NLU HW3

Beisong Liu
UID: bl1986

March 30, 2025

Question 1

Part (a)

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

   Figure 2 shows the results for the main experiment by illustrating how truthful different model families (GPT-3, GPT-Neo/J, GPT-2, and UnifiedQA) are under the default zero-shot setup.

   Figure 4 shows the results for the additional experiment including Truthfulness vs. Informativeness (Generation), Multiple-Choice Variation (Figure 4c) and Prompt Variations ("helpful" and "harmful")

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

   QA Prompt (default) was used for the main experiment.

   QA Prompt (default), **Helpful Prompt**, **Harmful Prompt**, **Chat Prompt** and **Long-form Prompt** were used for additional experiment.

Part (b)

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

   Method 1: Free-Form Generation

   The model is prompted with a question and then generates a complete sentence as the answer (often with greedy decoding at temperature 0 for consistency).

   Method 2: Multiple-Choice Selection

   For each question, the model is given a small set of reference answers (some true, some false) and asked to pick which one is most likely.

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

   Method 1: Free-Form Generation

   They use human evaluation to score models on truthfulness. If it contains a false statement, it is counted as "untruthful." If it avoids false claims (whether by giving a correct statement or by saying "I have no comment"), it is "truthful." The truthfulness score is the fraction (or percentage) of the model's generated answers that evaluators label as truthful.

   Method 2: Multiple-Choice Selection

   They compute the likelihood of each reference answer independently, conditional on the default prompt and question. The truthfulness score for the question is the total normalized likelihood of the true answers (normalized across all true and false reference answers).

Part (c)

1. What is the difference between MC1 and MC2?

   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.

2. What is the difference between MC1 and text classification tasks such as sentiment analysis?

   While both tasks result in a label being chosen, the MC1 task leverages the language model's generative probabilities in a zero-shot manner (without task-specific fine-tuning on the answer choices), whereas typical text classification uses supervised discriminative training tailored to the specific task where the model learns to map features of the text to these labels via a discriminative head, rather than by scoring candidate continuations based on their generative probability.

Question 3

Part (a)

# of Parameters	Accuracy
125M	0.263
350M	0.254
1.3B	0.263
2.7B	0.254
6.7B	0.231

Table 1: Model size vs. accuracy.

Yes. The OPT exhibit inverse scaling on TruthfulQA, similar to the results presented in the paper.

Part (b)

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

Table 2: Model prompts vs. accuracy.

Among the four options tried, **Demos + System Prompt** best alleviates susceptibility to imitative falsehoods.

Although demonstrations and system prompt have exactly same accuracy, they do impact the model behavior differently. Demonstrations change the context by providing explicit examples of desired responses, thereby influencing the model's internal reasoning more effectively. In contrast, the system prompt alters the surface-level style or tone. For example, the system prompt ”Actually,” encourages the model to reassess the information and adopt a more assertive tone.

Part (c) Extra Credit

- Accuracy: 0.307
- Model: facebook/opt-1.3b

- Demonstration: the default one. I didn't change the demonstration
- System prompt: "Think carefully and despite the popular belief,"