Capstone Project: Experiment With Various Models

High-Quality Text Summarization Using Large Language Models: Experiments with Prompt Tuning, Model Selection, and Performance Evaluation

This work demonstrates how to use large language models (LLMs) like GPT for generating high-quality summaries of long texts (e.g., original documents like TDoc). The process of generating summaries from the original text, formatted as a Word document, involves experimenting with different GPT models, prompt customizations, and parameter adjustments. Prompts and parameters are carefully tuned to prevent overfitting or underfitting of the LLM, ensuring balanced and relevant summaries.

Summary performance is evaluated using a range of metrics, including semantic similarity and n-gram-based similarity. The goal is to identify an effective prompt and parameter configuration that produces easy-to-follow summaries, with the right level of detail, and achieves high semantic scores (e.g., 8 out of 10).

Completed Tasks:

1. **Prepare LLM Input**: Convert the original docx file into text suitable for LLM input.
2. **Generate Summary**: Use various LLM models, such as GPT-4o, GPT-4o-mini, and GPT-3.5-turbo, to generate summaries from the input text.
3. **Customize Prompt and Avoid Overfitting/Underfitting**: Tailor the prompt for different roles (e.g., 'system', 'assistant', 'user') to ensure balanced model performance. Avoid overfitting by refraining from overly specific prompts or rigid constraints, and avoid underfitting with prompts that are too generic or lack focus.
4. **Analyze Parameter Effects**: Study how LLM parameters, like temperature, impact summary generation and tailor them for optimal performance.

5. **Evaluate Summary Performance**: Rate the summary using semantic similarity (e.g., GPT-4o's rating) and assess it with other metrics like BERT and ROUGE scores to compare generated summaries against the original text.
6. **Discussion and Future Fine-Tuning**: Review summary quality and explore data collection strategies to further fine-tune the model for improved performance.

Model performance: Semantic similarity of the proposed method

The proposed solution requires careful design of both the prompt and parameter values. The recommended prompt is available in the Google Colab.

Additionally, the performance of the solution across various ‘temperature’ values is illustrated below.

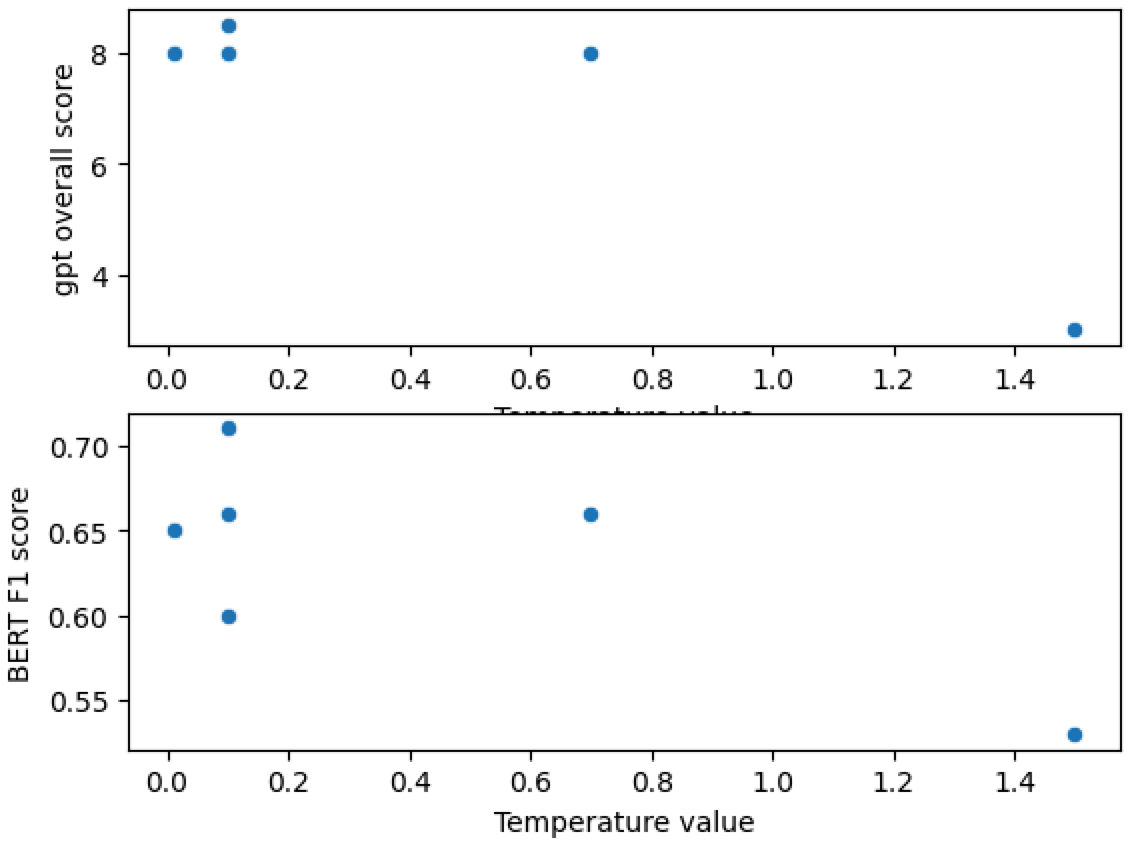

Fig. gpt semantic overall rating and BERT F1 score against the temperature parameter

Discussion on the proposed solution

Summary Generation: OpenAI's LLMs, such as GPT-4o, GPT-4o-mini, and GPT-3.5-turbo, have been tested for effectiveness in generating text summaries. Performance of these summaries was evaluated using GPT-4o semantic ratings, BERT score (another semantic similarity metric), and ROUGE score (an n-gram-based metric).

Docx File Processing for LLM Input: The original content in Word (docx) format is first converted to plain text. This text is then passed to the OpenAI API as input for summarization.

Model Impact: Among the tested models, GPT-4o generated the most accurate summaries. GPT-4o-mini struggled with prompt interpretation, while GPT-3.5-turbo encountered input token limitations (16,385 tokens max). Careful prompt design, including role specifications like system, assistant, and user, is essential. Lower temperature values (e.g., 0.01, 0.1, 0.7) are preferred for consistent, accurate summaries. Prompt quality is crucial to avoid overfitting or underfitting in the LLM.

Performance of the Proposed Approach: The LLM-based summarization approach achieved high semantic similarity scores, such as the GPT-4o rating and BERT score. Given the preference for abstract over extractive summaries in this project, semantic-based scores were prioritized. As anticipated, the non-semantic ROUGE score did not reflect summary quality as accurately, indicating that a semantic-based metric is more appropriate for evaluating the TDoc summarization task.

Final Model Selection: GPT-4o, paired with a well-constructed prompt, consistently produces high-quality summaries, achieving a GPT-4o semantic rating of 8 or higher and a BERT score of 0.6 or higher.

Future Fine-Tuning with User-Rated Summaries: Since reference summaries are currently unavailable, semantic scoring cannot rely on sample reference or human summaries. The proposed next step is to gather generated summaries with user ratings to create a dataset of user-rated summaries, which can then be used for further fine-tuning of the LLM.