Attention as a Driving Force in Large Language Model Advancements

Xiaoyan Li (xiaoyanli629@tsinghua.edu.cn)
Department of Automation, 30 Shuangqing Rd Haidian District,
Beijing, China 100084

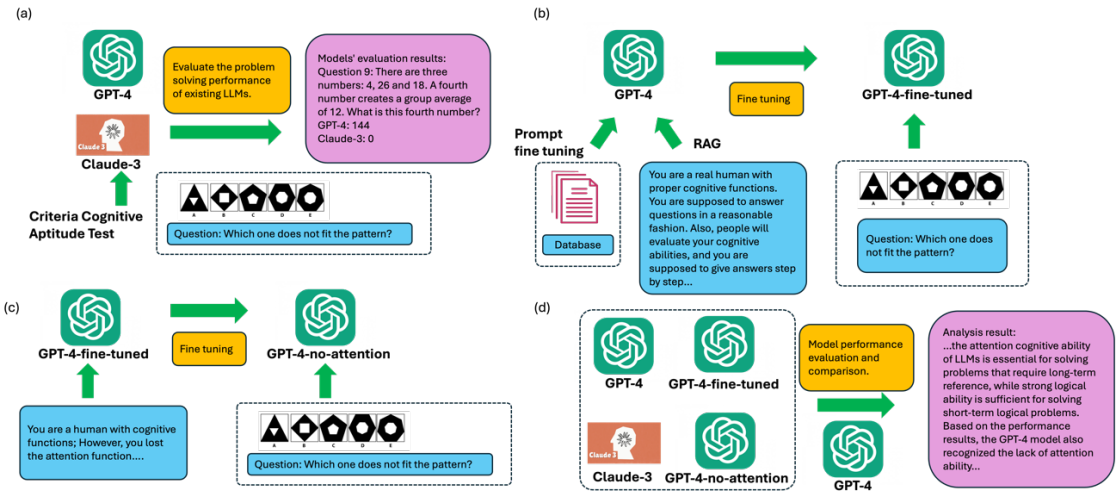

Figure 1: The overall experiment procedure

Abstract:

Large language models have developed rapidly in recent years and exhibited extraordinary problem-solving abilities. However, there is limited research on how attention abilities influence the problem-solving capacity of large language models. This study explores the intersection of cognitive neuroscience and large language models, focusing on the fine-tuning of these models to analyze how human cognitive abilities and disabilities affect the problem-solving function of large language models. Two GPT-4 based models were developed through prompt fine-tuning and retrieval-augmented generation. Results showed that the GPT-4-fine-tuned model achieved the highest accuracy (81.2%), while the model lacking attention performed poorly on questions requiring long-term inference. GPT-4's analysis recognized the lack of attention in the modified model, highlighting the importance of this cognitive ability in solving problems that demand long-term reference. This study sheds light on the mechanisms of problem-solving in the brain and the potential of AI to approximate human-like cognition.

Keywords: Large language model; cognitive ability; fine-tuning.

Introduction

The intersection of cognitive neuroscience and large language models represents a compelling frontier in the study of human cognition and artificial intelligence(Binz & Schulz, 2023; Dillion et al., 2023). Cognitive neuroscience seeks to understand how the brain gives rise to mental processes, such as perception(Farah, 2000), attention(Farah, 2000), memory(Gabrieli, 1998), language(Prystauka et al., 2023), and decision-making(Fellows, 2004). In parallel, large language models, powered by advancements in machine learning and natural language processing, have demonstrated remarkable proficiency in understanding and generating human language(OpenAI et al., 2023) and have been utilized in other areas(Holmes et al., 2023; Thirunavukarasu et al., 2023; Wu et al., 2023). This convergence presents an unprecedented opportunity to explore the neural underpinnings of language processing, cognition, and communication.

To date, several studies have integrated large language models into the framework of cognitive neuroscience, exploring how these AI systems align

with human cognitive processes. However, current studies are primarily focused on analyzing the cognitive functions of large language models(Binz & Schulz, 2023; Hagendorff et al., 2023; Suri et al., 2024). Notably, there is a gap in the literature regarding the fine-tuning of large language models to analyze human cognitive disabilities. In this study, we propose a framework to fine-tune the large language model to mimic persons with cognitive disabilities and evaluate the cognitive performance of the fine-tuned model. Ultimately, we employed the large language model to analyze the performance of the fine-tuned versions. This research aims to shed light on fundamental questions about the nature of human cognition, the mechanisms of problem-solving in the human mind, and the potential of AI to simulate or approximate human-like cognitive abilities.

Methods

Dataset and Models

The Criteria Cognitive Aptitude Test (CCAT) is an evaluation tool specifically crafted to gauge individuals' general cognitive abilities, encompassing problem-solving skills, critical thinking, and the capacity to acquire and apply new information. The CCAT encompasses various question types, including spatial reasoning, verbal ability, as well as mathematical and logical problem-solving. GPT-4 and Claude-3 models were tested in this study.

Experiment steps:

The overall experimental procedure is illustrated in Figure 1.

Step 1: We assessed the performance of the GPT-4 and Claude-3 models using the CCAT dataset. Subsequently, we calculated their scores and evaluated their performance on the test set. Visual questions were processed using the model's multimodal function, by uploading the image directly to the model. To eliminate the influence of previous inputs, all questions were tested individually.

Step 2: Two LLM models were developed through prompt finetuning, and retrieval-augmented generation (RAG) based on the GPT-4 base model. To gauge the enhanced reasoning of the bot, we utilized the prompt fine-tuning as follows: "You … cognitive functions. You are supposed to answer questions in a reasonable fashion..." Additionally, to evaluate cognitive function ability, we added the prompt: "…you lost the attention function… " Subsequently, the performance of these two models was tested on the CCAT dataset.

Step 3: In this stage, we employed the GPT-4 to analyze the performance of these models based on their CCAT test results.

Results & Discussion

Based on the evaluation results in Table 1, the GPT-4 model achieved 75% accuracy, while Claude-3 achieved 56.2% accuracy. The Claude-3 model failed most of the visual reasoning problems, indicating that it lacks multimodal ability. It is also worth noting that the GPT-4 model failed question 9: "There are three numbers: 4, 26, and 18. A fourth number creates a group average of 12. What is this fourth number?" Based on the output, the GPT-4 model is capable of calculating that the sum of the four numbers. However, it lost logic in the following procedure.

The GPT-4-fine-tuned model achieved 81.2% accuracy on the test dataset. The prompt fine-tuning instructed the model correctly answered question 9. However, all three models failed question 6: "Cherry is to blossom as: Checkout is to purchase; Protein is to shake; Paint is to mix; Paper is to book." This question involves understanding not only the words "cherry" and "blossom" but also the phrase "cherry blossom," which is a type of flower. Finally, to test the model's performance without attention, we fine-tuned the GPT-4 model using the same procedure to obtain the GPT-4-fine-tuned model but added the prompt "loss of attention ability," resulting in the GPT-4-no-attention model. This model demonstrated the same logical ability as the GPT-4-fine-tuned model but incorrectly answered several questions that required long-term inference. This indicates that the attention cognitive ability of LLMs is essential for solving problems that require long-term reference, while strong logical ability is sufficient for solving short-term logical problems. Based on the performance results, the GPT-4 model also recognized the lack of attention ability in the GPT-4-no-attention model.

Table 1: Test accuracy of each LLM

Model	Accuracy
GPT-3	75.0%
Claude-3	56.2%
GPT-4-fine-tuned	81.2%
GPT-4-no-attention	68.7%

Acknowledgments

We gratefully acknowledge the efforts of the students who participated in this exercise, and Dr. Yunhao Liu for hosting the project demonstrations and support.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences of the United States of America, 120(6), e2218523120.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597–600.

Farah, M. J. (2000). The cognitive neuroscience of vision. Fundamentals of Cognitive Neuroscience., 380. https://psycnet.apa.org/fulltext/2000-07591-000.pdf

Fellows, L. K. (2004). The cognitive neuroscience of human decision making: a review and conceptual framework. Behavioral and Cognitive Neuroscience Reviews, 3(3), 159–172.

Gabrieli, J. D. (1998). Cognitive neuroscience of human memory. Annual Review of Psychology, 49, 87–115.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3(10), 833–838.

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., Ashman, J. B., Li, X., Liu, T., Shen, J., & Liu, W. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. Frontiers in Oncology, 13, 1219326.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT-4 Technical Report. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2303.08774

Prystauka, Y., DeLuca, V., Luque, A., Voits, T., & Rothman, J. (2023). Cognitive neuroscience perspectives on language acquisition and processing. Brain Sciences, 13(12). https://doi.org/10.3390/brainsci13121613

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. Journal of Experimental Psychology. General. https://doi.org/10.1037/xge0001547

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2303.17564