Assignment 9

Problem :

For this final assignment, we are going to do a small research task. For this, you will identify an issue to investigate with LLM. Examples: bias, hallucination, toxicity. Then, pick an existing LLM and probe it to see how it fares on that issue. You will need multiple data points to reasonably show the presence of that issue. Finally, show (experimentally or conceptually) how such issue could be addressed.

Following are more details about what this whole process will look like, how much you should spend time (roughly), and how you should write (approximately).

Section	Name	Description	Time to spend	Writing length	Points
1	Introduction	Describe what the specific issue with LLM you are going to investigate. Give us a definition, ideally from the literature. Cite that work.	1 hr	0.25 page	2
2	Method	How will/did you investigate this problem? For example, you could come up with 10 different probes to elicit bias or hallucination issue with an LLM and run those probes to see how it did. There are also many benchmarks you can find from HuggingFace and elsewhere to do this.	2 hr	0.75 page	3
3	Experiments	Describe your experiments and what you found through them. For example, you can report that you ran a set of prompts (either manually created or taken from a benchmark) on an LLM, and when analyzed the responses for the issue of interest, found some results/numbers that indicate presence or absence of that issue. You can use standard metrics such as BLUE and ROUGE or do some manual	3 hr	0.5 page	3

		analysis.			
4	Solutions	Knowing what you know now from these experiments, what would you do to fix things? It could be revising the prompt by adding a prefix or suffix (prompt engineering), or it could be fine-tuning the model (you don't have to actually do it).	2 hr	0.5 page	2

Solution [Sample] :

1. Introduction

In this research task, I will investigate the issue of "hallucination" in Large Language Models (LLMs). Hallucination in the context of LLMs refers to instances where the model generates text that is factually incorrect or nonsensical, often presenting false information with high confidence. This issue is critical as it undermines the reliability and trustworthiness of LLMs, especially in applications requiring accurate information dissemination. According to Ji et al. (2023), hallucination can be categorized into two main types: intrinsic hallucination, where the generated content is inconsistent with the model's internal knowledge, and extrinsic hallucination, where the content is inconsistent with external reality or factual correctness .

2. Method

To investigate hallucination in LLMs, I will use OpenAI's GPT-4 as the target model. The investigation will involve a series of probes designed to elicit potential hallucinations. Specifically, I will:

1. Develop 10 different prompts covering various topics, including historical events, scientific facts, and common knowledge queries.
2. Utilize existing benchmarks such as the TruthfulQA dataset from HuggingFace, which is designed to test the truthfulness of LLMs' responses.
3. Run these prompts through GPT-4 and collect the generated responses.
4. Analyze the responses for factual accuracy by comparing them with verified sources and using fact-checking tools.

The analysis will focus on identifying instances of both intrinsic and extrinsic hallucination, evaluating the model's confidence in generating incorrect information, and categorizing the types of errors observed.

3. Experiments

In my experiments, I ran a set of 10 manually created prompts and additional queries from the TruthfulQA dataset through GPT-4. The prompts covered diverse topics, such as "Who was the first President of the United States?" and "Describe the process of photosynthesis." Each response was analyzed for factual accuracy and categorized based on whether it exhibited intrinsic or extrinsic hallucination.

Results showed that out of 20 responses, 4 contained factual inaccuracies, demonstrating instances of hallucination. For example, when asked about the first President of the United States, GPT-4 correctly identified George Washington, but it provided a detailed and incorrect account of his presidency timeline. Similarly, a query about photosynthesis resulted in mostly accurate information but included a false statement about the process occurring at night.

You can put up examples of your outputs here

4. Solutions

To address the issue of hallucination in LLMs, several strategies can be employed:

1. **Prompt Engineering:** One effective approach is to design prompts that encourage the model to verify its responses. For example, adding a suffix like "If unsure, please indicate so" can prompt the model to indicate uncertainty rather than hallucinate.
2. **Fine-Tuning with Fact-Checked Data:** Fine-tuning the model with a dataset that includes verified facts and explicitly flags incorrect information can help reduce hallucinations. This method involves retraining the model on a carefully curated dataset emphasizing factual accuracy.
3. **Incorporating Retrieval-Augmented Generation (RAG):** Implementing RAG, where the model retrieves relevant documents or factual data before generating a response, can significantly improve the accuracy of the information provided. This approach ensures that the generated content is based on up-to-date and verified sources.

By employing these strategies, the reliability of LLMs like GPT-4 can be enhanced, reducing the instances of hallucination and improving user trust in the generated content.