Challenges of Large Language Models(LLMs)

Aakash Roy
aakashroyiitb@gmail.com
24nd October, 2024

Generative AI applications, driven by large language models (LLMs), are increasingly adopted across industries despite high operational costs and evident limitations. While these models have demonstrated significant potential, their lack of true intelligence and reasoning capabilities along with high energy consumption highlights the need for continued research to develop reliable and efficient AI.

LLM Agents

Effective collaboration, a principle rooted in Game Theory, is essential for addressing complex tasks. This principle suggests breaking complex tasks down into smaller, simpler subtasks and assigning them to domain-specific teams, where each team member specializes in that particular field. Current Gen AI applications follow a similar design, calling these specialized team members"agents."[1]
When team members at any level, are assigned a task, they further divide it into smaller subtasks and create an execution plan. They use various tools and skills as needed; for example, developers engage in deep-level thinking by selecting relevant articles and libraries, reasoning through them, and deriving new insights that support their hypothesis to accomplish the task effectively. These same steps are also expected from LLM agents to actually execute complex tasks.

But, LLMs can't reason!

Apple recently published the paper "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models"[2] which concluded, *"We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data"*. Which means it's a new form of database for retrieving information by generating. If so, then they can be finetuned for specific use cases. Try "How many r are there in there"

Specific Finetuning

Imagine deep neural networks as segmented structures where specific prompts activate different areas—for instance, healthcare prompts engage different segments than Maths prompts. Now, consider a 70-billion-parameter LLM, which you aim to fine-tune for a healthcare chatbot. Using *Parameter Efficient Fine Tuning Techniques (PEFTs)*[3] like *Low-Rank Adaptation (LoRA)*[4], all weights are updated, making the chatbot closely mimic the fine-tuning documents.
But what if we could fine-tune only the parameters associated with healthcare-specific tasks? This would significantly reduce the parameters that need to be updated, resulting in more efficient fine-tuning, eventually consuming less energy.

Training Data is now a constraint

Current generative AI models, including large language models (LLMs), rely heavily on extensive datasets to achieve high-quality outputs. However, experts forecast a potential plateau in available data for training purposes in the near future. While expanding the dataset has historically led to improved model performance, relying solely on data scaling may not be sustainable or efficient in the long term. Consequently, there is an increasing need to investigate advanced architectures that go beyond existing frameworks. Such architectures should not only incorporate attention mechanism[5] but also introduce innovative techniques that infer, deduce, and contextualize knowledge across multiple dimensions. By synthesizing new data insights from existing resources, these models could effectively generate richer training data, refining their own learning process without solely depending on exhaustive data use.

The objective shifts from using all available information indiscriminately to developing specialized, expert AI systems that can selectively focus on pertinent information. These systems would leverage both structural advancements and selective data utilization to achieve robust performance.

Energy Limitations

According to a recent Forbes article[6], “Arm’s executives also see data center demand rising significantly: CEO Rene Haas said that without improvements in efficiency, "*by the end of the decade, AI data centers could consume as much as 20% to 25% of U.S. power requirements. Today that’s probably 4% or less.*" ”
He also mentioned that “Nvidia’s upcoming Blackwell generation boosts power consumption even further, with the B200 consuming up to 1,200W, and the GB200 (which combines two B200 GPUs and one Grace CPU) expected to consume 2,700W. **This represents up to a 300% increase in power consumption across one generation of GPUs with AI systems increasing power consumption at a higher rate**.”
Since LLMs are becoming a necessary part of us, this issue needs to be addressed as soon as possible.

LLMs Hallucination

Although it consumes a lot of energy, the significant challenge faced by industries currently during deployment is hallucination[7]. Hallucinations occur when models produce factually incorrect or fabricated information, diverging from the knowledge in their training data. This issue is especially critical in fields like medicine, where accurate information is essential, and the risks associated with misinformation are high. Most companies working in diverse fields are already ready with their version of Gen AI applications, but it’s taking time for beta testing before handing it over to the masses. There are several other issues including malicious prompt injection for retrieving Public Health Information from training data of LLMs, also hallucinating it to give out pieces of information like making guns, bombs etc. which is just one touch away from children.

Conclusion

Although Large Language Models (LLMs) have an uncountable number of flaws but have transformed our way of interacting with electronic devices. People out there do code, summarization, explanations of complex theorems or ask to write a poem that would make a day better using GPT. It has truly become a companion. At the same time, it has shown the power of data and algorithms which can result in human-like behaviour, which, if nurtured, can become smarter and more helpful. Lastly, we’re in the era of the overflow of information and handling them, organizing their facts and reasoning on top of that is the next big step required to support intelligence.

References

[1] Talebirad, Y., & Nadiri, A. (2023). Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.03314

[2] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024, October 7). *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models*. arXiv.org. https://arxiv.org/abs/2410.05229v1

[3] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235. https://doi.org/10.1038/s42256-023-00626-4

[4]Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2106.09685

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

[6] Beth Kindig, “AI Power Consumption Rapidly Becoming Mission Critical,” Forbes, June 20, 2024. Available at: https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

[7] G. P. Reddy, Y. V. Pavan Kumar and K. P. Prakash, "Hallucinations in Large Language Models (LLMs)," 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2024, pp. 1-6, doi: 10.1109/eStream61684.2024.10542617.