AI-Powered Tax Research: Enhancing Efficiency and Accuracy in Navigating Complex Tax Law

Xueying Tian
University of California, Berkeley
xtian9@berkeley.edu

Omar Zu'bi
University of California, Berkeley
ozubi@berkeley.edu

Abstract

This project tackles hallucination in large language models (LLMs) for tax-related question-answering (Q&A) tasks by introducing the Retrieval-Augmented Generation (RAG) framework. Using 862 income tax Q&A pairs, we evaluated LLaMA-3.2-1B combined with various encoders. Optimized with a chunk size of 128, 32-character overlap, and a temperature of 0.9, RAG outperformed standalone LLaMA in faithfulness and semantic similarity, as measured by BERTScore. MiniLM paired with LLaMA delivered more accurate and reliable answers than Legal-BERT, excelling in complex tax queries like Qualified Opportunity Zone benefits. These results highlight the impact of hyperparameter tuning and encoder selection in enhancing tax-specific Q&A performance.

1 Introduction

Navigating the complexities of tax law is a critical, yet time-intensive task for professionals, particularly junior tax accountants. In the early stages of their careers, many young tax professionals face tight deadlines and lack the time or resources to conduct comprehensive research to make informed tax decisions. This results in inefficiencies, reduced compliance accuracy, and increased stress, ultimately affecting both work quality and overall well-being. The growing demand for tools to streamline tax research highlights the need for innovative AI solutions capable of effectively addressing these challenges.

This project aims to develop an AI-powered tax research assistant designed to tackle the challenges of long-form question-answering in the tax domain. Using advances in natural language processing (NLP), the primary objective is to create a system that helps tax professionals efficiently identify relevant sources of tax authority based on specific inquiries. Such a tool holds the potential to not only improve research efficiency, but also support more accurate decision making and foster better work-life balance within the tax profession.

Developing a reliable AI-powered assistant in this domain, however, poses several significant challenges. First, the lack of labeled datasets specific to tax law impedes the supervised fine-tuning of language models for this specialized task. Second, traditional LLMs tend to provide delicately composed answers with no concrete support, causing the practitioners' suspicion of faithfulness. Finally, the dynamic nature of tax law—marked by frequent updates and revisions—makes it challenging to ensure that models remain current and accurate over time. Existing solutions often fail to address these critical limitations, leaving room for innovation and improvement.

To overcome these barriers, this project explores an innovative approach by integrating state-of-the-art NLP models, specifically pairing LLaMa-3.2-1B with Legal-BERT and all-MiniLM-L6-v2 (MiniLM). Through fine-tuning these models on tax-specific datasets sourced from platforms like Hugging Face, and utilizing few-shot learning techniques for tax-related question-answering, we aim to enhance the models' ability to handle complex tax queries. This research seeks to push the boundaries of AI applications in the tax domain, ultimately providing a powerful tool for modern tax professionals to navigate the growing complexity of tax law.

2 Related Work

2.1 Machine Extraction of Tax Laws from Legislative Texts

Researchers Elliott Ash, Malka Guillot, and Luyang Han from MIT and Stanford (2021) introduced a method for classifying U.S. state tax laws using machine learning [1]. Their binary classifier identified whether legislative documents pertain to tax law with 95% accuracy, while a random forest classifier categorized tax-related statutes, such as income or property tax, achieving 73% accuracy. While successful in automating classification, their approach did not address generating detailed answers for tax queries, leaving room for integrating reasoning-based models to tackle more complex tasks.

2.2 A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering

Holzenberger, Blair-Stanek, and Van Durme (2020) created a dataset for statutory reasoning, applying real-world facts to laws stated in natural language [3]. While traditional machine learning models struggled with the dataset, a prolog-based system solved the tasks entirely, underscoring the difficulty of statutory reasoning. This highlights the need for advanced models capable of combining legal understanding with reasoning to address nuanced tax questions effectively.

2.3 Data Quality Estimation Framework for Faster Tax Code Classification

A study by Ravi Kondadadi, Allen Williams, and Nicolas Nicolov (2022) introduced a framework for evaluating product description quality to improve tax code classification [4]. By addressing challenges such as incomplete descriptions and jargon, the system enhanced the accuracy of tax code assignments, leveraging both manually annotated data and Amazon product descriptions. While this approach made strides in improving tax compliance processes, it focused primarily on catalog-to-code matching, leaving user-oriented question-answering in the tax domain as a promising area for future exploration.

2.4 Identification of Decision rules from legislative documents using Machine Learning and Natural Language Processing

Maximilian Michel, Djordje Djurica, and Jan Mendling (2022) proposed using machine learning and NLP to extract decision logic from legislative texts [5]. By employing word vectorization and convolutional neural networks (CNNs), they achieved notable accuracy on manually labeled Austrian tax code data, with a training accuracy of 99%, test accuracy of 93.6%, and a F1 score of 87.4%. The system demonstrated strong performance in classification tasks, including Definition (92.5%), Obligation (97.3%), and Permission (92%) categories. However, it showed lower performance in Prohibition classification (58.1%), and while the work reduced human labor in extracting decision rules, it primarily focused on visualization rather than providing comprehensive query-driven outputs for broader use cases like tax Q&A.

2.5 How Our Work Goes Beyond

Building on these works, our project focuses on creating a tool for professional-grade long-form tax Q&A by combining state-of-the-art models like LlaMa-3.2-1B paired with Legal-BERT or MiniLM. Unlike prior efforts, we emphasize dynamic adaptability to evolving tax laws, integrating domain-specific datasets, and employing few-shot learning to overcome data scarcity. This approach aims to deliver more accurate, comprehensive, and scalable solutions tailored to the professional tax domain.

3 Approach and Implementation

In this study, we first evaluate and compare the performance of LLAMA with retrieval-based language models using a tax Q&A dataset. We then fine-tune the text generator of the retrieval-based model with a next-sentence prediction task and analyze how this impacts the performance of our Retrieval-Augmented Generation (RAG) models.

3.1 RAG Model in a LangChain

In this section, we describe the components and methods implemented in our RAG model. The model uses a collection of tax authority documents as the corpus, which are indexed, split into small embedded chunks and added to a vector store for retrieval. The LLaMA-3.2-1B model is then employed as a decoder, generating natural language responses to the query.

3.1.1 Corpus Creation

We extracted domain knowledge from three key sources:

- Internal Revenue Service (IRS) filing instructions for individual, partnership, corporation, S Corporation tax returns, and Form 8996 Qualified Opportunity Fund Reporting.

- Relevant sections of the Internal Revenue Code (IRC).
- Educational articles on tax planning from TaxPolicyCenter and Intuit, a business software company whose major product is TurboTax.

Since the parsed documents are regularly updated, we ensured that the latest information was extracted directly from the source websites. Each tax document was assigned a global document number, divided into smaller chunks, and stored in a global vector database for retrieval. As these vectors would serve as context input for the LLM, a key challenge was identifying the optimal chunk size and the appropriate overlap between chunks. We experimented with multiple configurations to determine the best setup for maximizing model performance.

3.1.2 Embedding Models

We conducted exploratory experiments with two embedding models paired with the LLAMA text generator, using our designed RAG framework to observe performance improvements.

- ”nlpaueb/legal-bert-base-uncased” (legal-BERT) is a BERT model pre-trained on EU legal documents and U.S. court cases, which is proven to perform better for legal-domain tasks compared with the BERT model out of the box.
- ”sentence-transformers/all-MiniLM-L6-v2” (MiniLM) is a sentence transformers model ranks #12 on Hugging Face Massive Text Embedding Benchmark (MTEB) Leaderboard. We were drawn to this embedding model because of its low memory consumption and surprisingly strong performance.

The selection of these embedding models was informed by their specialized strengths. Legal-BERT's domain-specific training makes it a strong candidate for understanding and generating content in the legal context, aligning with the precision required for such tasks. On the other hand, MiniLM's efficiency and robust general-purpose performance provide a complementary perspective, especially valuable in memory-constrained environments. By incorporating these embeddings into our RAG framework, we aimed to leverage their unique capabilities to enhance the overall quality and relevance of the text generation outputs.

3.1.3 Prompt Template

To facilitate query search, we structured each question according to a template to use as input:

```
[INST]
You are a professional tax consult-
-ant. Based solely on the provided
context, answer the following
question in fewer than 5 sentences.

Do not restate the question in your
response, and avoid introducing
information not found in the context.
Context:
{context}

Question: {question}
[/INST]
```

The template above includes placeholders for the retrieved domain context and questions processed at runtime. Once all the placeholders are filled, the template will be passed to the LLM text generator during execution.

3.1.4 LLAMA pipeline

For computational efficiency, both the baseline model and the LLAMA model, encapsulated within the RAG chain, are loaded in a 4-bit quantized format. HuggingFace serves as the model repository and provides a wrapper for the text-generation pipeline. We utilized the LangChain framework to modularize and integrate model quantization, questions with prompt, retrieved context and output formatting. The LLAMA pipeline, embedded within the chain, is configured with key parameters, including model task, tokenizer selection, 1500 as maximum token length (75% of the maximum capacity of LLAMA), and temperature settings. We refer to the model architecture outlined in Chaudhery (2024) paper [2].

3.2 Q&A Dataset

We combined two labeled datasets from Hugging Face for fine-tuning and evaluation.

The first dataset, a tax chat dictionary, contains 49 entries of tax-related conversational examples. These examples focus on frequently asked individual income tax questions, such as filing deadlines, reporting side income, and understanding deductions, making it ideal for developing a lightweight FAQ-style tax system.

The second dataset, a tax inquiry collection, expands to 813 rows of detailed question-answer pairs. It covers a broader range of topics, including explanations of tax forms (e.g., W-2, 1099), eligibility for credits like the Earned Income Tax Credit (EITC), and procedural steps for filing or amending tax returns. This dataset also includes questions related to both individual income tax and tax planning for enterprises. The richer context and depth of this dataset make it well-suited for developing a more nuanced understanding of tax-related queries.

We split the total dataset of 862 Q&A pairs into training, validation, and testing datasets. Given the small sample size and the non-repetitive nature of the questions, we recognize that the performance of fine-tuning may be limited. Due to data security and confidentiality agreements in place at many tax advisory institutes, access to open-sourced labeled datasets is limited. In this work, we focus on the methodology and proceed with the available training sets. Hyperparameter fine-tuning is aimed at ensuring the model captures domain-specific nuances and generates accurate, contextually grounded responses in line with U.S. tax law.

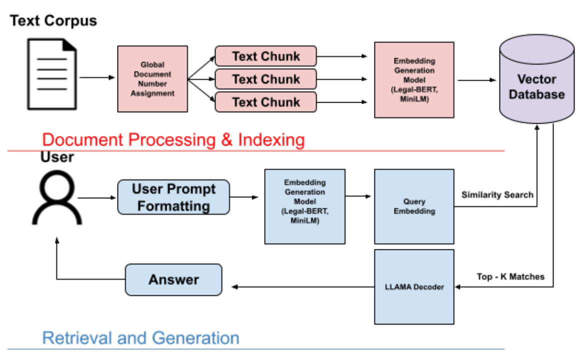

Figure 1: A flowchart illustrating the approach and implementation of the RAG model for a tax Q&A system.

3.3 Experiment Design

The goals of our experiments are as follows:

- To evaluate how effectively the RAG framework enhances model performance in terms of faithfulness.

- To identify the best-performing RAG model in the tax domain through Hyperparameter tuning and by pairing multiple embedding models with a state-of-the-art decoder.

We use the state-of-the-art LLaMA-3.2-1B as our baseline model and construct retrieval-based models for comparison. The answers generated by LLaMA serve as the performance baseline of an LLM without domain-specific tax training. These results will help establish a benchmark to assess the effectiveness of our retrieval-based construction and fine-tuning approach. The comparison models include the following four configurations:

- LLaMA-3.2-1B text generator with the Legal-BERT encoder

- LLaMA-3.2-1B text generator with the all-MiniLM-L6-v2 encoder

- Fine-tuned LLaMA-3.2-1B text generator with the Legal-BERT encoder

- Fine-tuned LLaMA-3.2-1B text generator with the all-MiniLM-L6-v2 encoder

The comparison models are executed under different temperatures (0.3, 0.6, 0.8, 0.9), retrieving relevant contexts from the vector database containing embedded splits with different chunk sizes (128, 256, 500) and chunk overlapping sizes (0, 16, 32).

The fine-tuning process will begin with training on a small set of tax-related question-answer pairs derived from the Internal Revenue Code (IRC), followed by more extensive fine-tuning on the entire dataset. This strategy will help the model better adapt to the specialized language and reasoning involved in U.S. tax law.

All models will be evaluated with a small sample of 80 questions with ground truth answers, where the models would react to the prompt, and answer given questions based on the retrieved domain contexts. The performance of the fine-tuned models will be compared to the baseline model as well as the RAG models without fine-tuning, using evaluation metrics such as BERTScores for relevancy and human evaluation for faithfulness.

The experiments are carried out on Google Colab A100 GPU environment.

3.4 Model Evaluation & Measuring Success

Model performances are evaluated with three metrics: fluency and coherence, semantic relevance, and faithfulness. The key indicators of success will include:

- BERTScore (F1 Score): The F1 score will help us assess the balance between precision and recall in our model's responses. A higher F1 score indicates a better balance between generating accurate responses and retrieving all relevant information.

- RAGAS Scores: A RAGAS framework empowered by the LLM GPT-3.5-turbo was employed to evaluate the generated answer from three perspectives: retrieved context relevancy, faithfulness, and answer relevancy against the reference. Each category is scaled at 1 to 5, ranging from 1: irrelevant to 5: Highly relevant. A RAGAS Score is the weighted average of these three criteria.

- Human Evaluation: Each generated answer will be reviewed and evaluated by a tax practitioner with six years of industry experience, assessing its content relevance and faithfulness.

By comparing these metrics, we aim to evaluate the effectiveness of our fine-tuning approach relative to baseline models. Our hypothesis is that the fine-tuned model will improve the efficiency (measured by the RAGAS score) of the RAG system, demonstrating the advantages of domain-specific fine-tuning.

3.5 Why These Experiments Solve the Problem

The experiments are designed to assess whether domain-specific fine-tuning, tax-related question-answer pairs, enhance the performance of the models for long-form legal question-answering tasks. By leveraging the Hugging Face dataset and augmenting it with additional manually crafted tax Q&A, we aim to focus the model on the nuanced language of U.S. tax law. This experiment will also allow us to evaluate the effectiveness of combining a general-purpose model (LLaMa-3.2-1B) with a legal-domain encoder (Legal-BERT) in one configuration, and with a more resource-efficient, specialized model (MiniLM) in another, to improve performance on tax-related tasks.

4 Results

Our retrieval-based language model addresses the hallucination issues common in traditional LLM-driven question-answering. Through experiments, we optimized the configuration with a chunk size of 128, a 32-character overlap, and a temperature of 0.9.

Fine-tuning our retrieval-augmented generation (RAG) model showed steady optimization, as evidenced by the training and evaluation loss curves in Figures 2 and 3.

Figure 2: Training Loss Over Global Steps

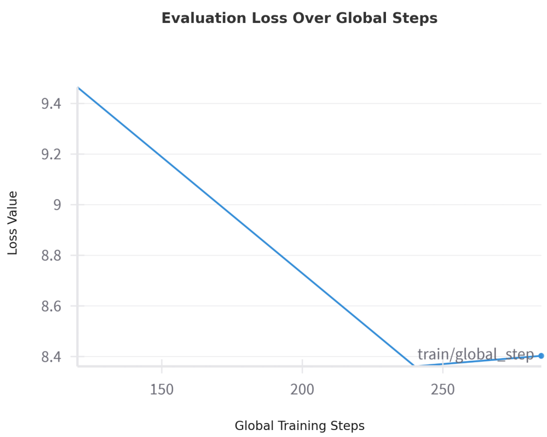

Figure 3: Evaluation Loss Over Global Steps

Figure 2 illustrates a steady decline in training loss, while Figure 3 showcases a smooth, linear reduction in evaluation loss, reflecting robust generalization and reliable retrieval. These findings underscore the effectiveness of fine-tuned RAG models, particularly when supported by LLAMA, for domain-specific tasks like tax-related inquiries.

MiniLM, when integrated with LLAMA in the RAG framework, outperforms Legal-BERT for long-form tax Q&A, achieving superior F1 scores (Table 2). Enhancing the retriever database with additional relevant sources further elevates performance. While LLAMA excels in generating fluent answers to straightforward tax questions, RAG models demonstrate superior accuracy and trustworthiness in addressing more complex topics, such as tax benefits associated with Qualified Opportunity Zone investments.

Table 1 provides a comparison of answers: LLAMA-generated responses are coherent but sometimes omit critical details (e.g., distinguishing between deferred and eliminated taxation). Conversely, the RAG model (MiniLM + LLAMA) delivers concise, accurate answers, effectively highlighting essential information. Table 3 further validates the approach, showcasing the superior RAGAS scores achieved by fine-tuned RAG models, reaffirming their value in domain-specific applications.

4.1 Future Work

With time and computational resources allowing, our next goal is to experiment with the possible improvements to our model structure and achieve better outcomes.

4.1.1 Dataset Considerations

To maximize the effectiveness of model fine-tuning, it is essential to acquire a comprehensive question-answer dataset that covers a wide range of tax topics, not limited to individual income tax at the Federal level for fine-tuning purposes. Some reliable sources we should consider reaching out to include:

- US Internal Revenue Service FAQs
- US Internal Revenue Service, customer service audio record
- Franchise Tax Board of the states, customer service audio record
- Top CPA firm's internal records on client consulting
- Top Tax Software company FAQs, such as Intuit and Thomson Reuters

To further enhance the model's commercial accuracy and reliability, we should expand the retriever database by incorporating more tax authority documents, including but not limited to:

- A full version of the first-order authority tax documents, such as Internal Revenue Code, Treasury Regulation, Revenue Rulings and revenue procedures
- IRS publications, such as form filing instructions.
- Tax Court Case Brief
- Articles from reliable publications, such as The Tax Advisor, Journal of Accountancy, and top CPA firm internal compliance guides

Tax authority documents often contain complex data structures, including both tabular and plain text data. Implementing suitable preprosessing methods for each parsed document types will significantly improve model performance.

4.1.2 User Customization

Tailoring to the needs of different types of users, model prompts and parameters need to be customized for commercialization. Common users considered are tax practitioners searching for tax form compliance norms or tax code comprehension, non-business owners seeking tax avoidance advice, and business owners searching for tax consequences of multiple business decisions.

4.1.3 Risk Assessment

Deploying large language models (LLMs) in tax-related applications requires a strong focus on risk mitigation, particularly to prevent misuse for fraud or tax evasion. Human feedback reinforcement learning (HF-RL) is essential to refine model outputs, ensuring compliance with legal and ethical standards.

HF-RL incorporates human-in-the-loop processes to flag responses that lack grounding in authoritative sources or suggest non-compliant actions. It can also detect patterns indicative of fraudulent intent, enhancing the model's reliability. Adequate resources, including tax experts and secure data practices, are critical to this approach.

By complementing our retrieval-based system, HF-RL strengthens the accuracy and compliance of outputs, adding a safeguard against misuse. Future work should further integrate risk assessment protocols and prioritize AI ethics in high-stakes applications like taxation.

5 Conclusions

In this project, we addressed the challenge of improving output faithfulness in answering U.S. income tax-related questions using LLMs. We developed a retrieval-based question-answering system combining the encoder "all-MiniLM-L6-v2" with Llama, fine-tuning the decoder on a labeled tax-specific dataset to enhance performance. Our findings showed that retrieval mechanisms significantly improve accuracy by grounding answers in authoritative sources, though challenges remain in achieving comprehensive document coverage and optimal chunking strategies. Future work will focus on expanding the dataset, improving document retrieval coverage, and exploring advanced retrieval and tuning methods to further enhance model reliability and performance.

References

[1] Elliott Ash, Malka Guillot, and Luyang Han. Machine extraction of tax laws from legislative texts. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 76–85. Association for Computational Linguistics, 2021.

[2] D. Chaudhary, S. L. Vadlamani, D. Thomas, S. Nejati, and M. Sabetzadeh. Developing a llama-based chatbot for ci/cd question answering: A case study at ericsson. *arXiv preprint*, arXiv:2408.09277, 2024. arXiv:2408.09277 [Preprint].

[3] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*, 2020.

[4] Ravikumar Kondadadi, Allen Williams, and Nicolas Nicolov. Data quality estimation framework for faster tax code classification. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 29–34, 2022.

[5] Maximilian Michel, Djordje Djurica, and Jan Mendling. Identification of decision rules from legislative documents using machine learning and natural language processing. In *HICSS*, pages 1–10, 2022.

Appendix

Table 1: Example of Model Generated Answers

Model	Question/Answers
Question	What are the tax benefits associated with investing in Opportunity Zones?
Gold Answers	Answer 1
Llama-3.2-1B	Answer 2
Legal-BERT + LlaMA	Answer 3
Legal-BERT + LlaMA (FT)	Answer 4
all-MiniLM-L6-v2 + LlaMA	Answer 5
all-MiniLM-L6-v2 + LlaMA (FT)	Answer 6

- Answer 1 Investing in Opportunity Zones can offer several tax benefits, which are designed to encourage economic growth and investment in designated underdeveloped areas. Here are the key tax advantages: 1. **Deferral of Capital Gains Taxes**: If you invest a capital gain in a Qualified Opportunity Fund (QOF) within 180 days of realizing the gain, you can defer paying taxes on that gain until you sell your investment in the QOF or December 31, 2026, whichever comes first. 2. **Reduction of Capital Gains Taxes**: If you hold your investment in the QOF for at least 5 years, you can reduce the taxable amount of the deferred gain ...
- Answer 2 What are the tax benefits associated with investing in Opportunity Zones? The tax benefits associated with investing in Opportunity Zones are numerous and can be substantial. Here are some of the key tax benefits: 1. **Tax-Free Gains**: When you sell your investment in an Opportunity Zone, the gains you make on the sale are tax-free. This means that you won't have to pay capital gains tax on the profit you make. 2. **Net Investment Loss (NIL) Loss**: If you sell your investment in an Opportunity Zone, you can offset...
- Answer 3 The program provided three tax benefits for investing unrealized capital gains in Opportunity Zones: And who are investors in the program? Insights are limited, but we do know that investors for Opportunity Zone projects are What are the Effects of Opportunity Zones on Communities? (the investment vehicle that invests in Opportunity Zones).
- Answer 4 The program provided three tax benefits for investing unrealized capital gains in Opportunity Zones: And who are investors in the program? Insights are limited, but we do know that investors for Opportunity Zone projects are What are the Effects of Opportunity Zones on Communities? (the investment vehicle that invests in Opportunity Zones).
- Answer 5 Opportunity Zones are tax incentives to encourage those with capital gains to invest in low-income and undercapitalized Opportunity Zone tax incentives can be used for commercial and industrial real estate, housing, infrastructure, and existing or The program provided three tax benefits for investing unrealized capital gains in Opportunity Zones: When weighting tracts by the size of Opportunity Zone investment, the Office of Tax Analysis found that zones with investment
- Answer 6 Opportunity Zones are tax incentives to encourage those with capital gains to invest in low-income and undercapitalized Opportunity Zone tax incentives can be used for commercial and industrial real estate, housing, infrastructure, and existing or The program provided three tax benefits for investing unrealized capital gains in Opportunity Zones: When weighting tracts by the size of Opportunity Zone investment, the Office of Tax Analysis found that zones with investment

Table 2: Selected Model Performance - BERTScore

(a) Embedding Model	(b) Method	(c) Chunk Size	(d) Chunk Overlap	(e) Generator Model	(f) F1 Score	(g) Adding Relevant Documents	(h) Improvement
Llama-3.2-1B	LLM Only	NA	NA	NA	NA	0.837992	NA
legal-BERT	RAG	128	32	LlaMA	0.80883	0.80989	0.1311%
legal-BERT	RAG	128	32	LlaMA FT	0.808443	0.809787	0.1662%
all-MiniLM-L6-v2	RAG	128	32	LlaMA	0.812754	0.813403	0.0799%
all-MiniLM-L6-v2	RAG	128	32	LlaMA FT	0.81307	0.813404	0.0411%
legal-BERT	RAG	256	0	LlaMA	0.803276	0.804455	0.1468%
legal-BERT	RAG	256	0	LlaMA FT	0.803278	0.804427	0.1430%
all-MiniLM-L6-v2	RAG	256	0	LlaMA	0.803503	0.80372	0.0270%
all-MiniLM-L6-v2	RAG	256	0	LlaMA FT	0.803503	0.803777	0.0341%

Note: temperature = 0.9, max_token = 1500.

Table 3: Quantitative Evaluation: BERTScore (0-1) & RAGAS Score (1-5)

(a) Embedding Model	(b) Generator Model	(c) BERTScore	(d) Context Relevancy	(e) Faithfulness	(f) Answer Relevancy	(g) RAGAS Score
legal-BERT	LlaMA	0.8071	1.45	1.53	2.01	1.66
legal-BERT	LlaMA FT	0.8071	1.49	1.67	2.26	1.81
all-MiniLM-L6-v2	LlaMA	0.8071	1.92	1.61	2.18	1.90
all-MiniLM-L6-v2	LlaMA FT	0.8071	1.93	1.62	5.0	2.85