UNIVERSITY OF INFORMATION TECHNOLOGY AND SCIENCES

Course Name : Industrial Attachment

Course Code : CSE 420

Student Name : Md Shariful islam sajib sarker

Student ID : 2125051016

Batch : 50

Section : 7A1

Semester : Autumn

Seminar Topic : prompt Engineering and Jailbreaking LLM Models

Keynote Speaker : Raihan Alam, Principal Software Engineering, Microsoft

prompt Engineering and Jailbreaking LLM Models

Introduction:

The seminar titled "Prompt Engineering and Jailbreaking LLM Models" was an insightful exploration into the complexities of large language models LLM like chatgtp. The event covered how these models operate, why they sometimes make unusual or unexpected errors, and how prompt engineering can impact their responses. There was also an interesting discussion on jailbreaking these models finding ways to bypass their limitations or alter their responses.

Keynote speaker background:

The seminar was led by Raihan Alam, Principal Software engineer at microsoft, who has extensive experience with artificial intelligence. he has worked closely with AI driven systems, developing strategies to make interactions with them more intuitive. His expertise in both the practical and ethical aspects of LLM made him an ideal speaker for this session.

Purpose and Goals of the Seminar:

1. Providing an understanding of how LLM generate responses and why they can sometimes fail at tasks like multiplication.
2. Teaching attendees the concept of jailbreaking LLM and the ethical considerations around it.
3. Introducing a game like approach to prompt engineering to encourage creative exploration of AI responses.

Key Points and Discussions:

1. **Understanding errors in llm**: presenter explained that LLM, like chatgtp may make basic errors in tasks such as multiplication because they rely on statistical prediction rather than true mathematical reasoning. He compared the process to how a child might

try to predict an answer based on previous patterns rather than understanding the exact calculation.

2. **Jailbreaking models**: he introduced the concept of jailbreaking LLM explaining how prompt engineering could be used to manipulate or alter the model's responses. this involves giving prompts designed to sidestep certain rules or limitations that the model might have highlighting both the creative possibilities and the ethical responsibilities of using this technique.
3. **Prompt engineering as a game**: To make prompt engineering more accessible Mr. Alam presented it as a kind of game with LLM where users can test different prompts to see how the model's output changes. This interactive approach encouraged attendees to explore AI capabilities and limitations creatively while staying within ethical guidelines.

New information or skills gained:

The seminar provided a better understanding of how prompt engineering can impact llm responses and the types of prompts that might lead to unexpected answers. we learned practical methods for experimenting with AI prompts and observed how LLM process language predictively rather than logically, which sheds light on why they sometimes produce surprising errors in tasks like multiplication.

Bridging the gap between academic learning and industry practices:

The seminar highlighted an interesting gap between academic training and real world ai applications. Traditional academic learning often assumes that models like chatgtp are inherently logical, while in reality, they are probabilistic and sometimes unpredictable. This seminar emphasized the importance of understanding these distinctions to work effectively with AI in an industry setting showing that practical knowledge of prompt engineering is just as essential as theoretical knowledge.

Insights for improving academic programs:

1. **Ethics of AI manipulation**: Including discussions on the ethics of jailbreaking and other model manipulation techniques would help us understand the broader implications of their work.

2. **Hands on AI experimentation**: Providing students with tools to experiment with LLM in a controlled environment would give them a better understanding of how these models operate in real world applications.

Seminar summary:

The Prompt Engineering and Jailbreaking LLM Models seminar gave attendees a comprehensive view of the inner workings of LLMs like chatgtp. Through presenters insights, participants gained a better understanding of how to structure prompts, the limitations and challenges of predictive AI, and the potential of prompt engineering to customize responses.

Final thoughts and overall significance:

This seminar was highly informative especially for those looking to work with LLMs in both technical and ethical capacities. It highlighted the unique challenges and opportunities in prompt engineering and the growing significance of understanding AI predictive limitations. this approach encouraged participants to think critically and responsibly when working with advanced AI technologies offering a well rounded perspective on both the technical and ethical aspects of prompt engineering. This seminar emphasized that AI is a tool that, when used thoughtfully, has immense potential for innovation and positive impact.