Optimising Retrieval for Linguistic Question-Answering in European Portuguese: A Benchmark on Ciberdúvidas Da Língua Portuguesa

Pedro Moura
ISCTE – Instituto Universitário de Lisboa, Portugal

Inês Gama
ISCTE – Instituto Universitário de Lisboa, Portugal

Fernando Batista
ISCTE – Instituto Universitário de Lisboa, Portugal
INESC-ID Lisboa, Portugal

António Lopes
ISCTE – Instituto Universitário de Lisboa, Portugal
ISTAR, Lisboa, Portugal

Abstract

Information retrieval for question-answering remains underexplored in specialised domains and under-resourced language variants such as European Portuguese. Existing benchmarks largely target general-domain English data and document-centric retrieval, failing to capture the semantic alignment required for linguistic consultation tasks over curated question–answer (QA) pairs. We address this gap by introducing a controlled evaluation framework for retrieval over the "Ciberdúvidas da Língua Portuguesa" corpus, comprising 29,145 expert-validated QA entries. Our approach systematically analyses the interaction between indexing strategies, encoder models, and retrieval paradigms, while modelling real-world query variability through a paraphrase-based benchmark of 600 queries across five user profiles, manually validated by a professional linguist to ensure semantic fidelity. Experiments show that dense retrieval with an IR-optimised monolingual encoder significantly outperforms both sparse (BM25) and hybrid methods, achieving a Mean Reciprocal Rank (MRR) of 0.93. Notably, hybrid retrieval underperforms due to lexical mismatch interference, challenging prevailing assumptions in the literature. Our contributions include a novel benchmark framework for linguistic QA retrieval, empirical evidence supporting monolingual IR-specialised models, and insights into retrieval robustness under paraphrastic variation, enabling improved QA systems for specialised and low-resource environments.

2012 ACM Subject Classification Information systems → Question answering; Information systems → Retrieval models and ranking; Information systems → Evaluation of retrieval results; Information systems → Digital libraries and archives

Keywords and phrases Information Retrieval, Question-Answering, European Portuguese, Sentence Encoders, Natural Language Processing

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.1

Funding This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: `https://doi.org/10.54499/UID/50021/2025`), UID/PRR/50021/2025 (DOI: `https://doi.org/10.54499/UID/PRR/50021/2025`), and UID/04466/2025 (DOI: `https://doi.org/10.54499/UID/04466/2025`).

1 Introduction

Information retrieval in specialised domains remains a central challenge in Natural Language Processing, particularly for languages that are under-represented in benchmark datasets and embedding model development. While retrieval systems have seen substantial progress in general-domain settings [17], their performance in certain fields and contexts remains underexplored. One domain where such systems can provide immediate value is language consultation. *Ciberdúvidas da Língua Portuguesa*[1] is a reference platform dedicated to clarifying doubts about the Portuguese language, covering topics such as spelling, phonetics, etymology, syntax, semantics, and pragmatics [7]. Over nearly three decades, the platform has accumulated more than 29,000 expert-curated responses, forming a rich and authoritative linguistic knowledge base. However, user interaction is currently limited to traditional keyword-based search, which constrains accessibility and makes it difficult to provide immediate, context-aware answers to recurring queries.

Deploying an effective natural language retrieval system over such a corpus presents distinct challenges. First, query formulation varies widely across users, where the same linguistic doubt may be expressed synthetically, formally, informally, or using domain-specific pedagogical terminology. In addition, the corpus consists of questions and answers in Portuguese, mostly in its European variant, a variety that has received comparatively little attention in the development of models [14]. Existing retrieval benchmarks predominantly target general-domain, well-formed queries in high-resource languages, offering limited guidance for specialised linguistic corpora in under-represented language variants. A controlled evaluation of how indexing strategies, embedding models, and retrieval paradigms interact in this setting is also notably absent from the literature.

To address this gap, we propose an evaluation framework for document retrieval that explicitly models real-world query variability through structured paraphrasing profiles. This work focuses on the design, systematic benchmarking, and evaluation of retrieval strategies over the "Ciberdúvidas da Língua Portuguesa" corpus. Rather than assuming a single optimal retrieval configuration, we conduct a controlled experimental analysis of how different preprocessing and indexing strategies, embedding models and retrieval paradigms affect the alignment between user queries and expert linguistic knowledge.

2 Related Work

The development of an effective retrieval system requires a comprehensive understanding of current paradigms in Information Retrieval. This section outlines the foundational literature and recent advancements pertinent to our study.

2.1 Retrieval Targets in Question-Answering Environments

In retrieval applied to question-answering environments, the literature considers different approaches to system design.

The first paradigm treats the corpus as generic passages and attempts to identify an answer to the user's query. This is predominantly seen in open-domain question-answering, where the goal is to search a large collection of documents to retrieve a small subset of candidate passages that might contain the factual answer [13], as presented in DPR. These architectures

[1] Ciberdúvidas da Língua Portuguesa. Available at: `https://ciberduvidas.iscte-iul.pt`

utilize a dual-encoder framework to map both the question and the text passages into a low-dimensional space, enabling the retriever to capture semantic relationships even when synonyms or paraphrases consist of completely different tokens [13]. In open-domain question-answering, these systems are commonly modelled under the retriever–reader paradigm, where a retrieval component first selects relevant documents and a reader subsequently extracts the final answer [28].

In contrast, question–answer retrieval systems typically do not require an explicit reader component, as they operate by ranking pre-existing QA pairs rather than extracting answers from raw text. A major challenge in this environment is the lexical gap which occurs in contexts where users' natural language queries are short and lack overlapping vocabulary with the pre-recorded FAQ pairs [12]. In [12], researchers have modeled the task as a relevance classification problem, employing a variety of semantic textual similarity features to match user queries with the corresponding FAQ. The research [5], applied to a FAQ corpus in European Portuguese, follows this line of development, directly calculating similarity scores between the user's query vector and the vectors of the queries in the corpus, and selecting the FAQ with the highest similarity.

Hybrid approaches such as the Multi-Field Bi-Encoder (MFBE) [3] exploit the multi-field structure of QA pairs, combining title, question body, and category, demonstrating that multi-field representation significantly outperforms approaches based solely on the title. Similarly, [16] proposes an initial re-ranking based on lexical similarity, followed by two BERT models dedicated respectively to query–question and query–answer matching, combining all re-rankers via late fusion.

2.2 Encoding Models and the Portuguese Language

The selection of the embedding model is critical to the performance of the retrieval system [8, 9], as it dictates how effectively the semantic meaning of user queries is mapped to the relevant documents in the knowledge base. The scope of this research is not to identify the absolute best model for this purpose; the experiments conducted are designed as a proof of concept. It will therefore be essential to examine the state-of-the-art (SOTA) for Portuguese language embeddings, as well as a multilingual baseline alternative based on SBERT. In [18], a performance comparison between monolingual and multilingual encoding models for Italian demonstrates that monolingual models exhibit greater sensitivity to language-specific syntactic and semantic patterns. Similar findings have been reported for Portuguese embedding models. In [10], for instance, Serafim outperformed all multilingual models included in the comparison. These results highlight the importance of exploring embedding models specifically designed for Portuguese and its linguistic variants.

The formalization of models focused on European Portuguese is an area that has been largely unexplored in the literature [14]. In the literature consulted, no surveys or comparative studies were found among the most recent models developed. Despite this, one main model with the potential to contribute to this research was noted. The Albertina family [20], provides open-source encoder foundation models for European and Brazilian Portuguese. Two models were created, with 100M parameters and 1.5B parameters, based on the DeBERTa architecture. These models were evaluated and obtained good results in STS (Semantic Textual Similarity) tasks, thus indicating high performance in embedding generation.

The Serafim family represents a set of open-source sentence encoders specifically developed for the Portuguese language, covering both the European (PT-PT) and Brazilian (PT-BR) variants [10]. Built upon foundation models such as Albertina and BERTimbau, Serafim establishes a new state of the art [10] for Portuguese in both Semantic Textual Similarity

(STS) and Information Retrieval (IR) benchmarks, significantly outperforming previous Portuguese-specific encoders and multilingual baselines. The 900M parameter version, based on Albertina PT-PT, demonstrated superior performance specifically on European Portuguese datasets [10]. The Serafim 900M model is available in two variants: a general-purpose sentence embedding model and an Information Retrieval (IR)-optimized version. For clarity and consistency throughout this article, the model identified as `serafim-900m-portuguese-pt-sentence-encoder` will be referred to as **Serafim900m**, while the model `serafim-900m-portuguese-pt-sentence-encoder-ir` will be referred to as **Serafim900m-IR**.

2.3 Retrieval Benchmarks

The development of robust evaluation frameworks has been central to progress in information retrieval and question-answering research [2].

The MS MARCO dataset [2], constructed from over one million anonymized queries from Bing search logs, has been widely used in the field. This dataset has even been used to train some state-of-the-art encoders, such as the MiniLM and E5 families [25, 26]. Although this dataset is in English, the mMarco version [4] performed machine translation of this dataset into 13 languages, one of them Portuguese (no variant specified). To address the limitations of evaluating information retrieval in narrow settings, the Benchmarking-IR (BEIR) benchmark was introduced [23], consisting of a selection of various documents from a range of domains, including question-answering. This benchmark was developed using datasets in English exclusively [23].

Multilingual efforts have been made, for example, in the MIRACL [27] dataset, which supports 18 languages (though Portuguese is not included). This dataset was specifically designed by human annotators to enable the evaluation of retrieval tasks in which the natural-language query and the corpus are in the same language.

Existing benchmarks are primarily designed for general-purpose information retrieval, where queries correspond to factual information needs and documents are treated as independent evidence sources [23, 2, 4]. In contrast, linguistic consultation tasks exhibit different properties: queries are often metalinguistic, involving questions about grammar, orthography, or semantic nuance, where relevance depends on alignment between a user question and a previously formulated expert explanation. Furthermore, existing benchmarks typically assume a document-centric retrieval paradigm [27, 2, 4], whereas in linguistic QA settings the corpus is composed of curated QA pairs, introducing a matching problem closer to question similarity and intent alignment than to traditional passage retrieval.

These limitations motivate the need for a dedicated dataset tailored to the specific characteristics of linguistic question-answering in European Portuguese.

3 Ciberdúvidas Dataset

The corpus was derived from the internal database of *Ciberdúvidas* platform, comprising consultations published between 1997 and January 2026. Due to the nature of our institutional partnership with the platform, this corpus and the generated test dataset are not publicly available. As the original content is proprietary to the platform, the dataset is subject to their internal data policies and licensing.

After extraction and preprocessing, the resulting corpus contains exactly 29,145 consultation entries. Each entry contains a unique ID, a title summarizing the subject, the user's question body (HTML), the expert's answer (HTML), and a timestamp.

The corpus exhibits significant variability in both question formulation and response length. Questions range from short lexical clarifications to longer contextualized inquiries about grammatical structures or language usage. Expert responses similarly vary in length, often including detailed explanations, historical context, and illustrative examples.

To better understand the linguistic phenomena present in the *Ciberdúvidas* collection, a latent topic analysis based on BERTopic was performed on the corpus. The distribution of queries reveals a predominant interest in etymology and word origins, as well as frequent concern with how particular sentences should be written. Furthermore, the corpus comprises a substantial volume of inquiries pertaining to phraseological semantics, idiomatic expressions, and syntactic analysis.

While the consultation service primarily produces content in European Portuguese, the dataset inevitably includes questions originating from other Portuguese variants. Users from diverse Portuguese-speaking regions may submit inquiries that exhibit lexical, orthographic, or syntactic patterns characteristic of Brazilian, African, or other Portuguese varieties. These instances are not explicitly annotated in the metadata, introducing a degree of linguistic variability in the corpus.

Considering token counts, the questions have an average length of 100.9 ± 124.2 tokens, ranging from 7 to 5542 tokens, while answers have an average length of 331.6 ± 319.3 tokens, ranging from 6 to 5756 tokens (all token counts were computed using the `Serafim900m-IR` model, after a basic HTML tag removal preprocessing).

4 Experimental Setup

To evaluate and choose the retrieval techniques to be used, a testing framework was developed to assess the robustness of different models, preprocessing methods, and approaches.

Figure 1 illustrates the complete experimental framework used in this study, which is suitable for generalization and application in other domains.

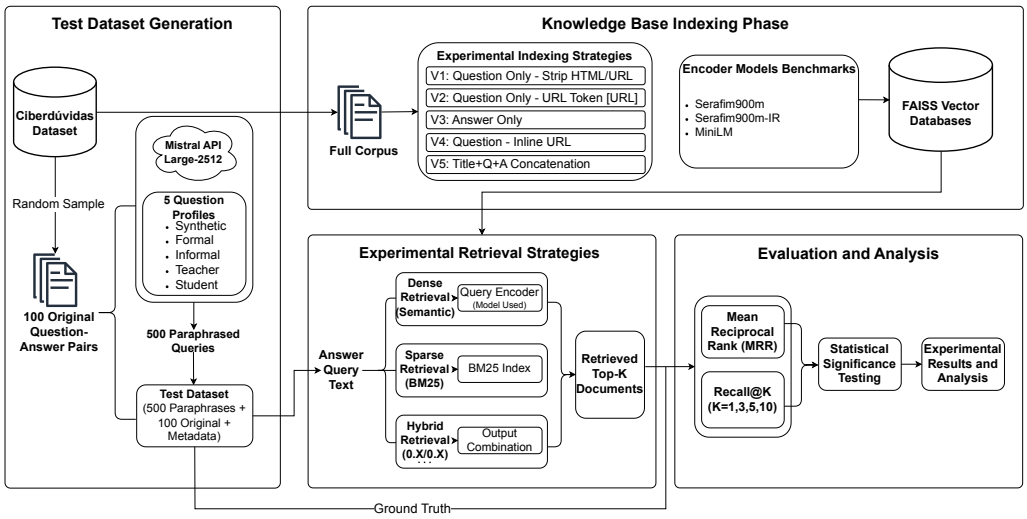

Figure 1 Overview of the experimental framework, illustrating the test dataset generation, knowledge base indexing phase, experimental retrieval strategies, and the evaluation pipeline.

4.1 Test Dataset Generation and Paraphrasing

In this context, it will be essential to correctly match each user's query with the appropriate document, regardless of the user's articulation, formality, or vocabulary, bridging "the gap between query and document vocabulary" [17]. To simulate real-world variance, an evaluation dataset was constructed by sampling 100 original questions from the Ciberdúvidas corpus, testing the retriever's ability to find documents related to the same query but phrased differently. This paraphrase-based testing approach was used in a similar context by [5], albeit on a small set of manually constructed questions.

Inspired by this approach, to generate the paraphrases, the "mistral-large-2512" model was used through the Mistral API. This model was chosen due to its native compatibility with the Portuguese language, performance, and capacity for structured outputs [1].

The generation of paraphrases was based on the creation of five different query profiles, which aim to portray different question approaches from the target audience of Ciberdúvidas:

- **Synthetic:** Short, direct, search-engine style queries.
- **Formal/Explicit:** Grammatically rigorous and highly detailed queries.
- **Informal/Natural:** Casual, conversational tone queries.
- **Portuguese Language Teacher:** Queries utilising technical pedagogical and linguistic terminology.
- **High School Student:** Direct, comprehension-focused queries.

This methodology yielded a final test suite of 600 queries (100 original + 500 paraphrased variations). Table 1 illustrates a representative example of how a single original consultation was transformed across the five distinct profiles.

Table 1 Example of paraphrases generated for a specific question.

Profile	Portuguese Query	English Translation
Original	Qual é o gentílico de Costa do Marfim?	*What is the demonym of Ivory Coast?*
Synthetic	Gentílico de Costa do Marfim?	*Demonym of Ivory Coast?*
Formal	Poderia indicar-me, com precisão, qual é o termo que designa os naturais ou habitantes da Costa do Marfim?	*Could you accurately indicate the term that designates the natives or inhabitants of Ivory Coast?*
Informal	Então, como é que se chama uma pessoa que é da Costa do Marfim? Não me lembro bem...	*So, what do you call a person who is from Ivory Coast? I can't quite remember...*
Teacher	No âmbito da onomástica geográfica, qual é o etnónimo correspondente ao topónimo 'Costa do Marfim', utilizado para denominar os seus cidadãos ou naturais?	*In the context of geographical onomastics, what is the ethnonym corresponding to the toponym 'Ivory Coast', used to denominate its citizens or natives?*
Student	Na aula falámos dos nomes das pessoas de cada país, mas não me recordo qual é o nome certo para quem nasce na Costa do Marfim. Pode ajudar?	*In class we talked about the names of people from each country, but I don't remember the right name for someone born in Ivory Coast. Can you help?*

4.2 Expert Validation of Dataset

To ensure the relevance and quality of the generated paraphrases, a manual validation was conducted. A specialised linguist from the Ciberdúvidas team performed a blind review of the entire evaluation dataset, comprising all 600 queries.

In this process, the evaluator was provided with the target expert answer alongside the corresponding six query variations (the original query and the five generated paraphrases), but was kept blind to the specific tags assigned to each generated variation. The main objective of this review was to assess whether the synthetic queries maintained semantic fidelity to the original problem without introducing misrepresentations. The linguist categorised each query into one of three classifications:

- **Valid:** The query accurately reflects the original intent and makes sense contextually and linguistically.
- **Valid but with adjusted scope:** The query maintains the core semantic value but introduces a shift in focus, such as deepening the theoretical premise, demanding a more elaborate explanation or using overly specific constraints compared to the original question.
- **Invalid:** The query fails to capture the original intent.

The outcome of the evaluation demonstrated high efficacy of the framework. Out of the 600 queries analysed, 573 (95.5%) were classified as Valid. A further 22 queries (3.6%) were classified as Valid but with adjusted scope. This occurred primarily in the "Teacher" profile, where there was a tendency to expand the query by including morphologically complex vocabulary and requesting additional details. Only 5 queries (0.8%) were deemed Invalid, underscoring the reliability of the dataset. In addition to the categorization, the expert provided positive qualitative feedback, noting that the dataset is highly relevant and presents a realistic challenge for the Ciberdúvidas, mirroring the diverse real-world formulations encountered by the Ciberdúvidas platform on a daily basis. Despite these minor deviations, all 600 queries were retained in the final benchmark to faithfully represent the inherent noise and edge cases present in real-world user submissions and automated query processing.

4.3 Data Indexing Strategies

To evaluate the impact of data representation on retrieval performance, several indexing strategies were explored by varying two key dimensions: the target text selected for embedding and the preprocessing applied to the textual content. Rather than present all possible configurations, we selected a subset of strategies that we considered most representative and relevant for the study.

Regarding the object of encoding, three main targets were explored: the Question, representing the user's original information need and serving as the primary retrieval anchor; the Answer, which emphasises the resolution or explanation and enables a query-to-answer matching paradigm; and a combined Title + Question + Answer, providing a more holistic representation by aggregating multiple semantic signals, albeit at the cost of increased verbosity.

In parallel, different preprocessing strategies were applied to assess their impact on embedding quality. Across all variants, a baseline normalization pipeline ensured textual consistency, including the conversion of typographical quotes, the collapsing of multiple whitespace characters, and the trimming of leading and trailing spaces. Beyond this baseline, variations focused on the treatment of URL elements: URLs could be completely removed from the text and stored in metadata, replaced with a placeholder token (e.g., `[URL]`) while preserving original links in metadata, or kept inline in a human-readable format (e.g., `text (url)`).

The combinations of these elements have given rise to various variants, described below:

- **V1** – Target is the Question. HTML and URLs are stripped and moved to metadata. *Primary Objective:* Baseline semantic matching.
- **V2** – Target is the Question. URLs are replaced with a `[URL]` token, and original links are moved to metadata. *Primary Objective:* Preserve link context without noise.
- **V3** – Target is the Answer. HTML and URLs are stripped and moved to metadata. *Primary Objective:* Shift semantic focus to resolution.
- **V4** – Target is the Question. HTML is stripped, but URLs are kept inline formatted as `text (url)`. *Primary Objective:* Retain explicit web references.
- **V5** – Target is the Title + Question + Answer. HTML and URLs are stripped and moved to metadata. *Primary Objective:* Holistic semantic representation.

4.4 Encoder Models

To evaluate the impact of embedding representations on retrieval accuracy, we benchmarked a diverse set of three encoder models. These models were selected to compare the efficacy of European Portuguese-specific training against state-of-the-art multilingual architectures, as well as to assess the impact of model size and task-specific fine-tuning. The evaluated models include:

- **Serafim Family (900M)**: As introduced previously, these represent the state-of-the-art for the Portuguese language. We evaluated both the base sentence encoder (`Serafim900m`) and its Information Retrieval-optimised variant (`Serafim900m-IR`) to isolate and measure the impact of IR-specific loss functions.
- **Paraphrase-Multilingual-MiniLM-L12-v2**: Selected as a lightweight baseline to determine the performance floor of a smaller model. For clarity and consistency throughout this article, this model will be referred to as MiniLM.

4.5 Retrieval Strategies

To benchmark performance, three distinct retrieval strategies were evaluated for the top k=10 documents:

Dense Retrieval (Semantic FAISS): This strategy utilises the high-dimensional embeddings generated by the models to perform semantic similarity searches. By mapping queries and documents into a shared vector space, it identifies relevant content based on linguistic meaning rather than exact word matches.

Sparse Retrieval (BM25): A traditional lexical matching algorithm that serves as a statistical baseline. utilising the `BM25Retriever`, this approach ranks documents based on term frequency and inverse document frequency, effectively capturing exact keyword overlaps and technical terminology.

Hybrid Retrieval: This approach employs an `EnsembleRetriever` that combines the outputs of both BM25 and Semantic FAISS via Reciprocal Rank Fusion, where different weight combinations were empirically tested by Grid search (e.g., 0.1/0.9, 0.2/0.8, ..., 0.9/0.1) to assess the relative contribution of the semantic (Dense Retrieval) and lexical (BM25) components, enabling the selection of the configuration that maximizes overall system performance.

4.6 Evaluation Metrics

Given the current context, we are dealing with a Single-Target Retrieval scenario, since for each query, there is (theoretically) only one correct document. Based on this premise, the system's main objective is to place the correct document as high as possible in the ranking. To evaluate this, we selected the following metrics:

- **Mean Reciprocal Rank (MRR):** Serves as the primary metric for this study. It is expressed as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

  where rank_i is the position of the first relevant document for the i-th query [24].
- **Recall@K (*R@k*):** This metric allows us to observe a Hit Rate scenario. It is expressed by:

$$R@k = \frac{|\text{Relevant Documents} \cap \text{Retrieved Documents@}k|}{|\text{Total Relevant Documents}|}$$

  If the document is in the top K, Recall is 1; otherwise, it is 0. This directly translates to the probability that the answer is among the first K results [11, 24, 6].

Thus, in light of the literature, we chose to exclude other prominent metrics, such as Precision@K, as we consider that these will not add essential information to the case.

To assess whether observed differences in retrieval performance are statistically significant, we adopt paired significance tests commonly used in Information Retrieval evaluation. Following the recommendations of [21], we employ Fisher's randomization test (permutation test), as it directly evaluates differences in the metric of interest without transforming scores into ranks. This approach avoids the loss of magnitude information associated with rank-based tests such as the Wilcoxon signed-rank test. Although the study [21] focused on the Mean Absolute Precision metric, the researchers note similar behaviour when applied to the Mean Reciprocal Rank.

All tests in these experiments are conducted over the full set of 600 queries, over Mean Reciprocal Rank, using paired observations per query. The permutation test was computed using 100,000 random shuffles and significance is determined at the $\alpha = 0.05$ level.

5 Experiments

To systematically evaluate the proposed architecture, we conducted a comprehensive series of experiments analyzing three primary dimensions: the impact of the indexing strategy, the performance of the encoder models, and the efficacy of different retrieval paradigms.

5.1 Indexing Strategy Study

The first phase of our evaluation aimed to determine the optimal textual representation for the vector database. To isolate this variable, we established a control environment utilising the IR-specialised encoder (Serafim900m-IR) and the dense retrieval method (Semantic FAISS). We evaluated the five preprocessing variants (V1 through V5).

To rigorously evaluate the marginal differences between V1 and V2, a Randomization test (Fisher's permutation test) was conducted across the 600 paired queries for comparison. The randomization test revealed no statistically significant difference in performance ($p = 0.12240$),

Table 2 Indexing Strategies Results.

Variant	MRR	R@1	R@3	R@5	R@10
V1	**0.9340**	**0.9200**	0.9450	0.9533	0.9567
V2	0.9320	0.9167	0.9467	0.9533	0.9550
V3	0.6217	0.5550	0.6767	0.7083	0.7383
V4	0.9276	0.9083	0.9450	0.9517	0.9550
V5	0.9314	0.9083	**0.9500**	**0.9617**	**0.9700**

confirming that the inclusion of the `[URL]` token does not mathematically degrade semantic retrieval. These results suggest that URL treatment has no substantial effect on semantic retrieval performance. This shows the model's ability to handle these nuances in the text. With regard to V3, it is interesting to note results that clearly differ from those observed in [19], since in this context this strategy proved to be of little use in all metrics. This may have been partly due to the truncation of documents, as the context was exceeded in 5,221 cases. The V1 variants had the highest MRR and Recall@1, with a very small difference to the V5 approach. A statistical comparison between these approaches also revealed no significant difference ($p = 0.93605$), leading to the conclusion that there is no statistical evidence to distinguish between the two approaches.

Given this, it is interesting to look at other metrics that allow for an objective conclusion about the strategy to choose.

Table 3 Token count statistics across the five indexing variants. All variants consist of exactly 29,145 documents. Tokens were calculated using the `Serafim900m-IR` tokenizer.

Variant	Average Tokens ($\mu \pm \sigma$)	Min Tokens	Max Tokens	>512 Tokens
V1	98.1 ± 123.7	5	5540	361
V2	98.1 ± 123.5	5	5540	360
V3	326.9 ± 316.6	4	5743	5221
V4	101.4 ± 128.7	5	5540	400
V5	446.7 ± 374.3	32	7839	8439

Analysis of token count statistics across the indexing variants shows that V1, V2, and V4 maintain relatively low average token lengths ($\sim$ 100 tokens), while V3 and V5 significantly increase the number of tokens due to the inclusion of expert answers and full document context.

Given the Serafim900m-IR model's limit of 512 tokens, we also concluded that some documents suffered from truncation, with part of the response being truncated (although stored in metadata for future context). This result is even more significant in V5, with 8439 documents undergoing this process.

Based on the limitations described, V2 will be evaluated in subsequent phases. Although the quantitative advantage does not show clear improvements over approach 1 and 4, the presence of a placeholder indicating a URL helps the model to syntactically structure the sentence, ensuring there is no abrupt break between words. In addition, this variant demonstrates excellent token efficiency when compared to V5.

5.2 Encoder Model Benchmark

Having established the V2 indexing strategy as the optimal baseline for semantic alignment, the second phase of our evaluation focuses on benchmarking the performance of different text-encoder models. To strictly isolate the embedding quality and semantic representation capabilities of each model, this comparative analysis evaluates performance exclusively under the dense retrieval paradigm (Semantic FAISS), omitting sparse lexical features.

Table 4 Global dense retrieval performance (Semantic FAISS) of encoder models using the V2 indexing strategy.

Encoder Model	MRR	R@1	R@3	R@5	R@10
Serafim900m	0.8705	0.8417	0.8900	0.9100	0.9267
Serafim900m-IR	**0.9320**	**0.9167**	**0.9467**	**0.9533**	**0.9550**
MiniLM	0.7052	0.6767	0.7217	0.7400	0.7717

As shown in Table 4, the Serafim900m-IR model delivered the best performance across all metrics, achieving an MRR of 0.9320. It is nevertheless important to note that the results obtained by MiniLM provide interesting data for a multilingual baseline, making it an acceptable lightweight alternative.

To ensure that the performance gap between the base Serafim900m model and its Information Retrieval variant was not due to statistical noise, another statistical test was conducted across the 600 paired queries. The test confirmed that the superiority of Serafim900m-IR over Serafim900m is statistically significant ($p = 0.00002$), validating the benefit of task-specific fine-tuning.

5.3 Retrieval Strategy Comparison

For the retrieval strategy analysis, we fixed the text representation to Variant 2 and analysed the performance using the top-performing model (`Serafim900m-IR`).

For this analysis, not only were simple retrieval strategies (Semantic and BM25) used, but a hybrid approach employing Reciprocal Rank Fusion was also adopted. To parametrise this approach, a grid search was carried out, as shown in Table 5.

Table 5 Results for Retrieval Strategies with (Semantic / BM25).

Retriever (Semantic / BM25)	MRR	R@1	R@3	R@5	R@10
0.0 / 1.0 (BM25)	0.6455	0.6267	0.6583	0.6683	0.6900
1.0 / 0.0 (Semantic)	**0.9320**	**0.9167**	**0.9467**	0.9533	0.9550
0.1 / 0.9	0.6957	0.6550	0.6883	0.6883	0.6933
0.2 / 0.8	0.7015	0.6650	0.6883	0.6883	0.6933
0.3 / 0.7	0.7015	0.6650	0.6883	0.6883	0.6933
0.4 / 0.6	0.7015	0.6650	0.6883	0.6883	0.6933
0.5 / 0.5	0.8073	0.6800	0.9283	0.9533	**0.9650**
0.6 / 0.4	0.9042	0.8633	0.9450	**0.9550**	0.9550
0.7 / 0.3	0.9061	0.8667	0.9450	**0.9550**	0.9550
0.8 / 0.2	0.9061	0.8667	0.9450	**0.9550**	0.9550
0.9 / 0.1	0.9124	0.8767	**0.9467**	**0.9550**	0.9550

The results, presented in Table 5, reveal a clear directional trend: retrieval performance scales almost monotonically with the semantic weight.

Configurations heavily reliant on lexical matching (semantic weights from 0.1 to 0.4) performed poorly, plateauing around a Mean Reciprocal Rank (MRR) of 0.7015. As the semantic weight increased past the 0.5 threshold, a significant performance leap occurred, jumping to an MRR of 0.9042 at the 0.6/0.4 split. The best hybrid performance was achieved with the maximum semantic weighting, in the 0.9/0.1 strategy, with an MRR of 0.9124.

As shown, the MRR of dense retrieval strategy significantly outperforms both the Sparse (BM25) and the best Hybrid approach across all metrics. BM25 establishes a baseline MRR of 0.6455, suffering visibly from the "vocabulary mismatch" problem when users formulate queries using terminology distinct from the indexed documents.

In contrast to prevailing trends in the literature, where hybrid retrieval with dense and sparse methods, is generally assumed to "capture different relevance features and (...) benefit from each other by leveraging complementary relevance information" [8], our ensemble approach consistently underperformed the the purely semantic strategy. The BM25 component frequently ranked lexically similar but semantically misaligned documents highly. When combined through weighted rank fusion, these noisy lexical rankings displaced correct dense-retrieval results, reducing the MRR.

It is also worth noting the interesting result obtained in R@10 with the (0.5/0.5) configuration. This showed a slight improvement over semantic retrieval, suggesting that further development of a reranker could potentially lead to an increase in MRR in these hybrid strategies.

6 Analysis and Discussion

To fully grasp the operational realities of deploying this architecture, it is necessary to examine how our chosen configuration interacts with the inherent semantic and structural complexities of natural user queries. Therefore, it is essential to explore the system's robustness across diverse user profiles, investigate specific linguistic patterns that influence retrieval difficulty, and critically assess the limitations of our current framework to contextualize the broader contributions and future directions of this research.

6.1 Robustness Across Query Profiles

To better understand the retrieval limitations, we disaggregated the performance metrics across the five synthesized query profiles. This analysis allows us to observe how different user articulations impact the system's ability to locate the correct document.

Table 6 Mean Reciprocal Rank (MRR) by Query Profile using Variant 2.

Query Type	Sparse	Semantic FAISS	
	BM25	**Serafim900m-IR**	**miniLM**
Original	1.0000	1.0000	1.0000
Formal	0.5654	0.9045	0.6230
Informal	0.6215	0.9575	0.8215
Synthetic	0.7041	0.9583	0.7482
Student	0.5636	0.9417	0.6934
Teacher	0.4185	0.8299	0.3449

It is interesting to note, first of all, that the differences between profiles in BM25 are much more pronounced than those obtained using semantic retrieval. This demonstrates that this type of retrieval is highly adaptable to vocabulary mismatches.

In addition, the query type that consistently yields the worst results is the "teacher" query. This can be seen in the analysis of an example query from the test set:

> **Original Question:** Qual é a diferença entre vem, vêm e veem? *[Eng: What is the difference between "vem" (coming), "vêm" (are coming) and "veem" (see)?]*

> **Teacher:** No âmbito do paradigma flexional do verbo vir e do verbo ver no presente do indicativo, como se articulam as formas "vem", "vêm" e "veem" em termos de número, pessoa gramatical e valor semântico, considerando as especificidades do português europeu? *[Eng: Within the inflectional paradigm of the verbs "vir" (come) and "ver" (see) in the present indicative, how do the forms "vem" (coming), "vêm" (are coming), and "veem" (see) relate in terms of number, grammatical person, and semantic value, considering the specificities of European Portuguese?]*

It is clear from the example above that, in some cases, the model not only expanded the vocabulary of the original question with specific technical terminology but also gave the query a more specific purpose, as previously highlighted by the linguistic expert. This is not a problem, since, in future real-world applications, it is desirable for questions (even those with different objectives but similar scopes), to serve as context for the user or system using this retriever. Nevertheless, this makes this paraphrase profile more challenging and explains why the results obtained, although high, are lower than those observed in other paraphrase profiles.

6.2 Query Difficulty and Patterns Observed

The results across the various query profiles show that query difficulty differs markedly from one profile to another. Consequently, it is crucial to analyze specific examples of the generated queries and determine the features that characterize the paraphrasing for each profile.

One of the challenges identified involves queries that focus on a direct quote, a specific word, or other elements that require reference throughout the paraphrases, without the possibility of using synonyms. For example:

> **Original Question:** No exemplo «Eu fui lá, à praia», estamos perante um complemento oblíquo («lá, à praia»), ou dois complementos ligados por uma vírgula? *[Eng: In the example "I went there, to the beach," are we dealing with an oblique complement ("there, to the beach"), or two complements connected by a comma?]*

> **Informal:** Olha, na frase «Eu fui lá, à praia», o «lá, à praia» é um complemento só ou são dois? Tipo, a vírgula separa mesmo duas coisas diferentes? – *[Eng: Look, in the sentence "I went there, to the beach," is "there, to the beach" a single complement or are there two? Like, does the comma really separate two different things?...]*

In this exact case, the question exhibited a 100% retrieval success rate. The inclusion of direct, word-for-word quotations certainly contributed to this, since the paraphrases, while adapting the context to the profile, had to retain a quotation to give the question its purpose.

The same holds in some examples that include large quoted passages from books or similar sources. This creates a dilemma: Leaving such examples in the test set could indirectly inflate performance metrics; on the other hand, they reflect real user questions that are likely to reappear, and their successful retrieval indicates that the system is functioning

correctly. Moreover, removing these questions from the test set would require a clear, specific rule, which inevitably introduces subjectivity: Should all questions containing quotations be removed from the test set? Only those with quotes longer than X words? On what basis?

6.3 Limitations and Future Work

The main limitation identified stems from the overestimation of the results obtained, particularly when examining the MRR metric. Given the nature of the proposed scenario, it was necessary to find the perfect balance between paraphrasing in a challenging way and altering the question so substantially that it no longer made sense for the retrieval evaluation. Consequently, in examples that relied on quotations or questions about exact words, these had to be maintained across the variations, substantially increasing the ease of retrieval. In addition, due to the syntactic nature of the dataset, it may not fully replicate certain characteristics of the end user, such as spelling errors.

Based on this analysis, an improved approach might involve creating a manual dataset based on an analysis of user interaction with the platform. This data could stem from interaction with a potential future conversational agent or natural language-based search system. These tools do not yet exist on the portal, but their future data could contribute to this development.

The observed performance gaps across query profiles and the underperformance of hybrid retrieval further motivate the exploration of advanced retrieval enhancement techniques. In the consulted literature, promising directions for future work emerge. First, the integration of reranking models [22] represents a natural extension of the current pipeline. In this approach, an initial set of candidate documents retrieved by the dense model would be re-evaluated and reordered using a cross-encoder or specialised reranker, potentially improving the precision of top-ranked results. Given that the current system already achieves high recall, reranking may be particularly effective in optimising ranking quality, especially in challenging or ambiguous queries. Also, query rewriting techniques [15, 6] could be explored to address the variability observed across query profiles, especially in more complex formulations such as the "Teacher" queries. By transforming user inputs into more standardized formulations, rewriting has the potential to further improve alignment between queries and indexed documents.

Given the solid foundation provided by the findings of this analysis, future work should also focus on evaluating solutions that enable the integration of this process into the platform. One of the key approaches to be evaluated involves the development of a Retrieval-Augmented Generation system, based on a similar process for searching and using these documents.

6.4 Contributions

First, this research provides a systematic analysis of the retrieval layer as the primary focus of study. Based on controlled experiments, we demonstrate how variations in indexing strategies, embedding models, and retrieval paradigms directly affect the system's ability to align queries with the knowledge base. Furthermore, this study proposes an evaluation framework that can be extended to other domains of question-answering systems. By constructing a test dataset based on paraphrases, this work establishes a realistic benchmark for retrieval evaluation.

The inclusion of specialised validation reinforces this approach and ensures the viability of the strategy employed. Also, this study enables a direct comparison of a state-of-the-art monolingual model for Portuguese with a robust multilingual baseline, revealing clear improvements that are supported by the relevant literature. At the same time, the study highlights the significant gains obtained through task-specific fine-tuning, as demonstrated by the performance improvements of the IR-optimised Serafim model.

From an artefact-oriented perspective, the results of this study enable an immediate improvement to the retrieval system in the "*Ciberdúvidas da Língua Portuguesa*" project, when compared to the current keyword-based search mechanism, enabling the future work discussed earlier. Consequently, these contributions establish a solid foundation for future advancements, especially in the direction of integrating more sophisticated retrieval and generation techniques, ultimately supporting the development of more accurate and user-aligned question-answering systems.

References

1 Mistral AI. Mistral large 3 - mistral ai | mistral docs. URL: `https://docs.mistral.ai/models/mistral-large-3-25-12`.

2 Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset. *CEUR Workshop Proceedings*, 1773, November 2016. `arXiv:1611.09268`.

3 Debopriyo Banerjee, Mausam Jain, and Ashish Kulkarni. Mfbe: Leveraging multi-field information of faqs for efficient dense retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13937 LNCS:112–124, March 2023. `doi:10.1007/978-3-031-33380-4_9`.

4 Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Zeta Alpha, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*, August 2021. `arXiv:2108.13897`.

5 Nuno Carriço and Paulo Quaresma. Sentence embeddings and sentence similarity for portuguese faqs. In *Proceedings of IberSPEECH 2021*, 2021. `doi:10.21437/IberSPEECH.2021-43`.

6 Daksh Chaudhary, Sri Lakshmi Vadlamani, Dimple Thomas, Shiva Nejati, and Mehrdad Sabetzadeh. Developing a llama-based chatbot for ci/cd question answering: A case study at ericsson. *Proceedings - 2024 IEEE International Conference on Software Maintenance and Evolution, ICSME 2024*, pages 707–718, 2024. `doi:10.1109/ICSME58944.2024.00075`.

7 Ciberdúvidas da Língua Portuguesa. Quem somos. URL: `https://ciberduvidas.iscte-iul.pt/quem-somos`.

8 Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, pages 166–169, December 2023. `doi:10.1109/AIxSET62544.2024.00030`.

9 Samira Ghodratnama and Mehrdad Zakershahrak. Adapting llms for efficient, personalized information retrieval: Methods and implications. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14518 LNCS:17–26, 2024. `doi:10.1007/978-981-97-0989-2_2`.

10 Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. Open sentence embeddings for portuguese with the serafim pt* encoders family. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14969 LNAI:267–279, 2025. `doi:10.1007/978-3-031-73503-5_22`.

11 Md Saleh Ibtasham, Sarmad Bashir, Muhammad Abbas, Zulqarnain Haider, Mehrdad Saadatmand, and Antonio Cicchetti. Reqrag: Enhancing software release management through retrieval-augmented llms: An industrial study. *Lecture Notes in Computer Science*, 15588 LNCS:277–292, 2025. `doi:10.1007/978-3-031-88531-0_20`.

12 Vanja Mladen Karan, Lovro Žmak, and Jan Šnajder. Frequently asked questions retrieval for croatian based on semantic textual similarity. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 24–33. Association for Computational Linguistics, 2013. URL: `https://aclanthology.org/W13-2405/`.

13 Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. Dense passage retrieval for open-domain question answering. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 6769–6781, 2020. doi:10.18653/V1/2020.EMNLP-MAIN.550.

14 Ricardo Lopes, Joao Magalhaes, and David Semedo. Glória: A generative and open large language model for portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 441–453, Santiago de Compostela, Galicia/Spain, March 2024. Association for Computational Linguistics. URL: https://aclanthology.org/2024.propor-1.45/.

15 Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 5303–5315, May 2023. doi:10.18653/v1/2023.emnlp-main.322.

16 Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. Unsupervised faq retrieval with question generation and bert. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 807–812, 2020. doi:10.18653/V1/2020.ACL-MAIN.74.

17 Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13:1–126, December 2018. doi:10.1561/1500000061.

18 Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. Multilingual vs. monolingual transformer models in encoding linguistic structure and lexical abstraction, 2025. URL: https://aclanthology.org/2025.clicit-1.78/.

19 Mindi Richia Putri, Ario Yudo Husodo, and Budi Irmawati. Simplification of embedding process in retrieval augmented generation for optimizing question answering chatbot model. *COMNETSAT 2024 - IEEE International Conference on Communication, Networks and Satellite*, pages 665–670, 2024. doi:10.1109/COMNETSAT63286.2024.10862926.

20 João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. Advancing neural encoding of portuguese with transformer albertina pt-*. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Proceedings*, volume 14115 of *Lecture Notes in Artificial Intelligence*, pages 441–453, Faial Island, Azores, Portugal, September 2023. Springer. doi:10.1007/978-3-031-49008-8_35.

21 Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. *International Conference on Information and Knowledge Management*, pages 623–632, 2007. doi:10.1145/1321440.1321528.

22 Hassan Soliman, Hitesh Kotte, Miloš Kravčík, Norbert Pengel, and Nghia Duong-Trung. Retrieval-augmented chatbots for scalable educational support in higher education. *CEUR Workshop Proceedings*, 2025. URL: https://www.scopus.com/pages/publications/105011268364?origin=resultslist.

23 Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Advances in Neural Information Processing Systems*, April 2021. arXiv:2104.08663.

24 S. Vijayakumaran, M. Vimaladevi, R. Thangamani, N. Vibeesh, M. Chandru, and Sathishkumar Veerappampalayam Easwaramoorthy. Revolutionizing legal access: An ai-driven rag chatbot for real-time judicial insights. *Proceedings - 3rd International Conference on Artificial Intelligence and Machine Learning Applications: Healthcare and Internet of Things, AIMLA 2025*, 2025. doi:10.1109/AIMLA63829.2025.11040439.

25 Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, February 2024. arXiv:2212.03533.

26 Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, February 2024. doi:10.48550/arXiv.2402.05672.

27 Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi:10.1162/TACL_A_00595.

28 Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, January 2021. arXiv:2101.00774.