Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5659

Evaluating Reranking Strategies for Portuguese Information Retrieval: Fine-Tuning, LLMs, and Sociocultural Aspects

Renato Okabayashi Miyaji [**Escola Politécnica da Universidade de São Paulo** | ***re.miyaji@usp.br***]
Pedro Luiz Pizzigatti Corrêa [**Escola Politécnica da Universidade de São Paulo** | ***pedro.correa@usp.br***]

Escola Politécnica, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 380 - Butantã, São Paulo - SP, 05508-010, Brazil.

Received: 09 March 2025 • **Accepted:** 11 December 2025 • **Published:** 05 May 2026

Abstract. Reranking plays a crucial role in improving Information Retrieval (IR) performance, particularly in low-resource languages, such as Portuguese. In this study, we evaluate different reranking strategies for Portuguese IR, comparing multilingual and Portuguese-specific models, as well as not-so-large language models and large language models (LLMs). We assess the performance of BM25 combined with ptT5 fine-tuned on multilingual and Brazilian Portuguese datasets, alongside multilingual state-of-the-art rerankers (BGE m3) and LLM as rerankers RankGPT (GPT-4) and Sabiá 3, a Portuguese-specific LLM. Additionally, we introduce a novel dynamic In-Context Learning (DICL) prompting strategy to enhance LLM performance. Experiments conducted on the Quati and Pirá 2.0 datasets show that fine-tuning on native Brazilian Portuguese data significantly improves retrieval effectiveness by up to 5 p.p. in nDCG compared to using translated multilingual datasets. Two fine-tuning approaches were tested: a binary classification strategy with 'true' and 'false' tokens and a relevance score-based training, both outperforming models fine-tuned on translated multilingual data. RankGPT achieved the best overall results, yet Sabiá 3 demonstrated competitive performance, particularly on queries related to sociocultural aspects. The DICL strategy further improved the results of both LLMs, significantly boosting their MRR@10. These findings highlight the importance of language-specific training and suggest that not-so-large language models can be viable alternatives for reranking tasks in Portuguese IR.

Keywords: Information Retrieval, Reranking, Fine-Tuning, Large Language Models, Not-so-Large Language Models

1 Introduction

The rapid advancements in neural reranking models have significantly improved Information Retrieval (IR) effectiveness by leveraging pre-trained language models to refine initial ranking outputs [Laitz *et al*., 2025]. These models have been extensively explored in English-centric datasets, such as MS MARCO [Nguyen *et al*., 2016], but their performance in low-resource languages, including Portuguese, remains a critical challenge. While multilingual models, such as mT5 [Xue *et al*., 2021], offer broad language coverage, their ability to capture linguistic and cultural nuances in Portuguese remains questionable [Bueno *et al*., 2024]. Consequently, there is an increasing demand for Portuguese-specific rerankers, particularly in domain-specific and sociocultural queries.

Recent efforts to develop Portuguese language resources have enabled the construction of more effective rerankers for the language. Notably, datasets such as Quati [Bueno *et al*., 2024] and Pirá 2.0 [Pirozelli *et al*., 2024] provide valuable benchmarks for evaluating retrieval models on native data. Additionally, models such as ptT5 [Carmo *et al*., 2020] have been fine-tuned for Portuguese, but their effectiveness in reranking tasks compared to multilingual counterparts remains an open question. Beyond language-specific adaptations, there is also a growing interest in smaller, computationally efficient rerankers, such as those proposed in InRanker [Laitz *et al*., 2025], which aims to balance performance and scalability.

A crucial aspect in reranking research is the impact of training data. Many existing Portuguese IR models rely on translated multilingual datasets, which may introduce biases and fail to capture native linguistic expressions. Fine-tuning on datasets generated by native speakers, such as Quati [Bueno *et al*., 2024], has been proposed as a solution, but the benefits of this approach over using translated multilingual corpora require further investigation. The presence of sociocultural biases in IR models is another concern, particularly for applications requiring culturally aware retrieval [Bueno *et al*., 2024].

This study systematically evaluates reranking strategies for Portuguese IR through a comparative analysis of different models and training approaches. Specifically, we assess two fine-tuning strategies for the ptT5 [Carmo *et al*., 2020] reranker, leveraging both the multilingual MS MARCO [Bonifacio *et al*., 2021] and the native Brazilian Portuguese Quati [Bueno *et al*., 2024] datasets. We then benchmark these specialized models against contemporary multilingual approaches, including the embedding-based BGE m3 [Chen *et al*., 2024] and the state-of-the-art listwise LLM-reranker, RankGPT [Sun *et al*., 2023], as well as the Portuguese-specific LLM, Sabiá 3 [Abonizio *et al*., 2024]. The performance of all models is evaluated on the Quati and Pirá 2.0 [Pirozelli *et al*., 2024] datasets using standard IR metrics (NDCG, MRR@10, and Recall@10). A key component of our analysis is a focused evaluation on sociocultural queries, designed to probe whether Portuguese-specific

rerankers more effectively capture subtle cultural and linguistic nuances.

One of the key contributions of this study is the first evaluation of Sabiá 3 [Abonizio *et al*., 2024] as a reranker, marking its debut in the literature. Furthermore, we propose and test a novel listwise reranking method that applies dynamic In-Context Learning to the reranking prompts for both RankGPT and Sabiá 3. This dynamic selection is achieved by calculating the similarity between the query being analyzed and a base of few-shot examples, allowing the model to leverage the most relevant examples for each specific reranking task. Alongside these LLM-based strategies, we also propose two distinct fine-tuning approaches for BM25 [Jones *et al*., 2000] + ptT5 [Carmo *et al*., 2020]: a sequence-to-sequence strategy using 'true' and 'false' tokens [Nogueira *et al*., 2020], and a relevance score-based training. Our findings demonstrate that fine-tuning on native Brazilian Portuguese datasets outperforms models fine-tuned on translated multilingual corpora, both on Quati [Bueno *et al*., 2024] and Pirá 2.0 [Pirozelli *et al*., 2024]. Additionally, while RankGPT achieves the best overall performance, Sabiá 3 attains comparable results on sociocultural queries, highlighting the potential of Portuguese-specific LLMs in reranking. This comparison is particularly noteworthy given the substantial disparity in scale and computational cost between the massive, general-purpose model powering RankGPT (GPT-4) and the more specialized and resource-efficient Sabiá 3, positioning the latter as a viable alternative for specific applications.

This research is based on the hypothesis that fine-tuning on Portuguese-specific datasets leads to improved retrieval performance compared to multilingual models, as it better captures the linguistic nuances and contextual intricacies of Brazilian Portuguese. Additionally, we hypothesize that Portuguese-specific rerankers are more effective in handling queries that involve sociocultural aspects, as they are trained on data that reflects native language usage and cultural references. Furthermore, we explore the trade-offs between using not-so-large language models, such as ptT5 [Carmo *et al*., 2020], and large language models (LLMs), such as GPT-4, as rerankers for Portuguese information retrieval (IR). In this context, we introduce a new hypothesis: that the zero-shot performance of LLMs can be significantly enhanced through our proposed dynamic few-shot learning [Brown *et al*., 2020] method, which provides contextually relevant examples at inference time. We hypothesize that while LLMs may achieve superior ranking performance due to their extensive pretraining, smaller models fine-tuned on domain-specific data could offer competitive results with lower computational cost. By testing these hypotheses, we aim to contribute to a better understanding of how training data, fine-tuning, and In-Context Learning strategies affect Portuguese IR performance.

The remainder of this paper is organized as follows: Section 2 discusses related work on reranking, multilingual models, and Portuguese-specific IR datasets. Section 3 details our methodology, including model selection, fine-tuning strategies, our proposed method, and evaluation metrics. Section 4 presents our experimental results and analyses, with a particular focus on sociocultural queries. Finally, Section 5 concludes the paper, summarizing key findings and proposing directions for future research.

2 Related Works

In this section, we review previous research relevant to our study, focusing on three key areas. First, we explore datasets for Portuguese Information Retrieval, highlighting both native datasets and multilingual datasets translated into Portuguese, which serve as benchmarks for evaluating retrieval models. Second, we discuss rerankers for Portuguese Information Retrieval, covering traditional BM25-based approaches, fine-tuned sequence-to-sequence rerankers, and embedding-based methods. Lastly, we examine the role of Large Language Models (LLMs) for reranking, detailing different ranking strategies. These discussions provide the necessary background to contextualize our experimental comparisons and contributions.

2.1 Datasets for Portuguese Information Retrieval

The availability of high-quality datasets is crucial for the development and evaluation of Information Retrieval (IR) models. In the context of Portuguese IR, datasets have historically been limited, with many retrieval models relying on translated multilingual corpora [Bonifacio *et al*., 2021] rather than datasets created by native speakers [Bueno *et al*., 2024]. However, recent efforts have sought to bridge this gap by developing dedicated Portuguese-language resources for passage ranking, question answering, and domain-specific retrieval.

One of the most notable contributions is Quati [Bueno *et al*., 2024], a Brazilian Portuguese IR dataset created using queries from native speakers. Unlike multilingual datasets that are simply translated into Portuguese, Quati [Bueno *et al*., 2024] was constructed and validated by Brazilian users, ensuring a more natural and representative query distribution. It consists of 200 queries and 10 million passages, covering diverse topics, and was designed to support document retrieval and passage ranking tasks [Bueno *et al*., 2024]. The dataset has been used to benchmark BM25 [Jones *et al*., 2000], and other commercial retrievers, such as mT5 [Xue *et al*., 2021].

The creation of the Quati dataset [Bueno *et al*., 2024] was a deliberate effort to address the shortcomings of translated corpora by constructing a benchmark rich with authentic Brazilian Portuguese nuances. Unlike mMARCO [Bonifacio *et al*., 2021], which relies on automated translation, Quati's queries were formulated and validated entirely by native speakers, ensuring they reflect genuine information needs and natural language patterns. This native-centric methodology embeds a wealth of linguistic and contextual intricacies, including colloquial query formulations, culturally-specific references (e.g., to local public figures, historical events, and social phenomena), and idiomatic expressions that are often flattened or lost in translation [Bueno *et al*., 2024]. Consequently, Quati provides a more rigorous and realistic testbed for evaluating a model's ability to move beyond literal term matching and achieve a true, culturally-grounded understanding of the Portuguese language as it is spoken and written in Brazil.

Another relevant dataset is Pirá [Paschoal *et al*., 2021], a bilingual Portuguese-English dataset designed for question answering (QA) about the ocean, climate change, and the

Brazilian coastline. The dataset was later extended into Pirá 2.0 [Pirozelli *et al*., 2024], providing benchmarks for specific tasks, such as question answering, machine reading comprehension, information retrieval, open question answering, answer triggering, and multiple choice question answering. This dataset is particularly relevant for domain-specific IR, providing a valuable benchmark for evaluating retrieval models in a scientific and environmental context. The benchmarks for Pirá 2.0 [Pirozelli *et al*., 2024] highlight the performance differences between monolingual and multilingual models, emphasizing the impact of domain-specific training data on retrieval effectiveness.

In addition to these general-purpose datasets, REGIS [Oliveira *et al*., 2021] provides a specialized test collection for geoscientific document retrieval in Portuguese. REGIS contains scientific papers and structured metadata, allowing for complex retrieval tasks in specialized fields. This dataset demonstrates the need for domain-specific IR benchmarks, as models trained on general-purpose corpora may not perform optimally in specialized domains [Oliveira *et al*., 2021].

Beyond Portuguese-specific resources, many IR models have leveraged multilingual datasets translated into Portuguese. A key example is mMARCO [Bonifacio *et al*., 2021], a multilingual adaptation of the MS MARCO passage ranking dataset [Nguyen *et al*., 2016]. While mMARCO provides a large-scale benchmark for training retrieval models in Portuguese, its reliance on automated translations raises concerns about fluency, cultural relevance, and query intent preservation [Bueno *et al*., 2024]. Similarly, datasets like RCV1 [Lewis *et al*., 2004] have been adapted for multilingual text classification and retrieval tasks, but they primarily focus on news categorization, limiting their applicability to open-domain IR.

These datasets collectively highlight the evolution of Portuguese IR benchmarks, from translated multilingual corpora to native speaker-generated datasets. However, a key research question remains regarding the advantage of Portuguese-specific datasets over translated corpora in IR tasks. Our study addresses this by evaluating reranking models fine-tuned on both native and translated datasets, comparing their performance in general and sociocultural IR tasks.

2.2 Rerankers for Portuguese Information Retrieval

Reranking models play a crucial role in modern Information Retrieval (IR) pipelines, refining the initial ranked list of documents retrieved by sparse or dense retrievers. A widely adopted two-stage approach [Nogueira *et al*., 2019] consists of: an initial retrieval step using models such as BM25 [Jones *et al*., 2000], which efficiently retrieves a set of candidate passages, and a reranking step, where a more complex neural model is used to refine the ranking based on deeper semantic understanding. This second step has been significantly enhanced by the introduction of pretrained transformer-based models fine-tuned for passage ranking tasks [Nogueira *et al*., 2019] [Bonifacio *et al*., 2021].

A key contribution in this area is the mT5-based reranker, fine-tuned on mMARCO [Bonifacio *et al*., 2021], a multilingual adaptation of the MS MARCO dataset. The mT5 model [Xue *et al*., 2021] follows a sequence-to-sequence paradigm, treating document ranking as a text generation task. Inspired by Nogueira *et al*. [2020], the fine-tuning strategy for mT5 trains the model to output a ”yes” token for relevant passages and a ”no” token for irrelevant ones. This formulation allows mT5 to learn context-aware relevance assessments.

An alternative fine-tuned model is ptT5 [Carmo *et al*., 2020], a variant of T5 specifically adapted for Portuguese. Similar to mT5, ptT5 has been fine-tuned on mMARCO [Bonifacio *et al*., 2021], leveraging translated multilingual training data for passage ranking. However, due to the linguistic differences between translated and native Portuguese corpora, fine-tuning on datasets created by native speakers, such as Quati [Bueno *et al*., 2024], may provide more natural query-document associations, enhancing ranking effectiveness in real-world IR scenarios.

Beyond T5-based rerankers, recent research has explored more efficient ranking approaches. InRanker [Laitz *et al*., 2025] introduces a distilled reranking model designed for zero-shot and few-shot IR tasks. Unlike full-scale LLM-based rerankers, InRanker is optimized for computational efficiency. By leveraging knowledge distillation, InRanker retains strong ranking capabilities while significantly reducing computational costs [Laitz *et al*., 2025].

The effectiveness of fine-tuned rerankers is grounded in transfer learning and domain adaptation theories. Following Howard and Ruder [2018], pretrained models can generalize well across tasks but require task-specific fine-tuning to optimize performance. In IR, fine-tuning on domain-specific datasets allows rerankers to better capture query intent and document relevance signals, especially in languages where multilingual models might not be fully optimized. Moreover, training on native speaker datasets rather than translated corpora is an important consideration, as it affects semantic nuances and retrieval accuracy.

Recent advancements in multilingual reranking have introduced BGE M3 [Chen *et al*., 2024], a self-knowledge distillation-based embedding model designed to support multilinguality, multi-functionality, and multi-granularity representations. Unlike traditional monolithic rerankers, which require fine-tuning for specific ranking tasks, BGE M3 leverages a self-distillation approach, where a larger model serves as a teacher to refine a smaller, more efficient student model. This process enables compact yet highly expressive embeddings, making BGE M3 effective across diverse retrieval, reranking, and embedding-based similarity tasks. The architecture is optimized to bridge the gap between sparse and dense retrieval by incorporating multi-granular representations, allowing the model to capture both sentence-level and passage-level relevance signals simultaneously [Chen *et al*., 2024].

Performance evaluations indicate that BGE M3 achieves state-of-the-art results in multilingual retrieval tasks [Chen *et al*., 2024]. The model has been benchmarked across multiple datasets, showing robust performance in both high-resource and low-resource languages, making it a promising alternative for cross-lingual IR applications. Notably, BGE M3 surpasses previous multilingual rerankers. Given these strengths, our study explores its potential in Portuguese-specific IR tasks, comparing its effectiveness with other

rerankers.

Given the aforementioned developments, our study evaluates the effectiveness of rerankers fine-tuned on Portuguese-specific datasets compared to those trained on translated multilingual corpora.

2.3 Large Language Models for Reranking

The advent of Large Language Models (LLMs) has introduced new possibilities for document ranking and reranking in Information Retrieval (IR). Unlike traditional neural rerankers, which are explicitly trained on ranking datasets, LLMs leverage their pretrained knowledge and instruction-following capabilities to perform zero-shot or few-shot ranking [Zhu *et al*., 2024]. Recent studies have investigated LLMs as reranking agents, analyzing their ability to assess relevance and reorder retrieved passages. These approaches primarily fall into three categories: pointwise, pairwise, and listwise methods [Zhu *et al*., 2024].

Pointwise reranking methods treat passage ranking as an independent relevance estimation problem, where an LLM generates a relevance score for each document-query pair separately [Nogueira *et al*., 2020]. A common technique is relevance generation, where the model is prompted to output a numerical score or a binary relevance label (e.g., ”relevant” or ”not relevant”) based on a given query [Guo *et al*., 2025]. Alternatively, LLMs can be used for query generation, where they rewrite or expand the input query to improve retrieval recall before reranking [Sachan *et al*., 2022]. These methods are computationally expensive but have shown strong performance in specialized domains.

Pairwise approaches aim to compare document pairs and determine which passage is more relevant given a query [Zhu *et al*., 2024]. This formulation is inspired by learning-to-rank frameworks and is particularly useful for LLMs, as it aligns with their strong reasoning capabilities. In these methods, an LLM receives two passages at a time and generates a preference judgment, which is then aggregated across multiple comparisons to produce a final ranked list [Qin *et al*., 2024].

Listwise reranking strategies evaluate the entire list of retrieved passages simultaneously, allowing the model to optimize the global ranking order rather than making isolated comparisons [Ma *et al*., 2023]. This is computationally demanding but can lead to better ranking coherence. Some instruction-tuned LLMs, such as RankGPT, have been designed to maximize listwise ranking effectiveness through innovative techniques [Sun *et al*., 2023].

RankGPT is a state-of-the-art LLM-based reranker that surpasses previous neural rerankers by introducing two key innovations. The first one is Instructional Permutation Generation, a method that prompts the model to generate ranked permutations of retrieved passages rather than scoring them individually [Sun *et al*., 2023]. By treating reranking as a text generation task, RankGPT effectively leverages its pretraining on ordered text sequences to improve ranking fluency and consistency.

The second innovation is the Sliding Window Strategy. Since LLMs have context length limitations, RankGPT uses a sliding window approach to process long document lists incrementally, ensuring that no relevant information is lost during reranking [Sun *et al*., 2023].

Empirical results demonstrate that RankGPT (powered by GPT-4) outperforms previous rerankers across multiple benchmarks [Sun *et al*., 2023]. It achieves state-of-the-art performance, particularly in zero-shot retrieval scenarios, reinforcing the viability of LLM-based rerankers as a competitive alternative to traditional fine-tuned models.

Given these developments, our study evaluates Portuguese-specific LLMs, such as Sabiá 3, against multilingual LLM-based rerankers, such as GPT-4, analyzing their effectiveness on Portuguese retrieval tasks. This comparison is particularly relevant considering the better capacity of Portuguese-specific rerankers to capture cultural and linguistic nuances compared to multilingual models. Moreover, it delves into the crucial trade-off between the immense computational cost and scale of a general-purpose model, such as GPT-4 and the potential cost-effectiveness of a specialized, smaller model, such as Sabiá 3, a key consideration for real-world applications.

3 Methodology

3.1 Reranking Strategies

Fine-Tuning Approach. Our methodology follows the widely adopted two-stage retrieval pipeline [Nogueira *et al*., 2019], which consists of an initial retrieval step followed by a reranking stage. In the first stage, we use BM25 [Jones *et al*., 2000] to retrieve an initial ranked list of candidate passages based on lexical matching between the query and document content. In the second stage, we apply a neural reranking model to refine the BM25-ranked list. Inspired by previous works [Nogueira *et al*., 2020] [Bonifacio *et al*., 2021], we fine-tune a sequence-to-sequence model, ptT5 [Carmo *et al*., 2020], using passage-query relevance labels. The goal of this reranking step is to leverage deep contextual understanding beyond simple term frequency matching, allowing the model to reweight and reorder candidate passages based on semantic relevance.

To fine-tune the ptT5 [Carmo *et al*., 2020] rerankers, we follow the methodology introduced by Nogueira *et al*. [2020]. We fine-tune the first ptT5 reranker using a traditional classification head that predicts relevance scores. For the second one, instead of using a traditional classification head, we fine-tune the model as a sequence-to-sequence task, where the model generates a token indicating the passage's relevance to the query. Specifically, during training, the model learns to generate either ”true” (indicating relevance) or ”false” (indicating irrelevance) based on query-document pairs. The query-document relevance score $g(q, d)$ can be computed using the log-likelihood values of the tokens ”true” and ”false”. This is achieved through the softmax function in Equation 1, where T_t and T_f correspond to the log-likelihood scores assigned by the reranker to the tokens ”true” and ”false”, respectively.

$$g(q, d) = \frac{\exp(T_t)}{\exp(T_t) + \exp(T_f)}, \tag{1}$$

This fine-tuning approach aligns with previous work on

document ranking using pre-trained transformers [Nogueira *et al*., 2020]. This method has shown state-of-the-art results in multilingual IR tasks, particularly in low-resource settings where translated datasets can be leveraged for fine-tuning [Bonifacio *et al*., 2021].

Figure 1 presents both fine-tuning approaches for ptT5 reranker: through a traditional classification head that predicts relevance scores and as a sequence-to-sequence task [Nogueira *et al*., 2020].

To ensure that our rerankers capture the linguistic and sociocultural nuances of Brazilian Portuguese, we performed fine-tuning using the training set of the Quati dataset [Bueno *et al*., 2024]. Quati was specifically created by native Brazilian Portuguese speakers, making it a more culturally relevant resource for evaluating retrieval models in Portuguese [Bueno *et al*., 2024]. By leveraging a dataset that authentically represents natural language usage, we aim to improve the model's ability to rank passages that reflect local idioms, discourse structures, and domain-specific knowledge that may not be adequately represented in machine-translated datasets.

For comparison, we also evaluated ptT5 rerankers fine-tuned on the multilingual mMARCO dataset [Bonifacio *et al*., 2021], a multilingual adaptation of MS MARCO that includes machine-translated passages across multiple languages, including Portuguese. While mMARCO provides a large-scale training resource, its translations may lack the linguistic richness and cultural context present in datasets curated by native speakers.

By adopting this approach, we aim to enhance ranking performance for Portuguese IR tasks while comparing models fine-tuned on multilingual translated datasets and those fine-tuned on datasets created by native Portuguese speakers.

BGE m3. In our evaluation, BGE m3 [Chen *et al*., 2024] serves as a powerful, state-of-the-art multilingual baseline. Unlike models that are explicitly fine-tuned for reranking, BGE m3 functions as a zero-shot reranker by leveraging its core capability as a dense embedding model. The process is executed by first encoding the input query and each candidate passage into high-dimensional vector representations using the BGE m3 model. Subsequently, the semantic relevance of each passage to the query is quantified by calculating the cosine similarity between their respective embeddings. The initial list of passages is then reordered based on these similarity scores in descending order, with higher scores indicating greater relevance. This approach relies on the model's pre-trained ability to map semantically similar texts to points that are close in the embedding space, thus providing an effective measure of relevance without requiring any task-specific fine-tuning [Chen *et al*., 2024].

Instructional Permutation Generation. List-wise reranking methods aim to optimize the ranking of an entire list of retrieved documents rather than considering individual documents in isolation (pointwise) [Guo *et al*., 2025] or comparing pairs of documents (pairwise) [Qin *et al*., 2024]. Large Language Models (LLMs) can be adapted to a list-wise ranking paradigm by treating ranking as a sequence generation problem, where the model directly outputs a reordered list of documents based on their estimated relevance to the query. Unlike traditional IR models, which rely on explicit relevance scores, LLMs leverage their contextual understanding and

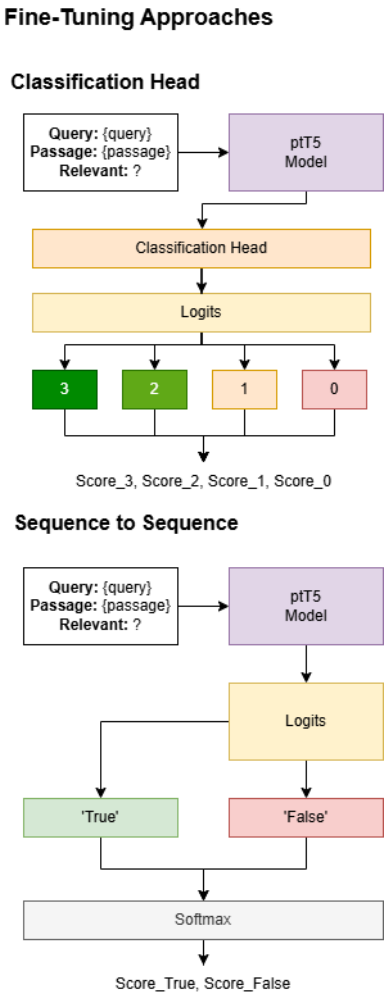

Figure 1. Comparison of the two fine-tuning strategies for the ptT5 reranker. The Classification Head approach frames reranking as a multi-class classification task over discrete relevance scores. The Sequence-to-Sequence approach reframes it as a text generation task, where the model learns to generate 'true' or 'false' tokens to indicate relevance. [Nogueira *et al*., 2020]

reasoning abilities to produce a more refined ranking order [Ma *et al*., 2023].

In this approach, the model takes as input a set of candidate passages retrieved by a first-stage retriever, such as BM25 [Jones *et al*., 2000], and generates a reordered ranking list by considering global interdependencies among passages rather than evaluating them independently. This holistic view of ranking allows LLM-based rerankers to account for aspects such as redundancy, diversity, and coherence in retrieved results [Zhu *et al*., 2024]. Recent work has shown that listwise rerankers using LLMs outperform traditional methods in certain tasks, particularly when handling ambiguous or complex queries, where contextual reasoning is crucial [Sun *et al*., 2023]. Figure 2 presents Instructional Permutation Generation approach.

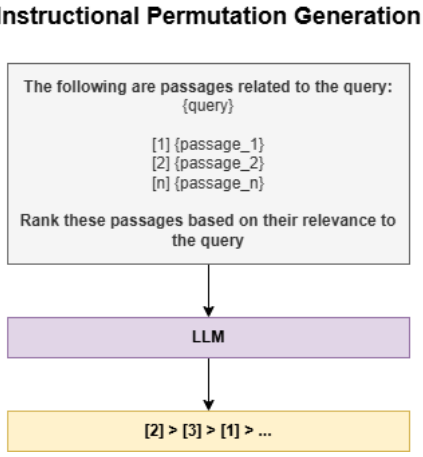

Figure 2. Instructional Permutation Generation [Sun *et al*., 2023]

RankGPT is a reranking approach based on GPT models that incorporates Instructional Permutation Generation (IPG) to improve passage ranking performance [Sun *et al*., 2023]. The core idea behind IPG is to explicitly instruct the LLM to consider multiple ranking orders of documents as training examples. This method enhances the LLM's ability to differentiate subtle relevance differences and prevents overfitting to a single ranking distribution [Sun *et al*., 2023].

Additionally, RankGPT employs a Sliding Window Strategy, which is crucial when dealing with length constraints inherent in LLMs. Since models such as GPT-4 have token limitations, the strategy processes ranked documents in overlapping segments, ensuring that all relevant passages are evaluated without truncating critical information. This incremental reranking approach allows RankGPT to handle longer candidate lists efficiently while preserving the global ranking consistency across segments [Sun *et al*., 2023]. Figure 3 illustrates the Sliding Window Strategy.

To assess the effectiveness of LLM-based reranking for Brazilian Portuguese information retrieval, we apply the Instructional Permutation Generation (IPG) and Sliding Window Strategy exactly as proposed in RankGPT, but using Sabiá-3 [Abonizio *et al*., 2024], a language model specifically trained for Brazilian Portuguese. This adaptation allows us to evaluate whether a Portuguese-specific LLM can outperform or match the performance of multilingual models in reranking tasks. By implementing these strategies, Sabiá-3 is used to consider multiple ranking permutations, leveraging its

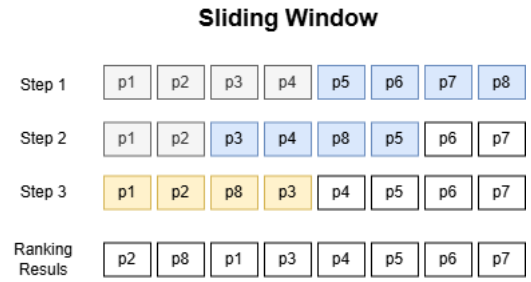

Figure 3. Sliding Window Strategy. An example demonstrating the re-ranking of eight passages using a sliding window approach with a window size of four and a step size of two. The blue-colored segments indicate the first two windows, while the yellow segment represents the final window. The sliding windows are applied in a back-to-first sequence, ensuring that the first two passages from the preceding window contribute to the re-ranking process in the subsequent window. [Sun *et al*., 2023]

ability to discern nuanced relevance relationships among retrieved passages. Additionally, the Sliding Window Strategy is crucial in handling longer passage lists while maintaining global ranking consistency. To ensure a comprehensive comparison, we benchmark Sabiá-3 against RankGPT using GPT-4, examining their reranking effectiveness on Brazilian Portuguese datasets, such as Quati [Bueno *et al*., 2024] and Pirá 2.0 [Pirozelli *et al*., 2024].

Dynamic In-Context Learning. Building upon the listwise reranking paradigm, we introduce a novel method to enhance the zero-shot performance of LLMs through dynamic In-Context Learning [Brown *et al*., 2020]. The core objective of this approach is to provide the model with highly relevant, task-specific examples at inference time, thereby guiding its reasoning process and adapting it to the nuances of the retrieval task without the need for expensive fine-tuning. Unlike static few-shot approaches, in which the same examples are used for all queries, our method customizes the prompt for each query. This strategy was applied to both RankGPT and Sabiá-3, allowing us to evaluate its effectiveness across different LLM architectures and its potential to improve performance on culturally-specific queries.

The dynamic nature of our method lies in the selection of few-shot examples. We first constructed a candidate pool of few-shot demonstrations from the Quati training set, where each demonstration consists of a query, a list of candidate passages, and the corresponding ground-truth ranked list. For each incoming query from the test set, we calculate the semantic similarity between its embedding and the embeddings of all queries in the candidate pool, using a pre-trained sentence-transformer model (MiniLM-L6 [Sentence-Transformers, 2021]). The top-k most similar demonstrations are then selected and formatted into the prompt. This process ensures that the examples provided to the LLM are contextually aligned with the current query, a technique hypothesized to produce more accurate and robust reranking performance. Figure 4 presents the proposed method.

3.2 Experimental Setup

Fine-tuning. Our fine-tuning experiments are based on the ptT5-base model [Carmo *et al*., 2020], a T5 variant specifically pre-trained on the large-scale Brazilian Portuguese Web as Corpus (BrWac). To adapt this model for reranking, we used the training set of the Quati dataset [Bueno *et al*., 2024],

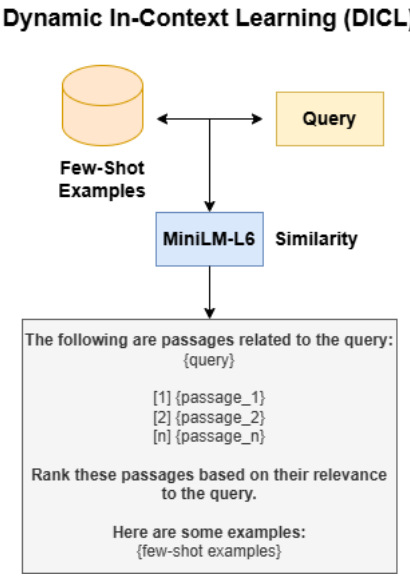

Figure 4. Proposed dynamic In-Context Learning (DICL) strategy. The process compares an incoming query against a pool of few-shot examples using a sentence-transformer model (MiniLM-L6) to find the most semantically similar demonstrations. These selected examples are then appended to the final reranking prompt, providing the Large Language Model with contextually relevant, in-domain examples to guide its ranking process at inference time.

a benchmark specifically designed for information retrieval in Brazilian Portuguese consisting of 200 queries each with 98 candidate passages, featuring relevance scores annotated by native speakers. The fine-tuning process was conducted using two distinct approaches. In the first approach, a classification head was used, where the original relevance scores from Quati were treated as classification labels, allowing the model to learn the association between queries and their respective relevance levels. In the second approach, a sequence-to-sequence task was implemented by adapting the original relevance labels into a binary classification scheme. In this case, documents with a relevance score of 0 were assigned the token ”false”, while those with a score greater than 0 were assigned the token ”true”, following the sequence-to-sequence fine-tuning strategy proposed by Nogueira *et al*. [2020], where models generate discrete tokens to indicate document relevance.

For the fine-tuning process, both approaches employed a batch size of 64, trained for 500 epochs, and used a learning rate of 0.001. The maximum sequence length was set to 512 tokens. Regarding the computational costs, fine-tuning the ptT5-base model on a NVIDIA T4 GPU took approximately 6 hours.

Datasets and Evaluation. To ensure a diverse evaluation of the rerankers under different retrieval conditions, we employed two information retrieval datasets in Brazilian Portuguese: Quati [Bueno *et al*., 2024] and Pirá 2.0 [Pirozelli *et al*., 2024]. The Quati dataset was developed by native Brazilian Portuguese speakers, focusing on open-domain question-answering and passage retrieval, with human-annotated relevance judgments. Meanwhile, Pirá 2.0 is a reading comprehension dataset covering topics such as the ocean, the Brazilian coast, and climate change [Pirozelli *et al*., 2024]. This dataset provides a domain-specific retrieval benchmark, allowing us to analyze reranker performance beyond general-purpose IR tasks. By leveraging both datasets, we ensure a comprehensive evaluation of the fine-tuned models, considering both general-purpose and specialized retrieval scenarios in Brazilian Portuguese.

We evaluated model performance on two distinct test sets: the Quati [Bueno *et al*., 2024] test set, which is composed of 50 queries with 98 passages each, and the Pirá 2.0 [Pirozelli *et al*., 2024] dataset, containing 40 queries with 40 passages each. To ensure a rigorous assessment of generalization, these evaluation sets are entirely disjoint from the Quati training data used for fine-tuning, with no intersection between them. The models’ performance was measured using three widely adopted information retrieval (IR) metrics: Normalized Discounted Cumulative Gain (nDCG), Recall@10, and Mean Reciprocal Rank (MRR@10). These metrics were chosen to provide a comprehensive assessment of the ranking effectiveness of each model across different evaluation criteria [Caseli and Nunes, 2023].

nDCG is a graded relevance metric that measures how well the ranking produced by a reranker reflects the true relevance scores of documents. It considers both the position of relevant documents in the ranking and the degree of their relevance, with higher-ranked highly relevant documents contributing more to the final score [Caseli and Nunes, 2023]. This metric is particularly useful in graded IR tasks, such as Quati [Bueno *et al*., 2024], where documents are not simply classified as relevant or non-relevant, but rather rated on a scale of relevance.

Recall@10 measures the proportion of truly relevant documents that appear within the top 10 retrieved results. A higher Recall@10 score indicates that the reranker is successful at capturing relevant documents early in the ranking, ensuring that users have quick access to relevant information. This metric is especially important in practical retrieval scenarios, where users typically do not browse beyond the first few results [Caseli and Nunes, 2023].

MRR@10 assesses the ranking position of the first relevant document within the top 10 retrieved results. It is computed as the reciprocal of the rank of the first relevant document, meaning that higher MRR values indicate that relevant documents appear earlier in the ranking. This metric is particularly useful for question-answering and passage retrieval tasks, where the goal is to quickly surface the most relevant document for a given query [Caseli and Nunes, 2023].

By evaluating the rerankers with these three complementary metrics, we ensure a robust comparison of their performance across different retrieval criteria, capturing their ability to both retrieve relevant documents and rank them effectively.

In addition to evaluating the metrics for the dataset as a whole, a separate assessment was conducted considering only the queries and passages related to sociocultural aspects of Brazil. To create this subset, we performed a qualitative analysis of the entire 50-query Quati test set, aiming to isolate queries whose relevance assessment depends on understanding local idioms, discourse structures, and culturally-specific knowledge. We deliberately excluded queries that required only factually direct answers or could be easily resolved through direct translation. This rigorous filtering process resulted in a curated subset of 25 queries, each retaining its original 98 candidate passages. These selected queries are strongly

associated with cultural aspects—such as sports, music, arts, fauna, flora, local history, and tourist attractions. Figure 5 illustrates examples of sociocultural and non-sociocultural query-passages from Quati dataset [Bueno *et al*., 2024]. Table 1 presents a thematic analysis of the 25 queries selected for our sociocultural subset that categorizes the types of knowledge they require.

4 Experimental Results

4.1 Quati dataset

Table 2 presents the average and standard deviation of the evaluated IR metrics (nDCG, MRR@10, and Recall@10) for the different models assessed on the full Quati dataset. The models include RankGPT [Sun *et al*., 2023] and Sabiá-3 [Abonizio *et al*., 2024] as LLM-based rerankers, BGE M3 [Chen *et al*., 2024] and mT5 [Xue *et al*., 2021] as multilingual neural models, and ptT5 [Carmo *et al*., 2020] fine-tuned on either multilingual datasets or Brazilian Portuguese-specific datasets. Figures 6 and 7 illustrate the distribution of nDCG and Recall@10 for full Quati test set.

Overall, RankGPT achieved the best performance across all three evaluated metrics, with 86.0% for nDCG, 64.5% for MRR@10, and 32.0% for Recall@10, confirming its status as the state-of-the-art method in the literature [Sun *et al*., 2023]. Among the LLM-based reranking approaches, RankGPT outperformed Sabiá-3 despite being a multilingual model. Another model that demonstrated competitive performance on the Quati dataset was BGE M3, which achieved 83.2% nDCG, 58.1% MRR@10, and 29.8% Recall@10. This result highlights the effectiveness of the self-knowledge distillation-based embedding model approach for language-specific IR tasks, such as Portuguese, as previously reported in the literature [Chen *et al*., 2024].

Regarding the ptT5 fine-tuned models, we observed that those fine-tuned on the Quati dataset outperformed the models specialized in the multilingual mMARCO dataset translated into Portuguese, particularly in nDCG and MRR@10. When comparing the classification head fine-tuning approach and the sequence-to-sequence task approach, we found that the nDCG values were nearly identical. However, the sequence-to-sequence task approach outperformed the classification head in MRR@10 by 4.7 percentage points (reaching 35.8%) and in Recall@10 by 1.3 percentage points (reaching 19.8%). These results support findings in the literature, indicating that this fine-tuning approach enables models to learn context-aware relevance assessments more effectively [Nogueira *et al*., 2020].

Thus, regarding the research hypotheses, we observed that, overall, LLMs as rerankers achieved superior IR performance compared to smaller models fine-tuned specifically for the reranking task. A possible explanation for this result is that their large-scale pretraining endows them with enhanced capabilities for text processing and reasoning, which may outweigh the benefits of task-specific fine-tuning in smaller models [Zhu *et al*., 2024] [Sun *et al*., 2023].

Notably, the first evaluation of the Sabiá 3 model, a Brazilian Portuguese-specific LLM as a reranker, demonstrated

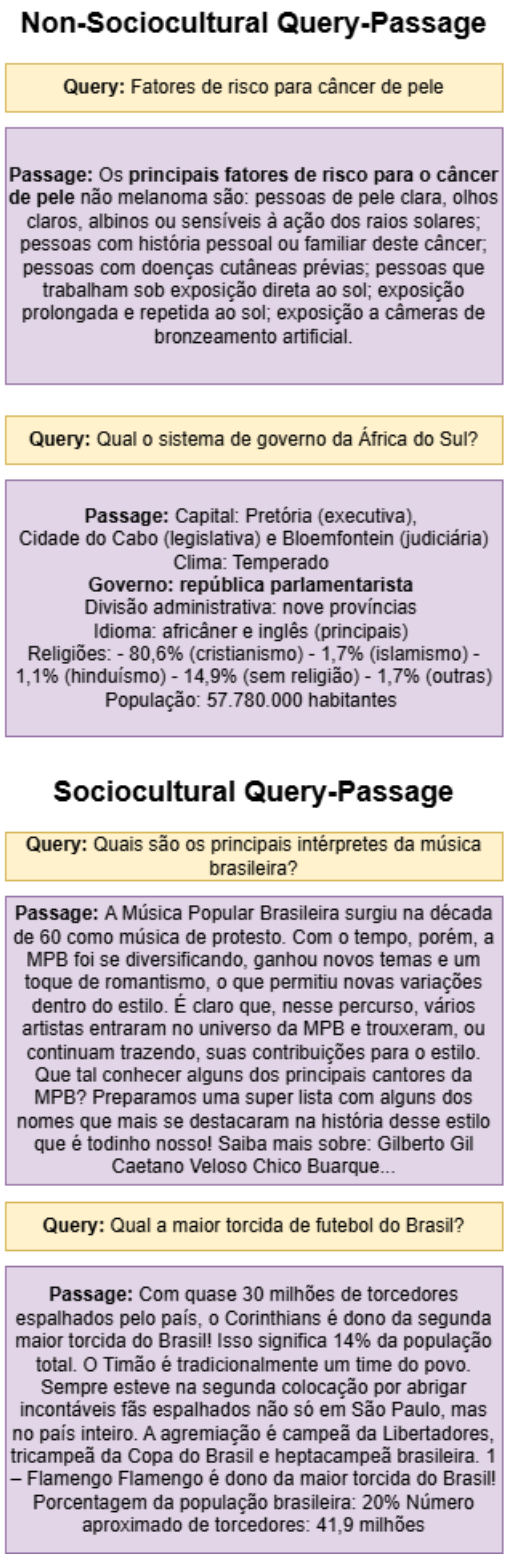

Figure 5. Examples illustrating the distinction between fact-based and culturally-contextual queries in the Quati dataset. The non-sociocultural queries (top) are objective and solvable through direct information extraction. In contrast, the sociocultural queries (bottom) are more complex, as their corresponding passages are richer in local idioms, employ more intricate discourse structures, and are dense with culturally-specific knowledge. This requires the model to go beyond simple fact-finding, forcing it to interpret cultural importance (”principais intérpretes”) or synthesize information from comparative statements (”Qual a maior torcida”), a process reliant on a deeper grasp of the local context. [Bueno *et al*., 2024]

Table 1. Thematic analysis of queries from the sociocultural subset of Quati. This categorization highlights the linguistic and contextual challenges inherent in the native dataset compared to translated, fact-centric corpora.

Category	Description, Analysis, and Query Examples
Cultural Figures & Entities	These queries require not just recognizing an entity, but understanding its cultural significance and hierarchy. A translated dataset might ask for a list of teams, but a native query asks for the one with the largest, most passionate fanbase—a concept deeply embedded in culture. The model must synthesize opinions, historical context, and social identity. • *"Qual a maior torcida de futebol do Brasil?"* (Which soccer team has the largest fanbase in Brazil?) • *"Quais são os principais intérpretes da música brasileira?"* (Who are the main interpreters of Brazilian music?) • *"Qual é a principal cidade do basquete brasileiro?"* (What is the main city for Brazilian basketball?)
Social & Historical Context	Relevance depends on understanding complex, multi-faceted narratives specific to Brazilian history. These are not simple date/event lookups; they require comprehending cause, effect, and socio-political interpretation. The language used in relevant passages is often analytical and interpretive rather than purely factual. • *"Por que os anos de 80 são considerados a década perdida no Brasil?"* (Why are the 80s considered the 'lost decade' in Brazil?) • *"Qual foi a importância da usina de Volta Redonda RJ para a industrialização brasileira?"* (What was the importance of the Volta Redonda plant for Brazilian industrialization?)
Regionalisms & Local Knowledge	These queries demand familiarity with uniquely Brazilian geography, biomes, and specialized vocabulary (e.g., *Cerrado*, *Caatinga*). Beyond terminology, they often seek practical, local advice that is absent from generic encyclopedic sources and is typically expressed in informal, colloquial language found in blogs or local guides. • *"Quais são os biomas do Brasil?"* (What are the biomes of Brazil?) • *"Qual a melhor cidade para ficar no Jalapão?"* (What is the best city to stay in Jalapão?) • *"Como podemos classificar o relevo brasileiro?"* (How can we classify the Brazilian relief?)
Implicit Context & Colloquial Phrasing	While not always containing dictionary-defined idioms, these queries (and their relevant passages) rely on understanding implicit, culturally-understood context. A query about taking children to a challenging natural landmark implies a need for information about safety, difficulty, and accessibility, often described using informal, colloquial language (e.g., *"é puxado," "é tranquilo"*). • *"É possível conhecer o Pico da Bandeira com crianças?"* (Is it possible to visit Pico da Bandeira with children?) • *"Como transformar uma cidade pacata em um polo turístico?"* (How to turn a quiet town into a tourist hub?)

competitive performance compared to state-of-the-art techniques in the literature, surpassing smaller models fine-tuned for the reranking task. Sabiá 3 achieved scores of 81.9% nDCG, 55.7% MRR@10, and 27.1% Recall@10, highlighting its effectiveness for Portuguese IR tasks.

Regarding the second research hypothesis, we observe that ptT5 models fine-tuned on the Quati dataset, a Portuguese-specific dataset, outperformed those fine-tuned on multilingual translated datasets such as mMARCO in IR evaluation metrics for the Quati dataset. This finding reinforces the importance of training rerankers on data that better captures the linguistic and sociocultural nuances of the target language [Bueno *et al*., 2024]. The distinction is particularly evident when analyzing the tasks illustrated in Figure 5. While non-sociocultural queries are objective and can be answered through straightforward information extraction, their sociocultural counterparts are inherently more nuanced. Their corresponding passages are often more complex, replete with local idioms, intricate discourse structures, and culturally-specific knowledge. Consequently, resolving these queries requires the model to move beyond simple fact-finding and demonstrate a deeper understanding of the local context, such as interpreting cultural importance or synthesizing information from comparative statements.

Table 2. Information Retrieval Metrics (nDCG, MRR@10, Recall@10) for full Quati test set.

Model	nDCG		MRR@10		Recall@10	
	Avg	Std	Avg	Std	Avg	Std
RankGPT	86.0	9.7	64.5	38.7	32.0	11.7
BGE M3	83.2	14.5	58.1	39.3	29.8	14.7
Sabiá	81.9	11.8	55.7	42.4	27.1	9.7
mT5	76.9	13.6	29.3	34.7	26.7	13.6
ptT5 Quati	71.9	15.2	30.1	38.9	18.5	8.5
ptT5 Quati S2S	71.9	15.8	35.8	39.7	19.8	8.0
ptT5 mMARCO	68.7	15.4	27.0	33.3	18.6	8.7

Table 3. Information Retrieval Metrics (nDCG, MRR@10, Recall@10) for sociocultural Quati test set.

Model	nDCG		MRR@10		Recall@10	
	Avg	Std	Avg	Std	Avg	Std
RankGPT	87.7	7.8	66.1	38.7	31.4	12.7
Sabiá	83.9	10.9	56.4	42.8	26.7	11.3
BGE M3	82.4	20.0	56.4	45.4	25.8	12.4
mT5	77.1	17.7	32.2	39.5	22.8	7.5
ptT5 Quati	74.2	17.0	30.1	35.6	19.4	7.6
ptT5 Quati S2S	74.6	14.1	39.3	37.3	22.9	9.6
ptT5 mMARCO	69.2	15.8	29.7	35.2	18.6	9.2

The careful separation of the fine-tuning and evaluation datasets provides a crucial lens for interpreting the source of the performance gains. This experimental design is significant because it shifts the focus away from simple factual memorization and towards a more profound form of adaptation rooted in the dataset's linguistic integrity. With no direct data overlap, the model is not merely learning what topics are relevant to Brazil; it is learning how relevance is expressed, structured, and sought in Brazilian Portuguese. By training on Quati, the model becomes attuned to the genuine syntactic patterns, colloquialisms, and culturally-embedded contexts that are inherently present in native-generated data. This suggests the learning is focused on acquiring a generalized capability to interpret the language's natural form, a more robust and transferable skill than simple topical alignment.

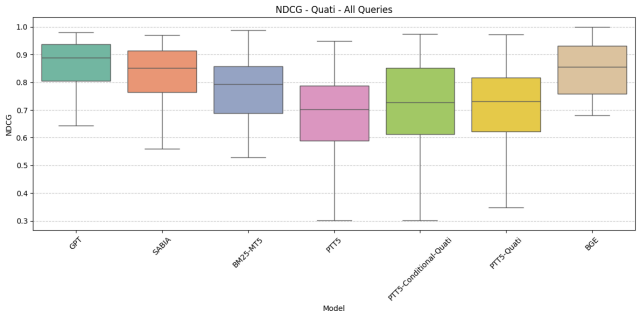

Figure 6. nDCG for full Quati test set [Bueno *et al*., 2024]

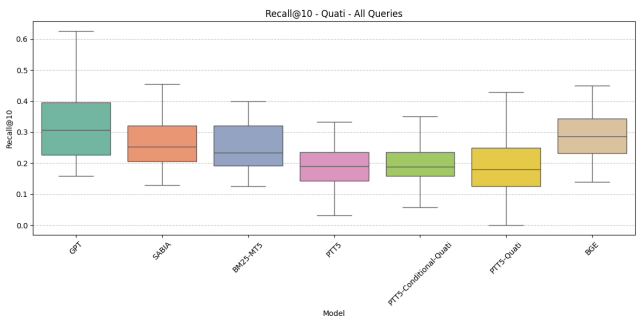

Figure 7. Recall@10 for full Quati test set [Bueno *et al*., 2024]

Table 3 presents the performance metrics of the same models but considering only the queries related to sociocultural aspects from the Quati dataset. Unlike the overall results shown in Table 2, we observe that Sabiá 3 outperforms state-of-the-art multilingual models, such as BGE M3, by 1.5 percentage points in nDCG, reaching 83.9%, and by 0.9 percentage points in Recall@10, achieving 26.7%. The MRR@10 obtained was the same for both models. These results suggest that Sabiá 3, a Portuguese-specific LLM, demonstrates a competitive advantage in handling queries that require deeper linguistic and sociocultural understanding. Figure 8 and 9 illustrate the distribution of nDCG and Recall@10 for sociocultural Quati test set.

Furthermore, the improvement in nDCG, MRR@10, and Recall@10 for models fine-tuned on Brazilian Portuguese-specific datasets, compared to those specialized in translated multilingual datasets, becomes even more pronounced when evaluated on queries related to sociocultural aspects. This finding reinforces the importance of training on native datasets to enhance the model's ability to capture linguistic nuances, local idioms, and culturally specific knowledge in information retrieval tasks [Bueno *et al*., 2024].

In this case, the ptT5 Quati models outperformed the ptT5 mMARCO model, achieving a 5.4 percentage point increase in nDCG (reaching 74.6%), a 9.6 percentage point improvement in MRR@10 (reaching 39.3%), and a 4.3 percentage point gain in Recall@10 (reaching 22.9%). These results further highlight the advantages of fine-tuning on native Portuguese datasets for improving retrieval performance in tasks involving sociocultural aspects [Bueno *et al*., 2024].

Table 4. Information Retrieval Metrics (nDCG, MRR@10, Recall@10) for Pirá 2.0 dataset.

Model	nDCG		MRR@10		Recall@10	
	Avg	Std	Avg	Std	Avg	Std
RankGPT	83.4	26.5	78.6	36.6	87.5	29.4
Sabiá	75.9	31.6	66.4	43.7	78.8	40.7
BGE M3	62.8	34.5	50.5	44.9	75.0	43.1
ptT5 Quati	81.0	28.9	74.2	39.0	90.0	30.4
ptT5 Quati S2S	80.6	30.0	70.4	45.8	84.6	50.4
ptT5 mMARCO	75.8	14.6	63.4	42.6	78.7	43.9

4.2 Pirá 2.0 dataset

Table 4 presents the same evaluation metrics (nDCG, MRR@10, and Recall@10) measured for the Pirá 2.0 dataset.

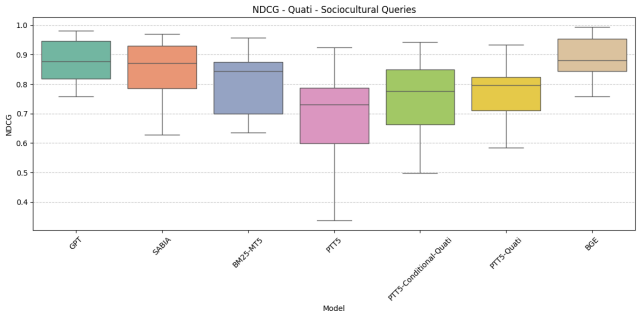

Figure 8. nDCG for sociocultural Quati test set [Bueno *et al*., 2024]

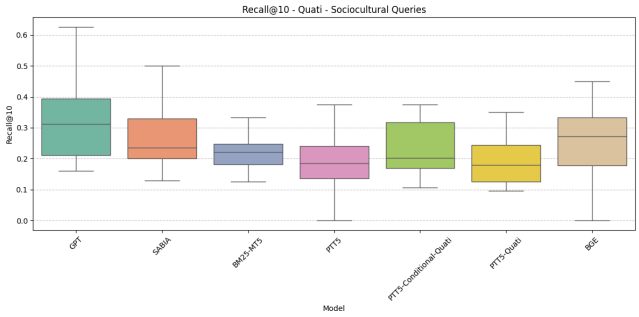

Figure 9. Recall@10 for sociocultural Quati test set [Bueno *et al*., 2024]

We employed this approach to verify whether the results obtained for the Quati dataset could be reproduced in other specific datasets for Information Retrieval. Similar to the results obtained for the Quati dataset, we observe the same pattern of better performance for RankGPT, an LLM-based reranker approach, utilizing a multilingual model. RankGPT achieved scores of 83.4% nDCG, 78.6% MRR@10, and 87.5% Recall@10.

Furthermore, once again, the Sabiá 3 model achieved competitive results compared to other state-of-the-art methods, even surpassing the BGE M3 model. Thus, we demonstrate that Sabiá 3 as an LLM-based reranker is a highly promising alternative for Information Retrieval tasks when dealing with datasets specific to Brazilian Portuguese. Sabiá 3 achieved scores of 75.9% nDCG, 66.4% MRR@10, and 78.8% Recall@10.

Regarding the ptT5 models fine-tuned on specific datasets for Information Retrieval tasks, the models fine-tuned on datasets produced in Brazilian Portuguese (Quati) once again outperformed the model specialized in the multilingual dataset translated into Portuguese (mMARCO). In particular, the model with the classification head fine-tuning approach outperforms in nDCG by 5.2 percentage points, reaching 81.0%, surpasses in MRR@10 by 10.8 percentage points, reaching 74.2%, and exceeds in Recall@10 by 11.3 percentage points, reaching 90.0%. This further reinforces the hypothesis regarding the importance of fine-tuning rerankers on data that better captures the linguistic and sociocultural nuances of the target language [Bueno *et al*., 2024].

In particular, in the case of the Pirá 2.0 dataset, the ptT5 models fine-tuned on the Quati dataset show competitive performance with the state-of-the-art method (RankGPT) and outperform others, such as Sabiá 3 and BGE M3, demonstrating the effectiveness of this approach. While the classification head approach of the ptT5 model demonstrated the strongest performance on Pirá 2.0, the sequence-to-sequence method (ptT5 S2S), although effective in general, did not perform as well on Pirá 2.0. This may be due to the distinct nature of Pirá 2.0, which, as a domain-specific dataset focused on oceanography, climate change, and the Brazilian coastline, requires a different type of understanding than Quati. It's possible that the more general and sociocultural training of the S2S approach of the model fine-tuned on the Quati dataset was less aligned with the narrower, more technical demands of the Pirá 2.0 dataset, leading to a competitive, but not superior result, comparing with a broader fine-tuned model. It reinforces the argument that effectiveness relies on the alignment between the nuances captured during fine-tuning and the specific retrieval context.

While RankGPT exhibits superior performance, the choice of Sabiá-3 over RankGPT hinges on a crucial trade-off between effectiveness and practical feasibility. RankGPT, driven by GPT-4, demands significant computational resources and thus incurs higher costs. Sabiá-3, conversely, offers impressive results with a smaller scale and lower associated costs, making it a more accessible option.

The decision to use Sabiá-3 instead of RankGPT hinges on a trade-off between performance and practical viability. Table 5 presents a comparison. RankGPT operates on a pay-per-use basis, where the costs per million input tokens ($2.50) and million output tokens ($10.00) [OpenAI, 2025] are notably high for large-scale operations. Conversely, the Sabiá-3 offers a highly competitive token-based pricing structure, with costs of just $0.94 per million input tokens and $1.88 per million output tokens [Maritaca AI, 2025]. This makes Sabiá-3 a more economical alternative for applications with high data volume, especially where cost optimization is a primary concern.

Table 5. Comparative analysis of RankGPT (GPT-4) and Sabiá-3 on practical viability metrics. Costs are estimated based on public pricing data from January 2025. [OpenAI, 2025] [Maritaca AI, 2025]

Metric	GPT-4	Sabiá-3
nDCG (Socioc. Quati)	87.7	83.9
Input Cost (per MM tokens)	$2.50	$0.94
Output Cost (per MM tokens)	$10.00	$1.88

Sabiá-3 demonstrates that achieving near-state-of-the-art performance is possible without relying on the largest and most expensive models available. This characteristic lends Sabiá-3 greater sustainability. For many applications, particularly those prioritizing cultural relevance and operational efficiency, a specialized model such as Sabiá-3 represents a more viable approach. Its ability to handle the cultural and linguistic nuances of Brazilian Portuguese, coupled with lower operational costs, makes it a strategic choice. In summary, Sabiá-3 emerges as an attractive alternative for information retrieval in Portuguese, especially when seeking a balance between high-quality performance, cultural relevance, and considerations of cost and efficiency. While RankGPT sets a high standard in terms of performance, Sabiá-3 exemplifies the possibility of achieving remarkable results with a more sustainable and accessible approach.

4.3 Dynamic In-Context Learning

To further investigate the capabilities of LLM-based rerankers, we evaluated our proposed prompting strategy: dynamic In-Context Learning (DICL). As shown in Table 6, applying this method yielded consistent performance improvements for both RankGPT and Sabiá 3 on the full Quati test set. The most substantial impact was observed in the MRR@10 metric, which saw a significant uplift for RankGPT (from 64.5% to 73.1%) and for Sabiá 3 (from 55.7% to 63.3%). This indicates that providing dynamically selected, in-domain examples is highly effective at helping the models place the most relevant document at a higher rank. While nDCG also saw modest but consistent gains for both models, Recall@10 remained unaffected.

This positive trend is further reinforced when analyzing the sociocultural subset, as detailed in Table 7. On these culturally-nuanced queries, the DICL method again provided a notable improvement, particularly for MRR@10. These results strongly suggest that dynamically providing contextually aligned examples helps LLMs better grasp the specific linguistic and cultural subtleties required for such queries.

Ultimately, the evaluation of the DICL method validates the hypothesis that the zero-shot performance of LLMs in reranking tasks can be significantly enhanced through advanced prompting techniques. This approach offers a powerful approach to improve ranking quality by providing relevant context at inference time. The consistent improvements across both RankGPT and Sabiá 3 highlight the robustness of the technique. These findings further strengthen the argument that a specialized, cost-effective model, such as Sabiá 3, when augmented with intelligent prompting strategies, represents a highly viable and compelling alternative for achieving state-of-the-art performance in Portuguese IR.

5 Conclusion

The findings of this study highlight the effectiveness of different reranking strategies for Portuguese Information Retrieval (IR), emphasizing the impact of fine-tuning on native datasets and the role of large language models (LLMs). Our results confirm that fine-tuning models on Brazilian Portuguese datasets, such as Quati, significantly improves retrieval performance compared to models trained on translated multilingual corpora. The ptT5 rerankers fine-tuned on Quati consistently outperformed those fine-tuned on mMARCO, reinforcing the importance of language-specific training to capture linguistic and sociocultural nuances.

Our evaluation also supports the hypothesis that LLMs, particularly RankGPT, achieve superior ranking performance in general IR tasks due to their extensive pretraining and instruction-tuned capabilities. However, our results also demonstrate that Sabiá 3, a Portuguese-specific LLM, is a viable alternative, particularly for queries involving sociocultural aspects. In sociocultural retrieval tasks, Sabiá 3 performed competitively against state-of-the-art multilingual models such as BGE M3, suggesting that Portuguese-specific LLMs can offer enhanced retrieval quality for culturally sensitive queries.

Furthermore, our introduction and evaluation of the dynamic In-Context Learning method revealed another powerful avenue for enhancing LLM-based reranking. This technique consistently improved the performance of both RankGPT and Sabiá 3, with particularly notable gains in MRR@10. This finding is crucial as it demonstrates that the zero-shot capabilities of LLMs can be substantially augmented through prompting strategies. By providing contextually relevant examples at inference time, DICL not only boosted overall performance but also further solidified Sabiá 3 as a highly competitive and resource-efficient alternative, particularly for culturally-specific tasks.

Table 6. Comparison of Information Retrieval Metrics (nDCG, MRR@10, Recall@10) for full Quati test set for the dynamic In-Context Learning (DICL) method.

Model	nDCG		MRR@10		Recall@10	
	Avg	Std	Avg	Std	Avg	Std
DICL RankGPT	87.4	10.2	73.1	35.4	31.9	13.1
DICL Sabiá	83.3	9.5	63.3	37.1	26.0	8.2
Standard RankGPT	86.0	9.7	64.5	38.7	32.0	11.7
Standard Sabiá	81.9	11.8	55.7	42.4	27.1	9.7

Table 7. Comparison of Information Retrieval Metrics (nDCG, MRR@10, Recall@10) for sociocultural Quati test set for the dynamic In-Context Learning (DICL) method.

Model	nDCG		MRR@10		Recall@10	
	Avg	Std	Avg	Std	Avg	Std
DICL RankGPT	88.9	7.3	75.3	34.2	30.5	9.1
DICL Sabiá	85.1	9.2	63.7	40.6	25.2	10.9
Standard RankGPT	87.7	7.8	66.1	38.7	31.4	12.7
Standard Sabiá	83.9	10.9	56.4	42.8	26.7	11.3

Additionally, the experimental results on the Pirá 2.0 dataset further validate our key hypotheses. The consistent advantage of fine-tuned Portuguese-specific rerankers over multilingual models highlights the potential benefits of domain-adapted training data. Interestingly, in specialized IR tasks, the ptT5 fine-tuned on Quati demonstrated competitive performance, achieving results comparable to LLM-based rerankers, which suggests that smaller models fine-tuned on high-quality native datasets can be an efficient alternative to computationally expensive LLMs.

Despite these promising results, several open questions remain. One limitation of this study is the computational cost associated with deploying LLM-based rerankers such as RankGPT. Future research could explore knowledge distillation techniques to transfer ranking capabilities from LLMs to smaller, more efficient models while maintaining retrieval effectiveness [Laitz *et al*., 2025] [Chen *et al*., 2024].

Another avenue for future work is the expansion of Portuguese IR datasets to include a broader range of query types, particularly in domains such as law, medicine, and finance, where domain-specific knowledge is crucial.

Furthermore, an important research direction is the integration of hybrid reranking strategies that combine the interpretability and efficiency of smaller fine-tuned models with the contextual reasoning capabilities of LLMs. Exploring ensemble approaches that leverage the strengths of both categories of models could lead to even more effective Portuguese IR systems.

Finally, given the observed improvements in sociocultural retrieval tasks with Portuguese-specific rerankers, further investigation is needed into how these models can be adapted for other low-resource languages with strong cultural nuances.

In summary, our study underscores the value of language-specific fine-tuning for IR in Portuguese and provides empirical evidence supporting the effectiveness of LLMs and not-so-large language models in reranking. By demonstrating the advantages of native dataset training and evaluating the trade-offs between model size and effectiveness, our findings contribute to the ongoing development of more efficient and culturally aware IR systems for the Portuguese language.

Declarations

Authors' Contributions

Experiment—R.O.M.; Writing—Original Draft Preparation—R.O.M.; Writing—Review & Editing—P.L.P.C.; All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was supported by the Amigos da Poli – project 2025_021. It was also financed in part by CAPES – Finance Code 001.

Funding

It was partly funded by CAPES – Funding Code 001.

Availability of data and materials

The datasets and/or softwares generated and/or analysed during the current study will be made available upon request.

References

Abonizio, H., Almeida, T., Laitz, T., Junior, R., Bonás, G., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *ArXiv*.

Bonifacio, L., Campiotti, I., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of ms marco passage ranking dataset. *CoRR*, 15413(1). DOI: CoRR abs/2108.13897.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bueno, M., Oliveira, E., Nogueira, R., Lotufo, R., and Pereira, J. (2024). Quati: A brazilian portuguese information retrieval dataset from native speakers. *Proceedings of the XV Brazilian Symposium on Information Technology and Human Language (STIL)*, 1.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *ArXiv*, 1.

Caseli, H. and Nunes, M. (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. Brasileiras - Processamento de Linguagem Natural.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Findings of the Association for Computational Linguistics: ACL 2024*.

Guo, F., Li, W., Zhuang, H., Luo, Y., Li, Y., Yan, L., Zhu, Q., and Zhang, Y. (2025). Mcranker: Generating diverse criteria on-the-fly to improve pointwise llm rankers. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *Annual Meeting of the Association for Computational Linguistics*.

Jones, K., Walker, S., and Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information processing management*, 36(6).

Laitz, T., Papakostas, K., Lotufo, R., and Nogueira, R. (2025). Inranker: Distilled rankers for zero-shot information retrieval. *Intelligent Systems. BRACIS 2024. Lecture Notes in Computer Science*, 15413(1). DOI: 10.1007/978-3-031-79032-$4_1$0.

Lewis, L., Yang, Y., Rose, T., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(1).

Ma, X., Zhang, X., Pradeep, R., and Lin, J. (2023). Zero-shot listwise document reranking with a large language model. *CoRR*.

Maritaca AI (2025). Maritaca ai documentation - models. `https://docs.maritaca.ai/pt/modelos`. Accessed on 15 January 2025.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *Annual Conference on Neural Information Processing Systems (NIPS)*, 1.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. (2020). Document ranking with a pretrained sequence-to-sequence model. *Empirical Methods in Natural Language Processing*.

Nogueira, R., Yang, W., Cho, K., and Lin, J. (2019). Multi-stage

document ranking with bert. *ArXiv*.

Oliveira, L., Romeu, R., and Moreira, V. (2021). Regis: A test collection for geoscientific documents in portuguese. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

OpenAI (2025). Openai documentation - pricing. `https://platform.openai.com/docs/pricing`. Accessed on 15 January 2025.

Paschoal, A., Pirozelli, P., Freire, V., Delgado, K., Peres, S., José, M., Nakasato, F., Oliveira, A., Brandão, A., Costa, A., and Cozman, F. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. *Proceedings of the 30th ACM International Conference on Information Knowledge Management*.

Pirozelli, P., José, M., Silveira, I., Nakasato, F., Peres, S., Brandão, A., Costa, A., and Cozman, F. (2024). Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 1(6):29–63.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, J., Liu, J., Metzler, D., Wang, X., and Bendersky, M. (2024). Large language models are effective text rankers with pairwise ranking prompting. *Findings of the Association for Computational Linguistics: NAACL 2024*.

Sachan, D., Lewis, M., Joshi, M., Aghajanyan, A., Yih, W., Pineau, J., and Zettlemoyer, L. (2022). Improving passage retrieval with zero-shot question generation. *Empirical Methods in Natural Language Processing*.

Sentence-Transformers (2021). all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. Pretrained sentence embedding model. Fine-tuned on 1B+ sentence pairs.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). Is chatgpt good at search? investigating large language models as re-ranking agents. *Empirical Methods in Natural Language Processing*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1.

Zhu, Y., Yuan, H., Wang, S., Liu, S., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J. (2024). Large language models for information retrieval: A survey. *ArXiv*.