SarangshoScribe (সারাংশscribe):

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

Abstract:

The focus point of this system is to generate better contextual summaries from news or long transcripts. After taking a news text or long transcript, we feed that into the finetuned model to generate a summary. However, analyzing multiple models pegasus-cnn-dailymail, EleutherAI/gpt-neo-125M, EleutherAI/gpt-neo-2.7B, we found a better model, Microsoft Phi-2, which is a LLM model. We used the model with prompt engineering techniques like zero-shot, few-shot and Retrieval-Augmented Generation. Using RAG on the models improves the generated summary significantly. Moreover, if we tune the parameters like temperature and num_of_tokens, we usually get even better summaries.

The method with System Diagram/Design Complexity

As mentioned in our title, we are utilizing Natural language Processing and implementing various models to generate the summary. After putting a text as input, the news transcript works through our models. We used several approaches. Among them, the Microsoft Phi-2 model is finetuned with zero shot, few shot, and RAG technique by prompt engineering, which works through our model and generates the summary of that particular transcript. Additionally, the model is perfectly analyzed according to the Rouge and Bleu score of the model.

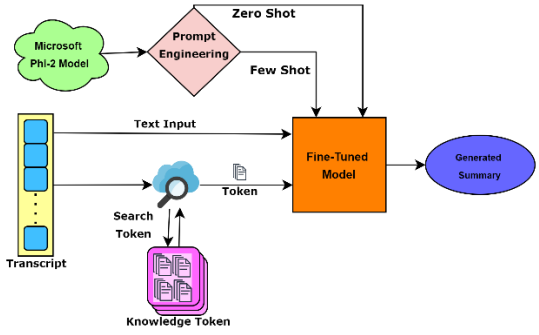

Figure: System Diagram

Novelty of project:

Plenty of research on summarization exists, but by analyzing Hemingway grade, ROUGE, and BLEU scores, we demonstrate superior performance compared to existing tools. Through rigorous analysis of multiple models, we identify the most effective ones for generating concise, coherent, and relevant meeting summaries by using Microsoft Phi-2 and RAG.

Impact on society:

Using technology to summarize transcripts will significantly affect society. Instead of doing lengthy text, reading short ones with most of the information seems to be the most practical thing to do. Our project helps people save time and work better together by summarizing long texts more effectively.

Business Model:

So far, our project has been implemented on various models for summarization from a transcript. Our business model involves deploying an unpaid version - basic summarization is free with limited customization and integration. News reporters can understand the news through generated summaries. Users can upgrade for more options and support and for the paid version - Users have to get a subscription for extra features like customizing, tuning, and better output summaries, advanced algorithms for better accuracy, can use locally, integration with popular tools, and priority support. As we grow, we'll expand our services and keep improving to stay ahead.

Conclusion:

After a thorough analysis, we find that the existing pre-trained models can perform even better if we implement prompt engineering and RAG on the model. It becomes more contextual and meaningful, which was our final goal to make a proper system that can produce an appropriate, meaningful short summary.