Higher cognition in large language models

Organizers
Nick Ichien (nichien@sas.upenn.edu) & Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA 19104

Invited speakers (in speaking order)
Anna Ivanova (a.ivanova@gatech.edu)
School of Psychology, Georgia Institute of Technology, Atlanta, GA USA 30332
Taylor W. Webb (twebb@ucla.edu)
Department of Psychology, University of California, Los Angeles, Los Angeles, CA USA 90095
Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology, Princeton University, Princeton, NJ, USA 08540
Marcel Binz (binz@tue.mpg.de)
Helmholtz Computational Health Center, Munich, Germany

Keywords: large language models (LLMs); artificial intelligence (AI); higher cognition; language processing; computational cognitive science

Introduction (Nick Ichien)

Large language models (LLMs) like OpenAI's GPT-4 (OpenAI et al., 2023), or Google's PaLM (Chowdhery et al., 2022) generate text responses to user-generated text prompts. In contrast to work that evaluates the extent to which model-generated text coheres with linguistic rules (i.e., *formal* competence) (Chomsky et al., 2023; Piantadosi, 2023), the present symposium discusses the work of cognitive scientists aimed at assessing the extent and manner in which LLMs show effective understanding, reasoning and decision making, capacities associated with human higher cognition (i.e., *functional* competence) (Binz & Schulz, 2023; Mahowald et al., 2023; Webb et al., 2023). Given both their expertise and their interest in clarifying the nature of human thinking, cognitive scientists are in a unique position both to carefully evaluate LLMs' capacity for thought (Bhatia, 2023; Han et al., 2024; Mitchell, 2023) and to benefit from them as methodological and theoretical tools. This symposium will thus be of interest not only to cognitive scientists concerned with machine intelligence, but also to those looking to incorporate advances in artificial intelligence with their study of human intelligence.

Anna Ivanova will begin the symposium talks with some conceptual framing, elaborating on the distinction between formal competence and functional competence. Under the broad umbrella of functional competence, Taylor Webb will go on to identify and clarify specific aspects of reasoning on which LLMs excel and others on which LLMs struggle, and propose a framework for incorporating LLMs into a larger architecture that selectively exploits their strengths. Next, Tom Griffiths will present analyses of LLMs that clarify their similarities and differences to human reasoning by attending to the problem they were trained to solve (i.e., rational analysis) and to the representations that mediate their overt behavior (i.e., representational analysis). Finally, Marcel Binz will end with a large-scale project that uses human behavioral data to fine-tune an LLM to better align model behavior with that of human reasoners and produce a useful methodological and theoretical tool for clarifying how humans think.

Distinguishing between formal and functional competence in LLMs (Anna Ivanova)

Today's LLMs routinely generate coherent, grammatical and seemingly meaningful paragraphs of text. This achievement has led to speculation that LLMs have become "thinking machines", capable of performing tasks that require reasoning and/or world knowledge. In this talk, I will introduce a distinction between formal competence—knowledge of linguistic rules and patterns—and functional competence—understanding and using language in the world. This distinction is grounded in human neuroscience, which shows that formal and functional competence recruit different cognitive mechanisms. I will show that the word-in-context prediction objective has allowed LLMs to essentially master formal linguistic competence; however, pretrained LLMs still lag behind at many aspects of functional linguistic competence, prompting engineers to adopt specialized fine-tuning techniques and/or couple an LLM with external modules. I will illustrate the formal-functional distinction using the domains of English grammar and arithmetic, respectively. I will then turn to generalized world knowledge, a domain where this distinction is much less clear-cut, and discuss our efforts to leverage both cognitive science and NLP to develop systematic ways to probe generalized world knowledge in text-based LLMs. Overall, the formal/functional competence framework clarifies the discourse around LLMs, helps develop targeted evaluations of their capabilities, and suggests ways for developing better models of real-life language use.

Evaluating, understanding, and improving reasoning in LLMs (Taylor W. Webb)

A major debate has recently emerged concerning whether LLMs are capable of human-like reasoning, with some

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

arguing that LLMs' apparent reasoning capabilities are based instead on mimicry of their vast training data, and others arguing that LLMs constitute an early form of artificial general intelligence (AGI). To shed light on these issues, I will discuss work that aims to systematically evaluate, understand, and improve the reasoning capacities of LLMs. First, I will argue that it is important not to treat reasoning as a monolithic capacity, but instead to carefully evaluate distinct modes of reasoning in isolation (e.g., analogical reasoning, mathematical reasoning, physical reasoning). I will present results that highlight interesting and surprising dissociations between these distinct modes, including findings suggesting that LLMs possess analogical reasoning capabilities on par with adults, while performing physical reasoning tasks below the level of young children. Second, I will discuss the mechanisms that enable these reasoning capacities, emphasizing the role played by architectural inductive biases, and drawing comparisons to traditional cognitive models. Finally, I will present an approach for improving reasoning and problem-solving in LLMs, introducing a modular architecture in which distinct LLM instances carry out specialized processes, inspired by the functional architecture of the human brain.

Using the tools of cognitive science to study LLMs (Thomas L. Griffiths)

LLMs present cognitive scientists with an interesting opportunity to study a new kind of intelligent system. Studying this system presents challenges — many LLMs are proprietary and hence don't allow access to internal states, so we just have to rely on their behavior. Of course, cognitive scientists have been dealing with a similar challenge for decades, developing methods for studying human cognition based just on behavior. I will outline how two methods from cognitive science can be used to gain insight into LLMs. The first is rational analysis, investigating the behavior of LLMs in a way that is guided by the computational problem they have to solve. The second is representational analysis, using similarity rating to explore the internal representations of these systems. Using these methods I will show that while LLMs have some parallels with human cognition, they are also shaped by the specific problem they solve and the data they use to solve it.

Towards foundation models of human cognition (Marcel Binz)

Most cognitive models are domain-specific, meaning that their scope is restricted to a single type of problem. The human mind, on the other hand, does not work like this – it is a unified system whose processes are deeply intertwined. In this talk, I will present our ongoing work on foundation models of human cognition: models that cannot only simulate, predict, and explain behavior in a single domain but instead offer a truly universal take on our mind. Together with a large international consortium, we have transcribed data from over 130 experiments – covering all major areas of cognitive psychology, including reinforcement learning, memory, decision-making, probabilistic reasoning, and many more – into a text-based form. We then used this data set to finetune an LLM, thereby aligning it to human behavior. The resulting model provides a window into human cognition and can be used for rapid prototyping of behavioral studies, to improve traditional cognitive models, and to generate new hypotheses about human information processing.

References

Bhatia, S. (2023). Inductive reasoning in minds and machines. *Psychological Review*. https://doi.org/10.1037/rev0000446

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. https://doi.org/10.48550/arXiv.2204.02311

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155. https://doi.org/10.1016/j.cogsys.2023.101155

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models* (arXiv:2301.06627). arXiv. http://arxiv.org/abs/2301.06627

Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, *381*(6654), eadj5957. https://doi.org/10.1126/science.adj5957

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* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, Lingbuzz*, *7180*. https://lingbuzz.net/lingbuzz/007180/v1.pdf

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), Article 9. https://doi.org/10.1038/s41562-023-01659-w