**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.

# 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.

# 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.

# 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.

A graph of a graph

Description automatically generated

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.

# 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](#LevelsOfSummarizationWithLangChain) 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](#LevelsOfSummarizationWithLangChain) 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.

# Comparing Suitable Datasets

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

|  |
| --- |
| [TIB dataset](#TIBDatasetSource)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](#QMSumDatasetSource)A benchmark dataset for Query-based Multi-domain Meeting Summarization with summarization queries and relatively short summaries. |
| [VT-SSum](#VTSSumDatasetSource)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.

# A graph of blue bars Description automatically generated with medium confidence

# 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.

# 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) |
| BERT Score |
| Flesch Kincaid Grade Level |
| Language Tool Python |

# Evaluating Closed Source Models

# Evaluating Open Source Models

# Fine-Tuning with LoRA

# Evaluating Fine-Tuned Models

# Conclusion: Comparing Trade-Offs

# Seeing It in Action!

**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>