International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Abstractive Long Text Summarization using Large Language Models

[1]Gunjan Keswani, [2]Wani Bisen, [3]Hirkani Padwad, [4]Yash Wankhedkar, [5]Sudhanshu Pandey, [6]Ayushi Soni

Submitted: 19/11/2023 **Revised:** 30/12/2023 **Accepted:** 11/01/2024

Abstract: Large Language Models (LLMs) have made significant strides in processing human-written texts. However, a major challenge persists - the retention of context over extensive texts or multiple documents. The current approach of LLMs to retain context is often inefficient, both in terms of storage and time. To address this issue, this paper proposes a novel approach for two key tasks - Summarization and Question Answering. The methodology ensures that the LLM is not overwhelmed with unrelated, repetitive, or redundant data, thereby saving considerable time and resources. This approach facilitates the generation of effective summaries and answers for the user, enhancing the overall performance and efficiency of the LLM.

Keywords: *Abstractive summarization; Large Language Models; LangChain; Natural Language Processing; Retrieval-Augmented Generation*

1. Introduction

Recent research has shown that in widely used summary datasets, human annotators prefer LLM generated summaries over the original reference summaries.The goal of text summarization, a Natural Language Processing (NLP) technique, is to extract the most important details and contextual meaning from a given input while minimizing the volume of text. It comes as no surprise that NLP-based automatic summarization has found widespread application in diverse scenarios, spanning a range of document lengths, owing to the significant time and resources required for manualsummarization. Thanks to recent breakthroughs in LLMs, abstractive summarization is poised to become even faster and more efficient than it was with earlier transformer models. This study delves into abstractive summarization of lengthy texts using LLMs. We define the functionality of LLMs in the context of abstractive summarization for long texts, with a specific focus on the Llama2 model, aiming to produce precise and con- textually relevant summaries. Section III provides a comprehensive exposition on the technologies utilizedin the learning process. Section IV contains a detailed methodology for the development of the final model. Section V is dedicated to discussing the results and conducting a comparative analysis with existing liter- ature. Finally, Section VI presents the conclusions and proposes potential avenues for future research.

[1]*Department of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: keswanigv@rknec.edu
[2]*Dept. of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: bisenwh@rknec.edu
[3]*Dept. of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: padwadhs@rknec.edu
[4]*Dept. of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: wankhedkaryn@rknec.edu
[5]*Dept. of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: pandeysv@rknec.edu
[6]*Dept. of Computer Science and Engineering Shri Ramdeobaba College of Engineering and Management (RCOEM) Nagpur, India*
e-mail: sonia@rknec.edu

2. Related Work

In the realm of abstractive long text summarization, several prominent transformer-based models have played a pivotal role in advancing the field. These models, including BERT, GPT-2, and XLNet, each bring their distinctive architectural strengths to the table. [1] Notably, Google's BERT (Bidirectional En- coder Representation from Transformers) has proven to be a versatile choice for various NLP tasks. BERToperates on the foundation of the Masked Language Model (MLM) technique, transforming text into word embeddings based on statistically derived similarity measures.[2]

To assess the quality of summaries generated by theBERT model, researchers often employ a set of standard evaluation metrics known as ROGUE, comparing the output with human-generated summaries. In contrast, BART (Bidirectional Auto-Regressive Transformer) has gained recognition for surpassing BERT in performance. BART introduces a novel pre-training method and architecture that enables it to functionas a sequence-to-sequence model (seq2seq model) for diverse NLP tasks. [3]

The Generative Pre-trained Transformer (GPT) family, initiated by OpenAI in 2018, has made remarkable strides in generating realistic text. Among its various iterations, including GPT-3 and GPT-3.5, the original GPT model was trained with 117 million parameters, aiming to comprehend language in a manner akin to human understanding. GPT-2, equipped with even more training parameters, was introduced to address the limitations of its predecessor. In 2023, the latest iteration, GPT-4, emerged, further building upon the strengths of GPT-3 and delivering improved performance.

3. Background

A. Vector database

Vector databases, a game-changing development in data storage and retrieval, have transformed how we manage complex data kinds including photos, audio, and textual embeddings. Unlike standard relational databases, which rely on organized data andpreset schemas, vector databases describe and index data points using the power of high-dimensional vector spaces.[4] This enables quick similarity searches and content-based queries, allowing applications such as recommendation systems, picture recognition, and natural language processing to improve accuracy and speed.

B. Llama2

LLMs are transformer-based neural networks that have been trained on vast amounts of data. Transformer models are specifically designed for tasks related to text completion and understanding.

Meta has introduced an open-source LLM called Llama 2, which includes two main variants:

Llama 2 - This is an updated version of the earlier Llama 1 model. It has been trained on a new combination of publicly available data. Notable enhancements include a 40% increase in the size of the pre-training corpus, a doubling of the model's context length. Llama 2 is available in several parameter sizes, including 7 billion (7B), 13 billion (13B), and an astonishing 70 billion(70B) parameters.

Llama 2-Chat - This is a fine-tuned version of Llama 2 that has been optimized for dialogue-related use cases. Similar to Llama 2, Llama 2-Chat is available in various parameter sizes, including 7B, 13B, and 70B.

These models have been made available to both theresearch community and commercial users for various applications. [5]

C. LangChain

LangChain is a versatile framework designed for creating applications empowered by language models. It empowers applications to:

1) Embrace Context-Awareness

LangChain offers a seamless integration of languagemodels with diverse sources of context through the useof modules like "chains" and "vector stores." These sources may include prompt instructions, few-shot examples, or relevant content, enabling the application to produce responses grounded in the given context

2) Facilitate Reasoning

Applications developed with LangChain harnessthe power of language models for various reasoning tasks. This encompasses utilizing the language model to make well-informed decisions regarding how to respond within the given context and to figure out the most suitable actions to undertake. These reasoning processes are facilitated through the use of modules such as "agents" within the LangChain framework. [6]

D. Retriever Answer Generator Model

The Retriever Answer Generator (RAG) Model merges retrieval-based and generative technologies for machine reading comprehension. It is widely used for development of all types of chatbots. It employs a twostep process, the first of which is retrieving relevant context from the dataset relating to the question that is asked. Once retrieved, the sources are passed to theLLM to generate the answer. This two step process al- lows the LLM to generate accurate answers, with rich detail and very little noise. [7]

4. Methodology

The project addresses two of the most complex aspects of NLP: Summarization and Question- Answering. Initially, the summarizer is employed to generate effective summaries of the input text. Subsequently, the Chatbot module is utilized for answering queries related to the input. Both these components are designed to handle a wide range of text sizes, fromsingle to multiple PDFs. This design ensures that the workload on the LLM is minimized, enabling it to operate effectively and efficiently.

A. Pre-processing

1) Chunking

The first step entails breaking down the data into smaller, more manageable segments, with each seg- ment limited to a maximum size of 500 characters and an overlap of 200 characters. Chunk overlap ensures that adjacent chunks share some common character preventing potential information loss at the boundary between the chunks.

This segmentation serves two purposes: it simplifies the handling of lengthy texts by splitting them into smaller, more easily processed parts, and it enables us to work within the token limitation of Llama2, which has a maximum token limit of 4096.

2) Knowledge base

A knowledge base comprises documents covering a wide range of subjects, products, or services that offer crucial context to support the functioning of a Language Model. It's important to highlight that LLMs may occasionally produce responses that are inaccurate or diverge from the intended context, often referred to as hallucinations. To ensure accurate responses, LLMs rely on context, which, in our scenario, is provided by the vector representation of the input. [8]

Embeddings, at their core, are numerical representations of text designed to capture its semantic meaning. In our process, all the chunks are transformed into embeddings. We employ the gpt4all embedding model, which generates embeddings with 384 dimensions. These resulting vectors are subsequently stored in the Qdrant vector database for future retrieval and application.

Fig 1: *A look at the clusters in 3D space*

B. Module 01: Summarization

In the field of Abstractive Long Text Summarization, one of the main hurdles is maintaining context throughout the text. For a summary to be truly abstractive, it's essential for the machine to grasp the complete text in its entirety. Yet, it is a common occurrence in literature for authors to initially diverge from the central theme and subsequently return to it later within the text. To effectively summarize these later parts, the machine needs to keep the embeddings of the earlier parts stored in memory. This poses a challenge as it either requires a significant amount of memory or a considerable amount of computational time, both of which are not always feasible in today's fast-paced technological landscape.

To address this issue, a unique approach has been developed. Here's a detailed breakdown of the methodology:

1) Clustering

The first step in summarizing the documents involves creating clusters. The cosine similarities between the vectors generated in the previous step is calculated and K-Means clustering is carried out to create clusters. K-Means is used as it is a centroid based clustering algorithm hence; it segregates objects into clusters based on their similarities, ensuring that objects within the same cluster are more alike while being dissimilar to objects in separate clusters. Selecting the appropriate number of clusters is vital to balance efficiency and result quality. It's essential to consider the token limit of Llama2-chat, which is 4096 tokens. To determine the optimal number of clusters, the elbow method is employed using the scikit-learn library. This technique involves varying the number of clusters, calculating the Within-Cluster Sum of Squares (WCSS), and identifying the "elbow" point on the plot, which represents the ideal cluster count. This approach ensures a meaningful data structure. The cluster count should fall within the range of 3 to 15, with 15 clusters being the maximum due to token limitations in LLM that prevent it from providing a summary beyond this threshold.[9] Figure 2 illustrates the plot of Within- Cluster Sum of Squares (WCSS) and the number of clusters. This plot was generated during the summarization of a research paper titled "*Neuralangelo: High-Fidelity Neural Surface Reconstruction.*"

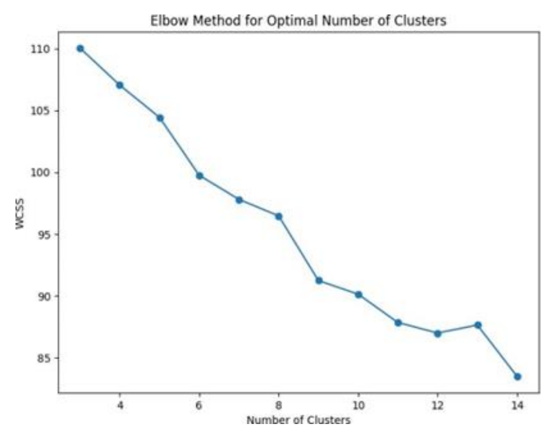

Fig 2: *WCSS vs Number of clusters plot*

2) Cluster Summaries

Within each cluster, every individual chunk shares a comparable semantic meaning, as evidenced by their vector representations being in close proximity. From each cluster, a single representative chunk is chosen, usually the one closest to the mean of the vectors, andthen it is inputted into the LLM. This chosen representative chunk effectively encapsulates the semantic essence of the entire cluster. For this experiment, a Llama 2 - 13b model is used. The LLM is tasked with summarizing each representative chunk. Aftercompleting this step, there are multiple chunk sum- maries, each representing a cluster. When combined, these summaries provide an overview of the entire document.

3) *Overall Summary*

Finally, the chunk summaries are now fed into the LLM for integration and further condensation, ensuring comprehensive coverage in a concise manner. This method allows us to create top-notch abstractive sum- maries, making it a practical solution for summarizing texts of varying lengths with comprehensive coverage.

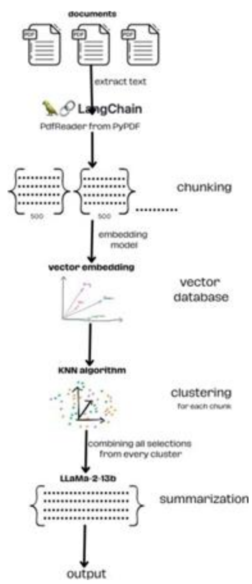

Fig 3: *Flowchart of Summarization Model*

C. *Module 02: Question-Answering*

Question-Answering (QA) stands as a significant subfield within NLP, with a primary focus on empowering machines to comprehend human questions and generate pertinent responses. Recent advancements in Natural Language Processing, such as Generative Pre- trained Transformer and Large Language Models, have greatly simplified this task, particularly for general use cases. [10]

A notable challenge in QA is that the answer toa question can be dispersed across various parts of the input text. The capacity to identify, grasp, and amalgamate these scattered fragments of information is pivotal for producing high-quality answers.

In this approach, the Vector Database Qdrant is leveraged, utilizing its cosine similarity search functionality to supply relevant information to the LLM from multiple documents, all based on the presented question. This methodology contributes to the development of a Retrieval-Augmented Generation (RAG) model, enhancing the effectiveness of question-answering.

D. *Generating the Answer*

In this phase, when the user poses a question, we first convert the question into a word embedding, as explained earlier. Next, cosine similarity is employed to search the Vector Database for chunks most closely related to the input question embedding. Cosine similarity quantifies the similarity between two non-zero vectors situated within an inner product space. It is computed as the cosine of the angle formed by these vectors, calculated through the dot product of the vectors divided by the product of their magnitudes. Crucially, cosine similarity is agnostic to the absolute magnitude of the vectors, focusing exclusively on their angular orientation [11]. This yields a similarity score ranging from -1 to 1: unit vectors or those pointing in the same direction achieve a similarity of 1, orthogonalvectors yield a similarity of 0, and vectors pointing in opposite directions result in a similarity of -1.

The four closest chunks are subsequently selected and provided to the LLM as context, alongside the query question. The LLM uses this context to generate a concise and relevant answer extracted from thesechunks. The choice to retrieve only four chunks is based on the desire for concise responses and balancebetween precision and speed in the question-answering system.

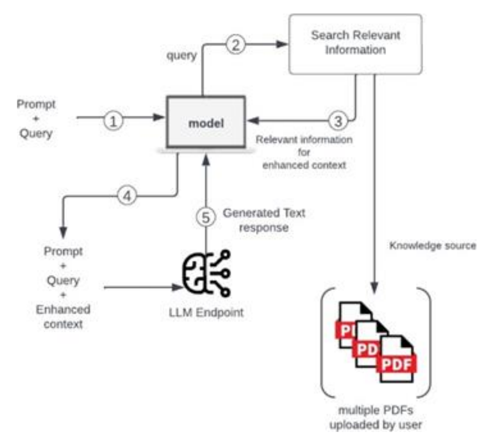

Fig 4: *Flow of QA module*

5. Result and Discussion

E. *RAG Evaluation*

A Retrieval-Augmented Generation (RAG) implementation comprises two essential components: Retrieval and Generation. The Retrieval process establishes the context, while the Generation process is executed by a LLM to produce the answer by utilizing the retrieved information. [11]

When assessing a Retrieval-Augmented Generation pipeline, it's crucial to evaluate both of these components separately and in conjunction to obtain an over- all score, as well as individual scores to pinpoint areas for improvement.

Ragas is a tool that employs Language Modelsto evaluate Retrieval-Augmented Generation (RAG) pipelines, providing actionable metrics with minimal reliance on annotated data.

1. *Retrieval:* Ragas assesses two key metrics – 'context relevancy' and 'context recall', which measure the performance of the retrieval system.

2. *Generation:* Ragas evaluates 'faithfulness', which assesses the consistency of information in the generated answer with respect to the provided context, and 'answer relevancy', which gauges how well the answeraligns with the relevance to the question.

a. *Faithfulness:* This metric measure the degree of information consistency in the generated answer compared to the provided context. It penalizes any claims made in the answer that cannot be inferred from the context. It is calculated using the 'answer' and 'retrieved context'.

b. *Context Precision:* This metric evaluates how relevant the retrieved context is to the question. Ideally, the context should contain only the information necessary to answer the question, and the presence of redundant information is penalized. It is calculated from the 'question' and 'retrieved context'.

c. *Context Recall:* Ragas measures the recall of the retrieved context using the annotated answer as a proxy for the ground truth context. It is calculated based on the 'ground truth' and 'retrieved context'.

d. *Answer Relevancy:* This metric quantifies the extent to which a response directly addresses and is appropriate for a given question or context. It doesn't consider the factual accuracy of the answer but penalizes redundant or incomplete answers in the context of the question. It is calculated from the 'question' and 'answer'.

These four aspects serve as a comprehensive measure of the QA system's performance, taking into account all the critical aspects. The testing examples for evaluation (Figure 5) were drawn from the RAGAS baseline experiments. In this case, the context from the dataset was passed as context for the pipeline, and the answers were subsequently evaluated after passing the questions. Figure 6 shows two PDFs passed asinput into the RAG pipeline along with a question, yielding their respective answers. Figure 7 illustrates the retrieved chunks that served as references for the pipeline while answering the question.

Metric	Score
context relevancy	0.4201680672268907
faithfulness	0.9999999999999998
answer relevancy	0.9305603597765211
context recall	0.5555555555555556

Table 1: *RAGAS Scores*

F. Summarizer Evaluation

The ROUGE score is a collection of metrics often employed to assess text summarization tasks, which involve the automatic creation of a brief summary for a longer text. ROUGE has been developed to gauge the effectiveness of machine-generated summaries by comparing them to reference summaries or documents. It evaluates the resemblance between a machine-generated summary and reference summariesby examining the shared word sequences, known asn-grams, within them. These n-grams, typically unigrams, bigrams, and trigrams, are compared to compute the recall in the machine-generated summary concerning the reference summaries. [12] [13] TheROUGE score is calculated using the formula:

$$\text{ROUGE} = \sum (\text{Recall of n-grams})$$

ROUGE scores are branched into ROUGE-N, ROUGE-L, and ROUGE-S.

a. *ROUGE-N:* ROUGE-N measures the overlap of n-grams (contiguous sequences of n words) between the candidate text and the reference text.

b. *ROUGE-L:* ROUGE-L measures the longest common subsequence (LCS) between the candidate text and the reference text.

c. *ROUGE-S:* ROUGE-S measures the skip-bigram (bi-gram with at most one intervening word) overlap between the candidate text and the reference text.

The ROGUE scores were computed by using both the final summary and the chunk-wise summaries of a research paper titled *"The Concept of EV's IntelligentIntegrated Station and Its Energy Flow"* as shown in Figure 8 and Figure 9 respectively. These scores offer a quantitative assessment, gauging the quality and relevance of the generated summaries (Final summary) concerning the source material (The chunk-wise summaries).

Metric	Score
Rouge1	0.5830508474576271
Rouge 2	0.42320819112627983
RougeL	0.4893617021276596
RougeLsum	0.4893617021276596

Table 2: *ROUGE Scores by Huggingface ROGUE Metrics*

6. Conclusion

In this research, a novel approach is presented, utilizing large language models (LLMs), vector similarity search engines, and clustering algorithms to abstractively summarize lengthy and complex PDF documents. Incorporating advanced technologies such as Llama2 and Langchain in conjunction with clustering algorithms like K-means and K-nearest neighbors (KNN) guarantees that the resulting summary is comprehensive and finely attuned to the document’s con- text. This innovative method has been demonstrated to generate exceptional abstractive summaries for extensive texts. Furthermore, it not only provides an engaging means for users to interact with the PDF document by utilizing everyday language in questions and responses but also extends this capability to thesimultaneous exploration of multiple papers. This approach enhances accessibility and understanding for a broader audience while simplifying the exploration of complex content, facilitating the seamless navigation and comprehension of multiple documents. Looking ahead, the future of LLM-based summarization promises several notable advancements. Foremost, domain-specific summarization will take center stage, with LLMs being finely tuned to serve specific industries such as medicine, law, and science. This will result in more precise and contextually tailored summaries, addressing the unique requirements of these specialized fields. Additionally, there will be a strong emphasis on advancing the capabilities of LLMs to deliver high-quality multilingual summarization. This will facilitate effective cross-language communication and support global applications that demand summarization services in multiple languages. Furthermore, the development of interactive summarization systems is anticipated. These systems will empower users to actively participate in the summarization process, allowing them to specify their preferences, including content inclusions and exclusions. This user-centric approach promises to enhance the quality and relevance of summarization result.

index	question	ground_truths	answer	contexts
0	How to deposit a cheque issued to an associate in my business into my business account?	["Have the check reissued to the proper payee.Just have the associate sign the back and then deposit it. It's called a third party cheque and is perfectly legal. I wouldn't be surprised if it has a longer hold period and, as always, you don't get the money if the cheque doesn't clear. Now, you may have problems if it's a large amount or you're not very well known at the bank. In that case you can have the associate go to the bank and endorse it in front of the teller with some ID. You don't even technically have to be there. Anybody can deposit money to your account if they have the account number. He could also just deposit it in his account and write a cheque to the business."]	You can't deposit a cheque made out to an associate directly into your business account. Instead, you need to have the associate endorse the cheque over to your business by signing it and providing ID at the bank. This is because the cheque is made out to the associate, not the business, so the bank won't allow you to deposit it directly. Additionally, if the cheque is for a large amount or you're not well known at the bank, they may require additional documentation or verification before allowing the deposit. It's best to have the associate go to the bank and endorse the cheque in front of a teller with their ID to avoid any issues.	['Just have the associate sign the back and then deposit it. It\'s called a third party cheque and is perfectly legal. I wouldn\'t be surprised if it has a longer hold period and, as always, you don\'t get the money if the cheque doesn\'t clear. Now, you may have problems if it\'s a large amount or you\'re not very well known at the bank. In that case you can have the associate go to the bank and endorse it in front of the teller with some ID. You don\'t even technically have to be there. Anybody can deposit money to your account if they have the account number. He could also just deposit it in his account and write a cheque to the business."I have checked with Bank of America, and they say the ONLY way to cash (or deposit, or otherwise get access to the funds represented by a check made out to my business) is to open a business account. They tell me this is a Federal regulation, and every bank will say the same thing. To do this, I need a state-issued ""dba"" certificate (from the county clerk\'s office) as well as an Employer ID Number (EIN) issued by the IRS. AND their CHEAPEST business banking account costs $15 / month. I think I can go to the bank that the check is drawn upon, and they will cash it, assuming I have documentation showing that I am the sole proprietor. But I\'m not sure.... What a racket!!"When a business asks me to make out a cheque to a person rather than the business name, I take that as a red flag. Frankly it usually means that the person doesn\'t want the money going through their business account for some reason - probably tax evasion. I\'m not saying you are doing that, but it is a frequent issue. If the company makes the cheque out to a person they may run the risk of being party to fraud. Worse still they only have your word for it that you actually own the company, and aren\'t ripping off your employer by pocketing their payment. Even worse, when the company is audited and finds that cheque, the person who wrote it will have to justify and document why they made it out to you or risk being charged with embezzlement. It\'s very much in their interests to make the cheque out to the company they did business with. Given that, you should really have an account in the name of your business. It\'s going to make your life much simpler in the long run.']
1	Can I send a money order from USPS as a business?	["Sure you can. You can fill in whatever you want in the From section of a money order, so your business name and address would be fine. The price only includes the money order itself. You can hand deliver it yourself if you want, but if you want to mail it, you'll have to provide an envelope and a stamp. Note that, since you won't have a bank record of this payment, you'll want to make sure you keep other records, such as the stub of the money order. You should probably also ask the contractor to give you a receipt."]	No, you cannot send a money order from USPS as a business. Money orders are only available for personal use and cannot be used for business purposes. If you need to make a payment to someone, you will need to open a business account with a bank or other financial institution in order to do so legally and safely. Additionally, it is not recommended to use money orders for large transactions or for payments to unknown parties, as they can be easily lost or stolen. It is always best to use a secure payment method such as a check or electronic transfer when making important business payments.	['Sure you can. You can fill in whatever you want in the From section of a money order, so your business name and address would be fine. The price only includes the money order itself. You can hand deliver it yourself if you want, but if you want to mail it, you\'ll have to provide an envelope and a stamp. Note that, since you won\'t have a bank record of this payment, you\'ll want to make sure you keep other records, such as the stub of the money order. You should probably also ask the contractor to give you a receipt."Lets say you owed me $123.00 an wanted to mail me a check. I would then take the check from my mailbox an either take it to my bank, or scan it and deposit it via their electronic interface. Prior to you mailing it you would have no idea which bank I would use, or what my account number is. In fact I could have multiple bank accounts, so I could decide which one to deposit it into depending on what I wanted to do with the money, or which bank paid the most interest, or by coin flip. Now once the check is deposited my bank would then ""stamp"" the check with their name, their routing number, the date, an my account number. Eventually an image of the canceled check would then end up back at your bank. Which they would either send to you, or make available to you via their banking website. You don\'t mail it to my bank. You mail it to my home, or my business, or wherever I tell you to mail it. Some business give you the address of another location, where either a 3rd party processes all their checks, or a central location where all the money for multiple branches are processed. If you do owe a company they will generally ask that in the memo section in the lower left corner that you include your customer number. This is to make sure that if they have multiple Juans the money is accounted correctly. In all my dealings will paying bills and mailing checks I have never been asked to send a check directly to the bank. If they want you to do exactly as you describe, they should provide you with a form or other instructions.""I have checked with Bank of America, and they say the ONLY way to cash (or deposit, or otherwise get access to the funds represented by a check made out to my business) is to open a business account. They tell me this is a Federal regulation, and every bank will say the same thing. To do this, I need a state-issued ""dba"" certificate (from the county clerk\'s office) as well as an Employer ID Number (EIN) issued by the IRS. AND their CHEAPEST business banking account costs $15 / month. I think I can go to the bank that the check is drawn upon, and they will cash it, assuming I have documentation showing that I am the sole proprietor. But I\'m not sure.... What a racket!!""]

Fig 5: *Evaluation Data from RAGAS Baseline Experiments*

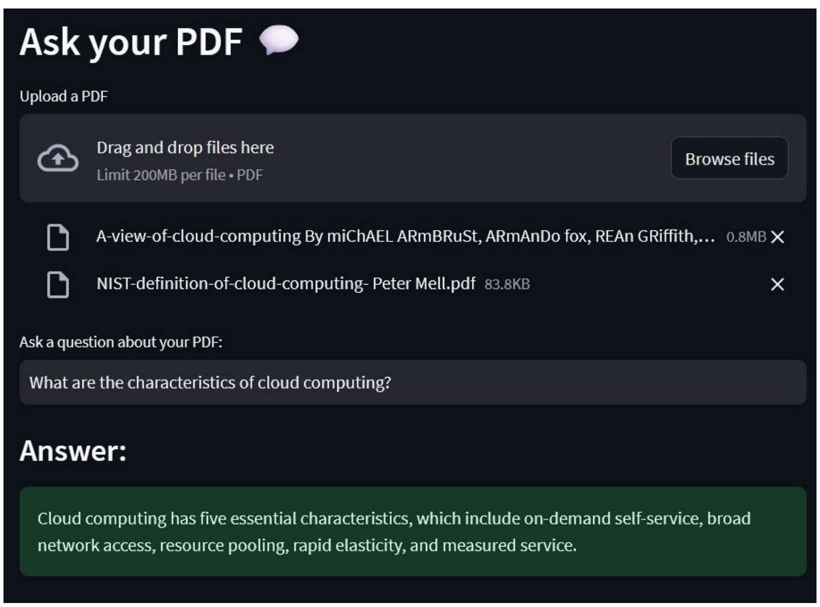

Fig 6: *Taking PDFs as input and answering question*

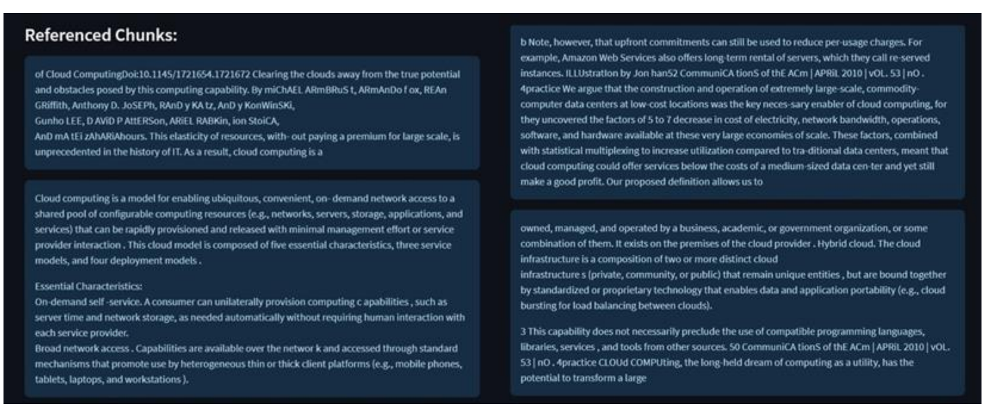

Fig 7: *Referenced Chunks*

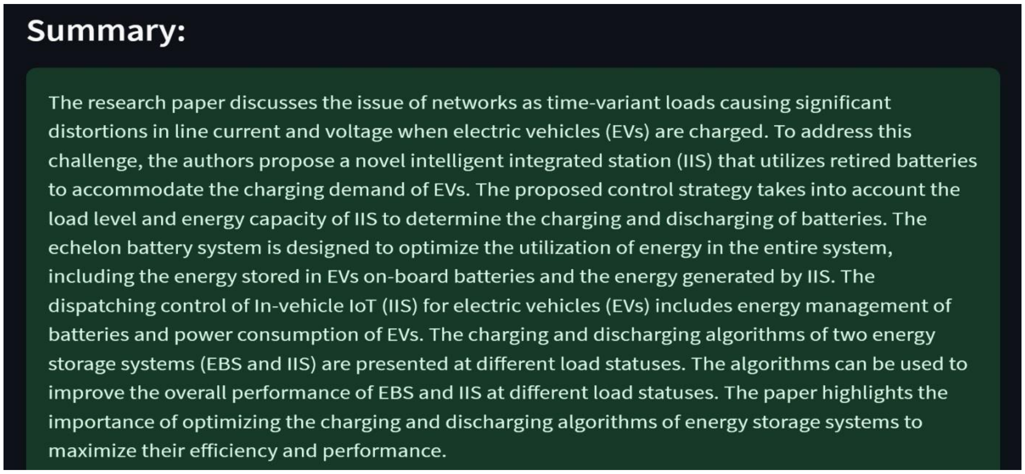

Fig 8: *Final Summary*

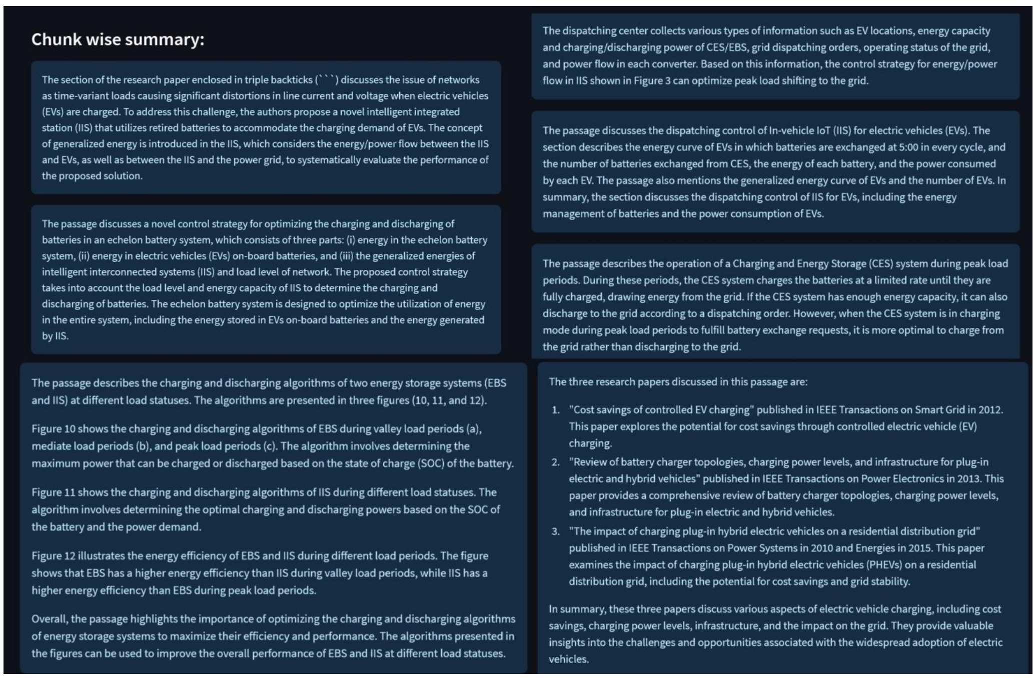

Fig 9: *Chunk-wise Summaries*

7. Statements

G. Competing Interest

None of the authors have any competing interests to declare.

H. Funding Information

No funding was received for this research.

I. Data Availability Statement

The data used in this study are available from the corresponding author upon reasonable request.

References

[1] Rani Horev (2018). BERT Explained: State of the language model for NLP. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[2] "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing." Google Research, 2 Nov. 2018, ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation,

Translation, and Comprehension". arXiv:1910.13461.https://arxiv.org/abs/1910.13461v1

[4] S. Pal, "The Why, What, Who and Where of Vector Databases", May 10, 2023. Available: https://www.eckerson.com/articles/why- what-who-and-where-of-vector-databases

[5] Touvron, H.et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models, arXiv.org. Available at: https://arxiv.org/abs/2307.09288

[6] Keivalya Pandya, Mehfuza Holia (2023). "Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations." arXiv: 2310.05421 [cs.CL].

[7] K.Martineau, "What is retrieval- augmented generation?," IBM Research Blog, Aug.2023,[Online]. Available: https://research.ibm.com/blog/retrieval-augmented-generation-RAG

[8] Atlassian, "What is a Knowledge Base?" Available:https://www.atlassian.com/itsm/knowledge-management/what-is-a-knowledge-base

[9] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, 2010, pp. 63-67

[10] Nan, F., Santos, C. N. dos, Zhu, H., Ng, P., McKeown, K., Nallapati, R., Zhang, D., Wang, Z., Arnold, A. O., Xiang, B. (2021). Improving factual consistency of abstractive summarization via question answering. arXiv. http://arxiv.org/abs/2105.04623

[11] P. Sitikhu, K. Pahi, P.Thapa, S. Shakya (2019). "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability." arXiv: 1910.09129 [cs.IR]. http://arxiv.org/abs/1910.09129

[12] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out, Barcelona, Spain, 2004, pp. 74-81. [Online]. Available: https://www.aclweb.org/anthology/W04- 1013

[13] Max Grusky. 2023. Rogue Scores. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1914–1934, Toronto, Canada. Association for Computational Linguistic.