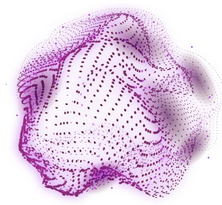
SarangshoScribe (সারাংশscribe):
To Explore How Good are Small LLMs in Summarization using Prompt Engineering

Supervisor: Dr. Mohammad Ashrafuzzaman Khan

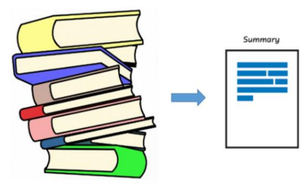
Sheikh Mohammed Wali Ullah, Zobaer Ahammod Zamil, Samia Sultana, Md Saiyem Raiyan

Department of Electrical and Computer Engineering, North South University



ABSTRACT

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.



SYSTEM DIAGRAM

Phi-2 by Prompt Engineering – Using zero shot, few shot, and RAG technique

Cross-match between the input transcript token and our knowledge-based token database, we gather the most similar token and pass it to our model along with the input transcript. Then, it generates the short contextual summary.

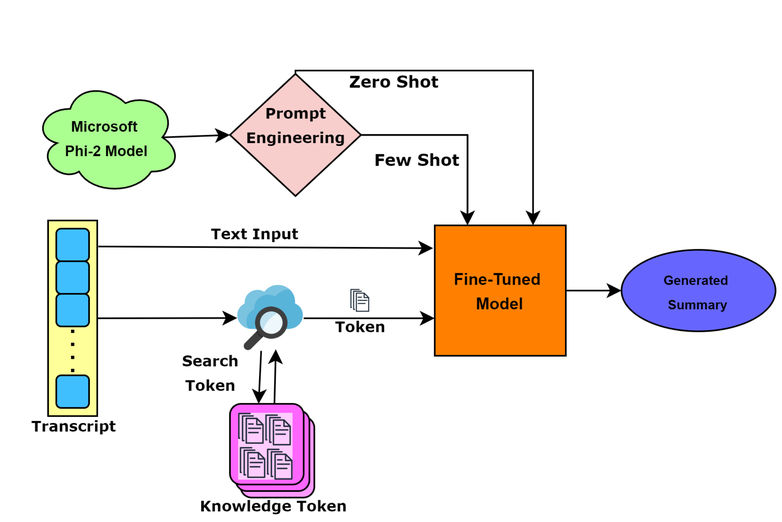


Figure: Microsoft Phi-2 with Prompt Engineering Technique

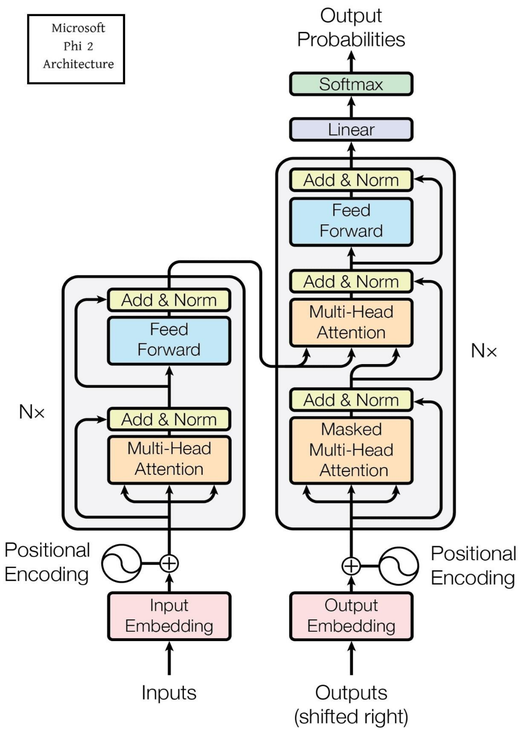


Figure: Microsoft Phi-2 Architecture

We started working on the CNN dataset on the Pegasus model, but after finding its drawback that it can't handle input script longer than 1000 words, we moved on to a new model named EleutherNetAI/GPT. We loaded this model using a happy transformer, but we couldn't be able to generate the summary from it of its hyperparameter issue. After that, we used a new model named Microsoft/phi-2 and using it, we can generate a good level of summaries compared to openAI summaries for long text. We used prompt engineering to use the phi-2 model. But, to get more accurate summaries, we used another model, named Zypher, which uses custom instruction to generate better output.

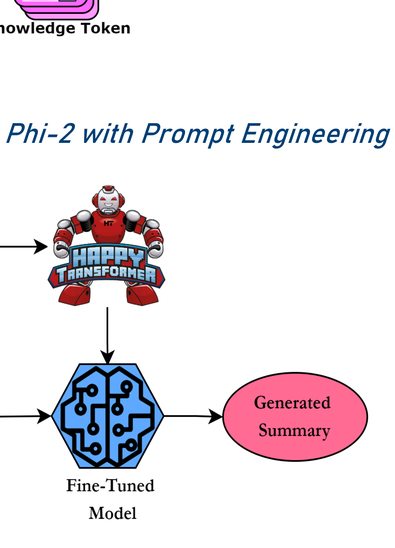


Figure: EleutherAI/gpt-neo 2.7B and 125M with Happy Transformer

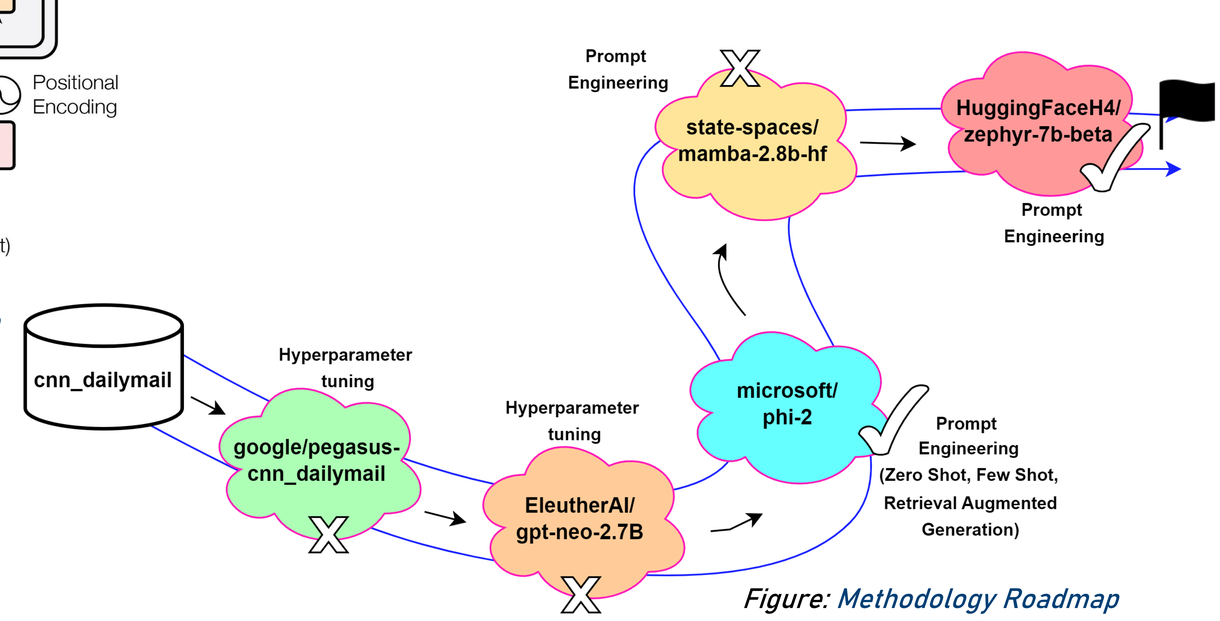
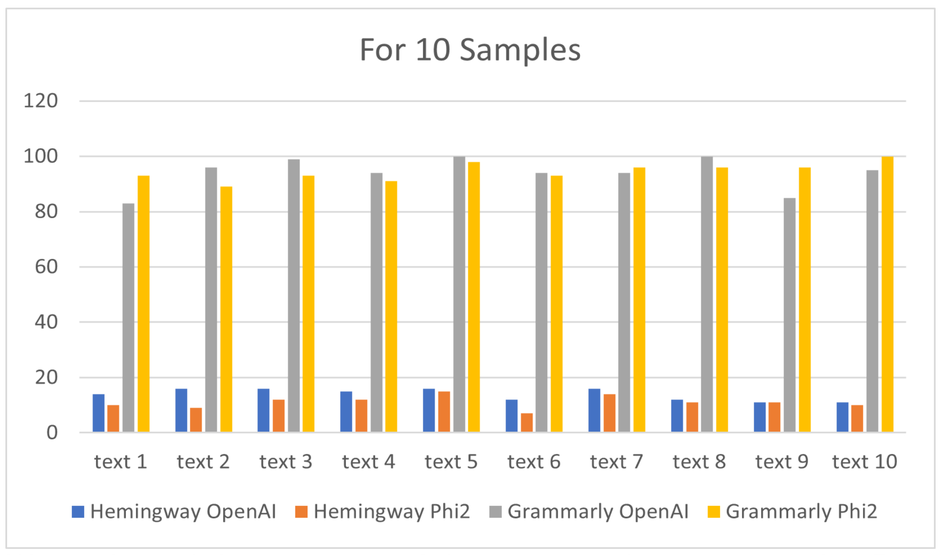
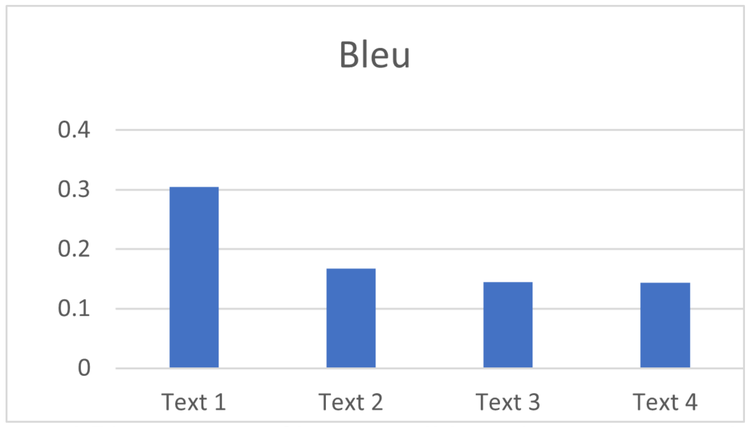


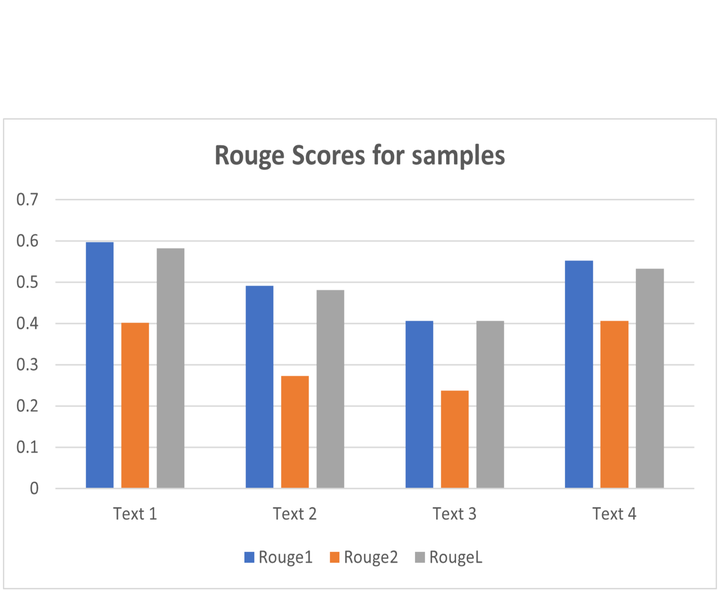
Figure: Methodology Roadmap

RESULTS



The bar chart shows the rouge score and BLEU score for some samples. As we don't have the summary data in CNN dataset we used openAI summary as ground truth and compared to the output of Microsoft/phi-2 using prompt engineering. We can see the bleu score and rouge scores for some sample text.





We evaluated the accuracy of the generated summary using Rouge and Bleu scores. Since we did not have the actual ground truth of the texts, we relied on the Hemingway score and Grammarly readability score to assess the generated texts from the Microsoft Phi-2 model and the OpenAI ChatGPT. Upon analysis, we found that the scores for the summaries generated from the Phi-2 model and ChatGPT are quite similar. This indicates we can achieve a proper contextual summary using a low-resource model like Microsoft Phi 2.

CONCLUSION

In conclusion, 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 model limitations where scripts longer than 1200 words 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.