Beyond Extraction: Contextualising Tabular Data for Efficient Summarisation by Language Models

Anonymous ACL submission

Abstract

The conventional use of the Retrieval-Augmented Generation (RAG) architecture has proven effective for retrieving information from diverse documents. However, challenges arise in handling complex table queries, especially within PDF documents containing intricate tabular structures. Our work introduces an innovative approach to enhance the accuracy of complex table queries in RAG-based systems. Our methodology involves storing PDFs in the retrieval database and extracting tabular content separately. The extracted tables undergo a process of context enrichment, concatenating headers with corresponding values. Furthermore, we enhance the tabular data with contextual understanding using the GPT-3.5-turbo through a one-shot prompt. This enriched data is then added to the retrieval database alongside other PDFs. Our approach aims to significantly improve the accuracy of complex table queries, offering a solution to a longstanding challenge in information retrieval.

1 Introduction

In the era of information retrieval and chat-bot, the Retrieval-Augmented Generation (Lewis et al., 2020) architecture stands as a robust framework for retrieving information from documents and interact with it. However, its effectiveness faces a substantial challenge with complex table queries, particularly within PDF documents housing intricate tabular structures. This ongoing issue has prompted the development of a new approach aimed at improving the accuracy of complex table queries within RAG systems. Some previous studies have been done, but no significant improvements have been made, especially when it comes to modifying the extracted data. This highlights the need for innovative solutions to address the limitations observed in existing methodologies.

Our approach begins by addressing the inherent limitations of RAG when dealing with tabular content. Instead of relying solely on textual retrieval, we advocate for a two-fold strategy. Firstly, the extracted data from PDF documents is stored in the retrieval Vector database, ensuring a comprehensive repository of the original data. Secondly, an extraction process is implemented to separately extracted and enrich tabular content. First we extracted the tabular content using Camelot library (Mehta, 2019). Then, the enrichment process involves combining the headers and their corresponding values within the tables of PDF documents. This concatenation ensures that the context within complex rows is preserved, creating a more cohesive representation of the tabular content. By linking headers and rephrasing it using GPT-3.5-turbo alongside a one-shot prompt, the augmented information becomes more structured and interpretable, allowing for improved understanding and accuracy in responding to complex table queries within the Retrieval-Augmented Generation architecture.

GPT-3.5-turbo, an advancement over GPT-3 (Brown et al., 2020) by OpenAI, offers improved efficiency, faster responses, and enhanced contextual understanding. It is optimized for interactive applications like chatbots and customer service tools. These enhancements make it more suitable for real-time conversational AI tasks.

For the summarisation part in RAG architecture, we integrate a fine-tuned (Dai and Le, 2015) version of the Llama-2-7B-base (Touvron et al., 2023), a large language model, specifically tailored for summarisation. This adaptation allows for an effective summarisation of the retrived content from the database.The fine-tuned Llama-2-7B-base model for summarisation ensures a specialised capability in distilling key information.

The augmented data, now possessing a refined contextual understanding, is seamlessly integrated into the retrieval database alongside the extracted data from the PDFs. This approach aims to significantly enhance the accuracy of complex table queries, addressing a long-standing gap in information retrieval methodologies. As we explore our methodology, this paper unfolds the layers of innovation driving a shift in the domain of document-based information retrieval.

2 Methodology

2.1 Model Used in RAG Architecture

Our framework uses a powerful approach called Retrieval-Augmented Generation (RAG) to find and understand information. We combine this with a state-of-the-art language model from Meta-AI called Llama-2-7B-base, which excels at summarizing information from various sources. Llama-2-7B-base is great at condensing large amounts of text into clear and concise summaries. This makes it a perfect fit for our RAG framework. During text generation, Llama-2-7B-base improves the framework's contextual understanding by selecting the most relevant information for summarizing from the knowledge list. The Knowledge list is retrieved from the database using cosine similarity (Singhal, 2001) between the query and extracted content stored in the database. This retrieved knowledge supplies important facts and language cues to the generative model, helping it create summaries that are both concise and accurate

Our methodology undertakes multi-step training of Llama-2-7B-base on large datasets to enable it to develop a comprehensive understanding of summarization across diverse topics and styles. The model is fine-tuned using supervised learning techniques that leverage human-written summaries as targets. This helps Llama-2-7B-base build advanced capabilities for identifying and connecting key information from retrieved knowledge while generating summaries reflecting the essence of source texts and the query from user.

2.2 Dataset

We built our research on a carefully selected dataset of policy documents and research papers from trusted sources, specifically chosen for their inclusion of both textual and tabular data. This dataset reflects the challenges of real-world information retrieval and covers a wide range of policy areas. To test how well our Retrieval-Augmented Generation (RAG) system works with this data, we designed 200 specific questions. These questions challenge the system to find information in both normal text and structured tables. We made sure to include a mix of both text-based and table-based questions to thoroughly evaluate the system's abilities

3 Experiment Setup

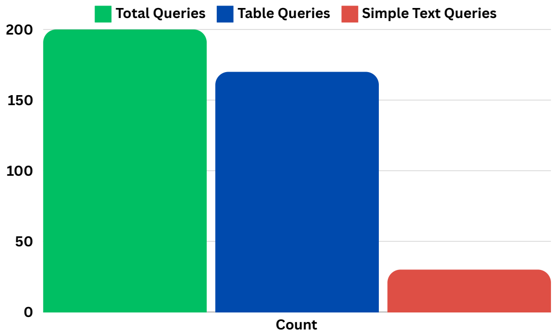

Figure 1: Visualizes query distribution in the experimental dataset: 170 table queries and 30 simple text queries out of a total of 200. Offers a balanced evaluation of Retrieval-Augmented Generation (RAG) architecture for both textual and tabular dimensions

3.1 Query Processing

Our experimental design included 200 diverse queries to reflect the complexity of real-world information retrieval. This set comprised 170 queries focusing on tabular data. Within this, we divided the queries further: 110 complex queries tested the architecture's ability to understand intricate relationships and patterns in tables, while 60 simpler queries assessed its handling of basic table structures. Additionally, 30 control benchmarks focused on non-tabular text queries, providing a solid baseline for evaluating the architecture's understanding of unstructured text. This comprehensive set of queries ensured a thorough evaluation of the architecture's capabilities across various information retrieval tasks.

Figure 2 shows how a question about the MPT model is effectively answered, even though the data is embedded in a complex table structure

3.2 Data Preparation

A crucial part of our methodology involves careful data preparation. Initially, we archived PDF documents in a retrieval database, creating a large pool of raw data. Next, we focused on a two-step strategy to extract tabular content using the Camelot library, resulting in a collection of complex tables. To enhance context, we combined column headings with their corresponding row values, giving the extracted tables more meaningful context.

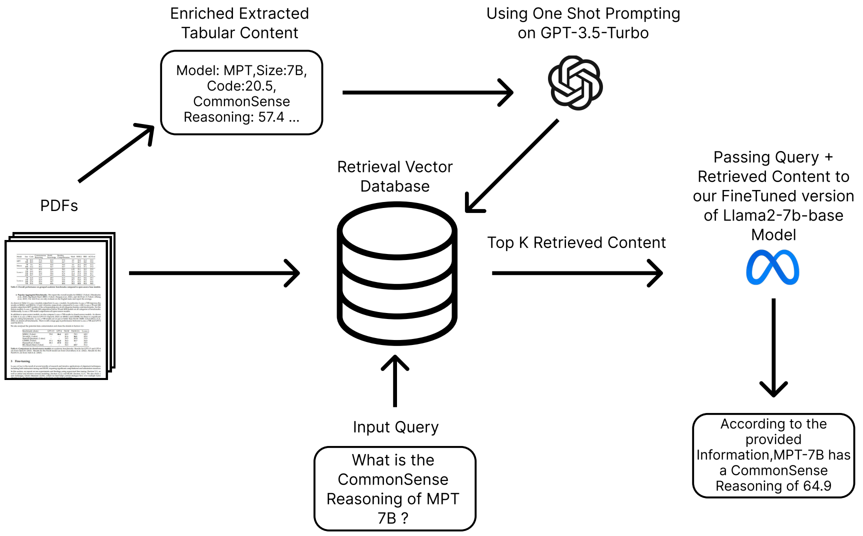
Enriched Extracted Tabular Content
Model: MPT,Size:7B, Code:20.5, CommonSense Reasoning: 57.4 ...
Using One Shot Prompting on GPT-3.5-Turbo
PDFs
Retrieval Vector Database
Top K Retrieved Content
Passing Query + Retrieved Content to our FineTuned version of Llama2-7b-base Model
Input Query
What is the CommonSense Reasoning of MPT 7B ?
According to the provided Information,MPT-7B has a CommonSense Reasoning of 64.9

Figure 2: Figure Represents presents the architectural diagram illustrating the experimental setup. The architecture is designed to demonstrate the workflow of our approach, showcasing the key components involved in enhancing the accuracy of complex table queries within the Retrieval-Augmented Generation (RAG) framework

3.3 Integration with GPT-3.5-turbo

In our approach, integrating advanced language models marks a significant innovation. The Llama-2-7B-base model, a core part of the RAG architecture, was crucial for summarization tasks, providing a refined understanding of tabular data. To further enhance this, we used the GPT-3.5-turbo API for additional context enrichment. A one-shot prompt refined the tabular data, making it more suitable for our summarization model. This enriched data was then stored in our retrieval database along with the content extracted from original PDFs using Camelot. This strategy was carefully designed to give our summarization model a deeper understanding, improving accuracy in information retrieval.

4 Results

This section presents the results of our experiments, where we compared three different methods for summarizing policy documents and research papers. We used a dataset of 200 questions to see how well each method could find relevant information in both text and tables. Table 1 shows the accuracy of each method for retrieving information from different parts of the documents.

4.1 Normal Existing Pipeline

The baseline pipeline extracts text from documents and feeds it into a retrieval database, achieving an accuracy of 86.6 for text-based queries. However, its performance is 48.2% for table-related queries, highlighting the limitations of text-only methods in understanding complex table data. Overall, the baseline approach has an average accuracy of 54%, showing the challenges of integrating tabular data into the summarization process. We used the Camelot library for table extraction, but despite its advanced features, the complexity of the tables still led to lower performance in table-related queries.

4.2 Table Extracting Separately and Context Enrichment

Our second approach focuses on improving how the system understands tables. We added a step that first extracts the tabular content from PDF using camelot and then combines table headers with their corresponding row values, providing more context. This enhanced method keeps the same high accuracy (86.6%) for text-based questions as our initial approach. More importantly, it significantly boosts the accuracy for table-based questions to 54.1%, leading to an overall accuracy of 59.4%. This jump clearly demonstrates the value of our strategy in tackling the complexities of retrieving information from tables.

4.3 Parsing Extracted Text to GPT-3.5-turbo API

The highlight of our methodology is the integration of the GPT-3.5-turbo API for parsing extracted tabular enriched text, resulting in a significant performance boost. Accuracy for text queries jumped to 93.3%, showing our approach's effectiveness in

Methodology	Simple Text Queries Accuracy (%)	Table Queries Accuracy (%)	Overall Accuracy (%)
Normal Existing Pipeline	86.6	48.2	54
Table Extracting separately & Context Enrichment	86.6	54.1	59.4
Parsing Enriched Extracted Text to GPT-3.5-turbo	93.3	61.1	66

Table 1: Summarizes experiment outcomes, evaluating three methodologies for information retrieval accuracy. Improved metrics observed, especially in handling complex table queries

dealing with unformatted text. This improvement highlights the robustness of our method in retrieving relevant information from less structured data. At the same time, accuracy for table queries increased to 61.1%, marking a major step forward in handling complex table queries. Combining strong text query accuracy and improved table query performance led to an overall accuracy of 66%. This achievement demonstrates a significant advance in our method's ability to extract valuable insights from both unstructured text and complex tables.

5 Conclusions

This study explores the effectiveness of the Retrieval-Augmented Generation (RAG) architecture in handling complex table queries in PDFs. Our approach combines model selection, dataset curation, and experimental design to retrieve and comprehend information from detailed tabular structures. Our findings show that traditional pipelines struggle with complex tables, but our method, which includes separate extraction and context enrichment for tabular data, significantly improves accuracy.

By integrating GPT-3.5-turbo for enriching the extracted tabular data, within the RAG framework, we achieve high levels of accuracy for complex table queries. This research not only enhances accuracy but also lays the groundwork for future advancements in information retrieval. Our approach demonstrates the potential of context-aware language models and RAG architectures in bridging the gap between human cognition and complex data structures.

6 Limitations

Despite advancements, our approach faces challenges with extremely intricate or non-standard table formats, such as multi-level headers and merged cells, which Camelot struggles to parse accurately. The integration of advanced language models like GPT-3.5-turbo and Llama-2-7B-base introduces substantial computational overhead, limiting scalability. The accuracy heavily relies on high-fidelity data extraction, and any errors can propagate through the pipeline. Additionally, the models may fall short in scenarios requiring deep domain-specific knowledge. GPT-3.5-turbo may not consistently convert extracted enriched text into meaningful sentences, especially when tables span multiple pages and lose header context. Our method's generalizability across different domains remains untested, and the multi-step process introduces latency, making real-time application challenging.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Informa-*

tion Processing Systems, volume 28. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Vinayak Mehta. 2019. Camelot: Pdf table extraction for humans.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.