



4 Hour AI Engineering Challenge

Objective:

Improve the performance of a given prompt used in a GPT-based text analysis tool. You will need to refine the prompt, develop a test suite to compare the old and new prompts, and provide an analysis of their testing.

Project Description:

Imagine you are working at a company called ExSum. ExSum is a tool that is mainly used in board meetings and its sole purpose is to create executive summaries of what has been said in these board meetings. This means that ExSum gets meeting transcripts as input and returns executive summaries back to its users.

Currently, ExSum uses a very simple prompt for generating executive summaries. ExSum received a lot of user feedback related to the quality of the AI generated summaries. A colleague of yours already noted down the most important insights from the feedback. Your goal is to improve the prompt so that you meet the insights generated from user feedback.

Main Challenges:

This challenge is divided into three sub challenges which aim to test your understanding of working with large language models and their usage for analytical production purposes.

Prompt Refinement

Description: You are provided with an existing prompt that is used to generate summaries of technical documents. Your first task is to propose improvements to the prompt to increase the relevance and conciseness of the summaries.

Task: Refine the prompt and explain your reasoning behind the changes made.

Current Prompt: Write an executive summary for the following meeting: {{TRANSCRIPT}}

Test Suite Development

Description: Develop a Python test suite that compares the performance of the old and new prompts. The suite should test the performance of the prompts given your evaluation criteria. You are responsible to create the evaluation criteria yourself.

Task: Implement a Python script that uses both the old and new prompts to generate summaries. Include evaluation criteria, how you created the evaluation criteria and what rationale you followed.



Analysis and Reporting

Description: Analyze the results generated by your test suite. Look for improvements or regressions in the performance of the new prompt.

Task: Write a brief report summarizing your findings. Include recommendations on whether the new prompt should replace the old one, supported by data from your tests. Also include limitations of your test suite and what you would improve if you have 40 hours instead of 4 hours.

Additional Material:

With this document you will find

- a "transcripts" folder containing a test dataset of 20 .txt files.
- a "Insights from User Feedback" document
- Secret API-key for Mingxi Zhang OpenAI access:
sk-proj-H7L0jx0s1B2NTNNv7Km5pjACBxcSo6hB64ZcG2PiIQpehz2gnhb2eHEndkoIJ
tX032VsBZ-L3IT3BlbkFJwXd2VkJgnNixBN2nL2ummZ3LdnZ7n8ElpVCQNJJpN_89q
EKel36ZxJkfXtzFwjRaaeyWqCR4A

Submission:

Please push your code as well as analysis to a GitHub repository. Use the README.md file for documentation of your decision making, assumptions that you made, limitations of the test suite, results and conclusions from your analysis.

In the follow-up interview, we will discuss your approach to the project, the decisions you made, and any challenges you encountered.

Please do not use more than 4 hours on the challenge.

Good luck, and we're looking forward to reviewing your work!