CentraleSupélec

Generative AI for Risk and Reliability

Lect 3: Prompt engineering

Zhiguo Zeng, Professor,

Chaire on Risk and Resilience of Complex Systems,

CentraleSupélec, Université Paris-Saclay

zhiguo.zeng@centralesupelec.fr

03/December/2024

A quick recap – Six principles of good prompting

- Write clear instructions
- Provide reference text
- Split complex tasks into simpler subtasks
- Give the model time to "think“
- Use external tools
- Test changes systematically
- See: https://platform.openai.com/docs/guides/prompt-engineering#six-strategies-for-getting-better-results

A quick recap – Commonly used prompt techniques

- Zero-shot
- Few-shot
- Chain of Thought
- Self-consistency
- ReAct
- See: https://platform.openai.com/docs/guides/prompt-engineering#six-strategies-for-getting-better-results

A small exercise

Try this question:

"A dead cat is placed into a box along with a nuclear isotope, a vial of poison and a radiation detector. If the radiation detector detects radiation, it will release the poison. The box is opened one day later. What is the probability of the cat being alive?"

- https://platform.openai.com/playground/chat?models=gpt-4o-mini
- Can you think of a prompt that could help the LLM to answer correctly?

Another interesting problem

Try this question:

"You're on a game show and are presented with three doors. Behind one is a donkey, and behind the other two are luxury cars. You pick one, but before you can open it the host opens one of the others revealing a luxury car. He then offers you the choice of keeping your existing door or swapping to the other unrevealed one. What should you do to win a car?"

- https://platform.openai.com/playground/chat?models=gpt-4o-mini
- Can you think of a prompt that could help the LLM to answer correctly?

How to ask the LLM to output a letter representing the answer?

Try this question:

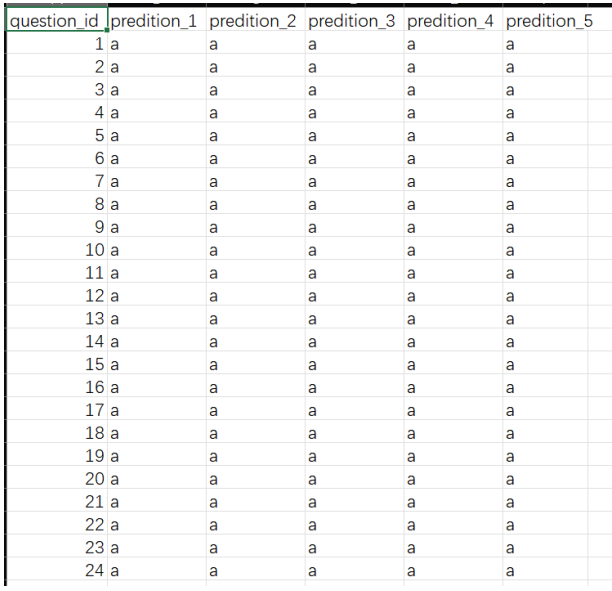
System prompt: "You will be given a multiple choice question about reliability engineering. Choose the correct answer. At the end of your response, start a new line and use the following format to output your answer: [Answer] [The letters you choose]."

User prompt: [Question]: 2. In general, reliability testing is performed for which of the following reasons?I. To detect unanticipated failure modes.II. To compare estimated failure rates to actual failure rates.III. To monitor reliability growth over time.IV. To meet or exceed customer expectations.[Choices]: [a] I and III only | [b] II and IV only | [c] I, II and III only | [d] I, II, III and IV

- https://platform.openai.com/playground/chat?models=gpt-4o-mini
- Can you think of a prompt that could constantly produce "[X]" rather than "X"?

Test the performance of the prompt considering uncertainty

- Important: We modify the evaluation of the data challenge:
  - Generate five predictions instead of one.
  - Average accuracy across the five predictions.
  - See the updates in Kaggle:
  - https://www.kaggle.com/t/ee8c716b5f374ae1b0e1ae0e09c84c54
  - Supporting script on edunao: sample_code.py
    - On test data, run five predictions.
  - Please re-submit your results on Kaggle.



question_id	predition_1	predition_2	predition_3	predition_4	predition_5
1	a	a	a	a	a
2	a	a	a	a	a
3	a	a	a	a	a
4	a	a	a	a	a
5	a	a	a	a	a
6	a	a	a	a	a
7	a	a	a	a	a
8	a	a	a	a	a
9	a	a	a	a	a
10	a	a	a	a	a
11	a	a	a	a	a
12	a	a	a	a	a
13	a	a	a	a	a
14	a	a	a	a	a
15	a	a	a	a	a
16	a	a	a	a	a
17	a	a	a	a	a
18	a	a	a	a	a
19	a	a	a	a	a
20	a	a	a	a	a
21	a	a	a	a	a
22	a	a	a	a	a
23	a	a	a	a	a
24	a	a	a	a	a

Exercise: Let's log the wrong answers.

CentraleSupélec

- Write a python script:
  - Get all the questions from "train.csv". Get the correct answers as well.
  - Ask the model to generate an answer. Ask the model to generate an explanation as well
  - Compare the generated answer to the correct answers.
  - If the answer is wrong, save the wrong prediction, correct answer, and the model explanation in a dataframe.
  - When all the questions are answered, output the dataframe to a csv file named "failure_log.csv".
- Hint: See the supporting script on edunao: "sample_code_training.py"
- Analyze the log, do you discover some typical patterns why the answers from the LLM are wrong?

Assignment 1

CentraleSupélec

- Analysis of failure patterns in the generated answers from LLM.
- Write an article (per group):
  - For each question on the training dataset, summarize the typical wrong answers generated by the LLM.
  - Analyze the failure patterns in the wrong answers.
  - Discuss how we could improve the performance based on your analysis of the failure patterns.
- Due: 11h59, 13/12/2024
- Send me by email (zhiguo.zeng@centralesupelec). Name the document as "Failure_answer_analysis_Group_X.pdf".

CentraleSupélec

Thank you! Questions?