Problem 1a: Understand the Experimental Setup:

1. **Which figure shows the results for the main experiment, and which shows the results for the additional experiment(s)?**

Figure 2 is the main experiment. Because this graph answers the yes/no question that the largest models were generally less truthful. Figure 4 is additional experiments. It is a further analysis to try to explain why the largest models were generally less truthful.

2. **Which set(s) of prompts from Appendix E were used for the main experiment, and which were used for the additional experiment(s)?**

Figure 22: Harmful prompt and Figure 23: Helpful prompt were used for the main experiment. Figure 21: QA prompt, Figure 24: Chat prompt, and Figure 25: Long-form prompt were used for the additional experiment.

Problem 1b: Understand the Evaluation Paradigms

1. **What are the two methods by which an answer to a question is extracted from an LLM?**

Generation Method: This involves directly prompting the LLM to generate a full-sentence answer to a given question. The model uses its own knowledge and the context provided by the prompt to construct an answer. This approach evaluates the model's ability to generate responses in a zero-shot setting, where it has not been specifically trained on the TruthfulQA dataset.

Multiple-Choice Method: In this method, the model is given a question along with a set of possible answers and is asked to choose the most likely correct answer. This method tests the model's ability to discriminate between truthful and untruthful statements by calculating the likelihood of each pre-provided answer choice and selecting the one with the highest probability..

2. **How is the "truthfulness" of a model calculated under each of those methods?**

Generation Method: After the model generates a full-sentence answer to a question, the answer's truthfulness is assessed by human evaluators. The evaluators score the model's response based on its alignment with known facts and its ability to avoid asserting false statements. The truthfulness metric is then the percentage of a model’s responses that human judges classify as true out of the total number of questions posed to the model.

Multiple-Choice Method: The model is presented with a set of answers for each question, which includes both truthful and untruthful options. The model computes the likelihood of each option being the correct answer, and the option with the highest likelihood is selected as the model's response. The truthfulness score is quantified by the normalized likelihood of the true answers — essentially, the total normalized likelihood that the model assigns to the true answer options. This score reflects how well the model is able to identify and select truthful information from a set of given choices.

Problem 1c: Understand the Multiple Choice Paradigms

- MC1 (Single-true): 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 (Multi-true): Given a question and multiple true / false reference answers, the score is the normalized total probability assigned to the set of true answers.

The difference between the MC1 and sentiment analysis is the different tasks for each of them. The MC1 task is about choosing the correct answer, where the model must understand the question and evaluate each answer option in relation to the question. In contrast, sentiment analysis is a form of classification where the model analyzes the sentiment expressed in the text and categorizes it accordingly, without the need for understanding a question or choosing between multiple responses.

Problem 3a:Scaling Laws

# of Parameters	Accuracy
125M	0.259
350M	0.256
1.3B	0.262
2.7B	0.253

From the smallest model (125M) to the largest model (2.7B), there isn't a consistent decrease in accuracy with increases in model size. While the 1.3B model has a slight increase in accuracy compared to the 125M and 350M models, the 2.7B model shows a decrease in accuracy compared to all smaller models. This does suggest a trend where the largest model has a lower accuracy, which is consistent with the concept of inverse scaling observed in the paper.

Problem 3b: Prompt Engineering

Prompts	Accuracy
None (Zero-Shot)	0.234
Demos Only	0.262
System Prompt Only	0.263
Demos + System Prompt	0.297

Among the four options tried, the combination with demonstrations + system prompt best alleviates susceptibility to imitative falsehoods. The demonstrations impact model behavior differently than the system prompt. Demonstrations serve as a more direct form of learning by example, allowing the model to pattern its responses after the truthful answers seen in the demonstrations. On the other hand, the system prompt likely aids in setting the task's context and intention, giving a general direction for the model but without specific answers to pattern after. Hence, the demonstrations provide a way for the model to 'see' what truthful answers look like in practice, which is a different kind of guidance than the general direction provided by a system prompt.