Legal Information Retrieval through Embedding Models and Synthetic Question Generation: Insights from the Philippine Tax Code

Matthew Roque, Nicole Abejuela, Shirley Chu, Melvin Cabatuan, Edwin Sybingco

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Legal Information Retrieval through Embedding Models and Synthetic Question Generation: Insights from the Philippine Tax Code

Matthew Roque[1], Nicole Abejuela[1], Shirley Chu[2], Melvin Cabatuan[1], Edwin Sybingco[1]
[1]Department of Electronics and Computer Engineering, [2]College of Computer Studies
De La Salle University
Manila, Philippines
{matthew_roque, nicole_abejuela, shirley.chu,
melvin.cabatuan, edwin.sybingco}@dlsu.edu.ph

Abstract

Legal information retrieval poses significant challenges, particularly in jurisdictions with limited technological resources. In this study, we compiled a manually annotated dataset consisting of 1,020 queries from bar exam reviewers and, modeled after these annotations, generated a synthetic dataset of 7,310 entries using Llama 3.1 8B Instruct. We conducted baseline evaluations of five embedding models—Word2Vec, SBERT, Jina Embeddings 2, Nomic Embed, and GTE—using the entire sections of the Philippine National Internal Revenue Code of 1997 (NIRC). Splitting the NIRC sections into smaller subsections yielded the most substantial improvements in retrieval accuracy, increasing Top-1 accuracy by up to 13% and Mean Reciprocal Rank (MRR) by up to 0.14. Among the models, GTE fine-tuned on the synthetic data and retrieving from the split NIRC achieved the best performance, with a Top-1 accuracy of 0.66 and MRR of 0.76. However, fine-tuning the models on the synthetic data with split NIRC sections resulted in little to no further enhancements, with improvements less than 2% compared to the pre-trained models on the split NIRC sections. Additionally, attempting to assist retrieval by matching input queries with synthetic queries did not contribute any improvements. These findings highlight that while section splitting can significantly enhance retrieval performance, the use of synthetic data to improve retrieval in highly nuanced and specialized domains like Philippine legal text remains limited.

1 Introduction

Legal information retrieval presents unique challenges, especially in jurisdictions with limited technological tools. In the Philippines, the absence of tailored retrieval systems for legal professionals or individuals seeking information from the National Internal Revenue Code of 1997 (NIRC) accentuates the need for specialized solutions. The specialized language of legal documents further complicates the quick and accurate retrieval of relevant information. Classical information retrieval techniques, such as vector space models and relevance feedback, laid the groundwork for automated search systems by representing and ranking text based on keyword matching and document structure (Ribeiro-Neto and Baeza-Yates, 2011). However, these methods often lack the semantic depth needed for nuanced legal documents, which has driven researchers toward embedding-based retrieval systems (Xiong et al., 2020).

Advances in natural language processing (NLP), particularly in embedding models, have paved the way for more precise information retrieval systems. Embedding models like Word2Vec (Mikolov et al., 2013a,b) and SBERT (Reimers and Gurevych, 2019) enable semantic understanding at the word and sentence levels, making them suitable for legal text retrieval tasks where traditional methods fall short. Word2Vec, a pioneering model in dense embeddings, has demonstrated the ability to capture semantic relationships through word co-occurrences but lacks contextual nuance, which models like SBERT address through sentence-level representations (Church, 2017). SBERT, by combining Siamese neural networks with BERT embeddings, achieves a contextual depth that has been shown to improve retrieval in various domains, including legal texts (Reimers and Gurevych, 2019). Despite the effectiveness of these embeddings in general NLP tasks, their adoption within the Philippine legal system has been limited, leaving a gap that this work aims to address.

Specialized models like LEGAL-BERT have further shown that domain-specific adaptations can significantly improve retrieval accuracy in legal contexts, as demonstrated in tasks involving complex legal documents (Chalkidis et al., 2020). In parallel, large-scale retrieval models have increasingly integrated synthetic data for model

fine-tuning, as seen in works like PAQ (Lewis et al., 2021), which generated millions of question-answer pairs for improved query relevance in question-answering tasks. By generating synthetic queries based on legal sections, we can similarly align models more closely with the dense, jargon-heavy language of the NIRC.

Recently, advanced long-context embedding models such as Jina Embeddings 2 (Günther et al., 2023), Nomic Embed (Nussbaum et al., 2024), and GTE (Zhang et al., 2024) have emerged to address limitations in sequence length, allowing for the processing of larger text segments. These models are capable of processing up to 8,192 tokens, overcoming constraints of traditional BERT-based embeddings. Jina Embeddings 2 extends its context capabilities with techniques like Attention with Linear Biases (ALiBi) (Press et al., 2021), while Nomic Embed employs a multi-stage training approach that uses a vast dataset of 235 million text pairs to capture complex dependencies. GTE, on the other hand, integrates a reranking system with contrastive learning to further improve retrieval precision. Together, these models facilitate the retrieval of information from extensive legal texts, making them well-suited for tasks involving lengthy documents, as required in legal retrieval.

Evaluation of these models on retrieval benchmarks, such as BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2022), has shown their effectiveness in both short and long-context retrieval tasks, providing a more comprehensive assessment of their abilities across varied retrieval scenarios. BEIR evaluates dense retrievers on a heterogeneous set of zero-shot retrieval tasks, while MTEB specifically measures embedding models across a wide range of tasks and sequence lengths, which is essential for understanding model performance on extended legal documents.

Complementing these embedding models, large language models (LLMs) such as GPT (Brown, 2020) and Llama (Touvron et al., 2023) have introduced new approaches for generating high-quality synthetic data. Among open-source LLMs, Llama 3.1 (Dubey et al., 2024), developed by Meta, offers promising capabilities for generating synthetic queries that could potentially enhance model alignment with real-world search tasks. Synthetic data generated by models like Llama 3.1 holds potential for improving retrieval performance while reducing the reliance on extensive manual annotation, an area currently under exploration in our work.

Retrieving From	Pre-trained	Fine-tuned
NIRC (311 sections)	Baseline	
Split NIRC (826 subsections)	Section Splitting	Fine-tuned
Split NIRC (826 subsections) + Synthetic Dataset (7310 entries)		Synthetic Query-Assisted

Table 1: Comparison of Models and Methods Used for Retrieval

This study evaluates the performance of five embedding models—Word2Vec, SBERT, Jina AI's Jina Embeddings 2, Nomic AI's Nomic Embed, and Alibaba NLP's GTE—in retrieving relevant sections of the NIRC from queries. We compiled a testing dataset from bar exam reviewers, which provided realistic queries tied to specific sections of the NIRC. To improve retrieval accuracy in the dense legal language of the NIRC, we explored section splitting, dividing lengthy sections into focused, content-specific segments. This helps the models target precise legal concepts and reduces the retrieval of unrelated information, aligning text structure with the models' strengths in representation. Fine-tuning the models with synthetic data generated by Llama 3.1 8B Instruct allowed us to simulate realistic legal questions, similar to the use of synthetic data in PAQ, enhancing query relevance without additional manual annotations.

Our experiments demonstrate that splitting large sections of the NIRC into smaller subsections significantly enhances retrieval performance, allowing models to focus on more granular legal text. However, fine-tuning with synthetic data yielded only marginal improvements in retrieval accuracy, suggesting that while synthetic data holds promise, current models may not fully capture the intricate language patterns of Philippine legal texts. The synthetic query-assisted retrieval approach also produced limited gains, highlighting areas where future research could refine embedding models for specialized legal applications. This work not only contributes to the development of legal retrieval tools in the Philippines but also underscores the potential and limitations of embedding models and synthetic data in complex legal NLP tasks.

2 Methodology

This section outlines the process of dataset creation, model training, and evaluation for the re-

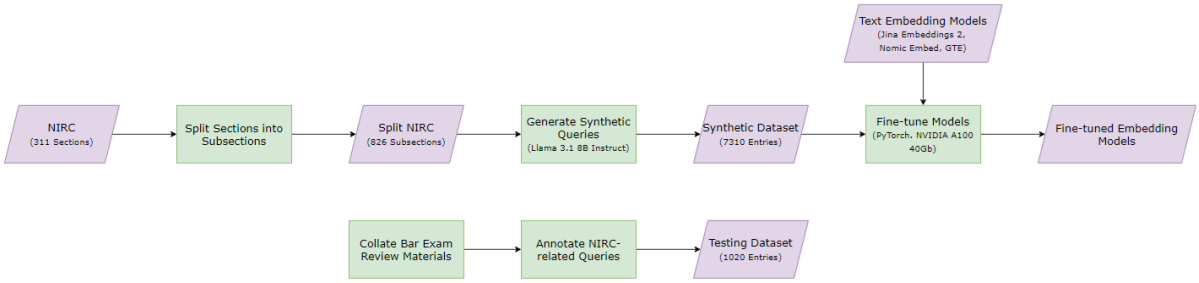

Figure 1: Synthetic Dataset and Testing Dataset Generation Process: Depicting the workflow from splitting the NIRC into subsections to generating synthetic queries with Llama 3.1 and annotating bar exam materials, followed by fine-tuning embedding models with the synthetic dataset.

trieval of relevant sections from the NIRC. The overall workflow for the creation of the datasets and the subsequent training process is illustrated in Figure 1. Two datasets were utilized: a manually annotated testing dataset compiled from bar exam reviewers and a synthetic training dataset generated using Llama 3.1. The synthetic dataset was used to fine-tune three embedding models with the goal of improving their retrieval accuracy.

The experimental setup, summarized in Table 1, involved testing the models' performance on legal queries from the bar exam reviewer dataset, both with and without fine-tuning on the synthetic data. Additionally, we implemented section splitting, dividing the NIRC into smaller subsections to enhance retrieval accuracy. This was done based on structural headings such as "(A)", "(B)", and similar markers found in the NIRC, where we kept any text that came before the heading all of the subsections. Afterwards, any sections found to be greater than 1,000 words were also split into two subsections at the period (".") closest to the center. We also evaluated the effectiveness of synthetic query-assisted section retrieval, which matched queries with synthetic questions to potentially improve retrieval outcomes. The experiments focused on two key performance metrics: Mean Reciprocal Rank (MRR) and Top-1 retrieval accuracy, with hyperparameter tuning for learning rate and threshold values.

2.1 Dataset Creation

To evaluate the performance of the embedding models on retrieving relevant sections of the NIRC, two datasets were utilized: a manually annotated testing dataset and a synthetic training dataset generated using Llama 3.1.

2.1.1 Manually Annotated Testing Dataset

The testing dataset was compiled from bar exam reviewers found in Philippine law school websites and libraries, consisting of legal questions that tax professionals and students commonly encounter when preparing for the Philippine Bar Exam. Each question in this dataset was paired with its corresponding section of the NIRC, as dictated in the bar exam review materials, which served as the ground truth for evaluating the retrieval accuracy of the models. These documents and the NIRC are written in English legalese. Augmentation to increase quantity and quality of the dataset was done by lawyers. This dataset provided a pragmatic basis to assess how well the trained embedding models could retrieve the correct section from the NIRC based on legal queries.

2.1.2 Synthetic Training Dataset

The synthetic training dataset was formatted akin to the manually annotated testing dataset, i.e., with Philippine taxation queries and the corresponding most relevant NIRC section. It was generated using Llama 3.1, was the sole data used for training the embedding models. Before generating the synthetic questions, the NIRC sections were first split into smaller subsections following the rules mentioned earlier. This allowed for a more granular breakdown of the legal text, ensuring that each part of the section was addressed individually. Once split, these subsections were fed into Llama 3.1 8B Instruct using a carefully crafted system prompt designed to generate diverse and insightful questions for each subsection. The prompt was as follows:

> You are an experienced law professor and tax consultant specializing in the Philippine Tax Code. Your goal is to help students, laypersons, and tax professionals

understand complex legal concepts by generating insightful and relevant questions that can be answered by the provided sections of the tax code. Focus on clarity, precision, and ensuring the questions test comprehension of the key legal principles. From the following text in the Philippine Tax Code, generate multiple potential queries regardless of the length of the section. These queries should include a mix of questions that both laypersons and tax professionals might ask. Ensure each part of the section is addressed with a relevant query. If the section is brief, provide both basic and more detailed queries. Avoid using the phrases 'Philippine Tax Code', 'this Code', 'this section', or 'this law' when referring to the section; use 'the law' if necessary. Format each query as 'Query 1: ... Query 2: ... Query 3: ...'

This prompt enabled Llama 3.1 to generate a variety of questions for each subsection of the NIRC, ensuring that both simple and complex aspects of the tax code were addressed. The synthetic dataset thus consisted of multiple queries for each subsection, covering the full spectrum of legal complexities found in the NIRC.

This synthetic dataset was used to fine-tune the embedding models, with the goal of improving their ability to match real-world user queries to the relevant sections of the NIRC. The training process aimed to enhance the models' retrieval accuracy by aligning them more closely with the structure and language of the NIRC, as reflected in the synthetic data.

2.2 Experimental Setup

The experiment was designed to evaluate the retrieval performance of five embedding models—Word2Vec, SBERT, Jina Embeddings 2, Nomic Embed, and GTE—on the NIRC. The setup involved two stages: pre-trained baseline testing and fine-tuning the models using synthetic data generated by Llama 3.1. Additionally, synthetic query-assisted section retrieval was explored to assess its effectiveness in improving retrieval accuracy.

2.2.1 Pre-training Baseline Setup

Initially, the models were tested without any fine-tuning to establish a performance baseline. In this stage, the pre-trained embeddings of Jina Embeddings 2, Nomic Embed, and GTE were directly used to retrieve relevant NIRC sections based on the bar exam reviewer queries. The entire NIRC was split into sections and further divided into subsections, each containing fewer than 2,000 words, to enable more granular retrieval. The queries from the manually annotated dataset were encoded using each model's pre-trained embeddings, and cosine similarity was computed between the query embeddings and the subsection embeddings. The resulting similarity scores were used to rank the relevant NIRC subsections for each query. This baseline evaluation was critical for comparing the performance gains achieved through fine-tuning with synthetic data.

Hyperparameter	Value
Seed	42
Max Sequence Length	2048 Tokens
Batch Size	Variable
Gradient Accumulation	42 / Batch Size
Learning Rate	Jina: 4e-7 Nomic: 4e-7 GTE: 1e-7
Optimizer	AdamW
Mixed Precision	Enabled
Loss Function	Cosine Similarity Loss

Table 2: Summary of Hyperparameters Used for Model Training

2.2.2 Fine-tuning with Synthetic Data

Following the baseline tests, the embedding models were fine-tuned using a synthetic dataset generated through Llama 3.1. The synthetic data consisted of multiple queries for each subsection of the NIRC, designed to simulate realistic legal questions that laypersons, students, and tax professionals might ask. Fine-tuning was carried out with the goal of aligning the embeddings more closely with the language and context of the NIRC, thus enhancing their ability to retrieve the correct sections when presented with bar exam reviewer queries.

All training was conducted on a single NVIDIA A100 GPU with 40GB of VRAM. Due to the limited memory capacity and to ensure training stability, a batch size of 3 was used for Jina Embeddings 2, Nomic Embed, and GTE, which take significantly more compute than Word2Vec and SBERT. However, to emulate a larger effective batch size, gradient accumulation techniques were employed (Piao et al., 2023). Specifically, the losses were accumulated over 14 training steps before updating the model weights, resulting in an effective

Section	Manually Annotated Testing Dataset Queries	Synthetic Dataset Queries
Sec. 38 (Losses from Wash Sales of Stock or Securities)	Are losses from wash sales of stocks or securities deductible to gross income?	What is the specific condition that prevents a taxpayer from deducting a loss from the sale of shares of stock or securities?
Sec. 58 (Returns and Payment of Taxes Withheld at Source)	Is the payee responsible for withholding the tax?	What entities, aside from withholding agents, can receive tax payments on behalf of the government?
Sec. 109 (Exempt Transactions)	Is a bar review center owned and operated by lawyers subject to VAT?	What type of educational institutions are automatically exempt from paying VAT, as per the law?

Table 3: Sample Queries from the Manually Annotated Testing Dataset and Synthetic Dataset, Tied to Corresponding Sections of the NIRC.

Statistic	Count
Manually Annotated Testing Dataset Entries	1,020
Synthetic Training Dataset Entries	7,310
Total Number of NIRC Sections	311
Total Number of NIRC Subsections (After Split)	826

Table 4: Dataset Statistics

batch size of 42. AdamW (Loshchilov and Hutter, 2017) was used as the optimizer to manage weight decay and improve generalization. This method allowed us to strike a balance between computational resource constraints and the need for larger batch sizes to stabilize training (Möller et al., 2021). Given that the longest NIRC section contains almost 7,000 words, we opted to train only with the subsectioned NIRC rather than experimenting with whole NIRC sections, as processing the entire sections would have exceeded the available memory capacity. This approach was consistently applied to the fine-tuning process across all three models: Jina Embeddings 2, Nomic Embed, and GTE.

The fine-tuning process involved standard training on the synthetic dataset, with the models learning to map queries to their corresponding NIRC subsections. The models were optimized using backpropagation, and during training, the embeddings were continuously adjusted to reduce the distance between the query embeddings and the target subsection embeddings. This stage aimed to increase retrieval accuracy by improving the models' understanding of the specific legal terminology and context of the NIRC. The hyperparameters are summarized in 2

2.2.3 Synthetic Query-Assisted Section Retrieval

To further investigate model performance, synthetic query-assisted section retrieval was tested. This mechanism was designed to determine whether a query in the test dataset had a strong similarity with a question in the synthetic dataset. If the similarity score between a test query and a synthetic question exceeded a certain threshold, the corresponding NIRC subsection from the synthetic data was ranked higher in the retrieval results, regardless of the cosine similarity score with the original section embeddings.

This system was introduced to explore whether the synthetic data could improve retrieval accuracy by providing an additional signal in cases where test queries closely resembled the synthetically generated questions. The mechanism's effectiveness was evaluated by tuning the threshold and observing its impact on retrieval metrics.

2.2.4 Evaluation Metrics and Hyperparameter Tuning

The primary evaluation metrics for the experiments were Mean Reciprocal Rank (MRR) and Top-1 retrieval accuracy. These metrics were used to quantify how well the models ranked the correct NIRC subsections relative to the bar exam reviewer queries. MRR was particularly important for measuring the rank position of the first correct answer, while Top-1 retrieval accuracy reflected how often the top-ranked subsection was the correct match.

Hyperparameter tuning was focused on two key areas: the learning rate and the threshold for synthetic data matching. Slower learning rates were selected, and the models were trained for only one epoch to preserve their pre-existing language capabilities from pre-training. The learning rates were tuned by sweeping through a range of values (1e-7 to 9e-7 in 1e-7 increments and 1e-6 to 9e-6 in 1e-7 increments) to identify the optimal setting for each model. This careful tuning process helped to re-

Model	Configuration	Top-1 Accuracy	MRR
Word2Vec	Baseline	0.34	0.46
	Split Sections	0.39	0.52
	Trained From Scratch	0.38	0.50
SBERT	Baseline	0.42	0.52
	Split Sections	0.54	0.64
	Fine-tuned	0.54	0.64
Jina Embeddings 2	Baseline	0.55	0.66
	Split Sections	0.62	0.73
	Fine-tuned	0.64	0.74
	Synthetic Query-Assisted (0.90)	0.63	0.73
	Synthetic Query-Assisted (0.95)	0.64	0.74
Nomic Embed	Baseline	0.51	0.60
	Split Sections	0.64	0.74
	Fine-tuned	0.66	0.75
	Synthetic Query-Assisted (0.90)	0.64	0.74
	Synthetic Query-Assisted (0.95)	0.66	0.75
GTE	Baseline	0.57	0.68
	Split Sections	0.66	0.75
	Fine-tuned	0.66	0.76
	Synthetic Query-Assisted (0.90)	0.66	0.76
	Synthetic Query-Assisted (0.95)	0.66	0.76

Table 5: Top-1 Accuracy and Mean Reciprocal Rank (MRR) for Each Model under Different Configurations

fine the models while preserving their pre-trained language understanding, ultimately improving performance without erasing their prior knowledge. The synthetic data matching threshold was similarly tuned by testing various possible thresholds to determine the point at which retrieval accuracy and MRR were maximized. This iterative tuning process was critical for refining the models and achieving optimal performance.

3 Results and Discussion

In this section, we present the outcomes of the experiments conducted to evaluate the performance of various embedding models in retrieving relevant sections of the NIRC. The results are discussed in the context of both pre-trained and fine-tuned models, with considerations given to the effects of section splitting and the introduction of synthetic data. The experiments aim to measure the retrieval accuracy and ranking effectiveness using two primary metrics: Mean Reciprocal Rank (MRR) and Top-1 retrieval accuracy.

The following subsections provide a detailed breakdown of the datasets used, the baseline evaluation of the models, the impact of section splitting, and the performance of the models under fine-tuning and synthetic query-assisted retrieval frameworks.

3.1 Datasets

The study utilized two primary datasets: a manually annotated testing dataset and a synthetic training dataset. The manually annotated dataset comprised 1,020 entries, featuring a variety of legal queries linked to specific sections of the NIRC. This dataset served as the foundation for evaluating the models' retrieval capabilities.

The synthetic training dataset, generated using Llama 3.1, included 7,310 entries designed to simulate diverse legal queries. This dataset covered 826 subsections of the NIRC, derived from splitting the original 311 sections. The section splitting enabled a more granular retrieval process, enhancing the models' ability to match queries with precise legal content.

3.2 Baseline Evaluation

The baseline evaluation assessed the pre-trained models' ability to retrieve relevant sections from the NIRC without any fine-tuning or section splitting. This provided an initial measure of their performance on legal queries, as summarized in Table 5.

Word2Vec and SBERT were included as initial baselines. Word2Vec achieved a Top-1 accuracy of 0.34 and an MRR of 0.46, indicating limited effectiveness in capturing semantic relationships within legal texts. SBERT performed better, with a Top-1 accuracy of 0.42 and an MRR of 0.52, demonstrating improved semantic understanding compared to Word2Vec.

Among the more advanced models, Jina Embeddings 2, Nomic Embed, and GTE showed superior performance. Jina Embeddings 2 achieved a Top-1

Query	Baseline Retrieval	Fine-tuned Retrieval
What would be an exception of a taxable trust?	Revocable trusts. - Where at any time the power to revest in the grantor title to any part of the corpus of the trust is vested (1) in the grantor either alone or in conjunction with ...	Imposition of Tax. (B) Exception. - The tax imposed by this Title shall not apply to employees trust which forms part of a pension, stock bonus or profit-sharing plan ...
What is the composition of gross income?	Period in which Items of Gross Income Included. - The amount of all items of gross income shall be included in the gross income for the taxable year in which received by ...	Gross Income. (A) General Definition. - Except when otherwise provided in this Title, gross income means all income derived from whatever source, including (but not limited to) ...
What is a general professional partnership?	Tax Liability of Members of General Professional Partnerships. - A general professional partnership as such shall not be subject to the income tax imposed under this Chapter ...	Definitions. - When used in this Title: (B) General professional partnerships are partnerships formed by persons for the sole purpose of exercising their common profession ...

Table 6: Example Queries and Retrieval Results comparing baseline pre-trained retrieval to results after section splitting and fine-tuning. Only the first part of the retrieved texts are shown for brevity.

accuracy of 0.55 and an MRR of 0.66, Nomic Embed attained a Top-1 accuracy of 0.51 and an MRR of 0.60, while GTE led with a Top-1 accuracy of 0.57 and an MRR of 0.68. These results indicate that the more sophisticated embedding models are better suited for handling the complexity of legal queries, providing a robust foundation for further enhancements through section splitting and fine-tuning.

Overall, the baseline results demonstrate that while simpler models like Word2Vec and SBERT offer a starting point, more advanced embedding models significantly enhance retrieval accuracy and ranking performance.

3.3 Section Splitting

Splitting the NIRC sections into smaller subsections was hypothesized to improve retrieval accuracy by allowing the models to pinpoint specific legal content more effectively. The results confirmed this hypothesis, with all models exhibiting noticeable improvements in performance after section splitting (see Table 5).

For example, Jina Embeddings 2 saw an increase in Top-1 accuracy from 0.55 to 0.62 and an MRR from 0.66 to 0.73. Nomic Embed improved from a Top-1 accuracy of 0.51 to 0.64 and an MRR of 0.60 to 0.74. GTE also benefited, with its Top-1 accuracy rising from 0.57 to 0.66 and MRR from 0.68 to 0.75. These enhancements suggest that section splitting enables more precise matching between legal queries and relevant portions of the NIRC by reducing ambiguity and allowing the models to focus on more specific text segments.

Table 6 illustrates examples where section splitting, combined with fine-tuning, led to correct retrievals where the baseline models failed. Specifically, queries related to the definitions of terms posed challenges for baseline models, as entire sections containing multiple definitions were treated as single embeddings. This often resulted in the retrieval of general sections rather than specific subsections containing the relevant definitions. By splitting sections into smaller, focused subsections, the models were able to accurately identify the correct portions of the NIRC, demonstrating the efficacy of the section splitting approach in enhancing retrieval precision.

3.4 Fine-tuning

Fine-tuning the embedding models using the synthetic dataset generated by Llama 3.1 was explored to potentially enhance retrieval performance. The best models were fine-tuned using learning rates of 4e-7 for Jina Embeddings 2 and Nomic Embed, and 1e-7 for GTE. These learning rates were optimized to achieve the best possible performance without overfitting the models to the synthetic data. However, the results showed only marginal improvements or negligible changes in Top-1 accuracy and MRR (see Table 5).

For instance, Jina Embeddings 2 experienced a slight increase in Top-1 accuracy from 0.62 to 0.64 and an MRR from 0.73 to 0.74. Nomic Embed showed a minor rise in Top-1 accuracy from 0.64 to 0.66 and an MRR from 0.74 to 0.75. GTE maintained consistent performance with minimal changes. These limited gains indicate that fine-

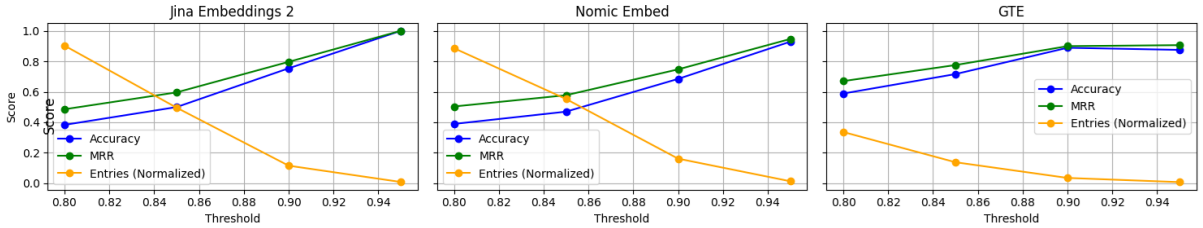

Figure 2: Performance of the models on synthetic data matching at different thresholds showing accuracy, MRR, and the percentage of entries from the testing dataset greater than each threshold.

tuning with the synthetic data did not significantly enhance the models' ability to retrieve relevant sections from the NIRC.

3.5 Synthetic Data Matching

We also evaluated the models' ability to match test queries with synthetic dataset questions based on similarity alone. The results indicated that synthetic query-assisted retrieval did not provide significant improvements over the fine-tuned models (see Table 5).

At higher similarity thresholds (0.90 and 0.95), the models achieved the best accuracy and MRR when matching the testing set queries with the synthetic data queries. However, these thresholds were met by fewer than 1% of the entries, limiting their practical utility. Lower thresholds allowed for a broader range of matches but resulted in decreased performance metrics, particularly for Jina Embeddings 2 and Nomic Embed. Although GTE maintained relatively stronger performance across all thresholds, the gains were not substantial. The added complexity and computational load of processing 7,310 synthetic entries did not translate into meaningful performance benefits, suggesting that the synthetic query-assisted approach may not be advantageous within the current framework.

4 Conclusion

This study evaluated embedding models for legal information retrieval within the Philippine National Internal Revenue Code of 1997 (NIRC). We started by compiling a manually annotated dataset of 1,020 queries from bar exam reviewers. Based on these annotations, we generated a synthetic dataset of 7,310 entries using Llama 3.1 8B Instruct.

Baseline evaluations were conducted using pretrained embedding models—Word2Vec, SBERT, Jina Embeddings 2, Nomic Embed, and GTE—on the full NIRC sections. Splitting the NIRC sections into smaller subsections yielded the most substantial improvements in retrieval accuracy, increasing Top-1 accuracy by up to 13% and MRR by up to 0.14.

We then fine-tuned the models on the synthetic data with split NIRC sections, but this resulted in little to no further enhancements, with improvements less than 2%. Additionally, attempting to assist retrieval by matching input queries with synthetic queries did not contribute any improvements.

These findings highlight that while section splitting significantly enhances retrieval performance, fine-tuning with synthetic data and synthetic query-assisted retrieval offer limited benefits in highly nuanced and specialized domains like Philippine legal text. Future work could explore more advanced models with greater capacity, such as Llama 3.1 405B, and incorporate larger, more diverse annotated datasets to improve legal information retrieval systems within the Philippine legal framework.

Acknowledgments

We thank Hans Nolasco, Jenine Valencia, and Michael Ng for their assistance with the manual testing dataset annotations. Their efforts contributed to the quality of this study.

References

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

XinYu Piao, DoangJoo Synn, Jooyoung Park, and Jong-Kook Kim. 2023. Enabling large batch size training for dnn models beyond the memory limit while maintaining performance. *IEEE Access*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Berthier Ribeiro-Neto and R Baeza-Yates. 2011. Modern information retrieval: the concepts and technology behind search.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.