TaxTajweez: A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG)

Mohammad Affan Habib
Habib University, Pakistan
Pakistan

Shehryar Amin
Habib University, Pakistan
Pakistan

Muhammad Oqba
Habib University, Pakistan
Pakistan

Sameer Jaipal
Habib University, Pakistan
Pakistan

Muhammad Junaid Khan
University of Central Florida, USA
USA

Abdul Samad
Habib University, Pakistan
Pakistan

Abstract

The advent of Large Language Models (LLMs) has heralded a transformative era in natural language processing across diverse fields, igniting considerable interest in domain-specific applications. However, while proprietary models have made significant strides in sectors such as medicine, education, and law through tailored data accumulations, similar advancements have yet to emerge in the Pakistani taxation domain, hindering its digital transformation.

In this paper, we introduce TaxTajweez, a specialized Retrieval Augmented Generation (RAG) system powered by the `OpenAI GPT-3.5-turbo` LLM, designed specifically for income taxation. Complemented by a meticulously curated dataset tailored to the intricacies of income taxation, TaxTajweez leverages the RAG pipeline to mitigate model hallucinations, enhancing the reliability of generated responses. Through a blend of qualitative and quantitative evaluation methodologies, we rigorously assess the accuracy and usability of TaxTajweez, establishing its efficacy as an income tax advisory tool.

Introduction

LLMs have transformed natural language processing, but despite their advantages, LLMs struggle with specialized fields with specific information, where precise expertise is crucial. Despite the prevalence of many chatbots in the medical (Van Doan et al. 2023), (Li and others 2023), (Wang and others 2023), (Xiong and others 2023), (Singhal and others 2023) and educational (Dan and others 2023), (Levonian and others 2023), (Baladón et al. 2023), (Zhang and others 2023) domains, there has not been much work done in the taxation domain.

Taxation is highly intricate, with regulations often spanning hundreds or thousands of pages with sophisticated and cryptic tax regulations. Consequently, individuals face significant obstacles in obtaining reliable taxation information and guidance (Amjad 2021) relying on human expertise in navigating tax functions. The problem is particularly prevalent in countries like Pakistan where tax laws can be obscure and poorly articulated.

A survey conducted between March and May 2023, involving over 1,800 legal and tax professionals in the U.S., U.K., and Canada, found that 82 percent of legal professionals and 73 percent of tax professionals see ChatGPT as applicable to legal or tax-related tasks (Reuters 2023).

However, this technology has its risks and limitations. Common issues with large language models like ChatGPT include presenting inaccurate or fabricated information and lacking transparency about information origins, hindering confident use in tax research (Alarie et al. 2023).

To address these challenges we developed Tax Tajweez to cater to Paksitan's complex tax dynamics using credible data from government sources followed by comprehensive testing and feedback.

Literature Review

The year 2023 saw many custom-fine-tuned and RAG-based LLM chatbots serviced towards different domains. Researchers have developed use cases for education, medical, legal, and emotional support domains.

In the educational domain, authors of (Levonian and others 2023) designed prompts that retrieve and use content from a high-quality open-source math textbook to generate responses using RAG to real student questions. The authors of EduChat (Dan and others 2023) utilized a combined approach incorporating both RAG and fine-tuning to provide personalized and intelligent educational support. The study presented in (Baladón et al. 2023) discusses the outcomes of involvement in the BEA 2023 shared task, focusing on generating AI teacher responses in educational dialogues. They experimented with various Open-Source LLMs and fine-tuning techniques, including Few-Shot and Chain-of-Thought approaches.

In the medical domain, authors of (Van Doan et al. 2023) and (Wang and others 2023) developed chatbots to address medical queries in the Vietnamese language and the Chinese language respectively. Similarly, (Wang and others 2023) ChatDoctor (Li and others 2023) created a specialized model fine-tuned on patient-doctor dialogues.

The paper (Cui et al. 2023) introduced ChatLaw, a Chinese legal Large Language Model (LLM) developed due to the absence of large-scale commercial models. ChatLaw incorporates legal-specific data and features modules to mitigate hallucination, extract legal words, calculate text similarity, and construct a legal exam testing dataset.

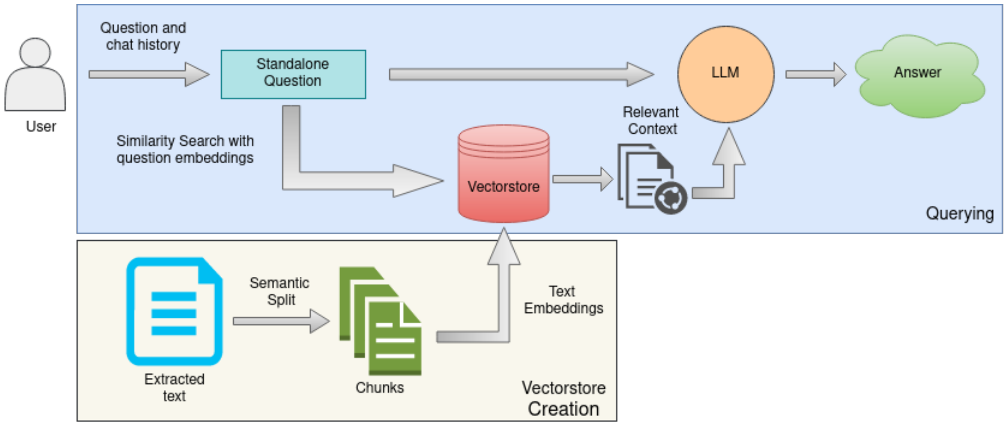

Figure 1: The overall pipeline of data processing, storage in vector database, and querying

Other notable studies presented in (Zhang and others 2023), (Zheng et al. 2023), (Xiong and others 2023), and (Singhal and others 2023), all utilized LLMs for various tasks including fine-tuning for particular tasks, emotional support, and healthcare.

To sum up, in this paper, we introduce TaxTajweez, an LLM-based domain-specific Chatbot that leverages our meticulously curated income-taxation dataset to provide accurate and tailored responses to user inquiries.

Dataset and Preprocessing

Income Tax Ordinance

In our pursuit of leveraging chatbots to facilitate easy access to information related to income tax in Pakistan, our initial focus was on finding a suitable dataset. The optimal corpus for this task would encompass all relevant laws and regulatory data. In our case, we utilized the latest version of the publicly accessible Income Tax Manual sourced from the Federal Board of Revenue website. The dataset encompasses diverse categories, including individual income tax, corporate tax, deductions, exemptions, and detailed accounting guidelines on the treatment of income. It provides a holistic view of income tax laws, rules, and regulations.

In addition to the Income Tax Manual, supplementary information has been incorporated by web scraping data from Taxationpk.com, including procedural data and best practices.

Preprocessing

During the initial phase, the content from PDF documents was extracted. This extracted text was then split into semantically similar chunks using a python library `Semantic-Split` (More and Brin 2023). `Semantic-Split` leverages the `spaCy` and `SentenceTransformers` libraries to first split the text into chunks and then creates embeddings for each chunk using the model `all-MiniLM-L6-v2`. These embeddings from the model are then grouped to form semantically similar text chunks by computing elementwise cosine similarity. As an illustrative example, consider the following input and output generated by Semantic-Split as given by (More and Brin 2023):

Input:

```
I dogs are amazing.
Cats must be the easiest pets around.
Robots are advanced now with AI.
Flying in space can only be done by
Artificial intelligence.
```

Output:

```
[ ["I dogs are amazing.", "Cats must
be the easiest pets around."], ["Robots
are advanced now with AI.", "Flying in
space can only be done by Artificial
intelligence."] ]
```

These groups of semantically similar chunks were then used to create vector embeddings using OpenAI's embeddings model `text-embedding-ada-002` (OpenAI 2024). The resulting vector embeddings were structured and stored in the `Facebook AI Similarity Search (Faiss)` vector store (LangChain 2024). This process enables efficient and improved representation and retrieval of information for subsequent analysis.

Approach and Implementation

Our exploration into crafting an intelligent and responsive tax-specific LLM chatbot hinges on the innovative principles embedded within the Retrieval-Augmented Generation (RAG) framework. The approach employed is shown in Figure 1.

Condensing Chat History and New Question: The process begins with a prompt that takes in the chat history and a new question, producing a standalone question. This standalone question is crucial as it serves as a refined query for the retrieval of relevant documents. We use the following prompt template to generate the standalone question: `"Given the following conversation and a follow up question, rephrase the follow up question to be a standalone question`

```
using only the relevant context from
the conversation, in its original
language.Chat History:{chat_history}
Follow Up Input: {question}".
```

Embedding Generation: Once the standalone question is formed, embeddings for the question are generated using OpenAI's embeddings model, specifically `text-embedding-ada-002`. These embeddings serve as a compact representation of the question in a vectorized form, facilitating efficient similarity searches.

Similarity Search for Document Retrieval: A similarity search is conducted between the generated embeddings for the standalone question and the embeddings stored in the Faiss vector store. This search retrieves relevant document chunks that align with the condensed query, ensuring that the information retrieved is pertinent to the current conversation context.

Leveraging OpenAI GPT-3.5-turbo: The retrieved chunks from the vector-store, along with the standalone question, are then sent to the OpenAI LLM `GPT-3.5-turbo` (OpenAI 2021) to generate a response. Feedback from user interactions is utilized to enhance prompt design, document retrieval, and response generation using `LangChain Chains ConversationalRetrievalChain` (LangChain Contributors 2023), ensuring the chatbot's adaptability to diverse user queries and evolving tax-related topics.

The described implementation forms the backbone of our model, TaxTajweez, offering a systematic approach to information ingestion, condensation, and generation.

Experimental Results

In this section, we explain our evaluation process, strategically divided into two components: a) quantitative evaluation; and b) qualitative evaluation. This division is inspired by the work of (Li and others 2023).

Quantitative Evaluation

Our quantitative evaluation process, inspired by (Wang and others 2023), is depicted in Figure 2.

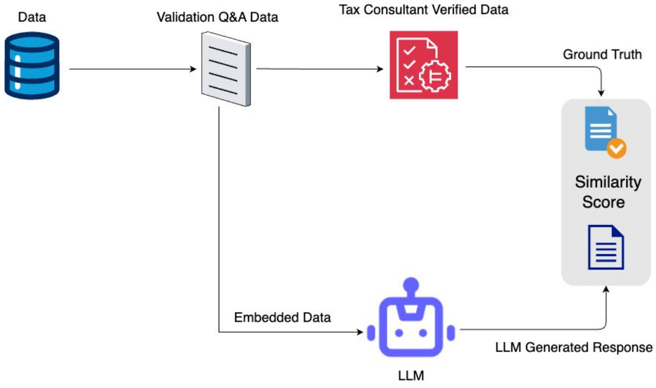

Figure 2: Quantitative Evaluation flow

Data Extraction: We extracted random sections of text from our dataset which comprised of income tax-related PDFs to construct Test examples using two different approaches to evaluate the performance of our `RAG` System.

Constructing Test Set (Manual Curation): With extracted sections in hand, we manually devised 75 specific questions and their corresponding answers. These question-answer pairs form our `Validation Q&A Dataset`.

Constructing Test Set (using RAGAs synthetic test data generation): We further expanded our Test Set i.e. `Validation Q&A Dataset` by employing the `Retrieval Augmented Generation Assessment (RAGAs)` framework's synthetic test data generation pipeline to generate 105 question-answer pairs from sections extracted from our dataset (Es and others 2023). These new pairs were then added to `Validation Q&A Dataset`.

Verification by Tax Consultant: The established question-answer `Validation Q&A Dataset` then went through meticulous verification by three trusted tax consultants, ensuring accuracy and compliance with established tax standards. Once verified, the `Validation Q&A Dataset` became ground-truths that we could evaluate our system against.

Incorporation into LLM Prototype: Questions from the `Verified Q&A Dataset` were then fed to our RAG system, leveraging the vector database (`Faiss vectorstore`) to generate answers for comparison.

BERT Similarity Score Calculation: Inspired by (Li and others 2023), we calculate the `BERT` (Bidirectional Encoder Representations from Transformers) similarity score between the ground-truths and those generated by the RAG system, quantifying the similarity. The results are shown in table 1.

To further evaluate our application-generated answers, we employed the RAGAS framework once again. This framework assesses the performance of the LLMs based on three metrics: context relevancy, faithfulness, and answer relevancy. These metrics quantify the similarity between the question and context, the context and answer, and the similarity between question and answer. The `RAGAs` framework is intended to be an open standard for the evaluation of RAG pipelines without the need to obtain ground truths. The framework utilizes OpenAI `GPT-4` to assign a score between 0 and 1 (inclusive) for each of the three metrics stated above. The final result `RAGAs` Score is obtained by taking the average of these assigned scores. The utilization of `ChatGPT4` and other language models for evaluation has been explored in (Wang et al. 2023). These results are shown in Table 1.

The high RAGAS score shown in Table 1 indicates the excellent performance of our RAG system. When compared to other applications involving the retrieval and generation of sensitive information through the RAG pipeline, such as the retrieval of pregnancy-related guidelines presented in (Al Ghadban et al. 2023), our results are particularly promising. Despite significant experimentation, (Al Ghadban et al. 2023) achieved only a RAGAS score in the range of $(0.1, 0.2]$, whereas we attained an impressive score of

Table 1: Quantitative Evaluation Results. We consider BERT score and RAGAS score ranging between 0 and 1.

Model	BERT Score	RAGAS			
		Context Relevancy	Faithfulness	Answer Relevancy	RAGAS Score
GPT 3.5 Turbo	0.89	0.94	0.82	0.94	0.90

0.90. Furthermore, a faithfulness score of 0.82 emphasizes the model adheres to the context retrieved and does not hallucinate.

Qualitative Evaluation

To assess our model's proficiency, we conducted a thorough qualitative evaluation using a novel CUUO framework, inspired by (Zheng and others 2023) and (Wang and others 2023), focusing on **C**oherence, **U**sability, **U**nsafety, and **O**verall performance.

Fifteen participants, all with backgrounds in taxation, evaluated the model's responses. Each engaged in a tax-related conversation with the chatbot, spanning at least 5 turns. Participants then rated the system's performance, choosing to continue or end the conversation. They responded to questions about coherence, usability, and overall performance.

1. Coherence: Were the responses coherent and accurate?
2. Usability: How easy was it to interact with TaxTajweez's chatbot interface?
3. Unsafety: At any instance, did the conversation contain unsafe content?
4. Overall Performance: Are you satisfied with the Taxation LLM chatbot's performance in addressing your income tax-related inquiries and needs?

Table 2: Combined Ratings

Rating	Coherence	Usability
	Number of Respondents	
5	9	5
4	5	7
3	1	3
2	0	0
1	0	0
Average	4.53	4.13

Based on survey evaluations, users provided insightful qualitative assessments. Coherence and usability were rated on a Likert scale (1 to 5), with results shown in Table 2. Most ratings leaned towards higher values, indicating reliability and ease of use.

Concerning safety, users reported no instances of offensive language, hate speech, or misleading opinions (Table 3), demonstrating satisfactory content moderation.

Table 3: Unsafety Responses

Response	Number of Respondents
Safe	15
Offensive Language	0
Hate Speech	0
Misleading Opinions	0

Table 4: Overall Performance Responses

Response	Number of Respondents
Satisfied	15
Unsatisfied	0

For overall performance, users expressed satisfaction (Table 4) with TaxTajweez's ability to handle income tax inquiries effectively. All 15 respondents reported satisfaction, indicating the chatbot's effectiveness in meeting users' needs.

Future Work

Our immediate next goal is to compare the RAG system presented in this paper to fine-tuned models on the same dataset. We also plan to explore the changes in the evaluation scores (RAGAS & BERT) in the following scenarios:

1. When we use different similarity quantifying algorithms to retrieve relevant context from the vector store.
2. When we use different LLMs to generate a response to the current query using the context retrieved.

Additionally, we plan to explore a combined approach that leverages both the RAG pipeline and a fine-tuned model to provide relevant, coherent, and accurate responses to tax-related queries.

Conclusion

In this paper, we propose TaxTajweez, a RAG-based Chatbot for income tax advisory. Additionally, we use a comprehensive evaluation mechanism incorporating quantitative frameworks as well as human evaluation to assess the effectiveness and accuracy of TaxTajweez. Our results indicate that TaxTajweez not only demonstrates robust performance in retrieving income tax information but also aligns well with user expectations showcasing the potential of large language models in enhancing efficiency and accuracy within this domain. These are valuable insights into the ongoing discourse on leveraging advanced language models for practical applications.

References

Al Ghadban, Y.; Lu, H. Y.; Adavi, U.; Sharma, A.; Gara, S.; Das, N.; Kumar, B.; John, R.; Devarsetty, P.; and Hirst, J. E. 2023. Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*.

Alarie, B.; Condon, K.; Massey, S.; and Yan, C. 2023. The rise of generative ai for tax research. *Social Science Research Network*. accessed Nov. 25, 2023.

Amjad, M. 2021. Tax policy in pakistan: Concerns and reforms.

Baladón, A.; Sastre, I.; Chiruzzo, L.; and Rosá, A. 2023. Retuyt-inco at bea 2023 shared task: Tuning open-source llms for generating teacher responses. *ACLWeb*. accessed Nov. 25, 2023.

Cui, J.; Li, Z.; Yang, Y.; Chen, B.; and Liu, Y. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv (Cornell University)*.

Dan, Y., et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv.org*. accessed Nov. 17, 2023.

Es, S., et al. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217 (2023)*.

LangChain Contributors. 2023. LangChain Documentation. `https://api.python.langchain.com/en/latest/chains/langchain.chains.conversational_retrieval.base.ConversationalRetrievalChain.html`.

LangChain. 2024. Facebook ai similarity search (faiss) - python documentation.

Levonian, Z., et al. 2023. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *Journal Name*. Accessed: Nov. 25, 2023.

Li, Y., et al. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv.org*.

More, A., and Brin, M. 2023. Semantic-Split: A Python library to chunk/group your text based on semantic similarity. `https://github.com/agamm/semantic-split/tree/main`.

OpenAI. 2021. Chatgpt: A large-scale generative model for open-domain chat. `https://github.com/openai/gpt-3`.

OpenAI. 2024. What are embeddings? - openai documentation.

Reuters, T. 2023. Chatgpt and generative ai in legal, corporate & tax markets. Accessed: Nov. 25, 2023.

Singhal, K., et al. 2023. Towards expert-level medical question answering with large language models. Submitted on 16 May 2023.

Van Doan, T.; Truong, Q.-T.; Nguyen, D.-V.; Nguyen, V.-T.; and Luu, T. N. 2023. Efficient finetuning large language models for vietnamese chatbot. *arXiv (Cornell University)*.

Wang, H., et al. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv.org*.

Wang, C.; Cheng, S.; Guo, Q.; Yue, Y.; Ding, B.; Xu, Z.; Wang, Y.; Hu, X.; Zhang, Z.; and Zhang, Y. 2023. Evaluating open-qa evaluation. Submitted on 23 Oct 2023, Affiliations: School of Engineering, Westlake University, China; Northeastern University, China; Amazon AWS AI; Fudan University, China.

Xiong, H., et al. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv (Cornell University)*.

Zhang, Y., et al. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv.org*. accessed Nov. 25, 2023.

Zheng, Z., et al. 2023. Building emotional support chatbots in the era of llms. *arXiv.org*.

Zheng, C.; Sabour, S.; Wen, J.; Zhang, Z.; and Huang, M. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. *arXiv.org*. accessed Nov. 25, 2023.