# DOG-RAG: A Galician Benchmark for Legal Retrieval-Augmented Generation

Pablo Rodríguez[1], Raquel Vázquez García[1], Silvia Paniagua Suárez[1], Alberto Botana López[1] and Pablo Gamallo[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)

**Abstract**

Retrieval-Augmented Generation (RAG) is essential for grounding Large Language Models in domain-specific knowledge, yet resources for low-resource languages like Galician remain scarce. This paper introduces DOG-RAG, the first curated benchmark for evaluating RAG systems in Galician legal texts. Constructed using a human-in-the-loop pipeline with documents from the Diario Oficial de Galicia, the dataset comprises 500 high-quality question-answer-context triplets. We evaluate 16 retrieval configurations, comparing lexical baselines (BM25) against multilingual dense retrievers and neural rerankers. Our results show that while BM25 remains surprisingly robust in legal contexts, multilingual dense models combined with neural rerankers provide superior semantic grounding. We release the dataset and evaluation tools to facilitate future research.

**Keywords**

Retrieval-Augmented generation, Galician, evaluation, dataset, open access

## 1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm to provide LLMs with external knowledge, reducing hallucinations and providing access to up-to-date information [1]. However, the effectiveness of a RAG pipeline is highly dependent on the quality of its retrieval component, which directly conditions the grounding and faithfulness of the generated answers [2].

Although research on RAG has accelerated, it suffers from a significant high-resource bias, where the dominant benchmarks and models are overwhelmingly English-centric [3, 4]. This creates a critical gap for low-resource languages like Galician, which lack the standard datasets and tools needed to evaluate and build effective RAG systems. This gap is particularly pronounced in specific domains, where creating datasets is very costly and evaluating tools is highly complex, despite their wide-ranging applications.

To address this gap, we created a new dataset using legal texts and analyzed how different configurations of the retriever step in a RAG system worked with it. In the absence of large-scale Galician native retrievers, we hypothesize that cross-lingual transfer from multilingual dense models is a viable and effective alternative. Furthermore, lexical baselines such as BM25 remain competitive thanks to the highly specific and structured terminology of legal texts. To test this hypothesis and facilitate future research, we present the following:

**A New Benchmark for Galician RAG:** We introduce a manually curated, synthetically generated dataset designed to evaluate retrieval and evidence-grounded answer generation for the Galician language.

**A Modular Evaluation Tool:** We release an open-source tool to automate the indexing, retrieval, and evaluation pipeline, facilitating reproducible research.

**A Comparative Analysis:** We show that existing multilingual dense retrievers are a viable option for Galician, complementing lexical baselines to recover more meaningful contexts.

This paper is organized as follows: Section 2 discusses related work in RAG evaluation, dataset creation and low-resource scenarios. Section 3 details the methodology behind the construction and manual curation of the Galician benchmark. Section 4 describes the experimental setup, including the specific retrieval architectures and evaluation protocols. Section 5 presents the quantitative results and provides a discussion of our findings. Finally, Section 6 concludes the paper and outlines limitations and avenues for future work.

The created dataset is available on HuggingFace [1], whereas the evaluation tool is available on GitHub [2].

## 2. Related Work

**RAG Evaluation** Evaluating a RAG system can be performed end-to-end or modularly. Frameworks like *RAGAS* [5] propose metrics for both stages, specifically targeting the retrieval component with *Context Precision*—measuring the proportion of relevant chunks retrieved— and *Context Recall* —assessing if the retrieved context contains the ground truth. Conversely, other frameworks prioritize the generation component; for instance, *RAGEval* [6] introduces schema-based metrics to evaluate completeness and hallucination. Additionally, traditional IR metrics such as precision, recall, and mean reciprocal rank (MRR) remain widely used [7, 8].

Complementing these metrics is the LLM-as-a-judge paradigm, where an LLM evaluates the quality of another model's response. This method has shown high correlations with human judgment [9, 10]. These evaluation methods are typically applied to datasets originally designed for standard Question Answering, such as HotPotQA [11] or Natural Questions [12]. However, specific RAG benchmarks have recently emerged to address the unique challenges of retrieval-augmented systems. Examples include MultiHop-RAG [13],

[1]https://huggingface.co/datasets/proxectonos/dog-rag
[2]https://github.com/proxectonos/nos-rag-eval

*SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.*
pablo.rodriguez@scc.uned.es (P. Rodríguez); silvia.paniagua.suarez@usc.es (S. P. Suárez); alberto.botana.lopez@usc.es (A. B. López); pablo.gamallo@usc.gal (P. Gamallo)
https://gramatica.usc.es/~gamallo/ (P. Gamallo)
0009-0008-7200-4197 (P. Rodríguez); 0009-0000-5260-869X (S. P. Suárez); 0000-0002-8981-3918 (A. B. López); 0000-0002-5819-2469 (P. Gamallo)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

which requires reasoning over specific data within retrieved fragments, and RGB [14], which explicitly tests robustness against noise and the informativeness of retrieved contexts.

**Dataset Creation** The creation of synthetic datasets has rapidly evolved towards LLM-driven pipelines. For general-purpose tasks, frameworks like Self-Instruct [15] demonstrated how LLMs can automatically generate diverse training samples by bootstrapping from a small set of seed queries. Building on this, methodologies such as Evol-Instruct [16] introduced techniques to progressively increase the complexity and diversity of these synthetic instructions, pushing models to handle more intricate reasoning. In the specific context of Retrieval-Augmented Generation (RAG), the focus shifts to extracting and grounding data directly from anchor documents. Methods like InPars [17] pioneered this by prompting models to synthesize questions for specific text chunks, an approach later expanded by evaluation frameworks like ARES [18] to automatically generate complete Question-Answer-Context triplets for rapid benchmark creation. However, while fully automated pipelines are highly scalable, they risk introducing translation artifacts and hallucinated interpretations when applied to low-resource languages or specialized domains [19]. To mitigate these issues, methodologies focused on complex domains rely on a hybrid, human-in-the-loop curation process [20] to guarantee both linguistic naturalness and strict domain accuracy.

**The Gap in Galician** Despite the growth of the Galician NLP ecosystem through initiatives like *Proxecto Nós* [21] and benchmarks like IberoBench [22] or Simil-Eval [23], resources have remained limited to general language understanding or semantic similarity. Currently, there is no dedicated benchmark for evaluating retrieval-augmented generation systems in Galician. Our work addresses this critical gap, providing the first native resource to evaluate how well RAG systems can locate and synthesize facts within Galician corpora.

# 3. A Galician RAG Benchmark

To address the lack of native evaluation resources, we constructed a new dataset from scratch rather than relying on translation. Automatic translation often introduces artifacts and fails to adapt cultural contexts, leading to questions about entities or events foreign to the target audience. Moreover, as this dataset is built directly from Galician legal sources, we can ensure linguistic naturalness and thematic relevance to the local sociocultural reality.

The dataset is composed of 500 entries containing triplets of questions, answers and contexts, as well as basic metadata such as the document ID and filename (Figure 1). In addition to the main triplet file, we also release complementary metadata files containing the full information used during dataset construction, including the source-document metadata required for traceability and verification. This traceability is important because the dataset is not intended to evaluate legal hierarchy reasoning or norm-conflict resolution. Rather, it targets retrieval-grounded QA over official *Diario Oficial de Galicia* (DOG) publications, with fact-based questions reflecting practical administrative information needs that ordinary users may have. The dataset results from a two-stage process based on a human-in-the-loop generation pipeline.

| | |
|---|---|
| **Question** | Por que se denegan as solicitudes da subvención do Programa de axuda ao alugamento de vivenda da convocatoria 2024 (VI483C)?<br>*Why are applications for the 2024 Housing Rental Assistance Programme grant (VI483C) being denied?* |
| **Answer** | Denéganse as solicitudes presentadas ou emendadas despois de esgotarse o crédito orzado e...<br>*Applications submitted or amended after the budgeted credit has been exhausted are denied and...* |
| **Context** | O IGVS, ditou as resolucións polas que se denega a subvención do Programa de axuda ao alugamento de vivenda na Comunidade Autónoma de Galicia (código de procedemento VI483C) a todas aquelas solicitudes presentadas ou ...<br>*The IGVS issued the resolutions whereby the grant for the Housing Rental Assistance Programme in the Autonomous Community of Galicia (procedure code VI483C) is denied to all applications submitted or ...* |
| **Metadata** | **ID:** 5 **filename:** AnuncioC3Q2-040325-0003_gl<br>**url:** https://www.xunta.gal/dog/Publicados/2025/2546...<br>**Category:** General |

**Figure 1:** Example entry from the Galician RAG benchmark. English translations are provided in *italics* below the original Galician text.

## 3.1. Data Sources

The dataset is based on publications from the *Diario Oficial de Galicia* (DOG), the Official Bulletin of the autonomous region of Galicia, composed of a wide variety of legal provisions (resolutions, decrees, orders, announcements, etc.). Due to the large volume of available material, we delimited a specific time frame to work with, namely the year 2025. From the 12,251 documents published in that year, we selected 122 texts considering the type of legal provisions and the frequency with which they appear in the DOG, as well as the month of publication, so that the resulting sample would be representative enough.

## 3.2. Data Creation Process

We followed a human-in-the-loop generation approach in which two main phases can be identified. Firstly, an LLM was used to synthetically generate the 500 entries of the dataset, and secondly, a comprehensive manual revision and validation of the output was conducted, in order to ensure that the final dataset had a relevant size and a strong focus on quality.

### 3.2.1. Synthetic Generation

To generate the dataset entries, the selected model was DeepSeek-V3 [24] due to its good linguistic and task-based performance [25]. We then designed a specific prompt with all our requirements and constraints (see Annex A). The main conditions enforced were the following:

- `Language`: The output required was standard Galician.

- `Entries structure`: Each entry had to be composed of two questions (one being a rephrasing of the other), two answers (one being a variant of the other), and a context field. This allowed us to optimize each call to the LLM, reducing the chances of an entry having no valid questions or answers.
- `Questions`: They had to be strictly fact-based, and avoid references to the title and date of legal provisions (asking about elements or procedures discussed in the text, rather than specific provisions).
- `Answers`: They had to avoid metatextual references (e.g., "According to the text).
- `Context`: It was to be provided in the form of fragments of text —consecutive or non-consecutive ones— containing the factual evidence for the answer and proving that the answer was specific to that question.

Some metadata fields were then automatically added to each entry: a unique identifier, the name of the document the entry was generated from, and the URL to that DOG publication. Furthermore, during the creation process, it was observed that some of the contexts used to support the answers could be grouped into different categories depending on their format or the type of information they contained. Although this information was not used during the evaluation of the dataset, it is useful for creating alternative evaluations to study the behavior of the systems in specific situations. The categories established were:

- `data_extracted`: In some cases, answering a question requires extracting a specific piece of data, a task that is more similar to text extraction than text generation.
- `sparsed_context`: When entries were based on very long documents, in which the distance between the answer and the subject matter of the document was significant, answers had to be justified using non-consecutive fragments of text. This can be problematic with some RAG techniques, which focus on retrieving consecutive fragments; this type of entries can penalize them.
- `table`: Due to the nature of the DOG's content, it is normal for some data to appear in the form of tables, which can be difficult to handle for less powerful models.
- `general`: Any other cases not covered by the previous categories.

### 3.2.2. Manual Curation

Once the entries were generated, they underwent a rigorous manual review by a total of nine native Galician experts, at two stages of the creation process: the first one was conducted half-way through to assess the quality of the output to that point, and the second, and most exhaustive one, at the end, once the target amount of 500 entries was reached.

The mid-way revision was conducted with a sample of 40 entries and four people (20 entries reviewed per couple). The main issues highlighted at this point were related to the order of elements in long sentences, which could be improved for readability purposes.

The final revision was split into two parts. A group revision, where a sample of 100 entries was reviewed by eight people (25 entries per couple); and an individual revision, where the remaining 400 entries were reviewed by the same person who designed the prompt, a native Galician linguist. This final revision also included the manual anonymization of entries with personal or company data for data privacy reasons. As a result, 11 entries were edited, and 43 more were replaced by new ones, following the same process described above (LLM generation followed by manual human validation).

Four aspects were reviewed in all cases:

- Linguistic Quality: Adherence to Galician grammar and orthography.
- Relevance of the Questions: Pertinence of the questions considering that any person could ask them, as there was no persona defined in the prompt.
- Validity of the Responses in relation to the Questions: Whether the information retrieved was indeed answering the question.
- Validity of the Context: Whether the fragments retrieved provided evidence for the answers, while also proving they were specific to the question and to the document they were extracted from.

Inter-annotator agreement in the group revision was assessed through percentage agreement and Gwet's AC coefficient, using AC1 for the two binary variables, linguistic quality (error vs. no error) and question relevance (relevant vs. irrelevant), and weighted AC2 for answer validity and context validity, which were annotated on a three-level ordered scale ranging from valid to partially valid to not valid. Gwet's coefficient was preferred for interpretation because it is more robust than Cohen's kappa in the presence of highly imbalanced category distributions [26], a pattern observed particularly in question relevance and, to a lesser extent, in linguistic quality and answer validity. Under this measure, raw agreement remained generally high for linguistic quality (72–92%), question relevance (80–100%) and answer validity (84–92%). The corresponding agreement coefficients indicated moderate to very high consistency for linguistic quality (AC1 = 0.52–0.91), high to perfect consistency for question relevance (AC1 = 0.76–1.00), and consistently high agreement for answer validity (weighted AC2 = 0.86–0.91). By contrast, context validity showed the lowest and most variable consistency overall, with raw agreement ranging from 8% to 76% and weighted AC2 values from -0.66 to 0.85, indicating that this was the most difficult aspect to assess consistently in the review process.

Most revision edits were related to two aspects. The first one was the context field, since the fragments retrieved were not always enough to justify the answer in relation to the question and the document. In some cases, fragments did justify the answer, but they did not mention the object of the question explicitly. They were so generic that they could have been extracted from many different sources. This is due to the fact that DOG publications have a specific structure, and therefore have many similarities in terms of what they say and how the information is phrased. This is why fragments had to be manually added or modified to ensure completeness. The second one was the autonomy of the questions and answers. This is related, on the one hand, to the requirement of avoiding references to specific legal provisions in questions (by using their name and date), and on the other hand, to the need of providing full context in the questions, so that they would stand alone, and avoiding references to the document in both questions and answers.

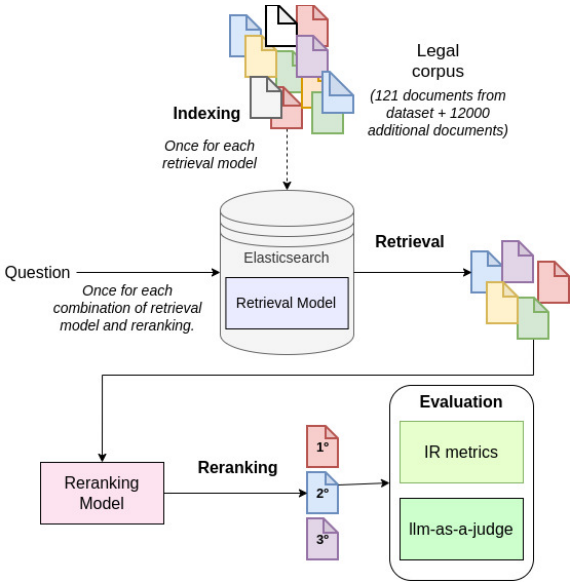


**Figure 2:** Overview of the evaluation pipeline. Paragraphs are indexed in Elasticsearch and retrieved using cosine similarity. Top-k candidates are optionally reranked by neural rerankers, and performance is assessed using both traditional IR metrics and an LLM-as-a-judge.

# 4. Experimental Setup

We designed experiments to systematically evaluate the retrieval stage of RAG in Galician legal texts using the created dataset. Specifically, we tested 16 distinct configurations, combining 4 retrievers with 4 reranking settings (no reranker versus 3 neural rerankers). Each configuration was executed 10 times, and results were averaged across runs. Multiple runs account for potential non-determinism in the indexing process or the LLM-as-a-judge evaluation. Evaluation employed both traditional IR metrics and an LLM-as-a-judge approach, as illustrated in Figure 2.

## 4.1. Retrieval and Reranking Models

All retrievers operate over chunks of 250 tokens indexed in Elasticsearch. In addition to the documents used to generate the dataset entries, more than 12,000 documents from the remaining 2025 DOG files are also indexed. Dense retrievers compute embeddings for both queries and fragments, using cosine similarity to rank candidates. Lexical retrieval is implemented via Elasticsearch's standard BM25 scoring.

We selected models to cover a spectrum of approaches, from classical lexical retrieval to multilingual dense embeddings and specialized neural rerankers. This setup allows us to study the effectiveness of cross-lingual transfer, the benefit of reranking, and the relative performance of different model families in a low-resource language setting.

**Retrievers**

**BM25 (Lexical):** A standard probabilistic IR baseline, relying on keyword matching to rank fragments.

**bge-m3:** A state-of-the-art multilingual dense retriever supporting dense, sparse, and multi-vector embeddings, enabling cross-lingual knowledge transfer [27].

**Qwen3-Embedding-0.6B:** A multilingual dense retriever derived from the Qwen LLM family, optimized for semantic similarity [28].

**EmbeddingGemma-300m:** A lightweight dense retriever fine-tuned from the Gemma architecture, designed for efficient embedding-based search [29].

**Rerankers**

**bge-reranker-v2-m3:** A lightweight multilingual cross-encoder optimized for fast and accurate reranking of candidate contexts.

**Qwen3-Reranker-0.6B:** A generative reranker leveraging the Qwen model to score relevance based on semantic content.

**jina-reranker-v3:** A high-performance multilingual reranker capable of handling long-context relevance and complex queries [30].

## 4.2. Evaluation Metrics

We evaluated all configurations using two complementary families of metrics, capturing both traditional retrieval performance and the quality of context grounding:

**Traditional IR Metrics:** We computed Precision@K, Recall@K, and Mean Reciprocal Rank (MRR) with $K = 3$ to assess the ability of each retriever to surface the correct documents. These metrics provide a standard measure of retrieval effectiveness and allow comparisons with established IR baselines.

**LLM-as-a-judge Metrics:** We used **Selene-1-Mini-Llama-3.1-8B** [31], a model fine-tuned specifically for evaluation, to compute Context Precision and Context Recall. These metrics assess whether the retrieved contexts actually contain the information needed to answer the question, capturing the semantic correctness of retrieval beyond simple keyword overlap. Detailed prompts for the LLM-as-a-judge evaluation are provided in Appendix B.

By combining these two families, we can measure both the surface-level retrieval accuracy and the deeper, context-aware relevance necessary for faithful RAG in a low-resource language like Galician.

# 5. Results and Analysis

Experimental results are reported in Table 1. Traditional IR metrics (P@3, R@3, MRR) are calculated at document level (checking if the retriever recovers the expected document), while the LLM-as-a-judge metrics (Context Precision and Context Recall) are calculated at context level (to see if the retriever recovers the expected fragments of the retrieved documents).

Contrary to expectations in general NLP tasks, the lexical baseline, `BM25`, achieves a high baseline performance, particularly in document-level ranking, with an MRR of 0,730 without reranking. When paired with neural rerankers, performance improves across all categories. The configuration `BM25 + jina-reranker-v3` achieves the highest Precision@3 (0,487) and Recall@3 (0,843), while `BM25 + bge-reranker-v2-m3` yields the best overall MRR (0,752).

Among dense retrievers, `bge-m3` consistently outperforms `Qwen3-Embedding` and `gemma-300m`, with the combination of `bge-m3` and a specific reranker performing better across all metrics than using that same reranker with `Qwen3-Embedding` or `gemma-300m`. Conversely, these lightweight models struggle to match the baseline lexical performance across both document and context-level metrics. Even when enhanced by neural rerankers, their highest

| Retriever Model | Reranker | P@3 | R@3 | MRR | Context Precision | Context Recall |
|---|---|---|---|---|---|---|
| BM25 | (None) | 0,465 | 0,840 | 0,730 | 0,430 | 0,589 |
| | bge-reranker-v2-m3 | 0,486 | 0,839 | **0,752** | 0,588 | 0,710 |
| | Qwen3-Reranker-0.6B | 0,470 | 0,837 | 0,718 | 0,472 | 0,636 |
| | jina-reranker-v3 | **0,487** | **0,843** | 0,748 | 0,585 | 0,724 |
| bge-m3 | (None) | 0,414 | 0,740 | 0,630 | 0,459 | 0,641 |
| | bge-reranker-v2-m3 | 0,449 | 0,792 | 0,707 | 0,595 | 0,721 |
| | Qwen3-Reranker-0.6B | 0,418 | 0,746 | 0,635 | 0,486 | 0,668 |
| | jina-reranker-v3 | 0,415 | 0,756 | 0,669 | **0,599** | **0,737** |
| Qwen3-Embedding-0.6B | (None) | 0,396 | 0,730 | 0,621 | 0,403 | 0,571 |
| | bge-reranker-v2-m3 | 0,427 | 0,760 | 0,684 | 0,552 | 0,662 |
| | Qwen3-Reranker-0.6B | 0,399 | 0,739 | 0,631 | 0,435 | 0,599 |
| | jina-reranker-v3 | 0,412 | 0,740 | 0,655 | 0,556 | 0,688 |
| gemma-300m | (None) | 0,389 | 0,675 | 0,585 | 0,413 | 0,580 |
| | bge-reranker-v2-m3 | 0,416 | 0,706 | 0,637 | 0,548 | 0,658 |
| | Qwen3-Reranker-0.6B | 0,406 | 0,693 | 0,603 | 0,451 | 0,614 |
| | jina-reranker-v3 | 0,403 | 0,704 | 0,631 | 0,548 | 0,668 |

**Table 1**
Traditional IR and llm-as-a-judge metrics results. Values use comma decimal separator. Best score in **bold**, second best underlined.

MRR scores remain significantly below the BM25 baseline, highlighting the difficulty of applying smaller dense models to specialized low-resource domains without fine-tuning.

Specifically, the combination of `bge-m3 + jina-reranker-v3` reaches the highest scores for semantic grounding, with a Context Precision of 0,599 and a Context Recall of 0,737. In this case, this combination also outperforms those using BM25 as a recuperator in these metrics, although `BM25 + jina-reranker-v3` is the second-best option for Context Recall and `BM25 + bge-reranker-v2-m3` is the third-best for Context Precision.

### 5.1. Analysis of Findings

The evaluation results reveal a nuanced interplay between lexical matching and semantic understanding in this specialized domain. By analyzing the performance variations across the tested configurations, three key insights emerge:

**1. Lexical robustness in legal texts** Unlike in generalistic domains, lexical retrieval via BM25 proves exceptionally robust for legal texts. This strong performance is attributable to the formulaic nature of DOG publications, which rely on a controlled vocabulary and unique identifiers (e.g., procedure codes like VI483C). These domain specific terms allow keyword-based systems to locate documents with high accuracy, bypassing the vocabulary mismatch problem often seen in other domains.

**2. Semantic vs. lexical precision** The data reveals a divergence between document-level and context-level performance. While BM25 finds the correct document, dense models like `bge-m3` demonstrate a superior ability to pinpoint the semantic boundaries required for an answer. The higher Context Precision and Recall scores for dense configurations suggest they are more effective at semantic alignment, which is critical for preventing hallucinations in the generation stage, a vital requirement when working in the legal domain.

**3. Rerankers as a strongly recommended component** Neural reranking provides a consistent performance multiplier. For BM25, the improvement in context-aware metrics is substantial (Context Recall jumps from 0,589 to 0,724 with `jina-reranker-v3`). These results confirm that a two-stage pipeline is beneficial regardless of the base retriever, as the reranker compensates for the lexical biases of BM25 and the document-level weaknesses of dense models.

## 6. Conclusions and Future Work

In this paper, we introduced DOG-RAG, a curated benchmark to evaluate Retrieval-Augmented Generation systems in Galician legal texts. Our evaluation of retrieval configurations showed that classical lexical baselines like BM25 remain highly competitive at the document retrieval level. However, multilingual dense retrievers yielded superior performance in pinpointing exact semantic boundaries. Furthermore, creating a two-stage retrieve-and-rerank pipeline using a neural reranker effectively balances lexical accuracy and semantic grounding.

This work also outlines clear avenues for future research. First, while our current experimental scope focused exclusively on the retrieval phase, future work will leverage the dataset's ground-truth answers to evaluate the generation component, assessing how effectively LLMs synthesize the retrieved Galician contexts. Second, the use of a generalist LLM-as-a-judge (*Selene*) may not capture all the linguistic and legal nuances of the Galician language. Future iterations should use more powerful LLMs to evaluate the results, and validate these automated metrics against human judgments.

Finally, we plan to expand the dataset to encompass multi-hop reasoning scenarios, where answers require synthesizing information across different legal provisions. This will allow us to evaluate RAG systems in more realistic administrative use cases, which often involve cross-referencing multiple documents. In addition, we plan to use the dataset creation methodology employed here to create evaluation resources in other fields, such as science and culture. This will allow us to validate the effectiveness of our hybrid pipeline in building reliable datasets for different specialized vocabularies, ultimately helping to bridge the resource gap for Galician in other domains.

# A. Dataset Creation Prompt

**Dataset Generation Prompt**

This is an example of a document from the DOGA (Diario Oficial de Galicia), an official publication of the Autonomous Community of Galicia containing legal provisions, regulations and announcements of general interest. From this publication:

- Generate a possible question that any user could ask to the system. Then generate a variant of that question, with the same meaning but different wording.
- Provide 2 valid answers to the same question. These answers must express the same factual content using different wording. All answers must be correct and directly found in the context. They must not include opinions or inferred interpretations. They must not be a copy of the text provided.
- Provide or copy the paragraph in which the answer is found precisely as the context.
- Provide a numeric ID (e.g., 0)

Then, deliver all the information extracted in a JSON format with the keys:
**id:** [numeric id]
**question:** [list of valid questions written differently]
**answer:** [list of valid answers written differently]
**context:** [paragraph in which the answer to the question is found]
Create 5 entries from this document following this structure.
**Constraints:**

- The output must be only in Galician (except JSON keys); avoid English, Spanish or Portuguese.
- Questions need to be strictly fact-based. Do not ask about the author's opinion or the meaning of the text.
- Do not ask about legal provisions; ask about the procedures and elements they mention.
- Questions must provide full context, so that the reader doesn't need to see the text to understand what they are asking about. For this purpose, include the name of the elements you are asking about (procedures, subsidies, etc.) instead of using demonstratives ("this resolution", "those subsidies").
- Avoid questions and answers involving personal names.
- Avoid explicit references to legal provisions in the questions.
- Avoid questions that refer to the content mentioned in previous questions.
- Avoid phrases like "according to the text", "in this provision", "contained in this resolution" or similar, both in the questions and the answers. Questions and answers must stand alone without requiring the reader to see the original text.
- For context, write only the full paragraph in which the answers are found, not the full text.
- When delivering the JSON format, escape the quotes inside the text.

# B. LLM-as-judge metrics prompt definition

**Context Precision Prompt**

You are tasked with evaluating whether a specific retrieved context is relevant to answering a Galician question, based on a binary scoring rubric. Provide comprehensive feedback strictly adhering to the rubric, followed by a binary Yes/No judgment. Avoid generating any additional opening, closing, or explanations.
**Note:** The context, question, and ground truth answer are written in Galician. Do not translate or modify the original language. Evaluate as-is.
**Here are some rules of the evaluation:**

1. You should prioritize evaluating whether the context clearly helps answer the question. Only respond "Yes" if the sentence is clearly relevant to producing or supporting the ground truth answer.
2. If the context does not help answer the question or is irrelevant, respond "No".

Your reply should strictly follow this format:
**Result:** <Yes or No>
**Here is the data:**
**Context:**
{context}
**Question:**
{question}
**Ground Truth Answer:**
{ground_truth}
**Score Rubrics:**
**Yes:** The context clearly helps answer the question and supports the ground truth answer.
**No:** The context does not help answer the question or is irrelevant to the ground truth answer.

**Context Recall Prompt**

You are tasked with evaluating whether a specific sentence from a Galician answer is supported by a retrieved context, based on a binary scoring rubric. Provide comprehensive feedback strictly adhering to the rubric, followed by a binary Yes/No judgment. Avoid generating any additional opening, closing, or explanations.
**Note:** The sentence and context are written in Galician. Do not translate or modify the original language. Evaluate as-is.
**Here are some rules of the evaluation:**

1. You should prioritize evaluating whether the factual content of the sentence is present in the retrieved context. The sentence should be considered supported if the same factual content exists in the context, without requiring translation or interpretation.
2. If the sentence cannot be confirmed based on the context, respond "No".

Your reply should strictly follow this format:
**Result:** <Yes or No>
**Here is the data:**
**Sentence:**
{sentence}
**Retrieved Context:**
{context}
**Score Rubrics:**
**Yes:** The factual content of the sentence is clearly present in the retrieved context.
**No:** The factual content of the sentence is not present or cannot be confirmed in the retrieved context.

# Acknowledgments

This paper was funded by MCIU/AEI (grants with references PID2024-161928OB-I00, and AIA2025-163322-C62), by the Galician Government (Research Center of Galicia accreditation 2024-2027 ED431G-2023/04 and GPCE ED431B 2025/16), by QUANTUM_IBER_IA (ERDF/POCTEP), and by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU — NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This

work has received financial support from the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04) and the European Regional Development Fund (ERDF), which acknowledges the CiTIUS - Centro Singular de Investigación en Tecnoloxías Intelixentes da Universidade de Santiago de Compostela as a Research Center of the Galician University System.

## References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[2] S. Filice, E. Haramaty, G. Horowitz, Z. Karnin, L. Lewin-Eytan, A. Shtoff, Generate but verify: Answering with faithfulness in RAG-based question answering, in: K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, D. P. Singh (Eds.), Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, 2025, pp. 1017–1037. URL: https://aclanthology.org/2025.ijcnlp-long.56/. doi:10.18653/v1/2025.ijcnlp-long.56.

[3] W. Liu, S. Trenous, L. F. R. Ribeiro, B. Byrne, F. Hieber, XRAG: Cross-lingual retrieval-augmented generation, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 15669–15690. URL: https://aclanthology.org/2025.findings-emnlp.849/. doi:10.18653/v1/2025.findings-emnlp.849.

[4] C. Creanga, L. P. Dinu, LLM-as-a-judge for low-resource languages: Adapting ragas and comparative ranking for Romanian, in: H. Hettiarachchi, T. Ranasinghe, A. Plum, P. Rayson, R. Mitkov, M. Gaber, D. Premasiri, F. A. Tan, L. Uyangodage (Eds.), Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 157–167. URL: https://aclanthology.org/2026.loreslm-1.15/. doi:10.18653/v1/2026.loreslm-1.15.

[5] S. Es, J. James, S. Espinosa Anke, Luis andº Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.

[6] K. Zhu, Y. Luo, D. Xu, Y. Yan, Z. Liu, S. Yu, R. Wang, S. Wang, Y. Li, N. Zhang, X. Han, Z. Liu, M. Sun, RAGEval: Scenario specific RAG evaluation dataset generation framework, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 8520–8544. URL: https://aclanthology.org/2025.acl-long.418/. doi:10.18653/v1/2025.acl-long.418.

[7] C. D. Manning, Introduction to information retrieval, Syngress Publishing,, 2008.

[8] E. M. Voorhees, et al., The trec-8 question answering track report., in: Trec, volume 99, 1999, pp. 77–82.

[9] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.

[10] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, R. A. Metoyer, Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, 2025, pp. 952–966.

[11] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, Hotpotqa: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2369–2380.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al., Natural questions: a benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466.

[13] Y. Tang, Y. Yang, Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, arXiv preprint arXiv:2401.15391 (2024).

[14] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17754–17762.

[15] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), 2023, pp. 13484–13508.

[16] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, D. Jiang, Wizardlm: Empowering large language models to follow complex instructions, in: ICLR 2024, 2024. URL: https://arxiv.org/abs/2304.12244.

[17] L. Bonifacio, H. Abonizio, M. Fadaee, R. Nogueira, Inpars: Data augmentation for information retrieval using large language models, arXiv preprint arXiv:2202.05144 (2022).

[18] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, Ares: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 338–354.

[19] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, et al., Mega: Multilingual evaluation of generative ai, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 4232–4267.

[20] C. Malaviya, S. Lee, S. Chen, E. Sieber, M. Yatskar,

D. Roth, ExpertQA: Expert-curated questions and attributed answers, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3025–3045. URL: https://aclanthology.org/2024.naacl-long.167/. doi:10.18653/v1/2024.naacl-long.167.

[21] I. de Dios-Flores, C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. García, P. Gamallo, E. Fernández Rei, A. Bugarín-Diz, M. González González, S. Barro, X. L. Regueira, The nós project: Opening routes for the Galician language in the field of language technologies, in: I. Aldabe, B. Altuna, A. Farwell, G. Rigau (Eds.), Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 52–61. URL: https://aclanthology.org/2022.tdle-1.6/.

[22] I. Baucells, J. Aula-Blasco, I. de Dios-Flores, S. Paniagua Suárez, N. Perez, A. Salles, S. Sotelo Docio, J. Falcão, J. J. Saiz, R. Sepulveda Torres, J. Barnes, P. Gamallo, A. Gonzalez-Agirre, G. Rigau, M. Villegas, IberoBench: A benchmark for LLM evaluation in Iberian languages, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 10491–10519. URL: https://aclanthology.org/2025.coling-main.699/.

[23] P. Rodríguez, S. P. Suárez, P. Gamallo, S. S. Docio, Continued pretraining and interpretability-based evaluation for low-resource languages: A Galician case study, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 4622–4637. URL: https://aclanthology.org/2025.findings-acl.240/. doi:10.18653/v1/2025.findings-acl.240.

[24] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, et al., Deepseek-v3 technical report, 2025. URL: https://arxiv.org/abs/2412.19437. arXiv:2412.19437.

[25] W. Xuan, R. Yang, H. Qi, Q. Zeng, Y. Xiao, A. Feng, D. Liu, Y. Xing, J. Wang, F. Gao, et al., Mmlu-prox: A multilingual benchmark for advanced large language model evaluation, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 1513–1532.

[26] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology 61 (2008) 29–48. doi:10.1348/000711006X126600.

[27] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[28] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, et al., Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025).

[29] H. S. Vera, S. Dua, B. Zhang, D. Salz, R. Mullins, S. R. Panyam, S. Smoot, I. Naim, J. Zou, F. Chen, et al., Embeddinggemma: Powerful and lightweight text representations, arXiv preprint arXiv:2509.20354 (2025).

[30] F. Wang, Y. Li, H. Xiao, Jina-reranker-v3: Last but not late interaction for listwise document reranking, arXiv preprint arXiv:2509.25085 (2025).

[31] A. Alexandru, A. Calvi, H. Broomfield, J. Golden, K. Dai, M. Leys, M. Burger, M. Bartolo, R. Engeler, S. Pisupati, et al., Atla selene mini: A general purpose evaluation model, arXiv preprint arXiv:2501.17195 (2025).