Introduction to Large Language Model

Q1: What is Large Language Model?

Large Language Model (LLM) is the latest development stage of language modeling (LM), which attempts to "model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens [1]" in computers. In other words, LLM is the state-of-the-art approach to enable machines to read, write and communicate like humans [1].

Being the latest development stage of LM, LLM is indeed an integration of the previous approaches to LM, which includes Statistical Language Model (SLM), Neural Language Model (NLM), and Pre-trained Language Model (PLM). SLM is an approach based on the idea of predicting the next word given the previous context, where NLM is based on the idea of utilizing neural networks, and PLM is based on the idea of pre-training the models on massive datasets before fine-tuning it for specific downstream tasks. In general, LLM grasps all these three main ideas and emerges surprising abilities after being scaled up in its model size.

In conclusion, LLM is the newest and the most integrated method to enable machines to understand human languages.

Q2: Why Large Language Model?

As mentioned earlier, language modeling (LM) attempts to enable machines to understand human languages. Along with the development stages in LM, the extent to which models can interpret human languages developed dramatically. During the Statistical Language Model (SLM) stage, models can only assist in specific language tasks like speech recognition, machine translation, text suggestions, etc. [2]. In the Neural Language Model (NLM) stage, models can perform typical Natural language processing (NLP) tasks like translation and speech recognition tasks [3]. Then, in the Pre-trained Language Model (PLM) stage, models can solve various NLP tasks like document intelligence, content creation, virtual assistant, and intelligent search engines [4]. Currently, in the LLM stage, models emerge increasing capabilities in solving real-world tasks as their model sizes are being scaled up. For example, tasks like text generation and code generation can be completed by LLMs given the real-world context [5].

In conclusion, LLM enables machines to understand and use human languages to not only solve NLP tasks, but also real-world tasks that could not be solved by previous LM approaches. Additionally, due to the emergent abilities in solving complex tasks discovered in LLMs after their model sizes being scaled up, the scaling effect on model capacities is being researched in the current development of LLMs.

Q3: What are the State-of-the-art LLMs?

The four most well-known state-of-the-art LLMs are o1-preview (OpenAI), Llama 3.1 (Meta), Gemini 1.5 Pro (Google), and Claude 3.5 Sonnet (Anthropic) [6].

According to Artificial Analysis, an LLM leaderboard, o1-preview from OpenAI has the best result in benchmarks like Reasoning & Knowledge (MMLU), Scientific Reasoning & Knowledge (GPQA), Quantitative Reasoning (MATH), Coding (HumanEval), and Communication (LMSys Chatbot Arena ELO Score). However, among the four state-of-the-art LLMs, the price per tokens of o1-preview is the highest, while the output speed is the second lowest. It is believed that the "long chain of thought before responding to the user [7]" in o1-preiview is the main cause of these results. According to OpenAI, o1-preview is "trained with reinforcement learning to perform complex reasoning [7]", such that the model can mimic how human think before answering a question. As a result, o1-preview has state-of-the-art performance on various benchmarks, but it is expensive to use and has a relatively slow output speed compared to the other three state-of-the-art LLMs.

In terms of context window and efficiency, Gemini 1.5 Pro from Google has a context window of 2 million tokens [8] while having the second lowest price per tokens. It is believed to be due to the Transformer and MoE architecture that Gemini 1.5 Pro is built on, where MoE models are learnt to "selectively activate only the most relevant expert pathways in its neural network [9]".

Llama 3.1 from Meta and Claude 3.5 Sonnet from Anthropic has their own characteristics: Llama 3.1 is a "open model" that not only it is free to use but also its model weights are available to the public [10]; while Claude 3.5 Sonnet is a relatively well-rounded model.

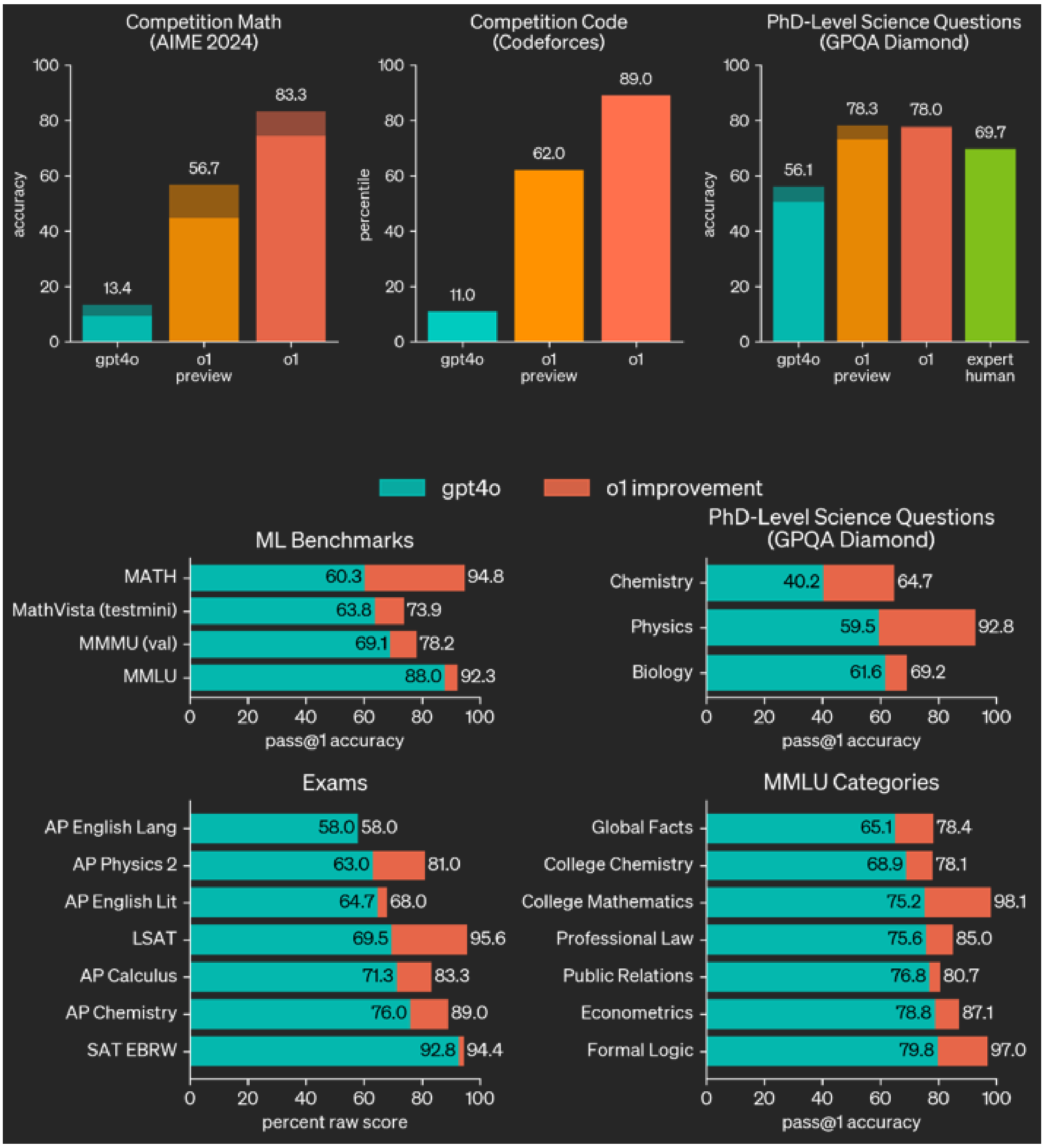

Figure 1: Performance Figures of o1-preview from OpenAI [11]

In conclusion, the state-of-the-art LLMs have minor differences in terms of performance in benchmarks, price per token, output speed, context window, etc. However, as shown in Figure 1, LLM that rivals the performance of human experts in many reasoning-heavy benchmarks [11] currently exists.

Q4: What are the Strengths and Limitations?

Being an approach to language modeling (LM), LLMs are designed to be capable of handling language generation tasks, including language modeling tasks (e.g., predicting the last word of sentences based on the given context), conditional text generation tasks (e.g., machine translation, text summarization, question answering, etc.), and code synthesis [1]. These areas, which only require basic language understanding and generation, are considered places where LLMs are expertise in.

As LLMs are scaled up, they present increasing ability in knowledge utilization [1]. They seem to be able to understand the knowledge carried out by human languages that they are trained in. Tasks like closed-book question answering, open-book question answering, and knowledge completion can be completed by LLMs to a certain extent. However, problems like hallucination (e.g., generating untruthful information) and knowledge recency (e.g., producing outdated information) occur sometimes in LLMs.

As LLMs are further scaled up, ability in complex reasoning has emerged from them. They seem to be able to utilize their knowledge and solve concrete problems step-by-step, especially when chain-of-thought prompting strategies are used correctly. Various tasks involving knowledge reasoning, symbolic reasoning, and mathematical reasoning [1] can be addressed by LLMs with human guidance. That being said, problems like reasoning inconsistency and numerical computational errors often occur in LLMs.

To conclude, LLMs have the best performance in tasks that only require basic language understanding. For tasks that require further language interpretation and even problem-solving skills, LLMs could return unintended output (especially for the more complex tasks).

Q5: Key Takeaways?

Although the fundamental concepts of LLMs are clear, the cause of those emergent abilities of LLMs that are beyond basic language understanding and generation remain mysterious. It is believed that as LLMs are scaled up, the models not only learn the human languages, but also the ideas that are carried by human languages. As a result, far more complex tasks compared to simple NLP tasks could be performed by LLMs.

However, different problems arise from LLMs when handling more complex tasks remind us that much more research is needed in addition to blindly scaling up the current state-of-the-art LLMs.

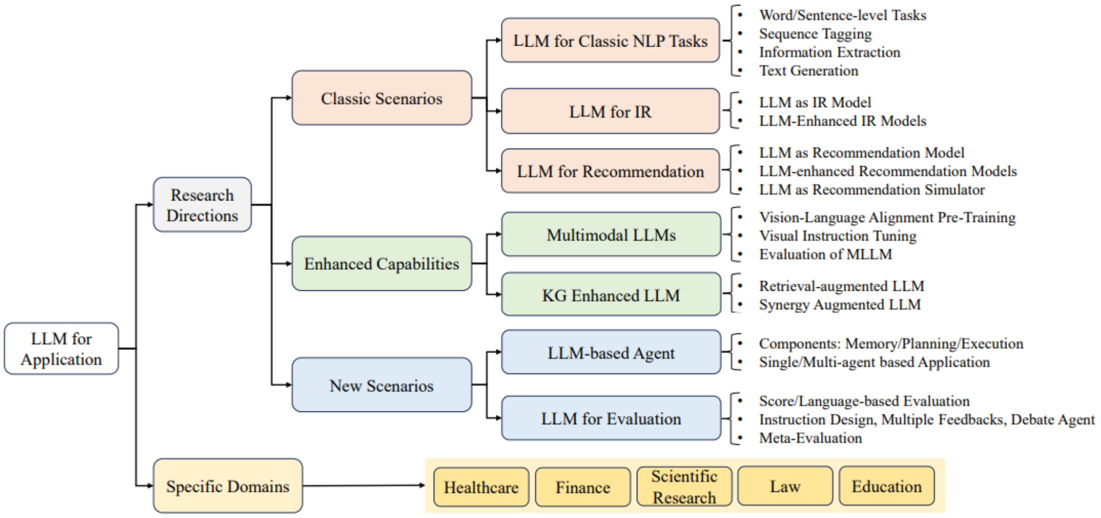

Figure 2: Applications of LLMs in representative research directions and downstream domains [1]

In summary, the field of LLM is currently still in the stage of development, but LLMs have already shown potential in solving tasks beyond NLP tasks. For example, LLMs are likely to be applied to various areas shown in Figure 2 in the future [1].

Reference

[1] W. X. Zhao *et al.*, “A Survey of Large Language Models,” *arXiv > Computer Science > Computation and Language*, Oct. 2024. Accessed: Nov. 10, 2024. [Online]. Available: https://arxiv.org/abs/2303.18223

[2] Engati Technologies Inc., “Statistical language modeling,” Engati, https://www.engati.com/glossary/statistical-language-modeling (accessed Nov. 10, 2024).

[3] K. Jing and J. Xu, “A Survey on Neural Network Language Models,” *arXiv > Computer Science > Computation and Language*, Jul. 2019. Accessed: Nov. 10, 2024. [Online]. Available: https://arxiv.org/abs/1906.03591

[4] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, “Pre-trained language models and their applications,” *Engineering*, vol. 25, pp. 51–65, Jun. 2023. doi:10.1016/j.eng.2022.04.024

[5] IBM, “What are large language models (llms)?,” IBM, https://www.ibm.com/topics/large-language-models (accessed Nov. 10, 2024).

[6] Artificial Analysis, “LLM leaderboard - compare GPT-4O, Llama 3, Mistral, Gemini & other models: Artificial Analysis,” Artifical Analysis, https://artificialanalysis.ai/leaderboards/models (accessed Nov. 10, 2024).

[7] OpenAI, “OpenAI O1 System Card,” OpenAI, https://openai.com/index/openai-o1-system-card (accessed Nov. 10, 2024).

[8] L. Kilpatrick, R. Kofman, and S. B. Mallick, “Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today,” Google Developers Blog, https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/ (accessed Nov. 10, 2024).

[9] S. Pichai and D. Hassabis, “Our next-generation model: Gemini 1.5,” Google, https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/ (accessed Nov. 10, 2024).

[10] Meta, “Introducing Llama 3.1: Our most capable models to date,” AI at Meta, https://ai.meta.com/blog/meta-llama-3-1/ (accessed Nov. 10, 2024).

[11] OpenAI, “Learning to reason with LLMS,” OpenAI, https://openai.com/index/learning-to-reason-with-llms (accessed Nov. 10, 2024).