Performant Lightweight Encoders for Spanish in the Legal and Administrative Domains

Codificadores Ligeros y Eficaces para los Dominios Legal y Administrativo en Español

Sara Dueñas-Romero, Alba María Mármol-Romero, Adrián Moreno-Muñoz, Miguel Montoya-Gavilán, Jaime Collado-Montañez, Lucas Molino-Piñar, Isabel Cabrera-de-Castro, Samuel Sánchez-Carrasco, Fabián Suárez-Maroto, Sergio González-Del-Moral, M. Teresa Muñoz-Martín, M. Estrella Vallecillo-Rodríguez, M. Victoria Cantero-Romero, Manuel García-Vega, Miguel Ángel García-Cumbreras, M. Rosario García-Viedma, M. Dolores Molina-González, Fernando Martínez-Santiago, Manuel Carlos Díaz-Galiano, Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, L. Alfonso Ureña-López, Arturo Montejo-Ráez

CEATIC, University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{sduenas, amarmol, ammunoz, mmontoya, jcollado, lmolino, iccastro, scarrasc, fsmaroto, sgmoral, mtmunoz, mevallec, vcantero, mgarcia, magc, mrgarcia, mdmolina, dofer, mcdiaz, sjzafra, maite, emcamara, laurena, amontejo}@ujaen.es

Abstract: This work introduces two lightweight, domain-adapted models for Spanish legal and administrative information retrieval developed within the ALIA initiative: a bi-encoder model and a cross-encoder, both built on the MrBERT-es backbone. Training relies on a synthetic data generation pipeline grounded in legal and administrative sources, combined with a six-phase curriculum learning strategy of progressively increasing difficulty and positive-aware mining for hard-negative selection. Evaluation on both general-domain and domain-specific benchmarks using the MTEB framework shows that the bi-encoder achieves the best performance on legal-administrative retrieval datasets, outperforming larger general-purpose models such as bge-m3 and Qwen3-Embedding. The cross-encoder attains competitive results against substantially larger rerankers on specialized Spanish legal datasets. These findings demonstrate that targeted domain adaptation and careful training design can effectively compensate for model scale in specialized retrieval scenarios.
Keywords: Embeddings models, legal domain, administrative domain, information retrieval

Resumen: Se presentan dos modelos ligeros adaptados al dominio para la recuperación de información legal y administrativa en español, desarrollados en el marco de la iniciativa ALIA: un modelo de *embeddings* y un *reranker*, ambos construidos sobre MrBERT-es. El entrenamiento se basa en un flujo de generación de datos sintéticos fundamentado en fuentes legales y administrativas, combinado con una estrategia de aprendizaje curricular en seis fases de dificultad progresiva y minería de negativos difíciles. La evaluación en *benchmarks* tanto de dominio general como específico mediante el *framework* MTEB muestra que el *bi-encoder* alcanza el mejor rendimiento en los conjuntos de datos de recuperación legal-administrativa, superando a modelos de propósito general de mayor tamaño como bge-m3 y Qwen3-Embedding. El *cross-encoder* obtiene resultados competitivos frente a *rerankers* sustancialmente más grandes en conjuntos de datos legales especializados en español. Estos resultados demuestran que la adaptación al dominio y un diseño cuidadoso del entrenamiento pueden compensar eficazmente la escala del modelo en escenarios de recuperación especializados.
Palabras clave: Modelos de embeddings, dominio legal, dominio administrativo, recuperación de información

1 Introduction

Text embedding models map textual inputs into dense vector representations within a shared semantic space, enabling efficient similarity search that overcomes the lexical gap of traditional sparse methods such as BM25 (Karpukhin et al., 2020; Reimers and Gurevych, 2019a). These models have gained further importance as the retrieval backbone of Retrieval-Augmented Generation (RAG) pipelines (Lewis et al., 2020), where the quality of the retrieved passages directly conditions the system's accuracy and faithfulness.

The legal and administrative domains pose particular challenges for text representation: long and syntactically complex sentences, domain-specific terminology, cross-references between provisions, and subtle semantic distinctions with significant practical consequences. General-purpose models often fail to capture these nuances, and while domain-adapted models such as LEGAL-BERT have shown improvements on legal Natural Language Preprocessing (NLP) tasks (Chalkidis et al., 2020), most efforts have targeted English and classification rather than retrieval. Meanwhile, the demand for accurate legal information retrieval continues to grow, as exemplified by platforms such as Justicio[1] for Spanish legal documents.

This work is developed within the ALIA initiative `https://alia.gob.es`, a Spanish government project coordinated by the Barcelona Supercomputing Center (BSC-CNS) aimed at building transparent, open, and efficient language technologies for Spanish, reducing dependence on English-centric models and fostering technological sovereignty. We present two lightweight, domain-adapted models for Spanish legal and administrative retrieval: a bi-encoder and a cross-encoder reranker, both built on MrBERT-es (Tamayo et al., 2026). Training combines a synthetic data generation pipeline inspired by the Qwen3-Embedding model series (Zhang et al., 2025a) with a six-phase curriculum learning strategy (Bengio et al., 2009) incorporating Positive-Aware Mining (PAM) (Moreira et al., 2025). Evaluation on general-domain and domain-specific benchmarks using the MTEB framework (Muennighoff et al., 2022; Enevoldsen et al., 2025) demonstrates that our models achieve state-of-the-art performance on legal-administrative retrieval while remaining competitive on general tasks, despite being substantially smaller than existing alternatives.

To promote transparency and reproducibility, all resources developed in this work will be made publicly available through our Hugging Face space.[2] This includes the synthetic datasets generated for training and evaluation, the bi-encoder and cross-encoder models, and the full training and data collected.

2 Previous Work

This section reviews the current landscape of preexisting models, specialized datasets, and the training methodologies that define the current state of the art, with a particular focus on the Spanish language ecosystem.

2.1 Preexistent Models

The field of Information Retrieval (IR) and semantic text representation has seen significant advancement through two primary architectures: bi-encoders (embedding models) and cross-encoders (rerankers).

In the general-purpose bi-encoder landscape, the bge-m3 family (Chen et al., 2024; Xiao et al., 2024) stands out for its robust multilingual capabilities, while recently introduced models like Qwen3-Embedding have set new benchmarks in textual understanding. Furthermore, foundational architectures such as MPNet-base (Song et al., 2020) remain relevant for their effective combination of masked and permuted language modeling.

For the reranking stage, cross-encoders like bge-reranker-v2-m3 and the Qwen3-Reranker family are prioritized for their deep semantic matching and generalization capabilities. A growing trend involves adapting LLMs for this task, exemplified by Llama-Nemotron-Rerank-1b-v2,[3] which leverages massive generative architectures to refine retrieval predictions.

In the Spanish context, the MrBERT family represents a pivotal effort in domain and vocabulary adaptation. MrBERT-es serves as a strong general-purpose masked language model, while its derivative, MrBERT-legal,

[1] `https://justicio.es`

[2] `https://huggingface.co/collections/SINAI/alia-es-legal-administrative`

[3] `https://huggingface.co/nvidia/llama-nemotron-rerank-1b-v2`

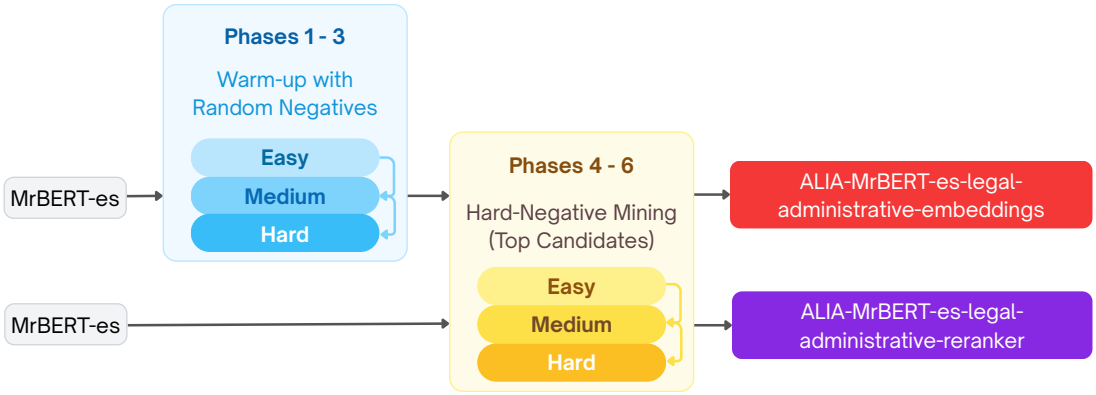


Figure 1: Progressive training pipeline architecture. The process uses MrBERT-es as the backbone and progresses through two macro stages of curriculum learning (Phases 1–3 and 4–6), each subdivided into incrementally increasing difficulty levels to optimize semantic convergence in the legal domain.

is specifically fine-tuned for Spanish legal corpora and evaluated on benchmarks like LexBOE and a legal subset of MTEB. These models provide the foundational backbone for recent efforts to achieve technological sovereignty in Spanish legal NLP.

2.2 Existing Datasets

Most legal datasets are predominantly English-centric, such as EURLEX, MultiEURLEX (Chalkidis, Fergadiotis, and Androutsopoulos, 2021), and the Pile of Law (Henderson et al., 2022), which cover broad legislative and court records. While MultiLegalPile extends this to 24 languages (Niklaus et al., 2024), many of these resources are designed for pretraining or classification rather than retrieval-specific tasks. More recently, benchmarks like ACORD (contract analysis) (Wang et al., 2025), CLERC (legal reasoning) (Hou et al., 2025), and the Massive Legal Embedding Benchmark (MLEB) (Butler, Butler, and Malec, 2025) have introduced the query-document structures necessary for training ranking models.

For the Spanish language, resources remain scarce. The Spanish Legal Domain Corpora (Gutiérrez-Fandiño et al., 2021) provides a comprehensive collection of constitutional and legislative texts. However, the Small Spanish Legal Dataset (SSLD) is one of the few specifically structured with query-positive-negative samples, making it directly suitable for training dense retrieval systems (Martel, 2025a).

2.3 Methodologies for Embedding Construction

Modern embedding construction is predominantly driven by *contrastive learning* (Weng, 2021), where models are optimized to align semantically related query–document pairs in a shared latent space. This paradigm, typically enhanced by *in-batch negatives* and *hard-negative mining* (Xiong et al., 2020), has become the standard for dense retrieval due to its scalability. Beyond standard objectives, recent research emphasizes knowledge distillation from cross-encoders and list-wise supervision to refine representation quality, as seen in the BGE series.

A key factor in the success of these models is the optimization of the training signal. Traditional training often struggles with noise or overly difficult negatives at early stages. To address this, curriculum learning has emerged as a strategy to present data in an increasing order of complexity, improving convergence and final generalization. Furthermore, while hard-negative mining is effective, it risks introducing "false negatives" that degrade the embedding space. Recent techniques like PAM refine this process by ensuring that selected negatives are truly informative without being semantically identical to the positive reference, thus stabilizing the contrastive objective.

Another major trend is the use of LLMs for representation learning through instruction tuning and multi-stage training, such as the Qwen model family (Yang et al., 2024; Zhang et al., 2025b), which achieves state-of-the-art results on benchmarks like MTEB.

These models often rely on synthetic data generation pipelines where LLMs generate diverse query–document pairs (Wang et al., 2024).

Additionally, Matryoshka Representation Learning (MRL) (Kusupati et al., 2024) allows vectors to be truncated while preserving semantic content. Finally, state-of-the-art retrieval systems typically adopt a two-stage architecture, combining efficient bi-encoder retrieval with high-precision cross-encoder reranking (Thakur et al., 2021).

3 Data

This section details the multi-stage process of corpus construction, the LLM-driven synthetic generation pipeline, and the strategic mining of hard negatives.

3.1 Training Dataset

The primary training data for our encoder-only models is a large-scale, heterogeneous data infrastructure designed for high-resource Spanish linguistic modeling. The plain text corpus serves as a consolidated repository of official administrative and legal documentation, comprising over 7 million unique instances and exceeding 5 billion tokens. To ensure high data fidelity, the raw documents underwent a rigorous technical curation pipeline using the datatrove library (Penedo et al., 2024), focusing on deduplication, quality filtering, and structural normalization. Finally, we developed a synthetic data generation and curation pipeline aimed at producing high-quality document-grounded queries and query-answer pairs for representation learning.

3.1.1 Sources

The corpus was constructed to provide a representative and balanced distribution of the Spanish institutional ecosystem. This includes a comprehensive collection of legislative, regulatory, and procedural texts from various administrative levels, ranging from national ministerial directives to regional parliamentary proceedings. By integrating these disparate sources into a unified, machine-readable format, the dataset enables the development of specialized language representations capable of capturing the nuanced legal-administrative register of the Spanish language.

Data Composition and Provenance The training set integrates documentation from a diverse array of Spanish official repositories. These sources are categorized into four primary thematic domains to ensure broad coverage of the regulatory landscape.

- **Official Bulletins**: We performed exhaustive web-crawling and massive downloading from the *Boletín Oficial del Estado* (BOE) for national legislation. This was supplemented by regional bulletins from *Andalucía*, *Castilla y León*, *Murcia*, *Cantabria*, *Islas Canarias*, and *Ceuta*, as well as provincial bulletins from *Granada*, *Jaén*, *Sevilla*, and *Córdoba*.
- **Specialized Registries and Codes**: To incorporate technical and commercial terminology, we included the *Registro Mercantil* (BORME), the *Biblioteca Jurídica* (legal codes), and the *Código Técnico de la Edificación*. Furthermore, the EuroPat parallel corpus was integrated to provide alignment with European patent standards.
- **Ministerial Documentation**: We targeted specific repositories from the Ministry for the Ecological Transition and the Demographic Challenge (covering climate change, energy, and environmental policy), the Ministry of Defense, the Ministry of Housing and Urban Agenda, and the Department of National Security.
- **Parliamentary and Financial Records**: The dataset includes the parliamentary proceedings from the Parliament of Andalusia (Spain), and the NORMA repository[4] for economic and financial regulations.

All data were retrieved from publicly accessible official sources under open-access protocols, ensuring the reproducibility of the collection process while maintaining a high standard of legal and linguistic relevance.

3.1.2 Data Collection and Processing

To ensure reproducibility and scalability, we implemented a standardized data acquisition and refinement pipeline. This infrastructure was designed to handle high-volume institutional data while maintaining strict standards for structural integrity and linguistic quality.

[4] `https://www.normadoc.gob.es/es-es/`

Quality Curation and Cleaning Pipeline Raw data from official repositories often contains structural noise, boilerplate text, and encoding artifacts. We deployed a multi-stage automated refinement pipeline to transform the raw collection into a high-quality modeling corpus.

- **Heuristic Language Filtering**: Each document is subjected to an automated language identification pass. Only documents meeting a strict confidence threshold for the Spanish language are retained, ensuring the linguistic purity of the encoder's pre-training signal.
- **Large-Scale Deduplication**: We utilized the MinHash algorithm to perform fuzzy deduplication at scale. By representing documents as sets of n-grams and hashing them into probabilistic signatures, the system identifies and removes near-duplicate content (e.g., repeated legislative headers or overlapping administrative circulars) that could otherwise lead to model overfitting.
- **Repetition and Structural Filtering**: To eliminate low-quality "web noise" and spam, we applied filters that analyze the internal distribution of n-grams and line-level patterns. Documents with excessive line repetitions, abnormal new-line densities, or disproportionate punctuation-to-token ratios are discarded.
- **Linguistic Quality Control**: The corpus is further refined by enforcing natural language constraints. This includes thresholds for average word length and the ratio of non-alphabetic characters. Furthermore, we mandate a minimum density of language-specific functional words (stopwords) to ensure that the training data consists of coherent prose rather than fragmented logs or semi-structured tables.

3.1.3 Synthetic Data Generation

The overall design follows the scaling and diversity principles introduced in Qwen3-Embedding, while adapting them to the legal and administrative domains.

The pipeline consists of three stages: persona assignment, multi-stage generation, and data curation.

First, we perform a semantic pre-filtering step to increase diversity in the generated data and expose the model to multiple plausible perspectives. For each source passage, we compute its embedding-based similarity against a large subset of user profiles extracted from PersonaHub (Ge et al., 2025) and assign the most semantically aligned personas. This provides a grounded contextual perspective that conditions the subsequent generation process.

Second, synthetic generation is carried out through a multi-stage procedure. Rather than directly prompting the LLM to generate the final outputs, we decompose the process into three sequential steps. The model first identifies suitable question types for the passage, such as factual, interpretative, summarization, or "yes/no" questions. Next, it selects the most relevant information spans or semantic foci within the passage. Finally, conditioned on the selected focus, the assigned persona, the target question type, and the difficulty level inferred by the model, it generates either a synthetic query or a synthetic query together with its corresponding answer. This staged design encourages structural diversity, enables variation in the complexity of the generated examples, strengthens document grounding, and reduces generic or weakly supported generations.

Third, we apply a curation stage to improve data quality and reduce redundancy. Since large-scale synthetic generation may produce repeated outputs, we remove only exact duplicates. We do not apply similarity-based filtering or near-duplicate removal at this stage. This step reduces repetition in the final corpus while preserving the full range of generated variations.

Overall, this workflow—persona assignment, question typing, focus selection, generation, and exact deduplication—enables the construction of a synthetic corpus with the diversity, difficulty, and grounding required for effective encoder training.

3.1.4 Hard Negatives Preparation

To enhance the discriminative capacity of the encoders, we construct hard negative examples that are semantically similar to the positive passages but do not satisfy the information need expressed by the query. Unlike random negatives, which are often trivial to distinguish, these examples provide a stronger learning signal by approximating re-

alistic retrieval scenarios in which candidate documents are topically related yet not relevant. This is particularly important in the legal and administrative domain, where documents often share terminology and structural patterns while differing in subtle but crucial aspects of meaning.

Hard negatives are mined from the same corpus used for synthetic data generation. For each synthetic query-positive pair, we encode all candidate passages using a dense embedding model implemented with the SentenceTransformers (Reimers and Gurevych, 2019b) framework. The resulting representations are indexed using FAISS (Douze et al., 2025) to enable efficient large-scale similarity search. Given a query embedding, we retrieve a ranked list of semantically similar passages and exclude the known positive passage. This produces an initial pool of candidate negatives that are close in the embedding space and potentially challenging for the model.

Since semantically similar passages in the legal-administrative domain may partially overlap in content or even express related regulations, we apply a filtering stage to reduce the risk of false negatives. Specifically, we retain only those candidates whose similarity to the query falls within a controlled range, discarding both very distant passages (which would be uninformative) and overly similar ones (which may correspond to alternative valid answers). In addition, we impose a margin-based constraint with respect to the positive passage to ensure that selected negatives remain sufficiently distinguishable while preserving their difficulty. This filtering strategy balances hardness and reliability, preventing the introduction of misleading supervision signals.

Formally, given a query q and a ranked list of candidate passages $\{d_i\}_{i=1}^{K}$ sorted by decreasing similarity $\mathrm{sim}(q, d_i)$, we define the set of admissible hard negatives as:

$$\mathcal{N}(q) = \{d_i \,|\, k_{\min} \leq i \leq k_{\max},\ \mathrm{sim}(q, d_i) \leq \tau\}\,,$$

where $k_{\min}$ and $k_{\max}$ define a bounded similarity window over the retrieval ranking, and τ is an upper similarity threshold that prevents selecting near-duplicate passages. In our setup, we use $k_{\min} = 10$, $k_{\max} = 50$, and $\tau = 0.8$.

To further increase diversity and avoid overfitting to a narrow set of near-duplicate candidates, we combine two complementary sampling strategies. First, we select top-ranked candidates from the FAISS retrieval results, which provide highly informative and challenging negatives. Second, we include randomly sampled passages from within the admissible set $\mathcal{N}(q)$, introducing variability and improving generalization. For each query, we retain a fixed number of five hard negatives. Preliminary experiments showed that using multiple negatives per query significantly improves model performance compared to a single-negative setup, providing a richer and more stable training signal.

Due to the scale of the corpus, the mining process is performed in chunks, allowing efficient processing without compromising coverage of the candidate space. The resulting data is organized into structured training instances in which each query is paired with its corresponding positive passage and an associated set of mined hard negatives. This representation is directly compatible with both bi-encoder contrastive objectives and cross-encoder reranking setups, facilitating seamless integration into the subsequent training stages.

Table 1 summarizes the resulting datasets for the legal-administrative domain, showing the number of mined hard negatives under different difficulty configurations. Difficulty levels correspond to progressively stricter similarity constraints, resulting in increasingly challenging negative examples. These datasets are later incorporated into the curriculum training strategy, where harder negatives are introduced in later stages to progressively increase task difficulty.

Difficulty	Strategy	Hard Negatives
Easy	Random / PAM	36,665
Medium	Random / PAM	842,911
Hard	Random / PAM	19,361

Table 1: Hard negative statistics for the legal-administrative domain. Difficulty levels correspond to progressively stricter similarity constraints during mining. PAM=Positive-Aware Mining.

4 Training

Training follows a curriculum learning strategy structured in six phases of progressively increasing difficulty, combined with contrastive learning and the PAM strategy (see Figure 1), which filters false negatives

by retaining only candidates whose similarity score falls below the positive score by a margin δ:

$$\mathcal{H}(q) = \{d \in \mathcal{C}(q) \mid s(q,d) < s(q,d^+) - \delta\}, \quad (1)$$

where $s(q,\cdot)$ denotes cosine similarity, d^+ is the positive document, and $\delta = 0.1$.

Both models follow the same six-phase schedule (Table **??**), differing only in their loss function. Phases 1–3 train on randomly sampled negatives of increasing query–document difficulty; phases 4–6 replace them with hard negatives mined via PAM. The cross-encoder is trained only on phases 4–6, since as a reranker it only needs to distinguish candidates already deemed relevant by the bi-encoder.

Phase	Negatives	Difficulty	Bi-enc.	Cross-enc.
1	Random	Easy	2 ep.	—
2	Random	Medium	2 ep.	—
3	Random	Hard	2 ep.	—
4	PAM	Easy	1 ep.	1 ep.
5	PAM	Medium	1 ep.	1 ep.
6	PAM	Hard	1 ep.	1 ep.

Table 2: Curriculum phases. PAM=Positive-Aware Mining.

The bi-encoder uses `CachedMultipleNegativesRankingLoss` with a `NO_DUPLICATES` batch sampler, which prevents repeated queries within a batch as required by the contrastive objective. The cross-encoder is trained as a binary classifier with `BinaryCrossEntropyLoss` over flattened query–document pairs labeled 1.0 (positive) or 0.0 (negative). Hyperparameters were selected via Optuna (see Appendix A).

5 Evaluation

The evaluation pipeline is built upon the MTEB (Massive Text Embedding Benchmark) framework (Muennighoff et al., 2022; Enevoldsen et al., 2025), which provides a unified and extensible interface for benchmarking text representation models across multiple task families. To align MTEB with the requirements of the ALIA project without overcomplicating the narrative with framework-specific details, we structured our adaptations into three core functional layers:

- **Task Abstractions (Retrieval & Similarity):** We implemented task-specific custom subclasses extending `AbsTaskRetrieval` (MTEB's abstract base class for information retrieval tasks) and `AbsTaskSTS` (the abstract base class for Semantic Textual Similarity). Extending these native interfaces allows our pipeline to run standard bi-encoder retrieval, reranking, and similarity evaluation logic directly on our domain-specific datasets.
- **Metadata & Data Ingestion:** Our custom subclasses dynamically generate `TaskMetadata`, which is the framework's native configuration class used to store essential task attributes like name, language, evaluation splits, and primary metrics. At the same time, these classes act as data ingestion layers, transforming local JSONL and Parquet resources into the internal dataset formats required by MTEB.
- **Pipeline Execution:** The actual evaluation is triggered via `mteb.evaluate(...)`, which is the core framework function responsible for running the model inference loop and orchestrating the benchmark. We invoke this function dynamically, passing it our custom-instantiated tasks alongside model-specific inference parameters.

The execution stage yields a robust suite of metrics tailored to each task family. For bi-encoder retrieval and reranking tasks, the pipeline records ranking-focused metrics including `NDCG`, `MAP`, `MRR`, `Recall`, and `Precision`, computed at multiple cutoff values. For semantic textual similarity, the pipeline captures correlation-based scores such as `cosine_spearman` and `cosine_pearson`, alongside related distance-based variants.

Beyond task-level integration, the evaluation scripts incorporate several architecture-oriented adaptations designed for our local infrastructure. Bi-encoder models are loaded locally through the `sentence-transformers` library, with optional reconstruction and persistence into a reusable `SentenceTransformer` format when the original checkpoint is not natively packaged. Cross-encoder models are likewise resolved from local storage and instantiated through the standard `CrossEncoder`

interface, ensuring a highly decoupled and reproducible benchmark environment.

5.1 Benchmarks

Rather than relying on the standard public datasets distributed with MTEB, we adapted the framework to operate on local ALIA evaluation resources. To ensure compatibility with the framework, we follow a two-stage protocol designed to reconcile heterogeneous source datasets with the input contracts required by the ALIA custom tasks. This protocol is motivated by the fact that several benchmark resources differ in split definitions, attribute naming, and relevance representation.

In the first stage, raw downloaded resources are transformed into project-consistent artifacts. For datasets originally provided with non-evaluation splits (i.e., train) or with schemas that do not directly match the evaluation protocol, the pipeline performs attribute selection, field renaming, structural harmonization, and, when required, score normalization. The outcome is a set of stored files with aligned semantics, suitable for deterministic loading in downstream evaluation.

The second stage is executed at runtime by the custom task classes. In this phase, the pre-formatted files are mapped to the exact structures expected by each evaluation task and by MTEB.

For the reranking task, when the top-ranked list is not directly available in the source resource, it is explicitly constructed during task loading to ensure cross-encoder compatibility. Concretely, for each query, the runtime formatter preserves the known relevant document and augments the candidate set with up to 100 randomly sampled distractor document identifiers from the same corpus (using a fixed seed for reproducibility). The resulting list is shuffled and exposed as `top_ranked`, allowing the cross-encoder to re-score and re-order a realistic candidate pool.

In this work, we use a combination of general-domain and domain-specific datasets to evaluate both the bi-encoder and cross-encoder models.

General-domain datasets. For the bi-encoder training, we use the following datasets:

- `miracl-es`[5] is a reformatted version of the original MIRACL dataset (Zhang et al., 2022), adapted to the format expected for MTEB reranking tasks and restricted to the Spanish language.
- We constructed a merged dataset combining three Spanish resources to evaluate Semantic Textual Similarity (STS): the test set of `paws-x`[6] (Yang et al., 2019), a cross-lingual adversarial dataset for paraphrase identification; the Spanish test set of the SemEval 2022 Task on Multilingual News Article Similarity[7] (Chen et al., 2022); and the Spanish subset of the SemRel2024 dataset, released as part of SemEval-2024 Task 1[8] (Ousidhoum et al., 2024a; Ousidhoum et al., 2024b).

For the cross-encoder models, we rely on the following widely used retrieval datasets:

- MIRACL (Zhang et al., 2023), a multilingual retrieval dataset covering search tasks across 18 languages.
- ESCIReranking (Reddy et al., 2022), a neural reranking dataset commonly used to train and evaluate models that refine an initial retrieval ranking using contextual representations from pretrained language models.

Domain-specific datasets. To adapt the models to the legal-administrative domain, we use both external and internally generated datasets.

- `justicio-rag-embedding-qa-tmp`[9] (Justicio) is a Spanish legal-domain dataset designed for RAG, containing question-answer pairs linked to legal documents.
- `small-spanish-legal-dataset` (Martel, 2025b) (SSLD) is a dataset composed of Spanish legal texts and semantically related sentence pairs, designed

[5] `https://huggingface.co/datasets/jinaai/miracl-es`

[6] `https://huggingface.co/datasets/google-research-datasets/paws-x`

[7] `https://huggingface.co/datasets/mteb/sts22-crosslingual-sts`

[8] `https://huggingface.co/datasets/SemRel/SemRel2024`

[9] `https://huggingface.co/datasets/dariolopez/justicio-rag-embedding-qa-tmp`

for evaluating semantic similarity and retrieval tasks in the legal domain.

- `Pairs1.5k` and `Pairs10k` for retrieval: These subsets comprise 1,500 and 10,000 synthetically generated evaluation pairs, respectively, following the methodology outlined in Section 3.1.3. The selection was conducted via a stratified approach, drawing an equal number of pairs from each source. The variation in sample sizes is intended to compel the model to operate within a broader search space, thereby systematically increasing the difficulty of retrieving the correct passage.
- `Pairs650` and `Pairs1.3k` for reranking: These subsets comprise 650 and 1,300 synthetically generated evaluation pairs, respectively, following the methodology outlined in Section 3.1.3. Passage selection employed the previously described stratified sampling approach, which randomly draws from the top-ranked passages. These reduced sample sizes were specifically chosen to ensure the computational overhead remained consistent with the other evaluations.

5.2 Results and Discussion

Table **??** and Table **??** report the performance of the evaluated models in terms of NDCG@10 across both general-domain and legal-administrative datasets. A detailed comparison of these models across all evaluation depths is provided in the supplementary tables of Annex D.

Overall, the proposed ALIA -MrBERT-es-legal-embeddings model achieves the best performance on the domain-specific evaluation datasets (*Pairs10k* and *Pairs1.5k*), outperforming all baseline models by a significant margin. This indicates that the domain-adapted embeddings are particularly effective for legal-administrative retrieval tasks.

In contrast, general-purpose embedding models such as bge-m3 and Qwen3-Embedding-0.6B obtain slightly better performance on general-domain tasks such as QA. However, their performance on domain-specific datasets is lower compared to the proposed model. It should be noted, that because the full training distributions of external proprietary or large-scale models are often undisclosed, we cannot definitively rule out data contamination. The possibility that these baseline models were exposed to some of our evaluation sources during their pre-training or fine-tuning stages introduces a potential bias that may affect the comparative integrity of the evaluation. These results highlight the benefits of domain adaptation when dealing with specialized legal-administrative information retrieval scenarios.

For the cross-encoder model (Table **??**), despite being substantially smaller, the proposed ALIA-MrBERT-es-legal-administrative-reranker achieves competitive performance compared to much larger general-purpose rerankers such as bge-reranker-v2-m3, Qwen3-Reranker-0.6B, and llama-nemotron-rerank-1b-v2, which contain billions of parameters. In particular, our domain-adapted model performs comparably or better on specialized Spanish legal datasets (e.g., *SSLD*, *Triplets650*, and *Triplets1.3k*). This suggests that domain adaptation can compensate for model scale when dealing with highly specialized retrieval tasks.

6 Bias and ethics

The models presented in this work are designed for information retrieval in the Spanish legal and administrative domains, and several ethical and bias-related considerations must be addressed accordingly.

All training data originates from sources carrying open licenses that permit research and commercial use (e.g., CC-BY 4.0 or equivalent), alongside official public-domain corpora from Spanish public institutions such as legislative repositories and administrative databases. Synthetic data generated during training is derived exclusively from these openly licensed sources, and no personally identifiable information is present, as the generation pipeline operates over anonymized document collections and produces fully artificial query-document pairs. This design ensures both legal compliance and alignment with best practices for open dataset construction.

Despite the use of openly licensed and official sources, the training corpus may reflect structural biases inherent to the Spanish legal and administrative system, including potential overrepresentation of certain legal traditions, administrative frameworks, or

Dataset	bge-m3 (0.6B)	qwen3-embedding (0.6B)	mpnet-base (0.3B)	ours (150M)
miracl-es	0.9839	**0.9869**	0.9419	0.9504
STS	0.4629	**0.5033**	0.4632	0.4787
Justicio	**0.6521**	0.6344	0.5396	0.6401
SSLD	0.6955	**0.7228**	0.3060	0.6409
Pairs10k	0.7200	0.7979	0.1655	**0.8716**
Pairs1.5k	0.8372	0.8874	0.2672	**0.9383**

Table 3: Bi-encoder performance (NDCG@10) across evaluation datasets. For the STS dataset, performance is reported as the cosine similarity of Spearman's rank correlation. The best results are shown in bold, and the second-best results are underlined.

Dataset	bge-reranker (0.6B)	qwen3-rerank (0.6B)	nemotron-rerank (1B)	ours (150M)
MIRACL	**0.6777**	0.6724	0.6258	0.5144
ESCI	**0.8321**	0.8133	0.8004	0.8018
Justicio	**0.8055**	0.7898	0.7998	0.7880
SSLD	0.8765	0.8700	**0.9198**	0.9140
Triplets650	**0.9814**	0.9812	0.9536	0.9811
Triplets1.3k	0.9820	0.9497	0.9807	**0.9840**

Table 4: Reranker performance (NDCG@10) across evaluation datasets. Best results are in bold; second-best are underlined.

institutional registers, as well as underrepresentation of regional legal contexts involving co-official languages such as Catalan, Galician, or Basque. Users should therefore be aware that model performance may degrade on legal texts from these jurisdictions or on documents that diverge stylistically from the training distribution. Furthermore, since the models are intentionally optimized for Spanish legal and administrative retrieval, their applicability to other languages, legal systems, or general-domain tasks should not be assumed. The use of synthetically generated training pairs also introduces a risk of distributional shift; although curriculum learning and PAM were applied to mitigate this, downstream users are encouraged to perform domain-specific evaluation before deployment.

Regarding responsible use, these models are intended as retrieval assistance tools for professionals navigating legal and administrative information, and they should not be used as substitutes for qualified legal advice. Any decision with legal consequences must involve human oversight, and deployment in automated decision-making pipelines without appropriate human review is strongly discouraged, particularly in high-stakes scenarios such as judicial processes or administrative resolutions. Finally, the training pipeline, synthetic data generation procedure, and evaluation benchmarks are described in sufficient detail to support reproducibility, and models will be released under permissive open licenses to foster a competitive ecosystem for Spanish legal NLP.

7 Conclusions and future work

This work introduces ALIA-MrBERT-es-legal-administrative, a suite of domain-adapted text encoding models comprising a bi-encoder for embeddings and a cross-encoder for reranking, optimized for Spanish legal and administrative information retrieval. Developed on the MrBERT-es backbone, these models are refined through a six-phase curriculum learning strategy and PAM for effective hard-negative selection. Our evaluations demonstrate that the bi-encoder achieves state-of-the-art performance on domain-specific benchmarks, outperforming significantly larger models such as BGE-M3 and Qwen3-Embedding. Additionally, the 150M parameter cross-encoder yields competitive results against rerankers with billions of parameters, confirming that targeted domain adaptation effectively compensates for model scale in specialized scenarios. In alignment with the ALIA initiative's commitment to transparency and technological sovereignty, we are making all model weights and training recipes publicly available under a permissive Apache 2.0 license. Regarding future work, our focus will shift to adapting these lightweight architectures to other specialized domains, specifically the biomedical and heritage sectors.

Acknowledgements

This work is funded by the Ministerio para la *Transformación Digital y de la Función Pública and PRTR*, funded by EU NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the EU NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHARLITI-101215052), funded by the EU under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP-PIDI-2024-00852) funded by the EU under the S4Andalucía 2021-2027.

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

Bengio, Y., J. Louradour, R. Collobert, and J. Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Butler, U., A. Butler, and A. L. Malec. 2025. The Massive Legal Embedding Benchmark (MLEB).

Chalkidis, I., M. Fergadiotis, and I. Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November. Association for Computational Linguistics.

Chen, J., S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Chen, X., A. Zeynali, C. Camargo, F. Flöck, D. Gaffney, P. Grabowicz, S. Hale, D. Jurgens, and M. Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States, July. Association for Computational Linguistics.

Douze, M., A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Enevoldsen, K., I. Chung, I. Kerboua, M. Kardos, A. Mathur, D. Stap, J. Gala, W. Siblini, D. Krzemiński, G. I. Winata, S. Sturua, S. Utpala, M. Ciancone, M. Schaeffer, G. Sequeira, D. Misra, S. Dhakal, J. Rystrøm, R. Solomatin, Ömer Çağatan, A. Kundu, M. Bernstorff, S. Xiao, A. Sukhlecha, B. Pahwa, R. Poświata, K. K. GV, S. Ashraf, D. Auras, B. Plüster, J. P. Harries, L. Magne, I. Mohr, M. Hendriksen, D. Zhu, H. Gisserot-Boukhlef, T. Aarsen, J. Kostkan, K. Wojtasik, T. Lee, M. Šuppa, C. Zhang, R. Rocca, M. Hamdy, A. Michail, J. Yang, M. Faysse, A. Vatolin, N. Thakur, M. Dey, D. Vasani, P. Chitale, S. Tedeschi, N. Tai, A. Snegirev, M. Günther, M. Xia, W. Shi, X. H. Lù, J. Clive, G. Krishnakumar, A. Maksimova, S. Wehrli, M. Tikhonova, H. Panchal, A. Abramov, M. Ostendorff, Z. Liu, S. Clematide, L. J. Miranda, A. Fenogenova, G. Song, R. B. Safi, W.-D. Li, A. Borghini, F. Cassano, H. Su, J. Lin, H. Yen, L. Hansen, S. Hooker, C. Xiao, V. Adlakha, O. Weller, S. Reddy, and

N. Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Ge, T., X. Chan, X. Wang, D. Yu, H. Mi, and D. Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, A. Gonzalez-Agirre, and M. Villegas. 2021. Spanish legalese language model and corpora.

Henderson, P., M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho. 2022. Pile of law: learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hou, A. B., O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, and B. Van Durme. 2025. CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7913–7928, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Karpukhin, V., B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Kusupati, A., G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, and A. Farhadi. 2024. Matryoshka representation learning.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Martel, W. 2025a. Small spanish legal dataset.

Martel, W. 2025b. Small spanish legal dataset.

Moreira, G. d. S. P., R. Osmulski, M. Xu, R. Ak, B. Schifferer, and E. Oldridge. 2025. Improving text embedding models with positive-aware hard-negative mining.

Muennighoff, N., N. Tazi, L. Magne, and N. Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niklaus, J., V. Matoshi, M. Stürmer, I. Chalkidis, and D. Ho. 2024. MultiLegalPile: A 689GB multilingual legal corpus. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand, August. Association for Computational Linguistics.

Ousidhoum, N., S. H. Muhammad, M. Abdalla, I. Abdulmumin, I. S. Ahmad, S. Ahuja, A. F. Aji, V. Araujo, A. A. Ayele, P. Baswani, M. Beloucif, C. Biemann, S. Bourhim, C. D. Kock, G. S. Dekebo, O. Hourrane, G. Kanumolu, L. Madasu, S. Rutunda, M. Shrivastava, T. Solorio, N. Surange, H. G. Tilaye, K. Vishnubhotla, G. Winata, S. M. Yimam, and S. M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Ousidhoum, N., S. H. Muhammad, M. Abdalla, I. Abdulmumin, I. S. Ahmad, S. Ahuja, A. F. Aji, V. Araujo, M. Beloucif, C. De Kock, O. Hourrane, M. Shrivastava, T. Solorio, N. Surange, K. Vishnubhotla, S. M. Yimam, and S. M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Penedo, G., H. Kydlíček, A. Cappelli, M. Sasko, and T. Wolf. 2024. Datatrove: large scale data processing.

Reddy, C. K., L. Màrquez, F. Valero, N. Rao, H. Zaragoza, S. Bandyopadhyay, A. Biswas, A. Xing, and K. Sub-

bian. 2022. Shopping queries dataset: A large-scale ESCI benchmark for improving product search.

Reimers, N. and I. Gurevych. 2019a. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Reimers, N. and I. Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Song, K., X. Tan, T. Qin, J. Lu, and T.-Y. Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Tamayo, D., I. Lacunza, P. Rivera-Hidalgo, S. D. Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, and M. Villegas. 2026. Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation.

Thakur, N., N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Wang, L., N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training.

Wang, S. H., M. Zubkov, K. Fan, S. Harrell, Y. Sun, W. Chen, A. Plesner, and R. Wattenhofer. 2025. ACORD: An expert-annotated retrieval dataset for legal contract drafting. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24739–24762, Vienna, Austria, July. Association for Computational Linguistics.

Weng, L. 2021. Contrastive representation learning, May.

Xiao, S., Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Xiong, L., C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Yang, A., B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. 2024. Qwen2 technical report.

Yang, Y., Y. Zhang, C. Tar, and J. Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Zhang, X., N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages.

Zhang, X., N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 09.

Zhang, Y., M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Zhang, Y., M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou. 2025b. Qwen3

embedding: Advancing text embedding and reranking through foundation models.

A Annex 1: Training Parameters

Hyperparameter tuning for both the bi-encoder and the cross-encoder was performed using Optuna (Akiba et al., 2019), a define-by-run optimization framework based on the Tree-structured Parzen Estimator (TPE) sampler. To limit computational cost, each trial was trained for a single epoch on a subset of 5.000 training examples for the bi-encoder and 1.000 for the cross-encoder, with a fixed random seed of 42 to ensure reproducibility. A 5% held-out split of the training set was used as the validation corpus for each trial.

Each Optuna study ran for 20 trials, with a Median Pruner configured to allow 3 startup trials before beginning to prune unpromising candidates. The search space explored for both models is summarized in Table **??**.

All models share AdamW, `bf16` precision, a 2048-token sequence limit, and gradient checkpointing. Retrieval and reranking quality during validation was measured with NDCG@10, MRR@10 and MAP@10, computed by an information-retrieval evaluator with a batch size of 128. The trial that maximized NDCG@10 on the validation set was selected as the final configuration for full training.

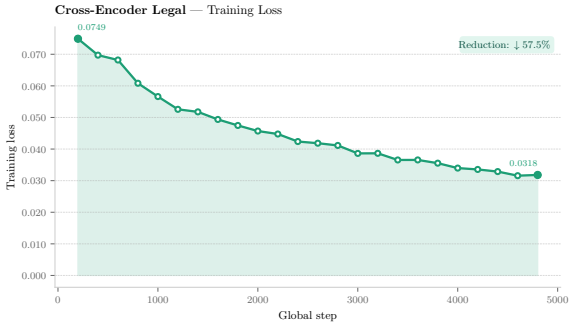
Results for each study were persisted to a SQLite database, allowing trials to be resumed or inspected after the fact without rerunning the optimization.

Final model-specific values are given in Tables **??** and **??**.

B Annex 2: Training Curves

Figure 2 shows the evolution of the training loss during the optimization process for both the bi-encoder and cross-encoder models used in our legal retrieval framework. As illustrated in Figure 2a, the bi-encoder training loss decreases steadily throughout the training process, indicating stable convergence of the embedding model on the legal-domain data. The loss reduction suggests that the model progressively learns more discriminative representations for query-document similarity. Similarly, Figure 2b presents the training dynamics of the cross-encoder reranker. The loss exhibits a smooth and consistent downward trend, demonstrating stable optimization and effective learning of fine-grained relevance signals between queries and candidate documents.

(a) Bi-encoder training loss.



(b) Cross-encoder training loss.

Figure 2: Training loss of the legal retrieval models.

Overall, both models show stable convergence behaviour without signs of instability or divergence during training, supporting the effectiveness of the training configuration adopted in this work.

B.1 Computational Resources

All experiments were conducted on a single compute node equipped with two AMD EPYC 7282 processors (2.80 GHz), 256 GB of DDR4 RAM, and four NVIDIA Tesla A100 GPUs (40 GB, Ampere architecture). Training the bi-encoder required approximately 77 hours to complete, while the cross-encoder converged in approximately 14 hours.

C Annex 3: Synthetic Data

This appendix documents the prompts used in our synthetic data generation pipeline for legal and administrative retrieval data. Following the staged prompting strategy used in Qwen3 Embedding, we decompose generation into intermediate configuration steps before producing the final synthetic query or

Hyperparameter	Bi-encoder	Cross-encoder	Type
Learning rate	$[10^{-6},\ 5\times10^{-5}]$	$[10^{-6},\ 5\times10^{-5}]$	log-uniform
Warmup ratio	$[0.05,\ 0.20]$	$[0.05,\ 0.20]$	uniform
Weight decay	$[0.00,\ 0.10]$	$[0.00,\ 0.10]$	uniform
Mini-batch size	$\{1, 4, 8, 12\}$	—	categorical
Gradient accum. steps	—	$\{4, 8, 16, 32\}$	categorical

Table 5: Hyperparameter search space used in the Optuna studies.

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size	256
Cache mini-batch	4
Warmup ratio	0.160
Weight decay	0.020

Table 6: Final hyperparameters for the bi-encoder.

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size	32
Gradient accum. steps	8
Warmup ratio	0.110
Weight decay	0.036

Table 7: Final hyperparameters for the cross-encoder.

query–answer pair.

Our pipeline is organized into four prompting stages: question-type induction, configuration selection, query generation, and query–answer generation. This design improves controllability, diversity, and document grounding during synthetic data construction.

C.1 Prompt for Question-Type Induction

This prompt receives a passage and returns a small set of plausible `Question_Type` values suitable for retrieval-oriented supervision.

```
The Passage is an excerpt from a
↪ {domain_language} {domain_type}
↪ document,
taken from {source}, and will be used in a
↪ Retrieval-Augmented Generation (RAG)
system to answer user queries.

Based on the Passage, generate a list of 4
↪ to 5 Question_Type values that are
contextually relevant to this kind of
↪ document and suitable for RAG retrieval
↪ scenarios.

Question Type Definitions:
- yes_or_no: ...
- background: ...
- acquire_knowledge: ...
- keywords: ...
- summary: ...
- procedural: ...
- compliance_check: ...
- interpretation: ...
- deadline: ...
- requirements: ...
- eligibility: ...
- consequences: ...
- application_process: ...
- clarification: ...

Important instructions:
Respond only with a single JSON object in
↪ English, containing a key
"Question_Types" with a list of question
↪ type strings.

Do not include explanations, markdown
↪ formatting, or extra text.
If unsure, respond with an empty list.

Passage:
{passage}
```

C.2 Prompt for Configuration Selection

After obtaining candidate question types, we prompt the model to select a configuration composed of `Character`, `Question_Type`, and `Difficulty`.

```
Given a Passage and Character, select the
↪ appropriate option from three fields:
Character, Question_Type, Difficulty, and
↪ return the output in JSON format.

First, select the Character most likely to
↪ be interested in the Passage.
Then select the Question_Type that the
↪ Character might ask about the Passage.
Finally, choose the Difficulty of the
↪ possible question based on the Passage,
the Character, and the Question_Type.

Character: Given by input Character

Question_Type:
- yes_or_no: ...
```

```
- background: ...
- acquire_knowledge: ...
- keywords: ...
- summary: ...
- procedural: ...
- ...

Difficulty:
- high_school: ...
- university: ...
- phd: ...

Here are some examples:
...

Now, generate the output based on the
↪  Passage and Character from user.
The Passage will be in Spanish and the
↪  Character will be in English.
Ensure to generate only the JSON output with
↪  content in English.

Passage:
{passage}

Character:
{character}
```

C.3 Prompt for Query Generation

Once the configuration tuple has been fixed, the model generates a synthetic query from the perspective of the selected character.

```
Role: You are a specialized model for
↪  generating realistic search queries in
↪  the
{domain_type} domain for RAG systems.

Task: Generate a natural, contextual query
↪  from a Character's perspective that
would retrieve the provided Passage. The
↪  query must meet the specified
↪  Requirements.

Key Principles
1. Character Authenticity: ...
2. Information Gap: ...
3. Retrieval Alignment: ...
4. Requirement Compliance: ...

Input Parameters
Character: {character}
Passage: {passage}
Requirements: Type: {type}; Difficulty:
↪  {difficulty}

Output Format
{
  "query": "Your generated query"
}

Generation Guidelines
- Use natural, conversational language
↪  appropriate to the Character's context
- Frame the query as the Character would ask
↪  it
- Ensure the query is answerable by the
↪  Passage content
- Match the difficulty level
- The query must be written in
↪  {domain_language}

Important: Generate only the query.
```

C.4 Prompt for Query–Answer Generation

For tasks requiring supervised query–answer data, we extend the previous prompt so that the model also produces an answer strictly grounded in the source passage.

```
Role: You are a specialized model for
↪  generating realistic search queries AND
↪  their answers in the {domain_type}
↪  domain for RAG systems.

Task: Generate a natural, contextual query
↪  from a Character's perspective that
↪  would retrieve the provided Passage, AND
↪  generate the answer to that query based
↪  strictly on the Passage content.

Key Principles
1. Character Authenticity: ...
2. Information Gap: ...
3. Retrieval Alignment: ...
4. Answer Fidelity: ...
5. Requirement Compliance: ...

Input Parameters
Character: {character}
Passage: {passage}
Requirements: Type: {type}; Difficulty:
↪  {difficulty}

Output Format
{
  "query": "Your generated query",
  "answer": "Your generated answer"
}

Generation Guidelines for Query
- Use natural, conversational language
↪  appropriate to the Character's context
- Frame the query as the Character would ask
↪  it
- Ensure the query is answerable by the
↪  Passage content
- Match the difficulty level

Generation Guidelines for Answer
- Base your answer strictly on the Passage
- Do not add external knowledge
- Keep the answer concise and directly
↪  relevant to the query

Important: Generate both the query and
↪  answer together.
```

D Annex 4: Full Results

Metric	bge-m3	qwen3-embedding	mpnet-base	ours
cosine_spearman	0.4629	**0.5033**	0.4632	<u>0.4787</u>
cosine_pearson	0.6349	<u>0.6491</u>	**0.6587**	0.6000
euclidean_spearman	0.4629	**0.5033**	<u>0.4846</u>	0.4914
euclidean_pearson	0.6168	0.6318	**0.6721**	<u>0.6178</u>
manhattan_spearman	0.4633	**0.5034**	0.4831	<u>0.4898</u>
manhattan_pearson	0.6165	0.6311	**0.6721**	<u>0.6177</u>
spearman	0.4629	**0.5033**	0.4632	<u>0.4787</u>
pearson	0.6349	<u>0.6491</u>	**0.6587**	0.6000

Table 8: Semantic textual similarity results across bi-encoders models. Best results are shown in bold and second-best are underlined.

Dataset	Metric@k	bge-m3	qwen3-embedding	mpnet-base	ours
miracl-es	NDCG@1	0.9614	**0.9676**	0.8873	0.9028
	NDCG@5	0.9834	**0.9869**	0.9384	0.9483
	NDCG@10	0.9839	**0.9869**	0.9419	0.9504
	MAP@1	0.9614	**0.9676**	0.8873	0.9028
	MAP@5	0.9781	**0.9828**	0.9249	0.9374
	MAP@10	0.9784	**0.9828**	0.9263	0.9383
	MRR@1	0.9614	**0.9676**	0.8873	0.9043
	MRR@5	0.9781	**0.9828**	0.9249	0.9382
	MRR@10	0.9784	**0.9828**	0.9263	0.9391
	Recall@1	0.9614	**0.9676**	0.8873	0.9028
	Recall@5	**0.9985**	**0.9985**	0.9784	0.9799
	Recall@10	**1.0000**	0.9985	0.9892	0.9861
	Precision@1	0.9614	**0.9676**	0.8873	0.9028
	Precision@5	**0.1997**	**0.1997**	0.1957	0.1960
	Precision@10	**0.1000**	0.0998	0.0989	0.0986
Justicio	NDCG@1	**0.5479**	0.5321	0.4185	0.5365
	NDCG@5	**0.6415**	0.6224	0.5229	0.6271
	NDCG@10	**0.6521**	0.6344	0.5396	0.6401
	MAP@1	**0.5479**	0.5321	0.4185	0.5365
	MAP@5	**0.6163**	0.5982	0.4931	0.6021
	MAP@10	**0.6206**	0.6032	0.5000	0.6075
	MRR@1	**0.5476**	0.5324	0.4185	0.5361
	MRR@5	**0.6160**	0.5988	0.4933	0.6020
	MRR@10	**0.6205**	0.6035	0.5001	0.6074
	Recall@1	**0.5479**	0.5321	0.4185	0.5365
	Recall@5	**0.7160**	0.6938	0.6121	0.7015
	Recall@10	**0.7489**	0.7311	0.6639	0.7415
	Precision@1	**0.5479**	0.5321	0.4185	0.5365
	Precision@5	**0.1432**	0.1388	0.1224	0.1403
	Precision@10	**0.0749**	0.0731	0.0664	0.0741
SSLD	NDCG@1	0.6221	**0.6310**	0.2456	0.5087
	NDCG@5	0.6844	**0.7111**	0.2939	0.6207
	NDCG@10	0.6955	**0.7228**	0.3060	0.6409
	MAP@1	0.6221	**0.6310**	0.2456	0.5087
	MAP@5	0.6669	**0.6898**	0.2798	0.5894
	MAP@10	0.6715	**0.6946**	0.2848	0.5978
	MRR@1	0.6221	**0.6310**	0.2456	0.5087
	MRR@5	0.6669	**0.6898**	0.2798	0.5894
	MRR@10	0.6715	**0.6946**	0.2848	0.5978
	Recall@1	0.6221	**0.6310**	0.2456	0.5087
	Recall@5	0.7363	**0.7741**	0.3359	0.7143
	Recall@10	0.7709	**0.8103**	0.3735	0.7766
	Precision@1	0.6221	**0.6310**	0.2456	0.5087
	Precision@5	0.1473	**0.1548**	0.0672	0.1429
	Precision@10	0.0777	**0.0810**	0.0374	0.0771

Table 9: Retrieval performance of bi-encoder models across general and domain-external datasets. Best results in bold and second-best underlined.

Dataset	Metric@k	bge-m3	qwen3-reranker	nemotron-rerank	ours
ESCI	NDCG@1	**0.7985**	0.7618	0.7331	0.7450
	NDCG@5	**0.8027**	0.7810	0.7631	0.7669
	NDCG@10	**0.8321**	0.8133	0.8004	0.8018
	MAP@1	**0.1215**	0.1157	0.1060	0.1095
	MAP@5	**0.3931**	0.3766	0.3614	0.3640
	MAP@10	**0.5802**	0.5590	0.5451	0.5451
	MRR@1	**0.8017**	0.7639	0.7380	0.7499
	MRR@5	**0.8720**	0.8495	0.8318	0.8406
	MRR@10	**0.8752**	0.8533	0.8355	0.8444
	Recall@1	**0.1215**	0.1157	0.1060	0.1095
	Recall@5	**0.4597**	0.4457	0.4412	0.4412
	Recall@10	**0.7096**	0.6965	0.6963	0.6945
	Precision@1	**0.7985**	0.7618	0.7331	0.7450
	Precision@5	**0.7245**	0.7091	0.6981	0.6976
	Precision@10	**0.6451**	0.6348	0.6314	0.6287
MIRACL	NDCG@1	**0.6451**	0.6370	0.5948	0.4846
	NDCG@5	**0.6357**	0.6280	0.5791	0.4642
	NDCG@10	**0.6777**	0.6724	0.6258	0.5144
	MAP@1	**0.2654**	0.2576	0.2366	0.1765
	MAP@5	**0.5274**	0.5198	0.4709	0.3549
	MAP@10	**0.5844**	0.5764	0.5277	0.4066
	MRR@1	**0.6564**	0.6370	0.6110	0.4992
	MRR@5	**0.7297**	0.7200	0.6963	0.5909
	MRR@10	**0.7345**	0.7243	0.7007	0.6003
	Recall@1	**0.2654**	0.2576	0.2366	0.1765
	Recall@5	**0.6390**	0.6341	0.5774	0.4696
	Recall@10	**0.7700**	0.7676	0.7141	0.6129
	Precision@1	**0.6451**	0.6370	0.5948	0.4846
	Precision@5	**0.3886**	0.3819	0.3488	0.2911
	Precision@10	**0.2485**	0.2465	0.2314	0.2019
Justicio	NDCG@1	**0.7775**	0.7637	0.7634	0.7593
	NDCG@5	**0.8013**	0.7861	0.7955	0.7834
	NDCG@10	**0.8055**	0.7898	0.7998	0.7880
	MAP@1	**0.7775**	0.7637	0.7634	0.7593
	MAP@5	**0.7951**	0.7805	0.7869	0.7767
	MAP@10	**0.7968**	0.7819	0.7886	0.7785
	MRR@1	**0.7795**	0.7640	0.7650	0.7610
	MRR@5	**0.7961**	0.7807	0.7878	0.7776
	MRR@10	**0.7979**	0.7821	0.7895	0.7794
	Recall@1	**0.7775**	0.7637	0.7634	0.7593
	Recall@5	0.8198	0.8027	**0.8212**	0.8034
	Recall@10	0.8326	0.8145	**0.8343**	0.8178
	Precision@1	**0.7775**	0.7637	0.7634	0.7593
	Precision@5	0.1640	0.1605	**0.1642**	0.1607
	Precision@10	0.0833	0.0814	**0.0834**	0.0818
SSLD	NDCG@1	0.7159	0.6969	**0.8238**	0.8190
	NDCG@5	0.8765	0.8700	**0.9198**	0.9140
	NDCG@10	0.8765	0.8700	**0.9198**	0.9140
	MAP@1	0.3582	0.3488	**0.4123**	0.4098
	MAP@5	0.8124	0.8037	**0.8756**	0.8654
	MAP@10	0.8124	0.8037	**0.8756**	0.8654
	MRR@1	**0.9870**	0.6969	0.9763	0.9008
	MRR@5	**0.9929**	0.8369	0.9872	0.9459
	MRR@10	**0.9929**	0.8369	0.9872	0.9459
	Recall@1	0.3582	0.3488	**0.4123**	0.4098
	Recall@5	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	Recall@10	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	Precision@1	0.7159	0.6969	**0.8238**	0.8190
	Precision@5	**0.3999**	**0.3999**	**0.3999**	**0.3999**
	Precision@10	**0.2000**	**0.2000**	**0.2000**	**0.2000**

Table 10: Reranking performance of cross-encoder models across general and domain-external datasets. Best results in bold and second-best underlined.