Large language models meet cognitive science
LLMs as tools, models, and participants

Mathew D. Hardy[1] (mdhardy@princeton.edu)
Ilia Sucholutsky[2] (is2961@princeton.edu)
Bill Thompson[3] (wdt@berkeley.edu)
Thomas L. Griffiths[1,2] (tomg@princeton.edu)
[1]Department of Psychology, Princeton University, Princeton, NJ 08544
[2]Department of Computer Science, Princeton University, Princeton, NJ 08544
[3]Department of Psychology, University of California, Berkeley, Berkeley, CA, 94720

Keywords: Artificial intelligence; machine learning; decision-making; language

Overview

Large language models (LLMs) like GPT-3 (Brown et al., 2020) are revolutionizing artificial intelligence, leading to breakthroughs in question answering, natural language understanding, and machine translation. Recent work in a variety of social science disciplines, including psychology (Hagendorff, Fabi, & Kosinski, 2022), economics (Horton, 2023), and political science (Argyle et al., 2022) has demonstrated remarkable similarity between the behavior of LLMs and human decision makers. At the same time, AI researchers and engineers struggle to understand these systems, leading to practical challenges and ethical questions about fair and safe deployment.

This workshop aims to bring together researchers to discuss work on using psychological methods to understand LLMs and LLMs as tools for understanding humans. Along these lines, we have invited leading researchers from cognitive science, psychology, and machine learning to present their work on topics that include: When and why do LLMs exhibit biased behavior? How do these compare to human biases? What sorts of psychological tasks do LLMs struggle with? Can we use psychological theory to structure this search? And how does the knowledge encoded in LLMs differ from human knowledge?

We expect these topics to be relevant to cognitive, developmental, and social psychologists, behavioral economists, sociologists, linguists, philosophers, computer scientists, and AI safety researchers. We also believe that the theme of this workshop is especially relevant to this year's focus on "Cognition in Context". LLMs are rapidly being deployed in many industrial settings and products, and psychological methods may be key to understanding these models (Binz & Schulz, 2022; Miotto, Rossberg, & Kleinberg, 2022; Hagendorff et al., 2022) and improving their performance (Prystawski, Thibodeau, & Goodman, 2022; Goyal & Bengio, 2022). Cognitive science may thus have a crucial role to play in the development of safe, robust artificial intelligence and systems, providing crucial guidance to government regulators and policymakers.

Workshop structure and speakers

Each in-person speaker will be given 30 minutes—25 minutes for their talk, followed by 5 minutes for Q&A. Virtual speakers will give 15-minute recorded talks. We will end the workshop with a one-hour panel discussion. We introduce the speakers below.

Using cognitive psychology to understand GPT-3

Marcel Binz is a postdoc in the Computational Principles of Intelligence Lab at the Max Planck Institute for Biological Cybernetics in Tübingen. His research focuses on using Bayesian modeling, deep learning, and other mathematical frameworks for studying learning and decision-making.

Talk title TBD

Judith Degen is an Assistant Professor in the Department of Linguistics at Stanford where she directs the interActive Language Processing Lab. Her research investigates how humans rapidly draw pragmatic inferences during language processing, using computational modeling, psycholinguistic experiments, and corpus analyses.

Language models show human-like content effects on reasoning

Ishita Dasgupta is a Senior Research Scientist at DeepMind New York City. Her research is at the intersection of computational cognitive science and machine learning. Ishita uses advances in machine learning to build new models of human reasoning, applies cognitive science approaches toward understanding black-box AI systems, and combines these insights to build better, more human-like artificial intelligence.

Assessing mixed-effects transformers as models of hierarchical adaptation in human language use

Robert Hawkins is a Postdoctoral Scholar at the Princeton Neuroscience Institute. He studies the computational principles underlying social interaction and communication in minds and machines.

Dissociating language and thought in large language models

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Anna Ivanova is a Postdoctoral Associate at MIT Quest for Intelligence. She obtained her PhD from MIT's Department of Brain and Cognitive Science, where she studied the neural mechanisms underlying language processing in humans. In addition to her neuroscience work, Anna is examining the language-thought relationship in large language models, using her cognitive science training to identify similarities and differences between humans and machines.

From Word Models to World Models – Rational Meaning Construction in a Probabilistic Language of Thought

Gabriel Grand is a PhD student in MIT CSAIL and BCS co-advised by Jacob Andreas and Josh Tenenbaum. His research explores programming as a framework for understanding human language and concept learning. Combining tools from program synthesis, machine learning, and probabilistic programming, he aims to develop AI systems that can fluidly translate between language and code, bridging the gap between human and machine cognition.

Lionel Wong is a PhD candidate in Brain and Cognitive Sciences at MIT. Their research focuses on computational and cognitive models that integrate structured conceptual reasoning with language, and that learn new concepts and abstractions from language.

The risk of bias transmission from large-scale language models to human users

Celeste Kidd is an Assistant Professor of Psychology at UC Berkeley. Her lab investigates learning and belief formation, including how algorithmic biases in emerging technologies like large-scale language models distort the beliefs of the humans who use them.

Integrating LMs into theory-driven, cognitive models – Moral judgment as a case study

Sydney Levine is a research scientist at the Allen Institute for AI with training in cognitive, social, and developmental psychology. She works at the intersection of moral psychology and AI development, using computational models of human moral cognition to help build AI systems that can predict and produce human-like moral judgment.

In search for the missing metaphor: octopuses, parrots, pianos, and autocomplete

Gary Lupyan is a professor of Psychology at the University of Wisconsin-Madison. His research explores the impact of language on cognition, perception, and memory, investigating its role in categorization, memory, and the relationship between linguistic and social structures.

Human induction and large language models

Andrew Perfors is a professor at the University of Melbourne, where he is the Director of the Complex Human Data Hub. He studies how reasoning and learning is shaped by the nature of one's input and the constraints and assumptions one operates under. By comparing humans and machines in terms of these factors and their ultimate behaviour, we can shed light on both.

Reflections on language and thought

Joshua Tenenbaum is a professor of computational cognitive science in MIT's Department of Brain and Cognitive Sciences, and a scientific director with MIT Quest. He is also a principal investigator at the Center for Brains, Minds and Machines and the Computer Science and Artificial Intelligence Laboratory. His research straddles cognitive science and artificial intelligence, where his goals are to reverse engineer human intelligence and to build machines that behave in human-like ways.

Emergent analogical reasoning in large language models

Taylor Webb is a postdoctoral scholar in the UCLA Department of Psychology, working with Keith Holyoak, Hongjing Lu, and Hakwan Lau. His research is situated at the interface between cognitive science and AI, with a particular emphasis on using neural network techniques to build cognitive models that are grounded in real-world perceptual inputs. He received his Ph.D. in cognitive psychology and neuroscience from Princeton University, where he studied with Jonathan Cohen and Michael Graziano.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Binz, M., & Schulz, E. (2022). Using cognitive psychology to understand GPT-3. *arXiv preprint arXiv:2206.14576*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, *478*(2266), 20210068.

Hagendorff, T., Fabi, S., & Kosinski, M. (2022). Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *arXiv preprint arXiv:2212.05206*.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

Prystawski, B., Thibodeau, P., & Goodman, N. (2022). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.