COMMENTARY

Promoting interactions between cognitive science and large language models

Youzhi Qu,[1] Penghui Du,[1] Wenxin Che,[1] Chen Wei,[1] Chi Zhang,[1] Wanli Ouyang,[2] Yatao Bian,[3] Feiyang Xu,[4] Bin Hu,[5] Kai Du,[6] Haiyan Wu,[7] Jia Liu,[8] and Quanying Liu[1,*]

[1]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[2]Shanghai AI Laboratory, Shanghai 200232, China
[3]Tencent AI Lab, Shenzhen 518057, China
[4]iFLYTEK AI Research, Hefei 230088, China
[5]School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
[6]Institute for Artificial Intelligence, Peking University, Beijing 100871, China
[7]Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau, Macau 999078, China
[8]Department of Psychology, Tsinghua University, Beijing 100084, China
*Correspondence: liuqy@sustech.edu.cn

Received: September 26, 2023; Accepted: January 8, 2024; Published Online: January 12, 2024; https://doi.org/10.1016/j.xinn.2024.100579

Citation: Qu Y., Du P., Che W., et al., (2024). Promoting interactions between cognitive science and large language models. The Innovation **5(2)**, 100579.

Large language models (LLMs) have made unprecedented progress, demonstrating human-like language proficiency and an extraordinary ability to encode complex knowledge. The emergence of high-level cognitive capabilities in LLMs, such as in-context learning and complex reasoning, suggests a path toward the realization of artificial general intelligence (AGI). However, we lack scientific theories and tools to assess and interpret such an emergence of the advanced intelligence of LLMs. Artificial intelligence (AI) has been extensively applied in various areas of fundamental science to accelerate scientific research.[1] Cognitive science has also broadened its focus to the cognition and intelligence of LLMs,[2] but the collaboration between cognitive science and LLMs has not been sufficiently emphasized. Here, we advocate for promoting interaction between cognitive science and LLMs (Figure 1), delve into challenges and solutions for such cooperation, and offer a perspective on future directions.

LLMs NEEDS COGNITIVE SCIENCE

The remarkable expansion in capabilities of LLMs raises pressing questions: do LLMs possess feelings, emotions, self-awareness, and free will? Could they potentially exhibit human-like personalities? Is it possible for LLMs to intentionally deceive or harm humans? Behind these questions are the safety concerns of LLMs. A comprehensive evaluation of LLMs' intelligence, with full consideration of the cognitive, moral, and ethical aspects, is urgently required.

Prior to the emergence of LLMs, cognitive science had already proposed methodologies for assessing the abilities of animals and humans. These well-established methodologies from cognitive science support a multidimensional assessment of intelligence: crystallized intelligence involving knowledge acquisition, fluid intelligence emphasizing adaptability in new situations, social intelligence focusing on understanding others, and embodied intelligence pertaining to interaction with the environment. Integrating methodologies from cognitive science (e.g., task paradigms and testing methods) can achieve a comprehensive assessment and understanding of the intelligence in LLMs. This integration can effectively guide the evolution of LLMs toward AGI.

The accelerated evolution of LLMs underlines the urgency for thorough assessments of their moral principles, personal values, and mental health. Our understanding of the reasoning processes and problem-solving strategy in LLMs is limited, potentially leading to risks. LLMs might learn biases from unfiltered data or intentionally deceive humans, causing catastrophic harm to human society. Such cognitive biases are difficult to identify through natural language tasks. They can be detected by experimental paradigms in cognitive science, such as the implicit association test, which detects subconscious associations between mental representations of concepts. Tasks from cognitive science can help assess mental state, personality, and multidimensional cognitive ability of LLMs, detecting potential risks in LLMs and thereby enhancing the safety of models.

COGNITIVE SCIENCE NEEDS LLMs

LLMs can serve as new study subjects in cognitive science, just as human subjects did. Human experiments in cognitive science are time consuming and tedious. Conducting experiments with LLMs can facilitate data collection, such as LLMs' behavioral data and artificial neuronal activity. Through training on substantial textual data, LLMs can exhibit human-like behaviors in cognitive psychology scenarios.[3] LLMs can effectively represent the mainstream opinions of the majority, thereby fostering rapid understanding and analysis of the public opinions and preferences. Moreover, LLMs can be customized using specialized textual datasets to reflect the cognitive perspectives of a specific population.

LLMs provide a unique experimental platform for investigating the evolution of intelligence. The intelligence of biology evolves over a billion years, making the observation of its evolutionary process impossible, while the intelligence of LLMs can be developed within a few months. This rapid training process mirrors the evolutionary process of intelligence. Observing LLMs at different training stages allows us to witness the evolution of intelligence. LLMs exhibit multidimensional intelligence, opening a window to uncover relationships between different dimensions of intelligence. LLMs possess greater transparency and manipulability compared to the biological brain. Modifying the architecture and parameters of LLMs is much easier than manipulating biological neural circuits, thereby offering a great opportunity to examine the causal effects on intelligence.

CHALLENGES

Barriers between academia and industry

Collaboration between academia and industry encounters barriers to resource sharing and teamwork strategies. Although AI companies has produced large models trained with extensive data, they tend to keep their models and data confidential to improve competitiveness and avoid public pressure. Training LLMs requires massive computational resources (around 135,168 GPU h on A100 GPU with 80 GB RAM for a 13B-parameter LLaMA model) and large datasets (around 1.0 T token after tokenization for the same model).[4] There is a misconception in many academic laboratories that working with large models necessarily requires extensive resources. In fact, running and testing LLMs do not need substantial resources. The 13B-parameter LLaMA can execute successfully on a single V100 GPU with 32 GB RAM.[4] Such costs are affordable for cognitive science laboratories. We advocate for establishing a platform to streamline the collaborative process, which encompasses a standardized set of tools and methodologies for interdisciplinary research. The industry provides interfaces of LLMs, and academia provides theoretical knowledge and methods for evaluating the intelligence of LLMs from a cognitive perspective. Both parties jointly promote the intelligence and safety of LLMs.

Lack of interdisciplinary collaborations

The disparities between computer science and cognitive science pose challenges to interdisciplinary collaboration. Computer science emphasizes engineering skills, while cognitive science prioritizes scientific thinking. They use different terminology systems. For example, the term "attention" in cognitive science refers to the cognitive ability to select and focus on relevant stimuli, whereas in AI, it denotes the self-attention mechanism in transformer models. Researchers in computer science are keen to publish in AI conferences, while those in cognitive science prefer journals. We call for optimizing curriculum design, unifying terminology systems, hosting cognitive science tutorials and workshops in

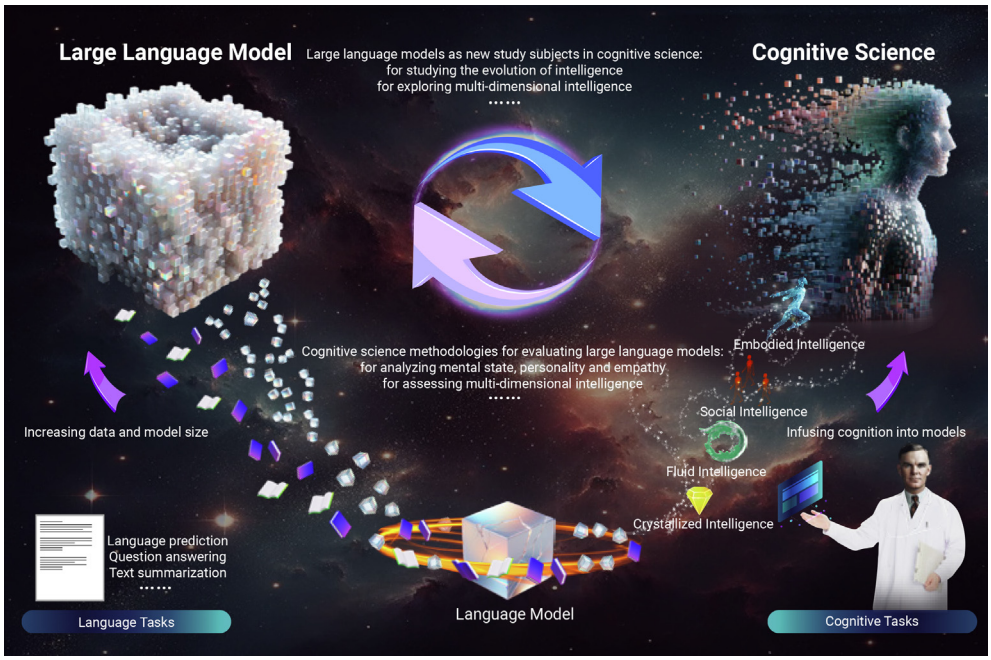

Figure 1. Promoting interaction between cognitive science and large language models (LLMs) We emphasize the reciprocal needs between cognitive science and LLMs, suggest bringing methodologies from cognitive science to evaluate the intelligence of LLMs, and recommend treating LLMs as study subjects in cognitive science.

AI conferences, and organizing AI hackathons and hands-on tutorials for cognitive science experts.

Insufficient funding supports

We emphatically advocate for concerted support from governmental agencies, nonprofit organizations, and foundations, as this is crucial for promoting interdisciplinary collaboration. Various entities have started providing financial support for interdisciplinary studies, such as the National Institutes of Health in the United States and the National Natural Science Foundation of China. However, these efforts are not sufficient to achieve comprehensive collaboration between the two fields, considering the substantial costs of equipment and personnel. The high computational costs make LLM training unfeasible in economically underdeveloped regions. Minority-language countries struggle to acquire sufficient textual data to develop LLMs. Therefore, we call for more attention and funding to improve the transferability of LLMs to minority languages and to help underdeveloped regions deploy and apply LLM technology, such as providing computing resources and collecting minority-language texts. We firmly believe that such support will alleviate the global imbalance in LLMs and cognitive science.

THE FUTURE OF LLMs AND COGNITIVE SCIENCE

We envision that interactions between LLMs and cognitive science will promote the integration of LLMs with human society, and LLMs will penetrate into all aspects of human society in the future.

As LLMs progressively integrate into human interactions and everyday applications, enhancing LLM security becomes substantial. By incorporating evaluation methods derived from cognitive science, a deeper understanding of the capabilities, personality traits, and mental states of LLMs can be attained. Based on the deficiencies identified through comprehensive assessment, the insufficient capabilities of LLMs can be improved by utilizing specific datasets or integrating appropriate external technologies, such as prompt engineering. Such improvements contribute to ensuring the safety and reliability of LLMs. We firmly believe that a promising future lies in strengthening the collaboration between LLMs and cognitive science, leveraging the expertise of each field to enhance the safety of LLMs and delivering benefits to human society.

LLMs may simply imitate language patterns like parrots without truly understanding or meaningfully applying them, a phenomenon referred to as "stochastic parrots."[5] The extensive unfiltered training data can lead to LLMs generating biased and inaccurate content. Increasing data size does not necessarily enhance the diversity of LLMs' outputs because textual data from the internet overrepresents the viewpoints of users from developed countries and regions while overlooking those of minority and marginalized populations.[5] Therefore, it is crucial to integrate cognitive tests for ethical standards and biases into the assessment of LLMs. Cognitive science can design specialized moral dilemma tests and simulations of various social scenarios for LLMs to assess their moral integrity, fairness, and empathy.

Aligning the behaviors of LLMs with human intentions and values is crucial for AI safety. As an AI alignment strategy, reinforcement learning from human feedback can reduce the generation of inaccurate and harmful content in LLMs based on human feedback on LLMs' output. However, aligning LLMs' output with humans only scratches the surface of the problem. For example, although LLMs can generate text with various emotions, they do not align with human emotions and lack the ability to genuinely understand or manifest human emotions. Deeper alignment calls for aligning with biological neural representations and human cognition.

REFERENCES

1. Xu, Y., Liu, X., Cao, X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. Innovation ***2***: 100179. https://doi.org/10.1016/j.xinn.2021.100179.
2. Binz, M., and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proc. Natl. Acad. Sci. USA ***120***: e2218523120. https://doi.org/10.1073/pnas.2218523120.
3. Dillion, D., Tandon, N., Gu, Y., et al. (2023). Can AI language models replace human participants? Trends Cognit. Sci. ***27***: 597–600. https://doi.org/10.1016/j.tics.2023.04.008.
4. Touvron, H., Lavril, T., Izacard, G., et al. (2023). Llama: Open and efficient foundation language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2302.13971.
5. Bender, E.M., Gebru, T., McMillan-Major, A., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. https://doi.org/10.1145/3442188.3445922.

ACKNOWLEDGMENTS

We thank Profs. Ruyuan Zhang and Zaixu Cui for their suggestion. This work was funded by National Natural Science Foundation of China (62001205), National Key R&D Program of China (2021YFF1200804), and Shenzhen Science and Technology Innovation Committee (2022410129, KCXFZ20201221173400001, and SGDX20201 10309280100).

DECLARATION OF INTERESTS

The authors declare no competing interests.