RESEARCH ARTICLE

Towards Economical Long-Form Summarization: A Chunk-Based Approach Using LLMs

Avishto Banerjee[1,*]

[1] SAP Labs India Pvt. Ltd., Bengaluru 560066, India

Abstract

In today's world anything almost everything related to literature can be achieved by LLMs. Be it summarization, abstraction, translation, transformation, etc. But not always is it possible to do those operations on extremely large content. Even with the large token output limits of newly launched advanced LLMs it is not always economically and technically feasible to perform such operations. To cater to such a problem this paper explores the idea of summarization of extensive contents by a chunk-based approach which is both efficient and economical. This approach also understands the drawback of loss of information while chunking and efficiently solves that issue. The usage of such a framework is highly demandable across various enterprise software industries as well as healthcare and financial industries to store, summarize as well as query various large contents which are sometimes challenging to maintain and query. To create a generic framework the approach used for the summarization is mainly zero-shot summarization.

Keywords: LLMs, summarization, chunking, generative AI, NLP.

1 Introduction

The invention of the transformer architecture laid the cornerstone for today's large language models (LLMs). LLMs have become a major buzzword in the AI community, primarily designed to answer queries and generate content based on user prompts. They are now widely applied across the software industry. However, operating LLMs remains expensive, especially when performing tasks such as summarization or information extraction on very large documents. Open-source pretrained language models (PLMs) such as BERT, BART, and T5 often fail to deliver sufficiently high accuracy for these tasks. In addition, many PLMs rely mainly on extractive rather than abstractive summarization. To address both the economic cost and the need for effective abstractive summarization of large PDF documents, this study employs an efficient chunking strategy that successfully overcomes these challenges.

2 Related Work

Currently a lot of works related to summarization using LLMs are being done. There are works and frameworks which are also focussing on the summarization quality as well. Many research related to topic segmentations of Video Lectures started using transformers and LLMs [1]. Alesh et al. [2] in their paper focuses on improving lecture video

Submitted: 27 April 2025
Accepted: 25 July 2025
Published: 17 November 2025

Vol. 1, **No.** 1, 2025.
10.62762/TLLM.2025.674475

***Corresponding author:**
✉ Avishto Banerjee
avishto2019@gmail.com

Citation
Banerjee, A. (2025). Towards Economical Long-Form Summarization: A Chunk-Based Approach Using LLMs. *ICCK Transactions on Large Language Models*, 1(1), 4–8.

summarization and segmentation. Since GPT 3.5 was expensive during that time, they used BART and LSG-BART and fine-tuned them. Some similar works related to the incoherence in the summarization of content and to solve it human summarized content is used to fine-tune different LLMs like Falcon-40B and LLama 2 [3]. In the early days a lot of comparison of different LLM models for abstractive and extractive summarization is used to understand their performances and it was usually found out that GPT was better at Abstractive Summarization where as Bert was better at extractive summarization [4]. The main metric for evaluation was ROUGE scores. In recent years, a substantial amount of research has focused on summarizing various types of financial and organizational data obtained from NASDAQ and related sources. The main algorithms used were LLAMA2, GPT and CLAUDE-2 [5]. There are also research on Timeline summarization which is challenging [6], their use of LLMs helps but still struggles to capture event progression across timelines effectively. Also the evaluation of the quality of summarization has been researched upon [7]. There has been research done on different comparisons of PLMS and LLMS and their quality of summarizations [8]. Also there has been comparison of summarizations of different languages like Bengali newspaper [9] and perisan content [10] and Arabic Content [11] and evaluation of those summarized contents. There has been a lot of research related to summarization using LLMs and transformer based models but very few has revolved around the concept of chunking and performing summarizations to optimise the content and also for cost effectiveness. This paper tries to follow that approach.

3 LLM Fundamentals for Summarization

3.1 Tokens vs Words

The term token is frequently encountered in discussions of large language models (LLMs) such as ChatGPT or Gemini. A token refers to the individual units into which text is segmented prior to model processing. These units can take various forms, including the following:

1. A word (ex: Hello)
2. Part of a word (ex: un, ing)
3. Punctuation (e.g., ".", "?", ",")
4. Whitespace or special characters

The number of tokens is usually more than or equal to the number of words, depending on the text. Table 1 shows an approximation of the tokens with words:

Table 1. Token to Word Approximation.

Text Type	Average Token to Word Ratio
Normal English text	∼1.3 tokens per word
Complex/technical	∼1.5 tokens per word
Code or symbols	∼2–3 tokens per word

Example:

Sentence: "ChatGPT is amazing!"

Words: 3 ("ChatGPT", "is", "amazing!")

Tokens: 5 ("Chat" + "GPT" + " is" + " amazing" + "!")

3.2 Cost of LLMs as per Tokens

Token play a crucial role in the cost of using any paid LLM. More the token used more the cost is required. Usually each LLM has the capability of handling very large number of token now a days but as the usage of tokens increase the cost of calling the LLMS also increases. The Table 2 explains the token limit as well as the token usage cost of LLMs:

3.3 Abstractive vs Extractive Summarization

Text summarization includes two main procedural approaches: extractions and abstraction. The process of abstractive summarization uses complex natural language generation methods to create new statements which represent the essential elements from original texts. The process duplicates human summary methods which involve transforming content through both paraphrasing and shortening the information. The extractive summarization process finds crucial sentences or phrases directly from original text documents while maintaining their original wording. The versatility of abstractive approaches allows for creating well-structured summaries but these methods require more complexity than extractive methods which maintain grammatical consistency. The abstractive summarization evaluation uses ROUGE (Recall-Oriented Understudy for Gisting Evaluation) which measures both n-gram agreements and overlapping sentences with reference summaries. The evaluation of extractive summarization depends on precision and recall together with the F1-score which measures overlap of sentences with the reference summary.

Table 2. LLM vs Token Limit vs Token Cost.

Methods	Token Limit	Input Cost (per 1k tokens)	Output Cost (per 1k tokens)
GPT-4-turbo	128K	$0.01	$0.03
GPT-4o	128K	~$0.005	~$0.015
GPT-4 (32K)	32K	$0.06	$0.12
GPT-3.5-turbo	16K	$0.001	$0.001
Gemini 1.5 Pro	128K (1M experimental)	$0.035	$0.15
Mixtral (MoE)	32K	~$0.50 (via third-party APIs)	~$1.50

4 Methodology

Our approach for LLM-based document summarization combines hierarchical summarization methods with semantic chunking processes. The initial step divides content into semantic sections which utilize paragraph or section boundaries to maintain the context throughout the entire process.

4.1 Chunking

Chunking is the process of dividing large texts into smaller, manageable parts to enable efficient processing by language models with limited context windows. Semantic chunking improves this by splitting the content with meaningful boundaries such as paragraphs, section headers, or newline characters, rather than fixed word counts. This process reserves the logical flow and context of the text within each chunk. Semantic chunking ensures that each chunk represents a self-contained idea or topic, which improves the quality of summarization or downstream tasks. It is commonly implemented by detecting paragraph breaks (`\n\n`), markdown-style headers, or using NLP techniques to track topic shifts. This method maintains context, reduces fragmentation, and is ideal for processing large contents.

4.2 Architecture

In this architecture, the LLaMA 3 model processes independent summaries through the local Ollama platform. The workflow begins with hierarchical summarization, in which initial summaries are merged and subsequently reprocessed by the pretrained language model (PLM) to generate context-aware outputs. This two-level organizational structure ensures precision at both high-level and fine-grained comprehension. The method further enables scalability, modularity, and robustness when summarizing large volumes of unstructured content, such as PDFs or legal documents, making it suitable for tasks such as knowledge distillation and automated report generation. Figure 1 presents a flowchart illustrating the overall architecture.

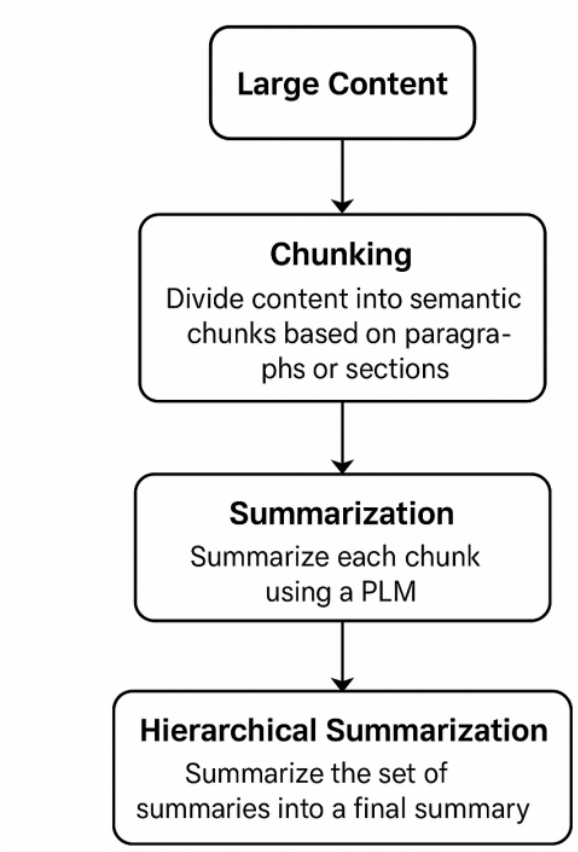

Figure 1. The architectural diagram.

4.3 ROUGE – L

The assessment of the summarization plays a very crucial roles in understanding the quality of the summarization [7]. **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) is a metric used to evaluate the quality of generated text, especially summaries, by measuring the longest common subsequence (LCS) between the

reference and the candidate. Unlike n-gram based metrics, ROUGE-L considers word order without requiring consecutive matches, making it better at capturing sentence-level structure. The metric evaluates how much of the reference summary is preserved in the candidate. The formulae for the ROGUE-L are as follows:

Let $\mathbf{X}$ be the reference and $\mathbf{Y}$ be the candidate summary

LCS (Length of Longest Common Subsequence)

Formula for RECALL:

$$ROGUEL = \frac{\mathbf{LCS}(\mathbf{X}, \mathbf{Y})}{|\mathbf{X}|} \quad (1)$$

Formula for PRECISION:

$$ROGUEL = \frac{\mathbf{LCS}(\mathbf{X}, \mathbf{Y})}{|\mathbf{Y}|} \quad (2)$$

Formula for F1-Score:

$$ROGUEL = \frac{2(\mathbf{P} \cdot \mathbf{R})}{\mathbf{P} + \mathbf{R}} \quad (3)$$

4.4 Results and Discussion

The architecture was executed on a large dataset, specifically a PDF book containing approximately 500 pages. The resulting ROUGE-L values, presented in Table 3, demonstrate the accuracy of the proposed solution.

Table 3. Results of the Summarization.

Text No.	ROUGE-L Recall	ROUGE-L Precision	ROUGE-L F1-Score
Chapter 1	0.146808511	0.56557377	0.233108105
Chapter 2	0.161764706	0.538461538	0.248788365
Chapter 3	0.145454545	0.341880342	0.204081628
Chapter 4	0.198555957	0.4296875	0.271604934

5 Conclusion

This paper investigates the concepts of chunking and hierarchical summarization and presents the corresponding experimental results. The findings indicate an overall improvement in summarization quality, with the F1-score increasing from 0.233 in Chapter 1 to 0.272 in Chapter 4, particularly for longer sections such as Chapter 4. Nonetheless, some variability remains—for instance, the decrease to 0.204 in Chapter 3—which is likely attributable to content-specific factors such as topic density. This variability highlights the need for more adaptive chunking strategies. Multiple architectural directions remain open for exploration. For example, [12] examines chunking techniques and the development of an optimized retrieval-augmented generation (RAG) pipeline, while [13] focuses on fine-tuning small LLMs for summarizing telephonic conversations. Similar approaches can be applied to further investigate alternative chunking mechanisms in the context of LLM-based summarization, RAG workflows, and evaluation methodologies.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Avishto Banerjee is an employee of SAP Labs India Pvt. Ltd., Bengaluru 560066, India. The author declares no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Soares, E. R., & Barrére, E. (2018, October). Automatic topic segmentation for video lectures using low and high-level audio features. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (pp. 189-196). [Crossref]

[2] Alesh, Y., Aoudia, M., Abdulghani, O., Al Ali, O., & Abu Talib, M. (2024, July). Abstractive Summarization of Lectures and Lecture Segments Transcripts with BART. In *International Conference on Artificial Intelligence in Education Technology* (pp. 43-55). Singapore: Springer Nature Singapore. [Crossref]

[3] Parmar, M., Deilamsalehy, H., Dernoncourt, F., Yoon, S., Rossi, R. A., & Bui, T. (2024). Towards enhancing coherence in extractive summarization: Dataset and experiments with LLMs. *arXiv preprint arXiv:2407.04855*.

[4] Kotkar, A. D., Mahadik, R. S., More, P. G., & Thorat, S. A. (2024, August). Comparative analysis of transformer-based large language models (llms) for text summarization. In *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)* (pp. 1-7). IEEE. [Crossref]

[5] Wilson, E., Saxena, A., Mahajan, J., Panikulangara, L., Kulkarni, S., & Jain, P. (2024, March). FIN2SUM:

advancing AI-driven financial text summarization with LLMs. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies* (pp. 1-5). IEEE. [Crossref]

[6] Sojitra, D., Jain, R., Saha, S., Jatowt, A., & Gupta, M. (2024, July). Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2657-2661). [Crossref]

[7] Provakar, M. M. (2024, October). Evaluating the Text Summarization Efficiency of Large Language Models. In *2024 2nd International Conference on Information and Communication Technology* (*ICICT*) (pp. 6-10). IEEE. [Crossref]

[8] Jiang, Z., Yang, J., & Rao, D. (2024, November). An Empirical Study of Leveraging PLMs and LLMs for Long-Text Summarization. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 424-435). Singapore: Springer Nature Singapore. [Crossref]

[9] Sultana, F., Fuad, M. T. H., Fahim, M., Rahman, R. R., Hossain, M., Amin, M. A., ... & Ali, A. A. (2024, December). How Good are LM and LLMs in Bangla Newspaper Article Summarization?. In *International Conference on Pattern Recognition* (pp. 72-86). Cham: Springer Nature Switzerland. [Crossref]

[10] VarastehNezhad, A., Tavasoli, R., Masumi, M., Majd, S. S., & Shamsfard, M. (2024, December). Evaluating LLMs in Persian News Summarization. In *2024 15th International Conference on Information and Knowledge Technology* (*IKT*) (pp. 195-201). IEEE. [Crossref]

[11] Aljohani, A., Alharbi, R., Alkhaldi, A., & Aljedaani, W. (2025, February). Evaluating LLMs for Arabic Code Summarization: Challenges and Insights from GPT-4. In *2025 8th International Conference on Data Science and Machine Learning Applications* (*CDMA*) (pp. 67-72). IEEE. [Crossref]

[12] Xiao, W., Liu, Y., Li, X., Gao, F., & Gu, J. (2024, December). TKG-RAG: A Retrieval-Augmented Generation Framework with Text-chunk Knowledge Graph. In *2024 25th International Arab Conference on Information Technology* (*ACIT*) (pp. 1-9). IEEE. [Crossref]

[13] Thulke, D., Gao, Y., Jalota, R., Dugast, C., & Ney, H. (2024, November). Prompting and Fine-Tuning of Small LLMs for Length-Controllable Telephone Call Summarization. In *2024 2nd International Conference on Foundation and Large Language Models* (*FLLM*) (pp. 305-312). IEEE. [Crossref]

Avishto Banerjee is an Associate Data Scientist at SAP, focused on SuccessFactors. He holds a B.Tech in Information Technology from WBUT and an M.Tech in Data Science from BITS Pilani. Since 2022, he has contributed to data science initiatives, internal research on Knowledge Graphs, and public sector analytics. His work centers on predictive modeling, AI-driven enterprise solutions, and large-scale business. (Email: avishto2019@gmail.com)