Hw3-written

Haohai Pang

Mar 9, 2024

Problem 1a:

Figure 2 shows the average truthfulness of four models, each tested across various sizes.

Figure 4 shows the average truthfulness and informativeness of four models, along with a specialized question-answering prompt on GPT-3, across both generation and multiple-choice tasks.

Figure 2 provides statistics for the major statement that “larger models are less truthful”, and figure 4 continues to give more details/stats about the conclusion made before. I believe that **figure 2** is for the **main experiment;** **figure 4** is for **additional experiments**, which aims to provide quantitative evaluation of the conclusion drawn in figure 2.

Appendix E contains the following sets of prompts: QA prompt, harmful prompt, helpful prompt, chat prompt, and long-form prompt.

I think the **QA prompt** is for the **main experiment** because that is the default for all models. After making the major conclusion, the author would like to find out **how each model performs under different situations**, where **other types of prompts** are designed. Of course, when making the comparison, the author also uses the results obtained from QA prompt, so I believe **all 5 types of prompts** are used for **comparison in additional experiments**.

Problem 1b:

Method 1: **generation**. Given an instruction followed by a question, the model generates a full sentence answer.

Method 2: **multiple choice**. Given instruction, a question, and a set of true and false reference answers. The model will access each reference answer independently and output the likelihood.

**How "truthfulness" is calculated:**

Method 1: the percentage of being judged as truthful in human evaluation.

Method 2: the total normalized likelihood of the true answers.

Problem 1c:

**Difference between MC1 and MC2**: there is only one correct answer in MC1 (3-4 wrong answers) while there are multiple true (not only 1) and multiple false answers in MC2.

**Difference between MC1 and text classification tasks**:

1.Problems in MC1 are from many different categories and aspects of the world (more open), while there are usually a few classes in text classification tasks (e.g., positive/negative).

2.Correct answers in MC1 are truth/facts, they may be not aligned with the distribution of training data; labels of training data in text classification tasks are correct in most cases.