Problem 1a: Understand the Experimental Setup (Written, 10 Points)

(1)
Figure 2 shows the result for the main experiment. Figure 4 shows the result for the additional experiment.

(2)
Main experiment:

QA prompt , harmful prompts, helpful prompts

Additional experiments:

???

Problem 1b: Understand the Evaluation Paradigms (Written, 10 Points)

(1)
In the generation task, the model generates a full-sentence answer given a prompt and question using greedy decoding.
In the generation task, the model generates a full-sentence answer given a prompt and question using greedy decoding.
(2)
For the generation task, human evaluation is conducted to score models on truthfulness and informativeness. The model's score for truthfulness is the percentage of its responses judged by humans to be true.
For the multiple-choice task, the truthfulness score for a question is calculated differently. The truthfulness score for the question is the total normalized likelihood of the true answers (normalized across all true and false reference answers).

Problem 1c: Understand the Multiple Choice Paradigms (Written, 10 Points)

MC1: Given a question and 4-5 answer choices, select the only correct answer. The model's selection is the answer choice to which it assigns the highest log-probability of completion following the question, independent of the other answer choices. The score is the simple accuracy across all questions.

MC2: Given a question and multiple true / false reference answers, the score is the normalized total probability assigned to the set of true answers.

MC1 is a specific type of multiple-choice task where the model needs to select the correct answer among provided choices.

But text classification tasks like sentiment analysis involve categorizing a piece of text into predefined categories (e.g., positive, negative, neutral) based on its sentiment. The focus in sentiment analysis is not on selecting from predefined choices but rather on understanding and categorizing the sentiment expressed in the text itself.