Enhancing Legal Question Answering in Brazilian Portuguese through Domain-Specific Embedding Models

Artur M. A. Novais[1], **David O. C. Ferreira**[1], **Josiel P. C. Silva**[1],
Matheus F. C. Brakes[1], **João P. C. Presa**[1], **Sávio S. T. de Oliveira**[1]

[1] Instituto de Informática – Universidade Federal de Goiás (UFG)

{artur.matos, oneil, josielpantaleao, brakes_fares}@discente.ufg.br

joaopaulop@egresso.ufg.br

savioteles@ufg.br

Abstract. *The increasing digitization of legal documents presents significant challenges for information retrieval. Traditional keyword-based search methods often fail to capture the semantic nuances of complex legal queries. Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for building Question Answering (Q&A) systems, but its effectiveness is highly dependent on the quality of its retrieval component. This paper addresses the problem of improving semantic search over legal texts from the Court of Accounts of the State of Goiás (TCE-GO). We introduce two specialized embedding models created by fine-tuning the state-of-the-art* `BGE-M3` *model on domain-specific corpora of jurisprudence and legislation, respectively. Our experimental results demonstrate that these specialized models significantly outperform general-purpose multilingual and Portuguese models in retrieval tasks, as measured by MRR@10 and Recall@10. Notably, our fine-tuned models, despite their moderate size, surpass much larger models, highlighting that domain specialization is a more parameter-efficient strategy than simply scaling model size for niche domains.*

1. Introduction

The vast and ever-growing volume of digital legal documents, including legislation, jurisprudence, and administrative acts, creates a significant information overload for legal professionals, public servants, and citizens. Effectively navigating this complex information landscape to find precise answers to specific questions is a formidable challenge. Traditional search systems, often based on keyword matching (e.g., BM25), struggle with the linguistic complexities of the legal domain, such as synonymy, polysemy, and intricate legal jargon [Chalkidis et al. 2020]. This limitation can lead to incomplete or irrelevant search results, hindering legal research and decision-making processes.

To overcome these challenges, modern Natural Language Processing (NLP) has introduced more sophisticated approaches. Retrieval-Augmented Generation (RAG) [Lewis et al. 2020] stands out as a promising architecture for developing advanced Question Answering (Q&A) systems. RAG combines the strengths of a dense retriever, which semantically searches a large corpus for relevant documents, and a generative language model, which synthesizes a coherent, human-like answer based on the retrieved information. The performance of a RAG system, however, is critically bottlenecked by the quality of its retriever [Gao et al. 2024]. If the retriever fails to find the correct documents, the generator, no matter how powerful, will produce irrelevant or incorrect answers.

The core problem this work addresses is the subpar performance of general-purpose embedding models when applied to the specialized domain of Brazilian administrative law, specifically the documents from the Court of Accounts of the State of Goiás (TCE-GO). These models, typically trained on broad web-scale corpora, lack the specialized vocabulary and contextual understanding required to accurately represent legal texts.

The primary objective of this research is to enhance the retrieval component of a RAG system for Q&A over TCE-GO's knowledge bases. To achieve this, we investigate the effectiveness of domain-specific fine-tuning of a state-of-the-art multilingual embedding model. The main contributions of this paper are threefold. First, we develop and release two novel, domain-specific embedding models for the Brazilian legal domain: `BGE-M3-DECISOES` and `BGE-M3-LEGIS`. Second, we create and release a high-quality evaluation benchmark of 10,000 question-document pairs tailored to the administrative law domain, featuring four distinct query profiles to ensure a comprehensive assessment. Finally, through extensive experiments, we demonstrate that targeted fine-tuning is a highly parameter-efficient strategy, enabling smaller models to outperform significantly larger general-purpose models in specialized retrieval tasks.

The remainder of this paper is organized as follows: Section 2 provides a theoretical background and reviews related work. Section 3 details the methodology, including our novel dataset generation process and the model fine-tuning procedure. Section 4 presents and discusses the experimental results in depth. Finally, Section 5 concludes the paper, summarizing our findings and outlining directions for future work.

2. Background and Related Work

This section provides an overview of the foundational concepts and prior research that underpin our work. We begin by introducing dense retrieval and the Retrieval-Augmented Generation (RAG) framework, which are central to our methodology. We then discuss the critical importance of domain adaptation for tailoring language models to specialized fields like law. Finally, we situate our contributions within the context of existing related work in the legal NLP domain, particularly for Brazilian Portuguese.

2.1. Dense Retrieval and RAG

Information Retrieval (IR) has evolved from sparse methods like BM25 to dense retrieval methods that leverage embedding models. These models, typically based on the Transformer architecture [Vaswani et al. 2017], map text into a low-dimensional vector space where semantic proximity can be measured by metrics like cosine similarity. RAG [Lewis et al. 2020] integrates this concept into a two-stage pipeline: a retriever fetches relevant documents from a knowledge base, and a generator (an LLM) synthesizes an answer based on the retrieved context. This approach grounds the LLM's response in factual data, mitigating hallucinations and enhancing trustworthiness [Gao et al. 2024], which is paramount in the legal domain. However, the effectiveness of RAG is susceptible to challenges such as the "lost in the middle" problem, where models struggle to utilize information located in the middle of long contexts [Liu et al. 2024]. This further emphasizes the need for a highly precise retriever that can rank the most relevant documents at the top.

2.2. Domain Adaptation for Language Models

General-purpose models often falter in specialized domains due to a mismatch in data distribution and vocabulary [Gururangan et al. 2020]. Domain adaptation aims to bridge this gap. While early approaches involved continued pre-training on domain-specific corpora [Chalkidis et al. 2020], recent work has focused on more efficient methods. Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) [Hu et al. 2021], have become popular for adapting LLMs. For embedding models, contrastive learning on domain-specific text pairs remains a powerful and widely used technique [Reimers and Gurevych 2019]. Our work employs a self-supervised contrastive approach, which does not require labeled query-document pairs, making it highly scalable.

2.3. Related Works

The application of NLP to the legal domain is a mature field. Seminal work like LEGAL-BERT [Chalkidis et al. 2020] established the value of domain-specific pre-training. For Brazilian Portuguese, BERTimbau [Souza et al. 2020] provided a strong monolingual baseline. More recently, research has focused on applying RAG to legal text. Studies have explored building legal Q&A systems for various sub-domains, such as Brazilian Transit Law [Rocho et al. 2024], and have proposed advanced RAG workflows to improve reliability in resource-constrained settings [Rocha and Pessoa 2024].

The evaluation of these systems is also a critical research area. Frameworks like the Massive Text Embedding Benchmark (MTEB) [Muennighoff et al. 2022] provide standardized leaderboards for embedding models, but often lack coverage for highly specialized domains like ours. Recognizing this, researchers are developing goal-oriented evaluation systems for interactive agents [Junior et al. 2024] and comparing different text representation models for specific tasks like app review classification [Araujo et al. 2020]. Our work contributes by providing both a new set of highly specialized models for Brazilian legal text and a bespoke, high-quality benchmark for their evaluation, addressing a clear gap in existing resources. We also align with works like [Evangelista et al. 2024] that analyze the trade-off between model performance and computational cost, a key consideration for practical deployments.

3. Methodology

This section outlines the research procedures, from our specialized corpus curation and benchmark generation to the model fine-tuning and evaluation protocol.

3.1. Corpus Curation for Fine-Tuning

To create specialized embedding models, we first curated two domain-specific corpora from the internal knowledge bases of the TCE-GO. The primary goal was to capture the two main types of documents legal professionals interact with: jurisprudence (court decisions) and legislation. The characteristics of these corpora are summarized in Table 1. The documents were preprocessed to remove formatting artifacts and ensure clean text content. For the fine-tuning process, these documents were segmented into smaller, meaningful chunks, such as paragraphs or sections, to serve as the base units for retrieval.

Table 1. Description of the domain-specific corpora used for fine-tuning.

Corpus Name	Documents	Size (MB)
TCE-GO Decisões (Decisions)	252,631	629
TCE-GO Legislação (Legislation)	3,697	4

3.2. Evaluation Dataset Generation

A robust evaluation requires a high-quality, diverse dataset of question-document pairs that reflect real-world information needs. To this end, we developed a systematic, semi-automated process for generating our evaluation set, as illustrated in Figure 1.

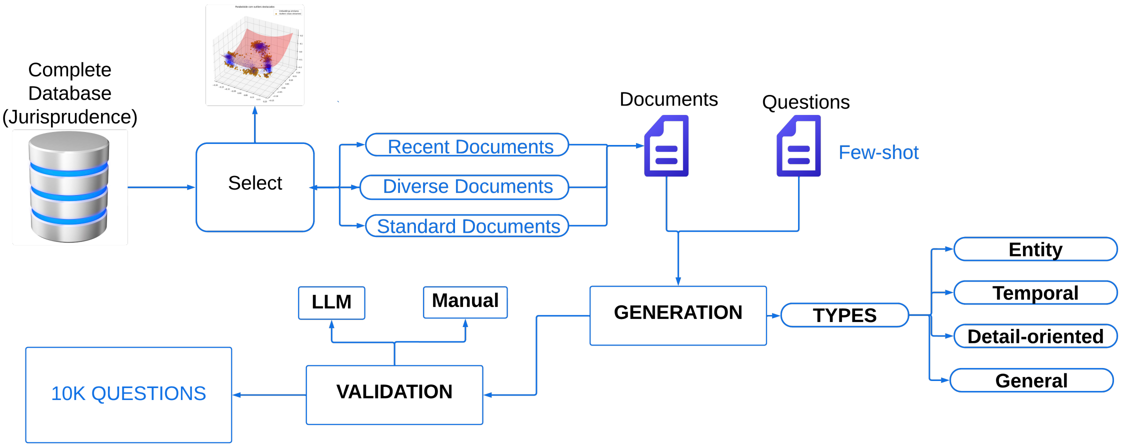

Figure 1. Flowchart of the semi-automated process for generating the evaluation dataset. The process begins by selecting a diverse set of documents that along with a few-shot set of example questions, are used by an LLM to generate new questions. The generated questions are categorized and then undergo a validation process, involving both an LLM and manual review, to produce the final set of high-quality questions.

The process began with the complete TCE-GO Decisions database. To ensure variety and challenge, we employed a strategic document selection phase. Using vector space analysis, we identified and sampled documents from different parts of the corpus: standard documents from dense clusters, outlier documents from sparse regions, and recent documents. This strategy ensures our evaluation covers a wide range of content, complexity, and temporal relevance.

Following selection, we initiated a few-shot generation process. Using a small set of existing, high-quality questions from the TCE as examples, an LLM was prompted to generate new questions based on the selected documents. This generation was guided by four distinct query profiles designed to test different retrieval capabilities. These profiles included Entity queries targeting specific names, organizations, or legal concepts; Temporal queries focusing on specific time ranges or years; Detailed queries requiring highly specific facts that challenge semantic retrievers; and General queries asking broader, more abstract questions.

Finally, all generated questions underwent a rigorous two-stage validation process. An initial automated validation was performed by an LLM to filter out malformed or irrelevant questions. This was followed by a crucial manual validation phase, where human experts reviewed the remaining questions and their corresponding documents to ensure relevance, coherence, and factual correctness. This semi-automated pipeline resulted in a high-quality benchmark of 10,000 validated question-document pairs, whose composition is detailed in Table 2.

Table 2. Composition of the final evaluation dataset by query profile.

Query Profile	Number of Questions
General	7,300
Detail-oriented	1,000
Temporal	950
Entity	750
Total	**10,000**

3.3. Model Fine-tuning

We selected `BAAI/bge-m3-base` as our foundational model due to its state-of-the-art performance and multilingual capabilities. We employed a contrastive learning framework using the Sentence Transformers library. In this setup, for each text chunk in our domain-specific corpora, we treat it as an ”anchor.” A ”positive” example is created by applying a minor, semantics-preserving augmentation to the anchor itself, while ”negative” examples are other randomly sampled chunks from the same batch. The model is then trained with a contrastive loss to minimize the distance between anchor-positive pairs while maximizing the distance to anchor-negative pairs.

This self-supervised approach effectively teaches the model the specific semantics of the legal domain without requiring labeled data. This process yielded our two distinct models: `BGE-M3-DECISOES`, fine-tuned on the jurisprudence dataset, and `BGE-M3-LEGIS`, fine-tuned on the legislation dataset. To promote reproducibility and foster further research, all models and source code developed in this study are publicly available[1].

3.4. Evaluation Protocol

To assess the performance of our models, we conducted a comprehensive benchmark against a wide array of publicly available embedding models. This diverse set was chosen to cover different architectures, sizes, and training specializations, providing a holistic view of the current landscape. The benchmark models included Portuguese-specific models: `serafim-100m-portuguese-sentence-encoder-ir`, `serafim-900m-portuguese-sentence-encoder` [Gomes et al. 2024]; general multilingual models like `gte-multilingual-base` [Zhang et al. 2024], `multilingual-e5-large-instruct` [Wang et al. 2024], `Solon-embeddings-large-0.1` [Muennighoff et al. 2022], `gte-qwen2-1.5B-instruct` [Li et al. 2023], and the very large `bge-multilingual-gemma2` [Jian-Xiang et al. 2024]; and another legally-oriented model, `bge-m3-portuguese-legal-v4`. We also included the original `BGE-M3`

[1] Anonymized for review. Available at: `https://anonymous.4open.science/r/tce_nlp-8FF0/`

model without fine-tuning, to serve as a direct baseline for measuring the impact of our specialization process.

We used two standard information retrieval metrics for evaluation. The first metric, MRR@10 (Mean Reciprocal Rank at 10), measures the average reciprocal rank of the first correct document in the top 10 results, thus heavily rewarding models that rank the correct answer higher. The second, Recall@10 (Recall at 10), measures the proportion of queries for which at least one correct document is found within the top 10 results, assessing the model's fundamental capability to find relevant information.

4. Results and Discussion

Our experimental results offer a clear and compelling narrative about the efficacy of domain specialization for semantic retrieval in the legal field. Table 3 provides a quantitative summary of the approximate performance metrics and parameter counts for all benchmarked models, ordered by descending MRR@10. The performance is also visualized in Figure 2 and Figure 3, which plot the metrics against the model sizes. This section provides a detailed analysis of these findings.

Table 3. Approximate Performance Metrics and Parameter Counts for Models. Our fine-tuned models are highlighted in bold. The table is ordered by descending MRR@10.

Model	Parameters	MRR@10	Recall@10
BGE-M3-DECISOES	570 M	**0.70**	**0.81**
BGE-M3 LEGIS	570 M	0.55	0.71
gte-Qwen2-1.5B-instruct	1.8 B	0.54	0.70
BGE-M3	570 M	0.53	0.70
bilingual-embedding-large	570 M	0.52	-
gte-multilingual-base	305 M	0.42	0.58
bge-multilingual-gemma2	9.2 B	0.40	0.60
Solon-embeddings-large-0.1	560 M	0.39	0.58
multilingual-e5-large-instruct	560 M	0.26	0.45
bge-m3-portuguese-legal-v4	570 M	0.16	0.27
serafim-100m-portuguese-pt-encoder-ir	109 M	0.12	0.19
serafim-900m-portuguese-pt-encoder	885 M	0.09	0.16

4.1. The Decisive Impact of Domain-Specific Fine-Tuning

The most significant finding is the commanding lead of our two fine-tuned models, BGE-M3-DECISOES and BGE-M3-LEGIS, over all other contenders. BGE-M3-DECISOES achieves the highest performance, with an MRR@10 near 0.7 and a Recall@10 exceeding 0.8. This success stems directly from the principle of data proximity: the model was fine-tuned on the jurisprudence corpus, the same domain from which our evaluation questions and documents were derived. This process allowed the model to internalize the specific vocabulary, entity relationships, and reasoning patterns prevalent in TCE-GO's decisions, leading to superior retrieval accuracy.

Following closely, BGE-M3-LEGIS is the second-best performer. Its strong performance is logical, as legal decisions are fundamentally built upon a foundation of legislation. By learning the language of laws, decrees, and regulations, the model gains a deep

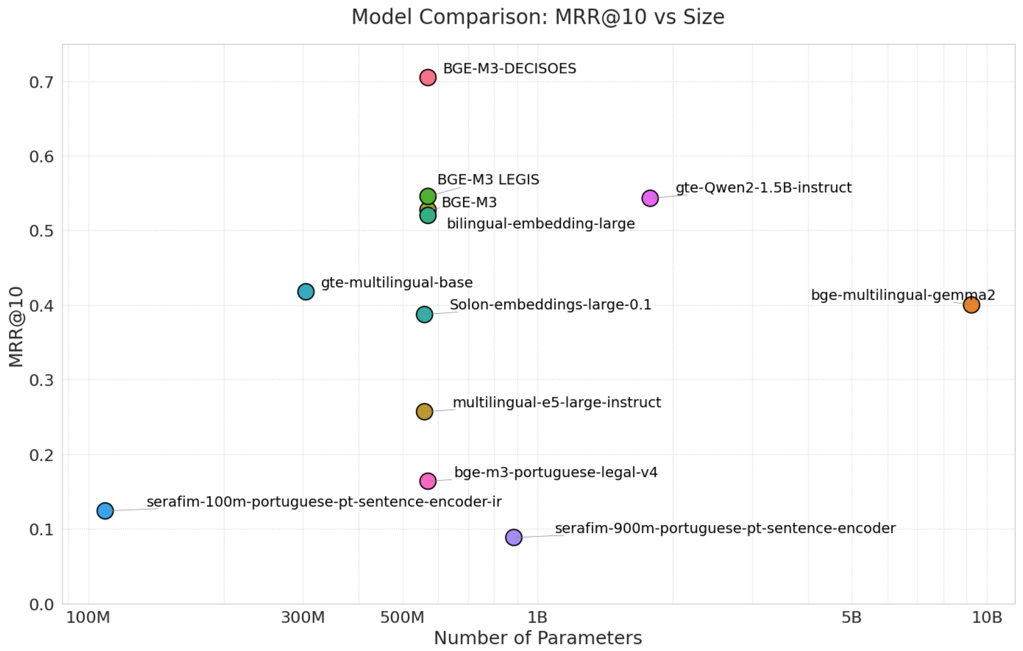

Figure 2. Model Comparison: MRR@10 vs. Number of Parameters. Higher is better.

semantic understanding of the concepts that underpin the court's jurisprudence. However, its performance is slightly lower than that of `BGE-M3-DECISOES` because while legislation provides the ”rules of the game,” the jurisprudence provides the ”recorded plays.” Our evaluation set, being derived from decisions, is better represented by a model trained on those very plays, which contain specific case details, party names, and contextual nuances not present in the abstract legislative text.

4.2. Comparative Analysis of Model Families

A deeper analysis of the benchmarked models reveals a clear performance hierarchy. Generally, larger pre-trained models, with more parameters, possess a greater capacity to learn complex patterns and store vast world knowledge. This increased capacity often translates to better performance on general tasks. However, our results demonstrate that for specialized domains, this principle is heavily conditioned by the relevance of the training data.

The group of general multilingual models, such as `gte-qwen2-1.5B-instruct` and `multilingual-e5-large`, forms a middle tier, showing that broad semantic knowledge provides a solid but insufficient foundation. In contrast, the monolingual Portuguese models from the `serafim` series perform poorly. This highlights a crucial insight: for this task, domain-specificity is more impactful than language-specificity, as the general Portuguese text they were trained on is too semantically distant from formal legal language. A particularly interesting case is the `bge-m3-portuguese-legal-v4` model, which, despite its ”legal” designation, fails to perform well. This suggests its legal training data did not align with our administrative law sub-domain, underscoring that ”the legal domain” is not a monolith and requires highly tailored solutions.

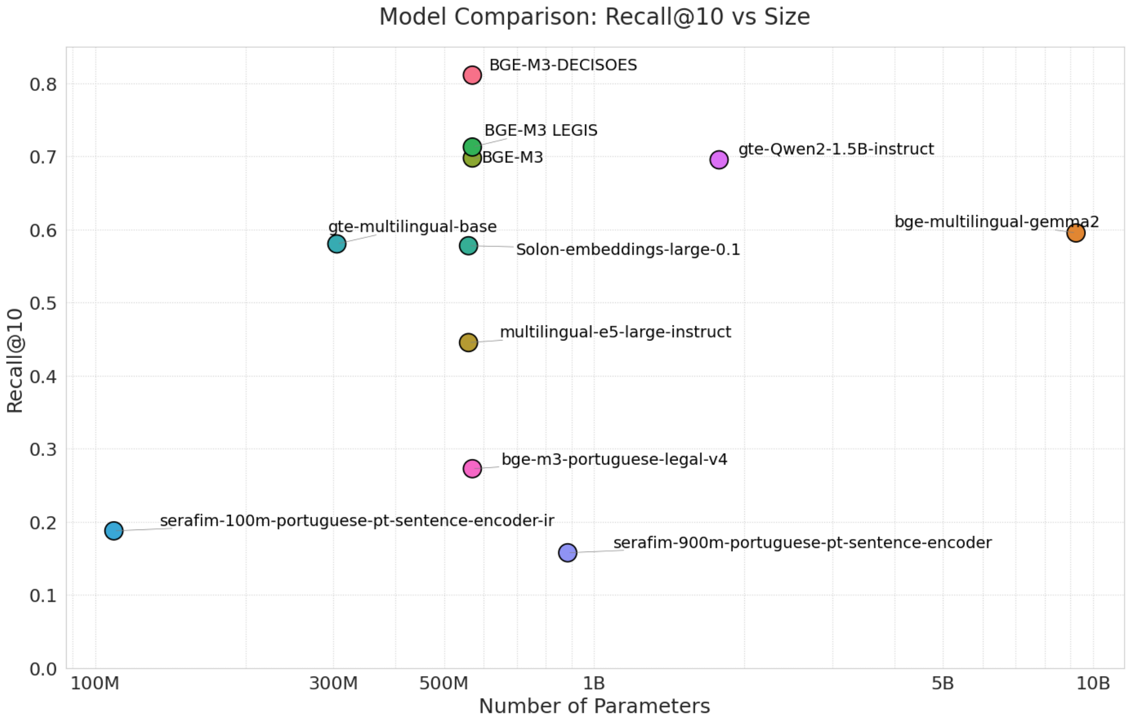

Figure 3. Model Comparison: Recall@10 vs. Number of Parameters. Higher is better.

4.3. The Parameter Efficiency of Specialization

The relationship between model size and performance offers the most compelling argument for our approach. The plots show no consistent correlation between size and performance among the generalist models. We observe a powerful counter-trend where focused specialization triumphs over scale. Our fine-tuned models, with approximately 540 million parameters, decisively outperform models that are orders of magnitude larger.

The most striking example is the comparison with the `bge-multilingual-gemma2` model, which has over 9 billion parameters. Our much smaller, specialized models achieve vastly superior results (e.g., an MRR@10 of 0.7 versus 0.4). This is because the additional parameters in a generalist model are not optimized for the specific semantic space of our target domain. They may store knowledge about a vast array of topics, from history to pop culture, which is irrelevant for interpreting legal texts. In contrast, fine-tuning reallocates the representational capacity of our moderately-sized model to be highly focused on the legal concepts, entities, and linguistic structures that matter. This demonstrates that for organizations aiming to deploy effective and efficient RAG systems, investing computational resources in targeted fine-tuning is a more fruitful and economically viable strategy than deploying massive, general-purpose models.

4.4. General Discussion and Implications

The results of this study carry significant implications for the practical application of NLP in specialized fields. Our core finding is that a modest investment in fine-tuning yields a disproportionately large return in performance. For the cost of training a moderately-sized model on a domain-specific corpus—a task that is becoming increasingly accessible—an organization can achieve retrieval accuracy that surpasses even the largest, most advanced general-purpose models available today. This makes the development of high-performance, in-house AI solutions a tangible goal for entities beyond large tech corporations.

Furthermore, the models developed in this work, particularly `BGE-M3-DECISOES` and `BGE-M3-LEGIS`, have strong potential for generalization within the Brazilian public sector. All Courts of Accounts in Brazil operate under a similar federal legal framework, including constitutional principles, national public finance laws, and bidding regulations. They share a common legal terminology and structure for their decisions and administrative acts. Consequently, a model fine-tuned on the data from one court, such as TCE-GO, is highly likely to serve as a powerful, zero-shot or few-shot retriever for other state and federal courts of accounts, drastically reducing the development time and cost for implementing similar Q&A systems across the country. This work, therefore, not only presents a solution for one institution but also provides a reusable and valuable asset for the broader Brazilian legal and administrative ecosystem.

5. Conclusion

In this paper, we addressed the challenge of improving semantic retrieval for a Q&A system operating on legal documents from the TCE-GO. We successfully developed two domain-specific embedding models, `BGE-M3-DECISOES` and `BGE-M3-LEGIS`, by fine-tuning a state-of-the-art base model on curated local corpora. Our contributions include not only these high-performing models but also a comprehensive 10,000-pair evaluation benchmark designed to test various retrieval capabilities in this domain.

Our extensive evaluation demonstrates that these specialized models establish a new state-of-the-art for this task, significantly outperforming a wide range of general-purpose models. The key takeaways are threefold: 1) domain adaptation through fine-tuning is the most critical factor for achieving high performance in specialized domains; 2) this approach is significantly more parameter-efficient than simply increasing model size; and 3) specialized domains like law are not monolithic, requiring models tailored to the specific sub-domain of interest.

While our results are promising, we acknowledge limitations. Our models were trained and evaluated on data from a single institution, and their generalization to other Brazilian legal domains remains to be tested. Furthermore, our work focused exclusively on the retriever; the end-to-end Q&A quality was not evaluated. Future work will proceed in several directions. We plan to expand our evaluation to other courts to test robustness and explore PEFT methods like LoRA for even more efficient fine-tuning. Finally, we will integrate our top-performing retriever into a full RAG pipeline to evaluate the end-to-end quality of the generated answers.

Acknowledgments

We acknowledge the support provided by the Center of Excellence in Artificial Intelligence (CEIA), affiliated with the Federal University of Goiás (UFG). We express our deepest gratitude to the teams and domain experts from the Court of Accounts of the

State of Goiás (TCE-GO), whose generous contributions of infrastructure, expertise, and datasets. This work has been supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Araujo, A., Golo, M., Viana, B., Sanches, F., Romero, R., and Marcacini, R. (2020). From bag-of-words to pre-trained neural language models: Improving automatic classification of app reviews for requirements engineering. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 378–389. SBC.

Chalkidis, I., Kamateri, E., Lazaridou, K., Aletras, N., Katakalou, M., and Krithara, A. (2020). LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Evangelista, G. A., de Oliveira, J. B., et al. (2024). Hybrid cnn-gnn models in active sonar imagery: an experimental evaluation. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 37–48. SBC.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gomes, L., Branco, A., Silva, J., Rodrigues, J., and Santos, R. (2024). Open sentence embeddings for portuguese with the serafim pt encoders family. In Santos, M. F., Machado, J., Novais, P., Cortez, P., and Moreira, P. M., editors, *Progress in Artificial Intelligence*, pages 267–279, Cham. Springer Nature Switzerland.

Gururangan, S., Marasović, A., Swaminathan, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. pages 8342–8360.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jian-Xiang, W., Shitao, X., Wang, Z., Jing-An, Y., Zhaoxu, D., Yu-Hong, L., Cun-Yue, G., and Shao-Dan, W. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through multi-objective training.

Junior, G. S. T., Peres, S. M., Fantinato, M., Brandao, A. A., and Cozman, F. G. (2024). A goal-oriented chat-like system for evaluation of large language models. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 743–754. SBC.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Ott, M., Chen, W.-t., Conneau, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rocha, L. M. and Pessoa, R. M. (2024). Advanced retrieval augmented generation for local llms. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 767–776. SBC.

Rocho, R. S. M., Perez, A. L. F., Farias, G. P., and Panisson, A. R. (2024). Integrating llms and chatbots technologies-a case study on brazilian transit law. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 731–742. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., et al. (2024). mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.