Unleashing the Power of Large Language Models: A Comprehensive Analysis on Long Document Transcripts Summarization

Jolene Chong
jolenechong7@gmail.com

Abstract

This report investigates the capabilities of large language models, specifically GPT and PaLM models, in *summarizing long-form webinar transcripts.* In this paper, we explore the **issues and challenges of long document summarization**. The computational and memory complexities of large transformer models meant the focus on BART models due to resource challenges with larger counterparts like **LongT5** and **LED**, our findings cater to researchers and professionals seeking informed decisions for diverse use cases. For evaluation and training, we used the **TIB** *dataset with Abstractive Summaries of Long Multimodel Videoconference Records*. Through rigorous experimentation, results show that GPT-4 consistently produced top-quality summaries with highest BERT Scores, while Bison text models excelled in speed and fine-tuned BART models offered a balanced, cost-effective solution.

1 Introduction

The need for effective and efficient webinar summarization has become paramount with the increasing number of long webinar content. Users want to know what the webinar is about before getting into it. This report embarks on a comprehensive exploration of the challenges and nuances associated with the task of summarizing lengthy documents.

As the volume and complexity of data grow, so do obstacles in distilling pertinent information concisely.

The choice of an appropriate dataset also plays a pivotal role in the robustness of this summarization study. We delve into the rationale behind selecting the TIB dataset for our research which offers a unique and diverse set of challenges reflective of real-world scenarios.

The selection of an appropriate evaluation method is also critical to discern efficacy of the summarization models. Before delving into the evaluation of various closed-sourced models and the fine-tuning of open-sourced counterparts, we will deliberate on the merits and demerits of different evaluation approaches such as ROUGE, BERT and Bleu Scores.

This report will also explore Parameter Efficient Fine Tuning techniques such as LoRA to overcome memory limitations during fine tuning selected open source models.

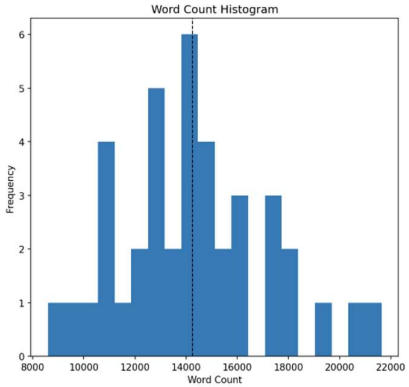
2 Overcoming Challenges with Long Document Summarization

Using LLMs for Summarization means using Abstractive Summarization instead of Extractive Summarization. This means it's more prone to hallucination but is more able to make coherent sentences.

2.1 Challenges Faced

There are max token limits on transformer models due to the architectures which LLMs are also predisposed to the issue.

With real-world data this model would be used on which are webinar contents of 40 videos of around 1-1.5 hours in length. On average, transcripts are 14k words in length.



Most LLM models are 4k, 8k and 16k in terms of the max token limits, which are all too small to fit average transcripts within 1 context window.

2.2 Methods to Overcome Challenges

The 2 main methods found to be commonly used were the Map Reduce as well as the Best Representation Vectors method.

The Map Reduce method consists of generating summaries of smaller chunks within token limits and then getting a summary of the summaries. This was the most popular and common way I found to be used. This in turn tends to be more computationally expensive.

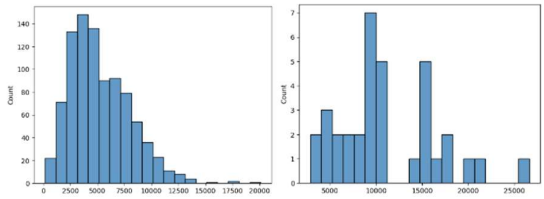
The Best Representation Vectors method relies on KMeans Clustering. The transcript is split into chunks within the context window limits which are then embedded as vectors. When text are similar, embeddings most likely to represent clusters (those closest to centroids) are selected and then summarized. This requires less computation in my experience but turned out to be slower with our use case.

3 Comparing Suitable Datasets

The goal was to find something similar to spoken language text data with human-written summaries.

TIB dataset A Dataset for Abstractive Summarization of Long Multimodal Videoconference Records. It focuses on long form transcription records of video conferences and abstracted human written summaries of it.
QMSum A benchmark dataset for Query-based Multi-domain Meeting Summarization with summarization queries and relatively short summaries.
VT-SSum A benchmark dataset with spoken language for video transcript segmentation and summarization including 125k transcript-summary pairs from about 9k videos. Unfortunately, it uses extractive summarization.

Other datasets suitable pre-trained models were trained on were the SamSum, CNN/DM and XSum datasets, where the XSum dataset proved to be unsuitable as the length of the summaries were too short.



The TIB dataset was chosen as the TIB dataset is much larger with a more suitable distribution of word lengths than the QMSum dataset which meant sampling is an option for fine-tuning and evaluating models.

4 Evaluating Suitable Evaluation Metrics

Some commonly used evaluation metrics were the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and the BERT Score. The Flesch Kincaid Grade Level Automatic Readability Score and Language Tool Python will also be used to get a gauge of the Readability and number of Grammatical Errors in the generated text.

Let's understand more about each evaluation metric for the task of summarization.

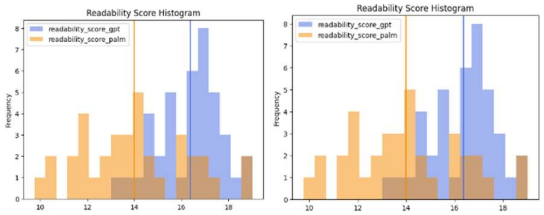
Recall-Oriented Understudy for Gisting Evaluation Score (ROUGE) A set of metrics used to evaluate quality of summaries generated by comparing it to a reference summary and **measuring the overlap in n-grams** (sequences of n words) Higher ROUGE Scores mean it captures more key information and overlaps more with the reference summary.
BERT Score Leverages pre-trained contextual embeddings from BERT models and matches word with reference summary by cosine similarity. Higher BERT Score mean **more similarity contextually** between the generated and reference summary.
Flesch Kincaid Grade Level A widely used readability formula which assesses the approximate reading grade level required to understand a text, considering factors such as sentence length, word complexity and syllable count. Higher scores indicate more complex text, requiring a higher level of education to understand.

Language Tool Python LanguageTool is an open-source grammar tool also known as the spellchecker in OpenOffice to detect grammar errors and spelling mistakes.

5 Evaluating Closed Source Models

The performance on the web interfaces of GPT-3.5-turbo and ChatGPT along with Google Bard and PaLM 2 models were different. Through automating summarization of the transcripts, there were certain scenarios where the language model was not able to provide a summary as a response, instead providing something like that "I'm just a language model, so I can't help you with that".

Comparing the differences in performance between the web interface versions ChatGPT and Bard, it was found that ChatGPT tends to have higher Flesch Kincaid Grade Level scores and slightly lower but close Grammar Error Counts.



5.1 Prompt Engineering

Comparing the performance of the models with different prompts.

Write a concise summary of the following: Readability Score: 15.30 Grammar Error Count: 1.60 ROUGE-1 F1 Score: 0.18 ROUGE-2 F1 Score: 0.05 ROUGE-L F1 Score: 0.17 BERT Score F1: 0.863640
TLDR in 200 words, without bullet points: Readability Score: 14.03 Grammar Error Count: 2.00 ROUGE-1 F1 Score: 0.19 ROUGE-2 F1 Score: 0.05 ROUGE-L F1 Score: 0.16 BERT Score F1: 0.861820

In evaluating the performance across these 2 prompts on GPT-4, the readability metric is deemed inconclusive due to the prevalence of null scores, particularly notable in the second prompt. The examination of grammar error count reveals the superiority of the first prompt, exhibiting lower counts and thus a higher linguistic precision.

Further emphasizing the significance of semantic alignment, the BERT F1 scores favor the first prompt, signifying a greater resemblance between its summaries and the gold standard.

Despite the second prompt garnering higher ROUGE scores, the prioritization of the BERT metric underscores the propensity of summaries from the first prompt to employ words more closely aligned with the gold standard summary.

5.2 Tuning Temperature

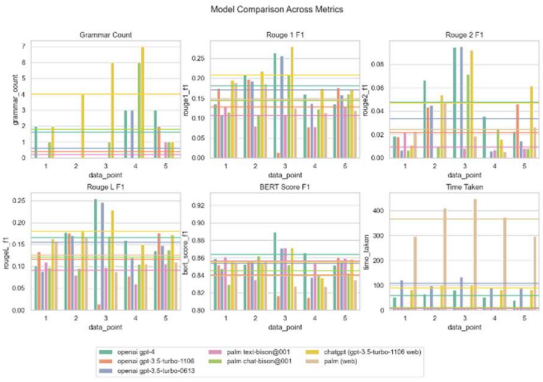
A higher temperature results in more creative, random, and imaginative text, while a lower temperature result is more accurate and factual text. Choosing the most probable word which helps to reduce the amount of variation in the generated text. Anything higher than 1.5 is usually not used due to the degree of randomness in its outputs.

With GPT-4, a temperature of 0 results in the best summaries with the highest BERT Score. The ROUGE Score also decreases as temperature increases as summaries start to use words that aren't in the test.

Temperature: 0 Readability Score: 15.30 Grammar Error Count: 1.60 ROUGE-1 F1 Score: 0.181107 ROUGE-2 F1 Score: 0.047568 ROUGE-L F1 Score: 0.166129 BERT Score F1: 0.863640
Temperature: 0.7 Readability Score: NaN Grammar Error Count: 2.40 ROUGE-1 F1 Score: 0.170811 ROUGE-2 F1 Score: 0.043525 ROUGE-L F1 Score: 0.151851 BERT Score F1: 0.857740
Temperature: 1 Readability Score: NaN Grammar Error Count: 2.00 ROUGE-1 F1 Score: 0.169784

ROUGE-2 F1 Score: 0.031849 ROUGE-L F1 Score: 0.154560 BERT Score F1: 0.857760

5.3 Closed-Source Model Comparisons



Only Bison models were available with the PaLM API on VertexAI, so only the text and chat versions of those were evaluated.

The GPT models perform better in terms of ROUGE scores meaning they use the same words commonly used in the gold standard summaries as well as the BERT Scores. It's interesting to note that the GPT-3.5-0613 (4k token limits) performs slightly better than its counterpart on the 16k token limits. ChatGPT is also no longer the top model in BERT Scores unlike in ROUGE Scores.

In terms of speed, the PaLM text model and the gpt-3.5-1106 (16k) *(as expected as this model can take in the full transcript within one context window)* were the fastest models.

It's important to note that research shows ROUGE tends to increase with longer summaries, since ChatGPT tends to produce longer summaries, it could be a cause for that. BERT is the recommended evaluation metric as a more reliable metric though slightly slower for model benchmarking, addressing the limitations associated with ROUGE.

Through this evaluation, GPT-4 and GPT-3.5-turbo-1106 provide higher quality summaries and palm text-bison@001 is a fast model with relatively suitable performance.

6 Evaluating Open Source Models

6.1 Choosing Suitable Pre-Trained Models

Before fine-tuning of open-source models for our summarization ask, a crucial preliminary step was to identify the most suitable model architectures.

Models such as LLaMA, though powerful were deemed as unsuitable due to extensive size, making it challenging to train on the limited resources available. Several other model architectures are commonly employed in summarization tasks such as the T5, Pegasus, Bart, LongT5, and LED. However, constraints in RAM availability posed challenges in training models with larger token limits, specifically LongT5 and LED, which required substantial resources. While the token limit of T5 was found to be restrictive at 512 tokens, both Pegasus and Bart, with a limit of 1,024 tokens, emerged as viable candidates for fine-tuning. Unfortunately, despite the appeal of LongT5 and LED, their training remained unfeasible due to the substantial RAM requirements associated with their 16,000-token limit.

The Pegasus model tended to perform more extractive summarization, while BART models trained on SAMSUM datasets tend to be able to pick up individual speakers and CNN/DM trained BART models performed well.

Pegasus Large Moderator: Good afternoon, everyone, and welcome to today's webinar on the fascinating and rapidly evolving topic of Artificial Intelligence. The future of Al holds immense promise, but it also presents important ethical and societal challenges that we need to address. As Al technologies continue to advance, it's essential that we consider the ethical implications. However, we must be cautious about data privacy and the need for responsible Al implementation in the healthcare sector.
BART Large CNN SAMSUM Today's webinar is on the topic of Artificial Intelligence. Dr. Emily Rodriguez, a renowned Al researcher and professor, and Dr. James Chen, a pioneer in Al ethics, are the first speakers. Sarah Patel, an expert in Al and its applications in healthcare, is the second speaker. Finally, Dr. Michael Johnson talks about the economic implications of Al.
BART Large CNN Artificial Intelligence has witnessed

remarkable growth over the past few decades. It's now ingrained in our daily lives, from voice assistants in our smartphones to self-driving cars. The future of Al holds immense promise, but it also presents important ethical and societal challenges that we need to address.
DistilBART CNN Artificial Intelligence has witnessed remarkable growth over the past few decades. It's now ingrained in our daily lives, from voice assistants in smartphones to self-driving cars and even in healthcare diagnostics. The future of Al holds immense promise, but it also presents important ethical and societal challenges that we need to address.

The Bart SAMSUM one performs well in getting multi-speaker sumamrizations but isn't as well phrased as the CNN model. Lastly the DistilBart model is the fastest and smallest model. After comparing performance of BART and DistilBART CNN/DM models, BART was chosen as not to sacrifice performance with much higher BERT and ROUGE Scores for a faster model.

The final decision to concentrate efforts on fine-tuning BART models was due to its practical token limits, performance reliability and widespread adoption, finding an optimal balance between model capability and resource constraints.

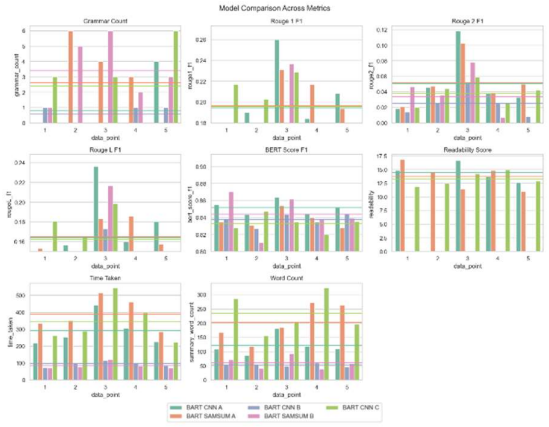
6.2 Fine-Tuning with LoRA

With the TIB dataset, sampling was done to fine-tune several models to get variations. All datasets are about 200 rows which needed about 5 to 6 hours to fine-tune. Dataset A, the first dataset had summaries with 150-300 words and text of more than 80000 words. Dataset B had longer summaries with 300-500 words with text that's more than 5000 words. Finally, with Dataset C with 150-300 word summaries and text that's less than 2500 words.

To optimize fine-tuning procedures in large language models, LoRA is a PEFT (Parameter Efficient Fine Tuning) technique. It involved representing weight updates using two smaller matrices, achieved through low-rank decomposition. These update matrices are then trained to adapt to new data while minimizing overall changes, ensuring efficiency. The original weight matrix remains frozen. Both the original and adapted weights are then used to generate the results.

Leveraging LoRA's quantization capabilities proved instrumental in reducing memory usage, enabling the model to operate seamlessly on a CPU with 32GB RAM. This choice was pivotal given the constraints of the available resources, comprising a CPU with 32GB RAM and a GPU with only 2GB VRAM—insufficient even for standalone PyTorch operations.

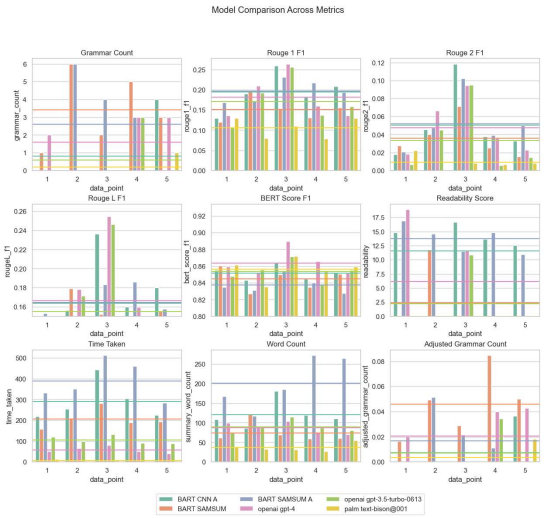
6.3 Evaluating Fine-Tuned Models



Dataset A and C result in the better performing models. With BART SAMSUM A performing the best in terms of ROUGE Score and BART CNN A performing the best in terms of readability and BERT scores and the fastest performing between the better performing models.

It's interesting to note that though the SAMSUM models are more able to capture multi-speaker summaries, the summaries generated tends to have more grammatical errors.

7 Conclusion: Comparing Trade-Offs



The BART SAMSUM A model performs the best according to ROUGE Score with GPT-4 performing slightly better in rougeL. Interestingly, GPT-4 performs a lot better in terms of BERT Score, with PaLM text-bison and gpt-3.5-turbo-0613, with BART CNN A following closely behind it. This shows that the GPT-4 model tends to use differing as those in the gold standard summary as compared to the BART CNN A and other open-source models.

PaLM text-bison and OpenAI GPT-4 models were the fastest models, with open-source models performing a lot slower. Under a minute vs 3-6 minutes.

Taking a closer look at the generated summaries, we realize that though SAMSUM models seem to perform worse according to the evaluation metrics, it still performs relatively well in terms of mult-speaker aspects.

GPT-4 (78 words)
In a webinar on Artificial Intelligence (AI), experts discussed the latest AI developments and their impact on life. Dr. Emily Rodriguez spoke about the rapid growth of AI technologies, Dr. James Chen emphasized the need for ethical guidelines, Dr. Sarah Patel highlighted AI's role in healthcare and the importance of data privacy, and Dr. Michael Johnson discussed AI's economic implications, including job displacement. The panel concluded that AI presents both immense promise and significant ethical and societal challenges.

GPT-3.5-turbo-0613 (74 words)
The webinar discussed the rapid growth of AI and its integration into daily life. Ethical and societal challenges, such as biases and privacy concerns, were emphasized. The potential of AI in healthcare and the need for responsible implementation and data security were highlighted. Economic implications, including upskilling the workforce and fostering innovation, were also discussed. The importance of considering the ethical and societal dimensions of AI for its positive impact on society was emphasized.

Dataset A SAMSUM Fine-Tuned Model (109 words)
Artificial Intelligence (AI) is a rapidly evolving technology that has profound implications for society, industry, and the economy. It has the potential to revolutionize many aspects of our lives, but it also presents important ethical and societal challenges that we need to address. In this webinar, we will hear from Dr. Emily Rodriguez, a renowned AI researcher and professor, Dr. James Chen, a pioneer in AI ethics, and Dr. Sarah Patel, an expert in AI and its applications in healthcare, who will discuss the ethical, societal, and economic implications of AI. Dr. Michael Johnson, a leading expert in the field of AI-related industries, will also discuss the economic implications.

Dataset A CNN Fine-Tuned Model (119 words)
Artificial Intelligence (AI) has become ingrained in our daily lives, from voice assistants in our smartphones to self-driving cars, and even in healthcare diagnostics. The future of AI holds immense promise, but it also presents important ethical and societal challenges that we need to address. This webinar will present the latest developments in AI and its impact on various aspects of our lives, including the ethical implications of the technology, as well as the economic and societal implications of AI. We will hear from a panel of experts who will share their insights on the current state of the art in the field of AI, including pioneers in AI ethics such as Dr. James Chen and Dr. Emily Rodriguez.

Dataset C CNN Fine-Tuned Model (121 words)
This webinar will shed light on the latest developments in AI and its impact on various aspects of our lives. The future of AI holds immense promise, but it also presents important ethical and societal challenges that we need to address. The panelists will discuss the role of AI in healthcare, the impact of AI on the economy, and the regulatory framework and ethical guidelines that are crucial to navigate this evolving landscape and strike a balance between innovation and safeguarding societal values. While AI has the potential to boost productivity and drive economic growth, it also poses challenges in terms of job displacement and workforce adaptation. The role of governments, businesses, and educational institutions in upskilling and retraining the

workforce.

The SAMSUM pretrained model before fine tuning also has a high BERT score but produced very short summaries, where fine tuning it on Dataset A has made it more descriptive. The model trained on CNN resulted in a longer summary on Dataset C.

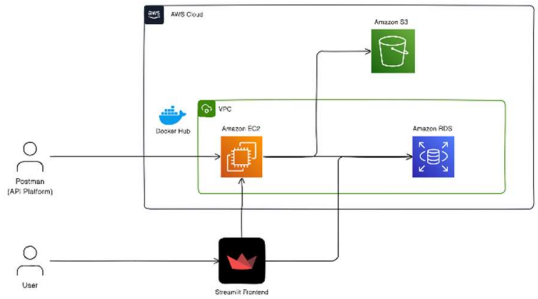
To wrap up, BART CNN has the best performance in terms of BERT and ROUGE Score among the open source models and has better summaries in terms of phrasing and its descriptiveness. However, the BART SAMSUM A model has a better performance with the multi-speaker aspect but still struggles a little with being more descriptive. GPT models also tend to return shorter summaries but have higher BERT Scores.

Looking at closed-source models, optimal choices depend on specific requirements and constraints. The GPT-4 model stands out as an ideal choice if budget considerations permit. Alternatively, the gpt-3.5-turbo-0613 model presents a viable option providing a balance between performance and pricings. If speed is a priority, the PaLM text-bison@001 model is the compelling choice.

Looking at open-source models tailored for multi-speaker environments, the Dataset A SAMSUM Fine-Tuned Model takes precedence, where it is honed for a nuanced understanding of multi-speaker interactions. If the emphasis is instead on generating summaries characterized by descriptiveness, quality and speed, the Dataset A CNN Fine-Tuned model proves advantageous or for slightly longer summaries the Dataset C CNN Fine-Tuned Model can also be considered.

8 Seeing It in Action!

To showcase a simplified set up for a clear understanding of the deployment process primarily designed for demonstration purposes.



The setup is streamlined, utilizing EC2 for backend hosting, which seamlessly connects to RDS and S3. The frontend is hosted on Streamlit. The EC2 instance, configured with Docker Hub, pulls the image after CI/CD pipelines build and push it via GitHub Actions. For obtaining summaries, Streamlit initiates a post request to the EC2 backend, where both the summary and the text are stored in RDS. The text transcript is concurrently saved to S3 as a text file, with the associated bucket name recorded in RDS for future reference. This deployment architecture ensures a cohesive and efficient flow of information.

References

Gigant, T., Dufaux, F., Guinaudeau, C., & Decombas, M. (2023, July 28). *Tib: A dataset for abstractive summarization of long multimodal videoconference records*. Accueil - Archive ouverte HAL. https://hal.science/hal-04168911

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X., & Radev, D. (2021, April 13). *QMSum: A new benchmark for query-based multi-domain meeting summarization*. arXiv.org. https://arxiv.org/abs/2104.05938

Lv, T., Cui, L., Vasilijevic, M., & Wei, F. (2021, July 15). *VT-SSUM: A benchmark dataset for video transcript segmentation and Summarization*. arXiv.org. https://arxiv.org/abs/2106.05606

gkamradt. (2023, October 2). *5 Levels Of Summarization: Novice to Expert*. GitHub. https://github.com/gkamradt/langchain-tutorials/blob/main/data_generation/5%20Levels%20Of%20Summarization%20-%20Novice%20To%20Expert.ipynb

Lin, C.-Y. (2004). *Rouge - a hugging face space by evaluate-metric*. ROUGE - a Hugging Face Space by evaluate-metric.

https://huggingface.co/spaces/evaluate-metric/rouge

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *Bert Score - a hugging face space by evaluate-metric*. BERT Score - a Hugging Face Space by evaluate-metric. https://huggingface.co/spaces/evaluate-metric/bertscore

Flesch reading ease and the Flesch Kincaid grade level. Readable. (2021, July 9). https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

Schluter, N. (2017). *The limits of automatic summarisation according to Rouge*. ACL Anthology. https://aclanthology.org/E17-2007/