**Experiment 0: BART Model training entire dataset**

**Rationale:**

BART is an ideal choice for our initial implementation due to its proven effectiveness in abstractive summarisation tasks and its flexibility in handling complex, unstructured input. As a sequence-to-sequence model that combines bidirectional encoding (like BERT) with autoregressive decoding (like GPT), BART excels at generating fluent, contextually accurate summaries—precisely what is needed for summarising long-form, dialogue-heavy interview transcripts.

Given its strong performance on established datasets such as CNN/DailyMail, BART offers a reliable baseline for evaluating summarisation quality on the MediaSum dataset. Its ability to capture conversational flow and retain key information makes it perfectly suited to our domain.

**Observations and Results:**  
In our first experimental attempt to fine-tune the pre-trained BART model on the MediaSum dataset, the process was halted due to GPU memory exhaustion. The MediaSum dataset, being substantially large, placed a heavy demand on computational resources. In this attempt, the entire dataset was unintentionally passed to the model at once, which overwhelmed the available GPU capacity and caused the training to fail.

**Recommendations for next experiment:**

Given the substantial size of the MediaSum dataset and the high computational requirements of the BART model, it is recommended that the next experiment utilize **DistilBART**, a distilled and more lightweight version of BART. DistilBART retains much of the performance capability of the original model while significantly reducing training time and memory usage, making it a more practical choice under current hardware constraints.

This shift will enable more efficient experimentation, faster iteration cycles, and better resource management, while still leveraging the architectural strengths of the BART family for abstractive summarisation tasks. The use of DistilBART will also help validate the feasibility of deploying compact summarisation models in real-world applications where scalability and performance must be balanced.

**Experiment 1: DistillBART Model training entire dataset**

**Rationale:**

Following the initial attempt with the full-scale BART model, it became evident that the combination of a large pre-trained architecture and the extensive MediaSum dataset posed significant challenges to available GPU resources. To address this, the first refined experiment will be conducted using DistilBART, a lighter and faster version of BART.

DistilBART preserves much of the summarisation capability of its larger counterpart while significantly reducing the model size and training time. This makes it particularly well-suited for environments with limited computational capacity. By adopting DistilBART, we aim to efficiently fine-tune a transformer model on the MediaSum dataset without compromising performance quality or summary coherence.

This experiment will serve as a foundational step to assess whether a compact model can handle large-scale dialogue summarisation tasks effectively, and it will guide future iterations involving either model scaling or dataset optimization.

**Observations and results:**

In this experiment, we attempted to fine-tune the DistilBART model on the MediaSum dataset as a lighter and more resource-efficient alternative to the full-scale BART model. The rationale was to mitigate the GPU memory issues encountered during the initial experiment by leveraging the smaller size and faster training capabilities of DistilBART.

However, despite the model’s reduced complexity, the experiment resulted in a similar outcome as the previous attempt. The training process was unable to complete due to GPU memory exhaustion, once again caused by attempting to load the entire MediaSum dataset into memory at once. While DistilBART did offer slight improvements in memory usage, the sheer volume of data in the dataset continued to overwhelm the available hardware resources.

This repetition of the issue reinforces the need to restructure the data pipeline to support mini-batch training or incremental data loading.

**Recommendations:**

To overcome the persistent GPU memory limitations encountered in the previous experiments, it is recommended that the next iteration involve training the **DistilBART model on a reduced sample of the MediaSum dataset**, specifically **10,000 records**. This subset size is expected to be manageable within current hardware constraints and will allow for the **successful completion of training**.

By using a representative sample, we can still evaluate key performance indicators such as ROUGE metrics, while significantly reducing computational overhead. This approach provides a practical pathway to validate the model’s capabilities and fine-tuning behavior before scaling up to the full dataset.

**Experiment 2: DistillBART Model training on a sample**

**Rationale:**

Given the recurring resource limitations encountered when training on the full MediaSum dataset, the next experiment will focus on training the DistilBART model using a subset of 10,000 records. This approach is designed to ensure that model training can be completed within existing GPU constraints, while still providing meaningful insights into performance and output quality.

Using a representative sample allows us to retain the diversity and complexity of the interview data without overloading system memory. This setup enables the evaluation of DistilBART’s summarisation capabilities—such as content accuracy, contextual understanding, and fluency—on a smaller scale, making it possible to generate measurable results, including ROUGE scores and qualitative assessments.

This experiment will act as a controlled proof of concept, allowing us to test and refine our training pipeline, validate model behavior, and establish a foundation for future scaling to larger datasets.

**Observations and results:**

This section outlines the key outcomes from each experimental phase of the project, focusing on the models’ performance in summarising interview transcripts from the MediaSum dataset. Four experiments were conducted: fine-tuning **DistilBART** (with two hyperparameter sets), **BART**, and **prompt-based summarisation using the ChatGPT API**.

**Experiment 2.1: DistilBART (Hyperparameter Set 1)**

The DistilBART model was successfully fine-tuned on a filtered 10,000-record subset of the MediaSum dataset. The smaller dataset helped complete training efficiently within GPU limitations.

**Evaluation Metrics (Set 1):**

| **Metric** | **Average F1 Score** |
| --- | --- |
| ROUGE-1 | 0.1206 |
| ROUGE-2 | 0.0422 |
| ROUGE-L | 0.0991 |

The model showed basic lexical overlap (ROUGE-1) but struggled with phrase continuity and semantic matching (ROUGE-2 and ROUGE-L). These results were expected, given that DistilBART was pre-trained on the CNN/DailyMail dataset, which differs in tone and structure from MediaSum’s dialogue-heavy format.

**Human Evaluation:**

When tested on an unseen transcript, the generated summary lacked detail, with key themes either omitted or misrepresented. This reinforced the model’s limitations in out-of-domain summarisation without extensive fine-tuning.

**Experiment 2.2: DistilBART (Hyperparameter Set 2)**

A second training run was conducted using a modified set of hyperparameters. Training was again successful, and slightly improved results were observed.

**Evaluation Metrics (Set 2):**

| **Metric** | **Average F1 Score** |
| --- | --- |
| ROUGE-1 | 0.2182 |
| ROUGE-2 | 0.0961 |
| ROUGE-L | 0.2018 |

This version showed enhanced performance across all metrics, indicating improved contextual alignment and summary quality. However, human evaluation remained consistent with earlier findings: summaries were fluent but lacked completeness and thematic depth.

**Recommendations for Next Experiment:**

Following the successful fine-tuning and evaluation of the DistilBART model on a 10,000-record sample from the MediaSum dataset, the next logical step is to apply the **same training strategy to the full-scale BART model**. With the data pipeline and resource management now optimized to handle the dataset efficiently, this experiment aims to leverage BART’s larger model capacity to potentially achieve higher summarisation accuracy and semantic fidelity.

Given BART’s superior representational power and its original design for abstractive summarisation tasks, we anticipate improved performance across ROUGE metrics, particularly in capturing contextual depth and sequence alignment. The same controlled dataset size (10,000 samples) will be used to maintain comparability with the DistilBART results, allowing for a meaningful evaluation of the trade-offs between model size and summarisation quality.

**Experiment 3: BART Model training on a sample**

**Rationale:**

Building on the successful fine-tuning of the DistilBART model, the next experiment will focus on applying the same training methodology to the full-scale BART model. While DistilBART demonstrated fluency and baseline performance, its limited capacity impacted the semantic fidelity and contextual depth of the generated summaries. BART, with its larger model architecture and richer representational capabilities, is expected to deliver improved performance, particularly in capturing nuanced content from dialogue-rich transcripts.

To ensure training efficiency and overcome prior GPU limitations, the experiment will again use a 10,000-record subset of the MediaSum dataset. However, in this iteration, an additional filtering step will be introduced: only samples with the lowest token counts will be selected. This approach reduces memory consumption during training and increases the likelihood of successful model execution without resource exhaustion.

**Observation and Results:**

n this experiment, the full-scale BART model was fine-tuned on a 10,000-record token-filtered subset of the MediaSum dataset. The training phase was successfully completed, demonstrating that the optimised data pipeline and reduced token count strategy effectively mitigated the earlier GPU memory constraints during model training.

However, the evaluation phase encountered a significant limitation. When attempting to run the model on the validation dataset, the GPU memory was exhausted before the process could be completed. Despite training and evaluation being attempted across multiple environments—including a system equipped with an NVIDIA RTX 4060 GPU and Google Colab Pro Plus—the available GPU memory consistently fell short by approximately 100–200 MB, preventing the validation from running to completion.

This challenge underscores the high memory demands of BART during inference, particularly in evaluation scenarios where batch-level decoding can be computationally intensive. The consistent failure across different platforms suggests that minor resource optimisations alone may not be sufficient, and future experiments may require either gradient checkpointing, reduced batch sizes, or dataset segmentation during validation to fully execute the model evaluation cycle.

Recommendations for Next Experiment:

Given the persistent GPU memory limitations during evaluation with large transformer-based models like BART, the next experiment will shift focus from model fine-tuning to prompt-based summarisation using the ChatGPT API. This approach removes the need for intensive computational resources and allows us to evaluate summary generation through prompt engineering.

We will structure the interaction with the model using a tailored prompt, positioning ChatGPT as a professional interview summariser. The prompt will instruct the model to read the input transcript, identify key themes, and generate a concise summary paragraph. The summarisation will be performed in a one-shot setting, leveraging the pre-trained knowledge and language generation capabilities of ChatGPT.

**Experiment 4: Prompt Engineering with ChatGPT API**

**Rationale:**

Due to the high GPU memory demands and recurring execution failures observed during transformer model evaluation, especially with BART, the next experiment will explore an alternative summarisation strategy that does not require model training or tuning.

By using the ChatGPT API, we can tap into a powerful, pre-trained language model capable of producing high-quality summaries via prompt engineering. This method significantly reduces the technical overhead, avoids memory bottlenecks, and allows for more agile experimentation. The prompt is carefully designed to reflect a real-world use case—summarising interviews with thematic clarity and brevity—mirroring our original project goals.

This experiment will help assess whether a large language model like ChatGPT, when given well-structured instructions, can match or surpass the output quality of traditional fine-tuned models, especially in terms of fluency, relevance, and contextual understanding. It also provides a path toward rapid deployment, making it a strong candidate for production use in media environments where speed and accuracy are equally critical.

Observation and Results:

This experiment utilized the ChatGPT API for summarising interview transcripts through prompt engineering, eliminating the need for model training. After testing multiple prompts, the following structure yielded the most effective results:

“You are a professional interview summariser. The following is the given interview transcript, identify key theme and generate concise summary. Output should just be the summary paragraph.”

Using this prompt, summaries were generated and evaluated using ROUGE metrics, with the following average F1 scores:

Metric Score

ROUGE-1 0.1874

ROUGE-2 0.0668

ROUGE-L 0.1592

These results showed improved performance over DistilBART, particularly in capturing content structure and main ideas. The summaries were fluent, contextually accurate, and generated without memory issues, making this a scalable and efficient alternative for summarising media interviews.