems development. Crafting effective prompts became central to system functionality and required rethinking how workflows are defined. This involved balancing clarity, depth, and consistency to achieve coherent responses, while accounting for how language models interpret structure and instruction. The process was iterative, revealing differences in how models respond in terms of interpretation, output, and cost. Prompt engineering therefore functions as an ongoing process of refinement shaped by both system requirements and model behavior.

In the development of systems of this kind, it becomes necessary to consider technological sustainability in relation to the institutional environment. The selection of tools, models, and architectures extends beyond immediate functionality to issues of maintenance, stability, and

dependency on external services. Technologies that rely on evolving interfaces or closed provisioning models may introduce constraints that become significant over time. At the same time, available infrastructure, including computational capacity and system integration, shapes technical decisions. This is accompanied by the need to account for production and usage costs, particularly where commercial models are involved. The merging of technical capacity, institutional resources, and cost structures supports a more stable framework for implementation and ongoing development.

A final consideration involves the implications of public deployment. Each query incurs a cost, as commercial language models are used in processing responses. Initial access was limited to the local community to ensure purposeful use. As broader availability was considered, it became necessary to ensure that the system supported legitimate research and information needs. This led to the implementation of a registration process requiring users to specify their objectives before accessing the system. The public version of LegMexIA is expected to be available by October 1, 2026, at https://legislacionmexicana.colmex.mx/.

Future work will focus on strengthening the system's capacity to support research tasks that require exhaustivity in retrieval, source verification, and follow-up questions. One priority is to refine the distinction between queries that call for illustrative responses and those that require exhaustive or near-exhaustive coverage of relevant documents. This is particularly important for historical and legal research, where partial retrieval can affect interpretation. Further development will also involve expanding the evaluation process with a larger group of researchers, subject specialists, and librarians to test the system across different types of questions, research strategies, and levels of familiarity with the corpus.

LIMITATIONS AND LESSONS LEARNED

The project revealed several limitations for future development. The first is query interpretation. Although the system distinguishes between retrieval-oriented questions and questions that require semantic interpretation, some queries are difficult to classify. When this occurs, results may be too narrow, too broad, or misaligned with the researcher's intention.

A second limitation concerns exhaustiveness and the role the system should play in research. Researchers noted that historical and legal work often requires comprehensive retrieval rather than a small sample of relevant documents. This raises a question of scope regarding whether the system should aim for completeness or function primarily as an exploratory tool that helps users identify documents, themes, and lines of inquiry. Because RAG systems usually depend on selecting a limited set of passages for response generation, they may support initial stages of research and synthesis, while still requiring additional mechanisms when researchers need to determine whether all relevant documents have been identified.

Development also showed that corpus structure affects system performance. The documents vary widely in length and often contain multiple topics, making uniform chunking insufficient. Addressing this required variable-length segmentation, stronger metadata control, and attention to citation and traceability.

The evaluation process reinforced the need to keep users at the center of the design process. Usability testing can help create a shared understanding between developers, librarians, and researchers about what the system does, what it does not do, and how its results should be interpreted. Sensemaking may be useful in this context because many institutions are still working

through the practical challenges of AI literacy, including how users understand AI-generated responses, how they evaluate source grounding, and how they decide when additional verification is needed.

FUTURE WORK

Additional work is needed to reduce variability in generated responses, improve query classification, and make retrieval decisions more transparent to users. Future development should also explore whether the chatbot can guide users more intuitively, reducing the need for separate instructional documentation. This would involve clearer prompts, better feedback when a query is too broad or ambiguous, and interface cues that help users understand when the system is retrieving documents, generating synthesis, or asking for refinement. Researchers also suggested expanding the corpus beyond Dublán and Lozano by incorporating other historical legal collections with similar evidentiary value. This expansion would require attention to corpus alignment, metadata normalization, document segmentation, citation practices, and the relationships among overlapping or complementary legal sources. Finally, future development will address issues of technological sustainability, including model selection, cost control, infrastructure requirements, and the conditions under which the system can be responsibly made available to broader research communities.

CONCLUSION

The development of the LegMexIA suggests one possible direction for how libraries can support interaction with large and complex collections. Systems such as LegMexIA go beyond simple retrieval and access, through generative AI's ability to understand user needs and craft optimized search strategies that could improve how libraries respond to users' information needs. This development extends existing practices, offers new opportunities to innovate library services, and raises new questions about how libraries define their role in research.

USE OF GENERATIVE AI

Generative AI tools (ChatGPT, OpenAI) were used to support translation of selected sections from Spanish to English, editing for clarity and concision, restructuring paragraphs, and refining technical descriptions in sections such as system architecture. Prompts were used to guide these tasks. All content was reviewed, revised, and validated by the authors, who retain full responsibility for the final text. All ideas, interpretations, and conclusions expressed in this work are those of the authors. To make this process transparent, the prompts used for translation, language editing, paragraph restructuring, technical description refinement, and system behavior definition have been documented in a public GitHub repository. This repository includes the system prompts and related prompt materials used during development, allowing readers to examine the instructions that guided both manuscript preparation and the design of the AI-enabled research assistant.[11]

Prompt

Translate the following academic article from Mexican Spanish into American English. This will be for peer-reviewed journal in Library Science.

Keep the original argument, meaning, structure, headings, citations, quotations, notes, and references. Do not summarize, remove, expand, or add additional information.

Avoid phrasing that follows Spanish syntax.

ENDNOTES

[1] Vannevar Bush, "As We May Think," *The Atlantic Monthly* 176, no. 1 (1945): 101–8, https://www.ias.ac.in/article/fulltext/reso/005/11/0094-0103; Hans Peter Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development* 1, no. 4 (1957): 309–7, https://ieeexplore.ieee.org/abstract/document/5392697.

[2] Franco Moretti, "Conjectures on World Literature," *New Left Review* 1 (2000): 54–68, https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature.pdf.

[3] Robert L. Collison, *Indexing Books: A Manual of Basic Principles* (London: Ernest Benn Limited, 1962), 26–27, 34–35, 50.

[4] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems* 33 (2020): 9459–74, https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

[5] Kathleen Kluegel, Catherine Sheldrick Ross, Jana Ronan, Kathleen Kern, and David Tyckoson, "The Reference Interview: Connecting in Person and in Cyberspace: Presentations and Responses from the RUSA President's Program, 2002 ALA Annual Conference, Atlanta, June 17, 2002," *Reference & User Services Quarterly* 43, no. 1 (2003): 37–51, https://research.ebsco.com/c/xwhikx/search/details/hqyzhc63ov.

[6] Frank Boateng, "The Transformative Potential of Generative AI in Academic Library Access Services: Opportunities and Challenges," *Information Services and Use* 45, nos. 1–2 (2025): 140–47, https://journals.sagepub.com/doi/10.1177/18758789251332800.

[7] Library Innovation Lab, Harvard Law School, "Open French Law RAG," https://lil.law.harvard.edu/open-french-law-rag/.

[8] Nicholas Pipitone and Ghita Houir Alami, "LegalBench RAG: A Benchmark for Retrieval Augmented Generation in the Legal Domain," arXiv, August 19, 2024, https://arxiv.org/abs/2408.10343; "LRAGE: Legal Retrieval Augmented Generation Evaluation Tool," arXiv, 2025.

[9] Keerthana Murugaraj, Salima Lamsiyah, Marten During, and Martin Theobald, "Topic-RAG for Historical Newspapers: Enhancing Information Retrieval in Humanities Research through Topic-Based Retrieval-Augmented Generation," *Computational Humanities Research* 1 (2025): e15, https://doi.org/10.1017/chr.2025.10018; P. Kelly, J. Schild, and A. Jafari, "FolkRAG: a Retrieval-Augmented Generation System for Cultural Heritage Materials," *Neural Computing and Applications* 37 (2025): 20281–97, https://doi.org/10.1007/s00521-025-11455-4.

[10] Stephen E. Robertson and Hugo Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval* 3, no. 4 (2009): 333–89.

[11] ColmexBDCV, "legmex_prompts," GitHub repository, https://github.com/ColmexBDCV/legmex_prompts.