Common sense in LLM's

Introduction

In recent years, AI has achieved remarkable feats, but these accomplishments are primarily attributed to large language models (LLMs), which come with significant challenges. The high cost of training LLMs limits access to a few tech companies, concentrating power and raising concerns about safety and accessibility. I came across a Ted talk on Youtube "**Why AI Is Incredibly Smart and Shockingly Stupid | Yejin Cho**, and she highlights the pitfalls of these LLM's. When developing this AI assistant, it is important to also consider the social aspect of the these systems. I will be diving into that ted talk and Yejin Cho research paper: **Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models**, where she touches upon the Theory of Mind.

Why AI Is Incredibly Smart and Shockingly Stupid

Yejin Choi, a respected AI researcher, emphasizes the absence of common sense in LLMs, despite their intelligence. These models often make simple mistakes that even children can avoid. Relying solely on brute-force scaling is not an optimal solution.

Choi outlines three societal challenges posed by extreme-scale AI models: concentration of power, lack of transparency, and the need for common sense. She advocates for democratizing AI, making it smaller, safer, and imbued with human norms and values.

To address these challenges, Choi proposes drawing inspiration from the David and Goliath story and adopting a strategic approach: scrutinizing AI, posing fundamental questions, and developing innovative data and algorithms. This approach can overcome the limitations of large language models, ensuring intelligent, ethical, and reliable AI systems.

Common sense is essential in AI development, as demonstrated by thought experiments where AI maximizes paperclip production without understanding human values. Although teaching common sense was once deemed impossible, recent advancements in LLMs offer both discouragement and hope, paving the way for innovative solutions.

Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models

1. **Theory of Mind (ToM) in AI**: The paper discusses the concept of ToM, which refers to the ability to attribute mental states to oneself and others, understanding that others have beliefs, desires, intentions, and perspectives that are different from one's own. This is an important aspect to consider when developing an AI assistant, as it would enhance the assistant's ability to understand and respond to user queries in a more human-like manner.
2. **Evaluation and Benchmarks**: The paper emphasizes the need for robust benchmarks to assess the ToM abilities of Large Language Models (LLMs). It suggests that these benchmarks

should be designed specifically for LLMs rather than using tests designed for humans. It also mentions that the performance of LLMs on these benchmarks can vary greatly depending on the type and complexity of the questions. Therefore, it's important to evaluate the AI assistant's performance across a variety of tasks and scenarios.

3. **Difficulty Level of Datasets**: The paper suggests that the difficulty level of a dataset can be evaluated by calculating the final score across different splits of the dataset. The difficulty level can then be determined based on the lowest score obtained among these splits. This could be a useful approach when training and testing your AI assistant.

4. **APIs and Funding**: The paper acknowledges the use of APIs, such as GPT-4 and AI21, and mentions funding sources for the project. This highlights the potential need for resources, both technical and financial, when developing an AI assistant.

5. **Existing Benchmarks & Variants**: The paper discusses several datasets used for testing ToM in AI, including Triangle COPA, SocialIQa, and ToMi. These datasets could potentially be useful resources for training and testing your AI assistant.

Conclusion

In conclusion, Yejin Choi's research emphasizes the challenges and potential solutions related to large language models (LLMs) in AI. The absence of common sense in LLMs, despite their intelligence, raises concerns about their reliability. Choi advocates for smaller, safer AI systems imbued with human norms and values to address the concentration of power and limitations of LLMs. By scrutinizing AI, posing fundamental questions, and innovating data and algorithms, the limitations of LLMs can be overcome. Theory of Mind (ToM) is crucial for developing AI assistants that understand user queries in a human-like manner. Evaluating ToM abilities through specific benchmarks and considering dataset difficulty levels are essential. Availability of APIs and funding sources highlights the technical and financial resources needed for AI assistant development. Leveraging existing benchmarks and datasets enhances training and testing of social reasoning abilities. By incorporating common sense, transparency, and democratization, we can create intelligent, ethical, and reliable AI assistants.

References

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models.