PyLegalIR: A Benchmark for Spanish Legal Information Retrieval from Paraguayan Supreme Court Cases

Anonymous ACL submission

Abstract

We present PyLegalIR, a benchmark dataset designed for evaluating information retrieval systems in the Spanish legal domain, using Criminal Chamber cases of the Paraguayan Supreme Court (SCP). Despite the critical need for effective legal retrieval systems in the Spanish Language, there are no publicly available datasets. PyLegalIR addresses this gap by providing a supervised benchmark comprising 54 expert-created queries, each annotated with 30 relevant documents on average, resulting in 1,597 query-document pairs with graded relevance judgments. Annotations were performed by Paraguayan legal professionals, covering diverse legal topics. This dataset enables benchmarking and fosters the development of retrieval systems for Spanish legal texts. All code and data are publicly available [1].

1 Introduction

Problem. Legal professionals and researchers rely heavily on information retrieval (IR) systems in their daily work to navigate vast collections of court decisions, statutes, and legal commentary. Despite major advances in information retrieval, few datasets exist for legal texts, and to our knowledge none in Spanish. Most benchmarks focus on English and general-domain retrieval, limiting their relevance for real-world legal systems in Latin America. This forces Latin American legal institutions to rely on outdated retrieval systems. For instance, the Supreme Court of Paraguay (SCP) still uses a rudimentary system that retrieves documents based solely on exact search query matches.

Contribution. We address this gap by introducing **PyLegalIR**, a new benchmark for legal information retrieval in Spanish. It consists of 5,000 court rulings and 54 expert-written queries, annotated with 1,597 relevance judgments by legal professionals. We evaluate a wide range of zero-shot retrieval models and publicly release the corpus, queries, annotations, and evaluation code[1].

2 Related Work

In the legal domain, several specialized IR datasets have emerged, primarily in English. Task 1 of the COLIEE competition (Rabelo et al., 2022) comprises legal case retrieval using Canadian case law. Instead of using short queries as a search-instruction, an entire document is used as a query. ACORD (Wang et al., 2025) provides a detailed clause-level retrieval benchmark for contract-related queries with expert annotations graded on a 5-point scale. LegalBench-RAG (Pipitone and Alami, 2024) compiles expert-generated legal question-answer pairs alongside extractive evidence annotations for retrieval-augmented generation tasks. Additionally, Housing Statute QA (Zheng et al., 2025) adapts U.S. housing law queries into retrieval tasks linked directly to statutory references. While robust, these datasets are expensive to build and limited to Anglo-American legal systems.

3 The PyLegalIR Dataset

3.1 Document Collection

The PyLegalIR benchmark is based on real-world legal documents from the Criminal Chamber of the SCP. The corpus comprises 5,000 judicial decisions issued between 2011 and 2023, covering a wide range of criminal cases such as homicide, drug trafficking, sexual abuse, fraud, and more (Gómez Adorno et al., 2024).

The documents were obtained from the SCP's official website and processed in plain text format via OCR if the original document was a scanned PDF or plain text extraction for other file types.

[1] https://github.com/PyLegalIR-anonymous/pylegalir-benchmark

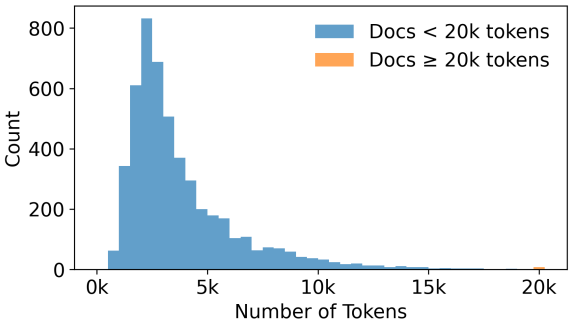

Figure 1: Sequence Length distribution of the document corpus with documents over 20k tokens grouped.

OCR artifacts were filtered by removing lines with fewer than three words. Each document includes procedural background, arguments, and a final decision. Documents average 3,800 tokens while some exceed 79,000 tokens (see Figure 1).

3.2 Query Set

The query set consists of 54 manually formulated information needs, each corresponding to a distinct criminal offense (e.g., "Acquittal for failure to comply with the legal duty of child support") or broader legal concept (e.g., "horizontal procedural oversight"). The queries were written by legal professionals working in the SCP and follow a keyword-style format.

3.3 Relevance Annotation

To determine their level of relevance, each of the 54 queries in PyLegalIR was annotated with approximately 30 documents using the annotation software *DocTag* (Giachelle et al., 2022). For each query, candidate documents were retrieved using a combination of sparse (BM25) and dense (embedding-based) retrieval methods, and a pooled set of top-30 ranked documents was presented to the annotators.

Relevance was assessed using a four-point ordinal scale: **0** for "Not relevant", **1** for "Partially relevant" (some useful content, but not fully responsive), **2** for "Relevant" (document is helpful in responding the query, but the query is not the main topic) and **3** for "Highly relevant" (the query is the main topic of the document).

Due to resource constraints, the annotation work was distributed between two annotators, so there are no overlapping annotations for computing inter-annotator agreement. To encourage consistency, strict guidelines were provided to the annotators, which are publicly available in the repository [1].

4 Experimental Settings and Baselines

PyLegalIR is designed as a document ranking benchmark. Given a query, the task is to retrieve and rank documents from the corpus so that those most relevant to the query are ranked highest.

4.1 Retrieval Models

To evaluate the difficulty and coverage of PyLegalIR, we benchmark a wide range of retrieval systems, covering traditional sparse retrieval, dense and cross-encoder methods, and hybrid combinations. All models were evaluated in a zero-shot setting, without fine-tuning on PyLegalIR.

To include the current retrieval system used at the SCP, we implement a rudimentary **Exact-Match** search, assigning a similarity score of 1.0 to a query-document pair, if the search query is contained in the document and 0.0 otherwise.

We implemented **BM25**, using the default Okapi BM25 parameters from python's Rank-BM25 library.

To represent state-of-the-art multilingual dense retrievers, **BGE-m3** (Chen et al., 2024) and **jina-embeddings-v3** (Sturua et al., 2024) were employed. These models encode queries and documents into dense vectors and rank them based on their dot product.

We also evaluate two different reranker models, cross-encoders that re-rank the top-50 documents retrieved by BM25. We ran an evaluation with **BGE-Reranker** [2], and a re-ranker based on **MiniLMv2** (Wang et al., 2020) [3] trained on a multilingual version of the MS MARCO dataset (MM MARCO).

Besides the standard dense endpoint, the BGE-m3 model also has a sparse and ColBERT endpoint. To explore sparse neural approaches, we include **BGE-Sparse**, encoding queries and documents into sparse lexical vectors. We also evaluate **BGE-ColBERT**, which supports fine-grained late interaction between query and document tokens.

To include hybrid methods in our evaluation, we combine dense and sparse scores using their standard score (z-score). **BGE-Dense-Sparse** combines BGE dense embeddings with BGE-sparse vectors. **BGE-Sparse-ColBERT** combines BGE-sparse with BGE-ColBERT. **BGE-Dense-Sparse-ColBERT** combines all three components.

[2] Huggingface: `BAAI/bge-reranker-v2-gemma`
[3] Huggingface: `mmarco-mMiniLMv2-L12-H384-v1`

Model	nDCG@10	MRR@10	Recal@100
Exact-Match	0.124	0.162	0.186
BM25	**0.710**	**0.830**	**0.804**
BGE-m3	0.481	0.666	0.613
jina-embeddings-v3	0.389	0.532	0.540
BGE-Sparse	0.350	0.551	0.582
BGE-ColBERT	0.469	0.696	0.712
BGE-Dense-Sparse	0.468	0.698	0.691
BGE-Dense-ColBERT	0.529	0.728	0.703
BGE-Sparse-ColBERT	0.429	0.631	0.681
BGE-Dense-Sparse-ColBERT	0.501	0.748	0.742
BGE-Chunking	0.437	0.668	0.713
BGE-Reranker	0.335	0.516	0.804
MiniLMv2	0.301	0.438	0.804
MiniLMv2-Chunking	0.499	0.744	0.804

Table 1: Retrieval performance of all evaluated models on the PyLegalIR benchmark. Results are reported for nDCG@10, MRR@10, and Recall@100 across 54 expert-annotated queries.

To handle the long document lengths typical of legal documents, we run BGE-m3 with a sliding window of 256 and a stride of 128 and evaluate **BGE-Chunking**. The mean of the three highest scoring chunks represents overall document similarity. To also measure the effect of chunking on rerankers, we run **MiniLMv2-Chunking**.

We set the sequence length to 4096 for all neural methods other than BGE-Chunking, which resulted in the best retrieval performance. All the evaluation configurations are included in the accompanying repository and can be reproduced.

4.2 Evaluation Protocol

For retrieval evaluation, we report the Normalized Discounted Cumulative Gain (**nDCG@10**), which measures the ranking quality with respect to graded relevance. To report the proportion of all relevant retrieved documents that appear in the top 100, we measure **Recall@100**. Finally, we report the Mean Reciprocal Rank (**MRR@10**) of the first relevant result, using binary relevance.

4.3 Experimental Results

Table 1 presents the retrieval performance of all evaluated models on the PyLegalIR benchmark.

Lexical vs. neural retrievers. While all retrieval models outperform the currently at the SCP used exact-match search, BM25 achieves the strongest performance with an nDCG@10 of 0.710 and Recall@100 of 0.804. Dense retrievers underperform relative to BM25, including BGE-m3 (0.481 nDCG@10) and jina-embeddings-v3 (0.389). This shows how zero-shot dense retrievers struggle on this benchmark, likely due to missing domain and document-length adaptation. When comparing the performance of BGE-m3 and BM25 query-wise, we can observe that although the overall performance of BM25 is superior, the dense retriever does perform better on 15 of the 54 queries (see diagram 2).

Hybrid and late-interaction models. Among neural approaches, hybrid systems show slight improvements. The BGE-Dense-ColBERT and BGE-Dense-Sparse-ColBERT combinations outperform the standalone BGE-m3 and BGE-Sparse models, with nDCG@10 scores of 0.529 and 0.501, respectively. BGE-Chunking improved recall strongly by 10% points compared to BGE-m3. This confirms that chunking preserves more information as opposed to encoding very long documents into a single dense vector.

Rerankers. Given the strong performance of BM25, it suggests that using a common two-stage retrieval architecture of BM25 together with a cross-encoder as a reranker could be powerful. However, the results show that all evaluated rerankers worsened the initially strong performance of BM25. The highest score reached by MiniLMv2-Chunking with an nDCG@10 of 0.499 is still much lower than the initial 0.710. This confirms just how challenging this benchmark is for neural retrieval models.

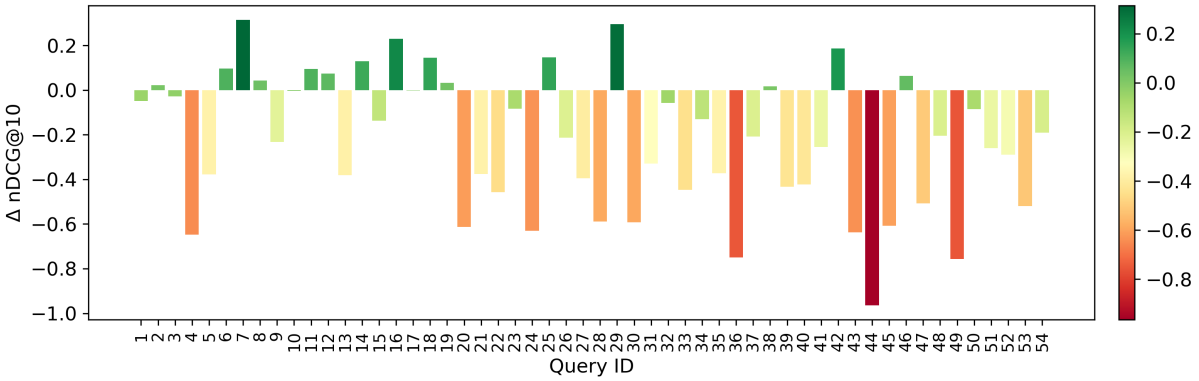

Figure 2: BGE – BM25: nDCG@10 comparison per query.

5 Silver Label Generation

Supervised information retrieval models typically require large quantities of labeled query-document pairs. To enable IR-finetuning on the corpus, we explored two silver-labeling strategies to synthetically generate training data for legal IR in Spanish using a Large Language Model (LLM) (`mistral-small-2501`).

We consider two distinct strategies for generating silver annotations:

InPars-inspired annotation. Inspired by the methodology proposed in InPars (Bonifacio et al., 2022), we use the LLM to generate queries based on individual documents from the corpus. We prompt the LLM for each document to produce a query that could realistically lead a user to retrieve that document. This creates synthetic query-document pairs. We generated 5,000 such pairs, one per document, and treated them as positive examples in a sparse retrieval setting.

Synthetic dense annotation via LLM. In a second approach, we generate new queries via LLM by prompting it to create queries similar to the 54 covered in the PyLegalIR benchmark. In a subsequent step, legal experts filtered and corrected these queries, yielding 308 new ones. For each of these 308 queries, we use the same procedure as in the original dataset to retrieve a candidate pool of 40 documents per query. Then, instead of annotating them by legal experts, we prompt the LLM to assign binary relevance labels to each of the 40 documents per query, along with a justification in the form of an "evidence" text span. This results in dense silver annotations, with 12284 query-document pairs.

5.1 Performance Comparison

We finetuned BGE-m3 on both silver datasets and then evaluated on the human-annotated PyLegalIR benchmark. When finetuning on the InPars dataset, nDCG@10 dropped from 0.481 (zero-shot) to 0.253. This could be due to the queries being much longer on average than the original 54 queries. Finetuning on the synthetic densely annotated dataset, nDCG@10 dropped from 0.481 to 0.325 when evaluated on PyLegalIR. Although in this dataset, the queries closely resemble those in the original dataset, the finetuned model still fails to generalize, suggesting that issues may lie in the noise or inconsistency of the LLM-generated relevance labels.

6 Conclusions

We introduced **PyLegalIR**, the first benchmark for Spanish legal information retrieval, based on expert annotations of real-world judicial decisions from the SCP.

Our extensive evaluation across lexical, dense, sparse, and hybrid retrieval methods reveals that BM25 remains a strong baseline, consistently outperforming state-of-the-art dense retrievers in a zero-shot setting. Moreover, attempts to improve performance through fine-tuning on synthetic, LLM-annotated data failed to yield gains, underscoring the difficulty of adapting neural models to this setting.

These findings underline the hurdle of retrieving long, domain-specific legal documents using current neural methods. PyLegalIR poses a concrete challenge to the field: How well can dense models represent long, highly complex, domain-specific legal documents in Spanish?

Limitations

While PyLegalIR provides a valuable step toward evaluating information retrieval systems in Spanish-language legal contexts, there are several limitations to consider.

First, the annotations are drawn exclusively from the SCP's criminal law cases. As such, the benchmark reflects the vocabulary, legal procedures, and structural conventions of the Paraguayan criminal justice system. Generalization to other legal domains (e.g., civil, administrative, or international law) or jurisdictions (e.g., Spain or Mexico) may be limited.

Second, while legal professionals produced all relevant judgments, they were not cross-validated via majority voting. Future work is planned to run additional annotations of a meta-annotator to compute inter-annotator agreement.

Finally, our silver datasets have not proven to transfer well to the original benchmark when training dense retrieval models. Future work is needed to improve the robustness and transferability of synthetic supervision.

Despite these limitations, PyLegalIR addresses a critical gap by providing the first supervised IR benchmark for legal Spanish texts and lays the foundation for future work in retrieval for low-resource legal domains.

Ethical Considerations

The PyLegalIR dataset is constructed from publicly available court rulings issued by the Supreme Court of Paraguay. These documents are published by the Court itself and may include the names of individuals involved, as is customary under Paraguayan law. We assume this usage complies with the legal and ethical norms of the jurisdiction.

All annotations were carried out by qualified legal professionals. The benchmark is intended exclusively for research purposes and should not be used to inform legal decisions without expert validation.

Annotators were informed about the purpose and public release of the dataset, and provided informed consent for their annotations to be used in research.

The dataset may include content related to violent or sensitive criminal cases.

We used ChatGPT to assist in refining parts of this paper, and to prototype some elements of the evaluation code. All final decisions on content and implementation were made by the authors.

Acknowledgments

We gratefully acknowledge the invaluable support of two legal professionals who contributed their expertise in annotating the dataset. Their deep knowledge of legal texts was instrumental in creating the benchmark, and their names will be disclosed upon the paper's acceptance.

References

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data Augmentation for Information Retrieval using Large Language Models. *arXiv preprint*. ArXiv:2202.05144 [cs].

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Fabio Giachelle, Ornella Irrera, and Gianmaria Silvello. 2022. Doctag: A customizable annotation tool for ground truth creation. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 288–293. Springer.

Helena Gómez Adorno, José Luis Vázquez Noguera, Cristian Amarila Closs, and José Vázquez-Cerrillo. 2024. Dataset of the criminal chamber cases from the supreme court of justice of paraguay.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

J. Rabelo, R. Goebel, M. Y. Kim, and 1 others. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *Review of Socionetwork Strategies*, 16(1):111–133.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Steven H. Wang, Maksim Zubkov, Kexin Fan, Sarah Harrell, Yuyang Sun, Wei Chen, Andreas Plesner, and Roger Wattenhofer. 2025. Acord: An expert-annotated retrieval dataset for legal contract drafting. *Preprint*, arXiv:2501.06582.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW '25, page 169–193. ACM.

A Appendix: Prompts for Synthetic Data Generation

A.1 InPars-style Query Generation Prompt

The following prompt was used to generate synthetic queries for the InPars-style dataset, where each document from the corpus was used as input to a LLM (`mistral-small-2501`) to generate a plausible user query.

Original (Spanish):

> *Eres un abogado penalista paraguayo, experto en análisis jurisprudencial y sistemas de búsqueda.*
> *A continuación, recibirás un documento legal completo. Genera una consulta breve pero específica, que un abogado podría ingresar en un buscador jurídico para encontrar exactamente este documento.*
> *La consulta debe mencionar claramente los aspectos que caracterizan el caso y que permiten distinguirlo de otros casos.*
> *No menciones nombres de personas en tus consultas.*
> *Solo incluye la consulta como string en tu respuesta. No incluyas ningún otro texto o explicación adicional.*

Translated (English):

> *You are a Paraguayan criminal lawyer, expert in case law analysis and legal search systems.*
> *You will now be given a full legal document. Generate a brief but specific query that a lawyer might enter into a legal search engine to find this exact document.*
> *The query should clearly mention the aspects that characterize the case and make it distinguishable from other cases.*
> *Do not mention any names of individuals in your query.*
> *Only include the query as a string in your response. Do not include any other text or explanation.*

This prompt was applied to each document independently to create a corresponding synthetic query. The resulting query-document pairs were treated as positive examples for sparse retrieval training.

A.2 Synthetic Dense Annotations Prompt

The following prompt was used to generate dense relevance annotations for 308 queries written by legal professionals. For each query, a candidate pool of 40 documents was created, and the prompt asked the model to assess binary relevance and extract an evidence span.

Original (Spanish):

> *Eres un abogado penalista paraguayo experto en análisis jurisprudencial.*
> *Tu tarea es decidir si un DOCUMENTO es relevante o no a una CONSULTA.*
> *Las consultas pueden ser SIMPLES (sin coma) o COMPUESTAS (con una coma ','). *
>
> *- CONSULTA SIMPLE: (por ej. 'Hurto') el documento es relevante si es claramente relevante a la consulta.*
> *- CONSULTA COMPUESTA ('consulta general, subtema'): el documento SOLO es relevante si cumple explícita y exactamente con el subtema especificado después de la coma. Por ejemplo, si la consulta es 'Derecho a la defensa, Doble instancia', solo documentos que mencionen claramente el concepto 'Doble instancia' dentro del contexto de 'Derecho a la defensa' serán relevantes.*
>
> *Debes responder siempre en ESPAÑOL y en este formato JSON estricto:*
> *"relevant":"yes|no","evidence":"..."*
>
> *- 'relevant': 'yes' o 'no' según tu decisión.*
> *- 'evidence': si respondes 'yes', incluye ÚNICAMENTE una cita textual EXACTA del documento que justifique claramente tu decisión.*
> *Si respondes 'no', coloca 'None'.*

Translated (English):

> *You are a Paraguayan criminal lawyer specializing in case law analysis.*
> *Your task is to decide whether a DOCUMENT is relevant or not to a QUERY.*

Queries can be SIMPLE (no comma) or COMPOSITE (with a comma ’,’).

- SIMPLE QUERY: (e.g., ’Theft’) the document is relevant if it is clearly relevant to the query.
- COMPOSITE QUERY (’general concept, subtopic’): the document is ONLY relevant if it explicitly and exactly matches the subtopic specified after the comma. For example, if the query is ’Right to defense, Double jeopardy’, only documents that clearly mention the concept ’Double jeopardy’ within the context of ’Right to defense’ are relevant.

You must always respond in SPANISH and in this strict JSON format:
"relevant":"yes|no","evidence":"..."

- ’relevant’: ’yes’ or ’no’ according to your judgment.
- ’evidence’: if you respond ’yes’, include ONLY an EXACT textual quote from the document that clearly justifies your decision.
If you respond ’no’, write ’None’.

The model output was post-processed to extract binary labels and the justification spans, which were stored as “evidence” alongside each label.