DAY 5: RISKS AND LIMITATIONS OF CHATGPT

Introduction

ChatGPT can introduce false information to copyright infringement, there are factors to consider before integrating it into our lives.

It does not truly think about the words it is producing because it's a trained language model, not a person.

large language models rely on word co-occurrence in their training data to guess what is the best thing to say next. So it is able to guess context by word association and not from the meaning of the words.

Misinformation, offensive content, and data breaches are serious threats to people relying on ChatGPT for content. Which could cause reputational harm or even legal liabilities in the right situation.

———————————————————————————————

ChatGPT and Misinformation

While ChatGPT can be a valuable tool for answering questions and generating text, it also has the potential to spread **misinformation.**

While the internet houses much of the world's information, ***not all of that information is true***. Any information coming from ChatGPT should be verified before being relied on.

Remember, ChatGPT doesn't actually understand the information it is writing! It generates the next word it thinks is the most likely to appear after what it has written so far.

sometimes AI will literally ***make up information***, such as a **hallucination**. Lacking nuanced knowledge, it simply produces a response that it deems likely to occur next.

———————————————————————————————

ChatGPT and Disinformation

ChatGPT can sometimes produce wrong information on its own, but it can also be

used to produce false information on purpose, such as **disinformation** intended to mislead people.

Consider people using ChatGPT to:

- Create many fraudulent positive or negative reviews
- Write many different articles all repeating the same false narrative
- Post many variants of the same comments on people's videos

Fraudulent content produced by ChatGPT would outpace true content produced by real people. They can take advantage of people who may believe that ChatGPT is smarter than people.

Researchers are trying to correct these problems within ChatGPT.

———————————————————————————————

ChatGPT and Bias

The internet and literature, which ChatGPT is trained on, contain many harmful biases. As a result ChatGPT can replicate these biases in its output.

Asking ChatGPT to produce descriptions of people in certain fields, or with certain traits can result in stereotypical responses. ChatGPT updates have closed many of the ways to do this, but people continue to find new prompts that get past its safeguards.

It perpetuates false harmful stereotypes, and, as mentioned in the last exercise, ChatGPT can act as a propaganda machine for bad actors.

Removing all bias from information is an open problem in creating large language models.

———————————————————————————————

ChatGPT and Data Security

Like any data-heavy application, ChatGPT is at risk for data breaches or misuse of user data. It is important to be careful when uploading sensitive information.

———————————————————————————————

Review

As with any technology, there is a need for caution and awareness when using ChatGPT or any language model.

- **Misinformation**: The tendency for ChatGPT to produce incorrect information
- **Disinformation**: The ability for ChatGPT to be used to produce information intended to mislead people
- **Bias**: ChatGPT presenting biases based on the training data it has received
- **Data Security**: The security and privacy issues stemming from the data ChatGPT is collecting

It is important to understand these limitations and to use ChatGPT responsibly to avoid spreading misinformation or causing harm.