# SummaryScribe: Transformative Meeting Summarization using Prompt Engineering with Retrieval Augmented Generation (RAG) Technique Based on Natural Language Processing

Sheikh Mohammed Wali Ullah

NORTH SOUTH UNIVERSITY, sheikh.ullah@northsouth.edu

Zobaer Ahammod Zamil

NORTH SOUTH UNIVERSITY, zobaer.zamil@northsouth.edu

Md Saiyem Raiyan

NORTH SOUTH UNIVERSITY, saiyem.raiyan@northsouth.edu

Samia Sultana

NORTH SOUTH UNIVERSITY, samia.sultana01@northsouth.edu

The model aims to generate a better contextual summary from news or long transcripts using a SLM model. It involves a thorough process of model evaluation and optimization. Initially, multiple models, including Pegasus-CNN-DailyMail and EleutherAI/gpt-neo-125M, were analyzed to determine their effectiveness in summary generation. Among these, we found a better model - the Microsoft Phi-2 model, a small language model (SLM), as the best performer. To enhance the quality of the generated summaries, we utilized various prompt engineering techniques such as zero-shot, few-shot, and Retrieval-Augmented Generation (RAG). The implementation of RAG (Retrieval-Augmented Generation) makes the summaries much better by adding more context and retrieving useful information. Furthermore, using low-cost fine-tuning and prompt engineering techniques, this model can preserve the semantics and meaning of the textual context properly.

CCS CONCEPTS • Deep Learning → NLP → Summarizer → Low Resource Model → Microsoft Phi-2 → Prompt Engineering

**Additional Keywords and Phrases:** NLP, Hyper-parameter Tuning, Zero Shot, Few Shot, RAG

## 1 INTRODUCTION

The focus point of this system is to generate better contextual summaries from meetings, news or long Transcripts. It involves a thorough process of model evaluation and optimization. Initially, multiple models, including Pegasus-CNN-DailyMail and EleutherAI/gpt-neo-125M, were analyzed to determine their effectiveness in summary generation. Among these, we found a better model - the Microsoft Phi-2 model, a small language model (SLM), as the best performer. To enhance the quality of the generated summaries, we utilized various prompt engineering techniques such as zero-shot, few-shot, and Retrieval-Augmented Generation (RAG). The implementation of RAG (Retrieval-Augmented Generation) makes the summaries much better by adding more context and retrieving useful information. Furthermore, using low-cost fine-tuning and prompt engineering techniques, this model can preserve the semantics and meaning of the textual context properly.

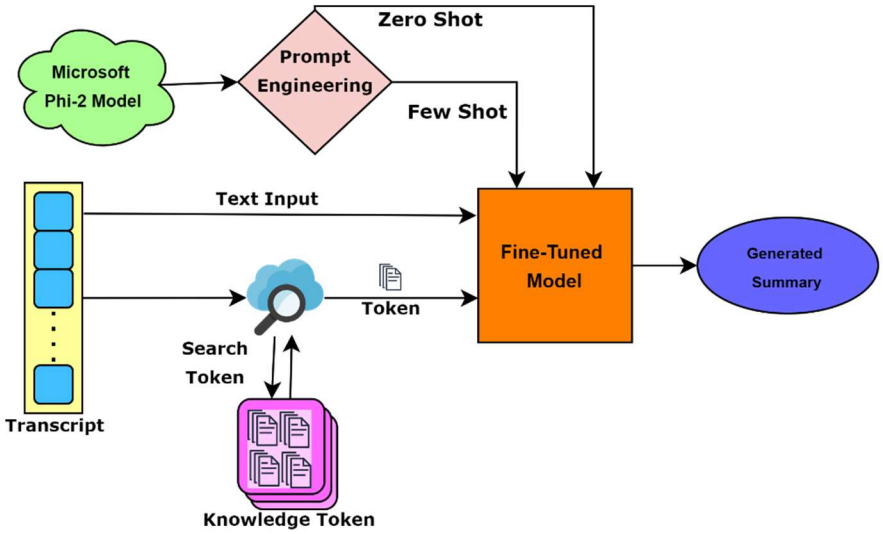



Figure 1.1: Microsoft Phi-2 Model with Zero shot, Few shot, and RAG

## 2 LITERATURE REVIEW

With the advent of deep learning, abstractive summarization techniques emerged, allowing models to generate summaries that are not just extracts but paraphrases of the original content. These methods have shown significant improvements in producing coherent and fluent summaries. A comprehensive review by Saiyyad and Patil (2024) [1], "Text Summarization Using Deep Learning Techniques: A Review" [1], highlights various deep learning techniques, including neural networks and transformers, which have revolutionized text summarization by enabling models to understand and generate human-like text.

Hanlei Jin, Yang Zhang (2024) [2] surveyed adopts a process-oriented schema to categorize text summarization methods, focusing on practical applications in real-world scenarios. It also explores the integration of Large Language Models (LLMs) in the automatic text summarization process.

Kumar, S., Solanki, A. (2023) [3], proposed an ATS model using a Transformer Technique with Self-Attention Mechanism (T2SAM) that model improves the performance of text summarization and is trained on the Inshorts News

dataset combined with the DUC-2004 shared tasks dataset. The performance of the proposed model has been evaluated using the ROUGE metrics, and it has been shown to outperform the existing state-of-the-art baseline models. This paper presents a novel approach to abstractive text summarization using transformer models, particularly focusing on their application to lengthy documents. The study demonstrates how transformer-based models can generate high-quality summaries by understanding and rewriting the content.

Basyal, L., & Sanghvi, M. (2023) [4],as shown remarkable promise in enhancing summarization techniques. Their paper embarks on an exploration of text summarization with a diverse set of LLMs,including MPT-7b-instruct, falcon-7b-instruct, and OpenAI ChatGPT text-davinci-003 models. The experiment was performed with different hyperparameters and evaluated the generated summaries using widely accepted metrics such as the BLEU Score, ROUGE Score, and BERT Score. According to the experiment, text-davinci-003 outperformed the others. This investigation involved two distinct datasets: CNN Daily Mail and XSum. Its primary objective was to provide a comprehensive understanding of the performance of LLMs when applied to different datasets.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023) [5], benchmarked several large language models, including different versions of GPT-3 and other models like Pegasus and BRIO, for their performance in news summarization. It evaluates the models using human ratings on criteria such as faithfulness, coherence, and relevance.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024) [6] provide a comprehensive overview of recent advancements in prompt engineering, categorized by application area. The paper details various prompting methodologies, their applications, the models involved, and the datasets used. It discusses the strengths and limitations of each approach and includes a taxonomy diagram and a table summarizing key points, models, and datasets for each technique. The survey categorizes and reviews techniques such as zero-shot, few-shot, and chain-of-thought prompting, and assesses their effectiveness in enhancing LLM performance across different tasks. This structured analysis aims to aid in understanding the field's rapid development and to highlight open challenges and future opportunities in prompt engineering.

## 3 METHODOLOGY

### 3.1 Dataset

In our quest to find the most influential text summarization approach, we conducted an extensive study utilizing the CNN/Daily Mail dataset. Due to our resource limitation, we use one-third of the dataset over the models.

### 3.2 Models

Our initial foray involved the Pegasus model, which, despite undergoing fine-tuning and training on our dataset, fell short of our expectations. The model's inability to handle input texts exceeding 1000 words proved a significant limitation, prompting us to explore alternative avenues.

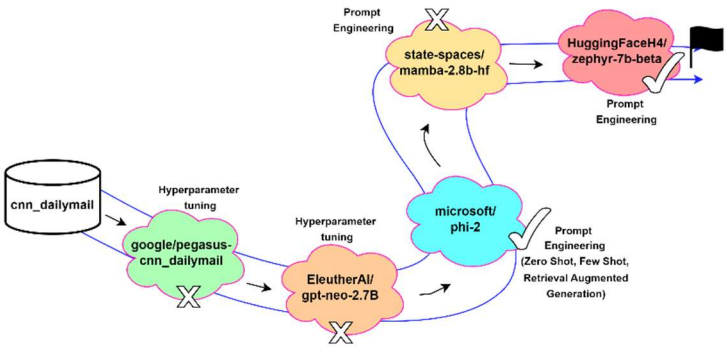



Figure 3.1: System Roadmap Diagram

Undeterred, we turned our attention to the EleutherNetAI/GPT model, loading it via the Happy Transformer library. However, our efforts were thwarted by persistent hyperparameter issues, resulting in unsatisfactory summarization outputs even after fine-tuning the model.

Determined to find a solution, we delved into the realm of Microsoft Phi-2 model, employing prompt engineering techniques to enhance its summarization capabilities. To our delight, this model excelled, generating high-quality summaries that rivaled those produced by OpenAI's models, establishing it as our study's most accurate option for lengthy text summarization.

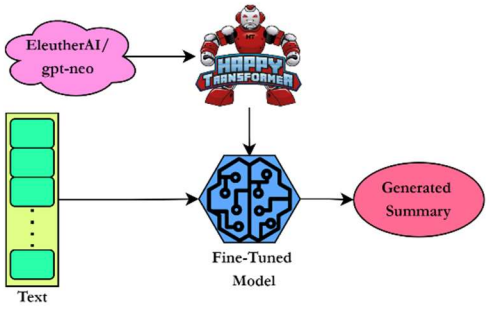



Figure 3.2: Hyperparameter tuning with EleutherNetAI/gpt-neo

But our pursuit for perfection didn't stop there. To further refine the accuracy of our summaries, we integrated the Zypher model, which employs custom instructions for optimization. While Zypher delivered superior summaries, its demanding computational requirements posed significant challenges given the constraints of our available resources. We also considered the Mamba model, but its resource-intensive nature rendered it impractical for our infrastructure.

As we reflect on our journey, the Microsoft/phi-2 model emerges as the optimal choice, striking a balance between accuracy and resource requirements for generating precise summaries of lengthy texts. Prompt engineering was pivotal in optimizing the performance of the models we evaluated. Although the Zypher model showcased the promising potential for even better results, its practical application was hindered by our computational resource limitations.

Looking ahead, future endeavors may focus on addressing these resource constraints, paving the way for the full exploitation of advanced models like Zypher unlocking their true potential in the realm of text summarization.

### 3.3 Algorithms

ALGORITHM 1: Implementation of Microsoft Phi-2 with Prompt Engineering

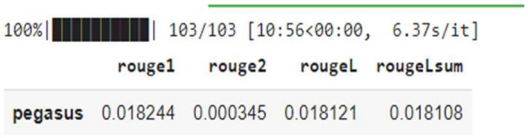
**Require:** Install transformers, accelerate
**Import** AutoModelForCausalLM, AutoTokenizer from transformers
**MODEL LOAD**
**Promt** = "PROMPT + TEXT + SHORT SUMMARY(?)"
**SET** MAX_NEW_TOKENS and TEMPERATURE
**GET** Knowledge Based Tokens from Database
**GENERATE** OUTPUT (SHORT SUMMARY)
Here, PROMPT and TEXT will be given. Default MAX_NEW_TOKENS = 512, TEMPERATURE = 0.3

## 4 EXPERIMENTAL RESULTS

According to our *pegasus cnn-dailymail* model, we fine-tune the model and get a result of dialogue summarization. From the samsum dataset, we take the input of the main text and summarize it in our model. After fine-tuning the model with the dataset the rouge values we found were unsatisfactory. Here, we trained the model for 15 epochs.


| | rouge1 | rouge2 | rougeL | rougeLsum |
|---|---|---|---|---|
| pegasus | 0.015495 | 0.000302 | 0.015503 | 0.015519 |

Figure 4.1: Before Train




| | rouge1 | rouge2 | rougeL | rougeLsum |
|---|---|---|---|---|
| pegasus | 0.018244 | 0.000345 | 0.018121 | 0.018108 |

Figure 4.2: After Train

But in EleutherNetAI/GPT model with the daily mail dataset, we got some summaries with less context. Though it gives some summaries, but it uses high resources and less contextual. Furthermore, with the Microsoft Phi-2 model, which is basically a small language model, gives more contextual summaries when we use prompt engineering.

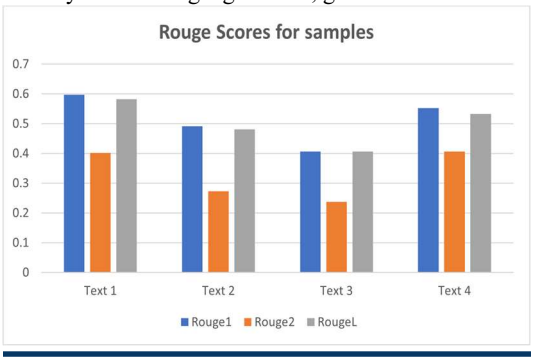



Figure 4.3: Rouge score for Phi-2

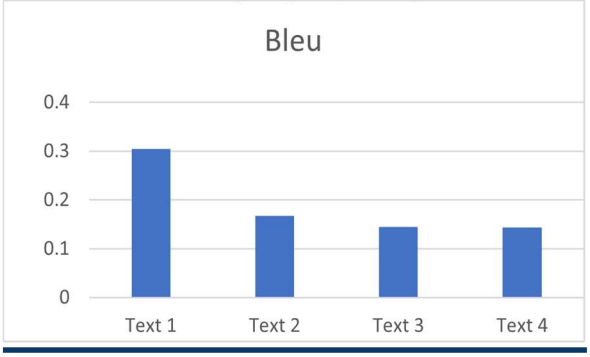



Figure 4.3: Bleu score for Phi-2

Which is much better than others. We also check the third-party scoring platform for our model generated summary and OpenAI: ChatGPT generated summary.

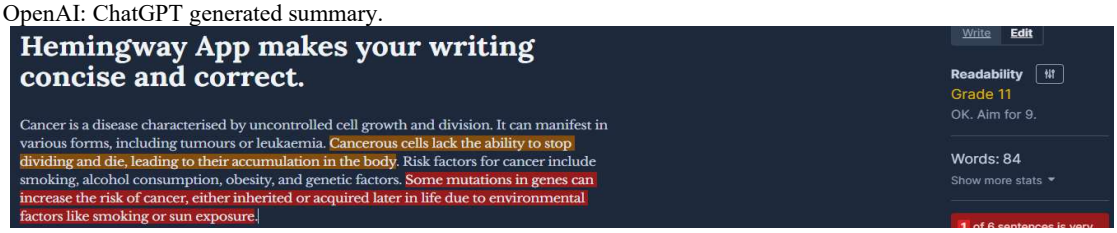



Figure 4.4: Summary and score for OpenAI

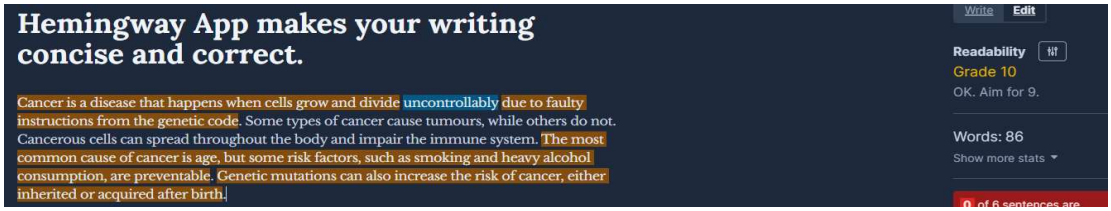



Figure 4.5: Summary and score for Microsoft Phi-2 with Prompt Engineering

We also utilized the Mamba and Zyphre models, but we were unable to load the actual models because of their large size and resource constraints. Here's the result: we have determined that with our model, we can achieve a summary that closely resembles the one generated by OpenAI.

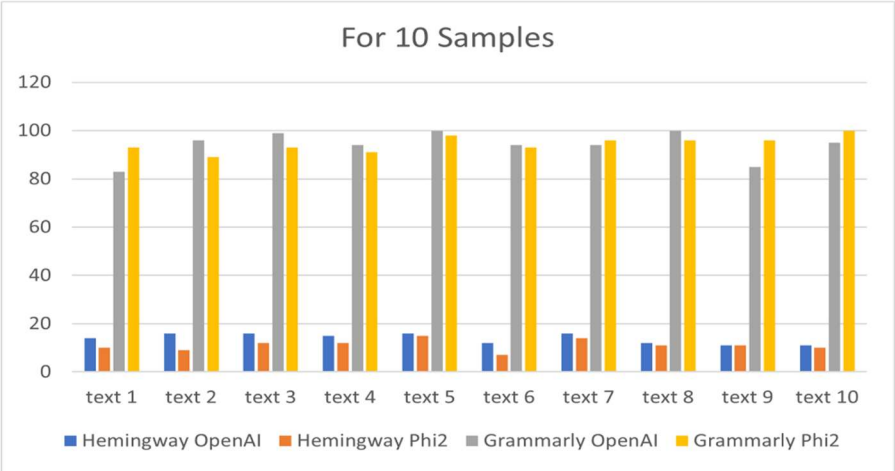



Figure 4.5: Hemingway and Grammarly score for 10 samples

## 5 DISCUSSION

The quest for an effective long-text summarization solution has been arduous, marked by triumphs and setbacks alike. Initially, we grappled with the limitations of the Pegasus model, its robust capabilities overshadowed by its struggle with lengthy inputs. The EleutherNetAI/GPT model showed promise, but persistent hyperparameter issues hindered its performance. A breakthrough emerged with Microsofts phi-2 model, which, through the artful application of prompt engineering, delivered high-quality summaries while balancing accuracy and resource efficiency. However, our pursuit of perfection led us to the Zypher model, whose custom optimization yielded superior summaries, yet its computational demands posed significant challenges. This experience underscores the delicate trade-off between performance and resource availability, compelling us to explore more efficient models and optimization strategies to unlock the full potential of highly accurate summarizers like Zypher for widespread practical use.

## CONCLUSION

To sum up, our project targets a broad audience by providing accessible summarization tools. Our findings not only enhance our current offerings but also provide valuable insights for future Researchers. In the future, we plan for local deployment and API integration, allowing flexible use across different platforms. While barriers in long text summarization, such as maintaining context in large discussion scripts with multiple contexts, and the low resource devices

may cause breakdowns. We continuously worked on overcoming these issues to deliver accurate and coherent summaries. As we expand, our commitment to innovation and improvement ensures that we remain at the leading edge of summarization technology.

## LIMITATION

Locally, our memory cannot load this. It at least needed 32GB external memory or multiple GPUs. As a result, it is quite impossible to go with this. LLM models need big resources and enough cost. For those drawbacks, we worked with the Small language Model (SLM) which worked so well. While the Microsoft Phi-2 model and prompt engineering techniques have shown promise, certain limitations must be acknowledged. Firstly, the quality of generated summaries heavily relies on the training data and the model's comprehension abilities. Biases or inaccuracies in the training data may propagate through the summaries, potentially spreading misinformation. Additionally, as a black box system, the Phi-2 model's decision-making processes lack transparency, raising concerns about accountability and potential unintended consequences, especially in sensitive applications.

Furthermore, the computational demands and resource constraints we faced highlight the challenges of scaling and deploying such systems in real-world scenarios. The trade-off between performance and resource availability may hinder widespread adoption, particularly in resource-constrained environments or applications requiring real-time summarization.

## IMPROVEMENT

To further improve our summarization approach, we could explore techniques like transfer learning and multi-task learning to leverage knowledge from diverse domains and tasks. Incorporating attention mechanisms and memory components could enhance the model's ability to capture long-range dependencies and contextual information. Ensemble methods combining multiple models could also boost performance. Additionally, developing more interpretable and explainable models would increase transparency and trustworthiness. Lastly, continual learning approaches could enable efficient model adaptation to new data distributions, reducing the need for computationally expensive retraining from scratch.

## ETHICS STATEMENT

In our relentless pursuit to push the boundaries of text summarization, we must remain steadfastly committed to upholding the highest ethical standards. It is our moral imperative to ensure transparency, accountability, and an unwavering dedication to integrity throughout this journey. We bear the profound responsibility of mitigating the insidious threats of bias and misinformation, fortifying our defenses through meticulous data curation, rigorous quality control, and continuous vigilance. Only through the development of interpretable models can we cultivate trust and facilitate responsible deployment. As we navigate the intricate tapestry of resource constraints and computational demands, we must embrace sustainable and eco-conscious approaches, aligning our endeavors with a broader vision of environmental stewardship. Ultimately, our quest must be guided by an unwavering commitment to ethical principles and a deep reverence for the well-being of society.

## REFERENCES


[1] Saiyyad, M. M., & Patil, N. N. (2024). Text Summarization Using Deep Learning Techniques: A Review. MDPI. https://doi.org/10.3390/engproc2023059194

[2] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang*, Jinghua Tan (2024). A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. (n.d.). Ar5iv. https://ar5iv.labs.arxiv.org/html/2403.02901

[3] Kumar, S., Solanki, A. An abstractive text summarization technique using transformer model with self-attention mechanism. Neural Comput & Applic 35, 18603–18622 (2023). https://doi.org/10.1007/s00521-023-08687-7

[4] Basyal, L., & Sanghvi, M. (2023, October 16). Text summarization using large language models: a comparative study of MPT-7B-Instruct, Falcon-7b-Instruct, and OpenAI Chat-GPT models. arXiv.org. https://arxiv.org/abs/2310.10449v2

[5] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023, January 31). Benchmarking large language models for news summarization. arXiv.org. https://arxiv.org/abs/2301.13848

[6] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024, February 5). A Systematic survey of prompt engineering in large language Models: Techniques and applications. arXiv.org. https://arxiv.org/abs/2402.07927

[7] OpenAI. (2023). ChatGPT [Large language model]. https://chat.openai.com