In this project, you will develop a machine-learning–based text summarization solution to automatically condense dialogue conversations into concise summaries. You'll work with the SAMSum dataset, which contains messenger-like conversations paired with their human-written summaries. Your task is to implement an encoder-decoder architecture using pre-trained BERT models and auto-regressive modeling with ChatGPT to generate accurate and concise summaries of these conversations.

Acme Communications, a leading messaging platform company, is facing a challenge with information overload in group chats. Users complain that important details get lost in lengthy conversations, making it difficult to catch up after being away. The product team has identified automated conversation summarization as a potential solution to improve user experience.

As a data scientist at Acme, you've been tasked with developing a proof-of-concept for a dialogue summarization feature. Your solution should be able to condense conversations into concise, accurate summaries that capture the essential information.

The business goals include:

* Reduce information overload for users.
* Improve user engagement by making conversations more accessible.
* Enhance the platform's capabilities with AI-powered features.

Your task is to develop a working prototype that demonstrates the feasibility of this feature, focusing on both technical performance and business value.

DELIVERABLES:

**Project Pitch Report (PDF, 3-5 pages)** including:

* Problem statement and approach
* Timeline
* Analysis of the dataset
* Description of model architecture
* Performance evaluation and results

**MVP Discussion**

In the data science workflow, receiving feedback on your approach and implementation is invaluable for improving your final product. For this discussion, you'll share your project at approximately 90% completion, with your core analysis, modeling, and initial conclusions in place, to receive constructive feedback from peers.

You will need to include the following in your discussion submission

* Project Brief
* Version 1 (Rough Draft) Project
* Written Reflection
* 2 Peer Project Reviews

**Jupyter Notebook**containing your complete implementation with:

* Code for data loading and preprocessing
* Model architecture implementation
* Training and optimization process
* Evaluation metrics and analysis
* Sample outputs and demonstrations

**Project Report (PDF, 3-5 pages)** including:

* Problem statement and approach
* Analysis of the dataset
* Description of model architecture
* Performance evaluation and results
* Challenges faced and solutions implemented
* Potential improvements and future work

**5-minute Video Presentation showcasing:**

* Overview of the problem and approach
* Demonstration of your solution
* Key findings and results
* Business implications and applications

INSTRUCTION

**Step 1: Dataset Exploration and Preperation**

* Load the SAMSum dataset and explore its structure.
* Analyze the characteristics of the dialogues and summaries.
* Prepare the data for input to the BERT and auto-regressive with ChatGPT models:
  + Implement appropriate tokenization.
  + Create training and validation splits.
  + Build data loaders for efficient model training.

**Step 2: Model Architecture Implementation**

* Implement an encoder-decoder architecture using BERT models and auto-regressive modeling with ChatGPT .
* Configure the model for the summarization task.
* Set up the necessary components:
  + Encoder (BERT-based)
  + Generation mechanism to include the decoder. A decoder example can be Chat GPT-2 or model on huggingface.   
    - Try to find a free model that will give you a proof-of-concept for text.
* Do the analog for the auto-regressive modeling with ChatGPT.

**Step 3: Training & Optimization**

* Implement the training loop.
* Set up appropriate loss functions and evaluation metrics.
* Configure optimization parameters.
* Implement early stopping and checkpointing.
* Monitor training progress.
* Manage computational resources effectively.

**Step 4: Evaluation & Analysis**

* Evaluate model performance using ROUGE scores.
* Analyze model outputs qualitatively.
* Compare generated summaries with reference summaries.
* Identify patterns in model successes and failures.
* Consider model limitations and potential improvements.

**Step 5: Documentation & Presentation**

* Document your approach and findings.
* Create visualizations to illustrate results.
* Prepare a concise presentation of your solution.
* Discuss potential applications and extensions.

CONSIDERATIONS

* The full SAMSum dataset is very large, so consider using a subset for initial development and testing.
* Training transformer models is computationally intensive, plan your resources accordingly.
* Balance training time with model performance.
* Consider implementing beam search or other decoding strategies for better summary generation.
* Document your design choices and the rationale behind them.

RESOURCES & TOOLS

**Datasets**

* [SAMSum dataset (available through Hugging Face datasets library)Links to an external site.](https://huggingface.co/datasets/knkarthick/samsum)

**Reference Materials**

* Research paper: ["SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization"Links to an external site.](https://aclanthology.org/D19-5409/)
* [Hugging Face Transformers documentationLinks to an external site.](https://huggingface.co/docs)
* [BERT and encoder-decoder architecture documentationLinks to an external site.](https://huggingface.co/docs/transformers/en/model_doc/encoder-decoder)