Cognitive-Science–Inspired Evaluation of Large Language Models

Organizers

Shuhao Fu (fushuhao@santafe.edu)[1], Claas Beger[1], Ryan Yi[1], Melanie Mitchell[1]
[1]Santa Fe Institute

Contributors

Michael C. Frank[2], Raphaël Millière[3], Robert Hawkins[2], Tomer D. Ullman[4]
[2]Stanford University, [3]University of Oxford, [4]Harvard University

Keywords: Large Language Models (LLMs); computational cognitive science; model evaluation

Introduction

Large language models (LLMs) can appear impressively capable across conversation, reasoning tasks, and standardized benchmarks. Yet merely appearing intelligent on traditional metrics is not the same as possessing robust, generalizable cognitive capacities. This disconnect highlights the need for more principled approaches to model evaluation. To address this gap, the symposium brings together four talks centered on developing a cognitive-science–inspired approach to assessing LLMs.

Many evaluation methods risk overstating competence by mistaking superficial success for genuine ability (Mitchell, 2023). In practice, benchmark performance is highly sensitive to prompt wording and other surface-level factors (Errica et al., 2025). Estimates of capability can also be inflated by training data contamination and approximate retrieval, which blur the line between genuine generalization and prior exposure (Deng et al., 2024). Even when models produce correct answers, they may do so by exploiting unintended shortcut heuristics (Beger et al., 2025; Du et al., 2023). More broadly, benchmarks often reward pattern completion over robust concept use and transfer to novel variants (Deveci & Ataman, 2025). Taken together, these issues call for careful experimental design, strong controls, and cautious interpretation.

This symposium brings together perspectives from cognitive science, developmental psychology, linguistics, and philosophy to advance principled approaches to evaluating large language models. It highlights controlled task perturbations as tools for discriminating candidate algorithms, draws on developmental psychology to support grounded comparative evaluation aligned with human data, and showcases applications of psycholinguistic frameworks for assessing LLMs as adaptive communicators in interactive, multi-turn settings. It also examines the epistemological limits of behavioral evidence for inferring computational strategies in LLMs—in contrast to human cognition—and the resulting need for complementary mechanistic approaches. By importing mature methodological lessons from cognitive science and related disciplines, the symposium aims to foster more principled evaluation methods and a more cumulative, scientifically grounded understanding of how "intelligent" AI systems really are.

Developmental Evaluation of "Baby" Language Models Using Egocentric Video

Michael C. Frank

Large language models and associated multimodal models provide exciting tools for describing human development, but this promise can only be realized if training and evaluation are aligned with human data. I will describe a program of research using the BabyView dataset, a large corpus of egocentric videos from children ages 6 months to 3 years. We use these videos as training data for small-scale language, vision, and vision-language models, and we assess the models' performance on a variety of tasks, including direct comparison with developmental outcomes. By contrasting models with different assumptions, we can learn about the minimal ingredients necessary to achieve human-scale learning.

A Bayesian Epistemology for LLM Evaluation

Raphaël Millière

Evaluating the capacities of large language models (LLMs) requires inferring latent abilities from observable task performance. I propose to formalize such inferences within a Bayesian framework, drawing on the epistemology of comparative cognitive science. The framework makes explicit how posterior credences about a system's abilities depend on prior probabilities over hypotheses about computational strategies, likelihood functions shaped by task demands, and marginalization over auxiliary factors that influence performance independently of the target ability. Within this framework, I argue that the relative evidential weight of behavioral versus mechanistic evidence for algorithmic-level hypotheses concerning the representations and computations underlying task performance depends systematically on background knowledge about the target system. For human cognition, decades of research in cognitive science, neuroscience, and evolutionary biology provide strong priors that constrain the space of plausible computational strategies, while well-characterized auxiliary factors (attention, motivation, working memory limits) allow informative marginalization. Behavioral paradigms consequently provide substantial evidential leverage for discriminating between algorithmic hypotheses. For LLMs, this evidential hierarchy shifts: weak priors about learned computational strategies and poorly characterized auxiliary factors (e.g., sensitivity to prompt formatting, tokenization artifacts,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

training distribution) render behavioral evidence systematically underdetermining across competing hypotheses. Mechanistic interpretability, by contrast, provides more direct access to representations and computations, offering evidence that can discriminate where behavioral evidence cannot. This analysis yields a methodological prescription: LLM evaluation should integrate mechanistic evidence not as a supplement to behavioral studies, but as an epistemically essential component for constraining algorithmic-level hypotheses.

Interactive Evaluations for Interactive Models

Robert Hawkins

Language use is fundamentally adaptive: speakers build common ground with their conversation partners, adjust their strategies based on feedback, and coordinate on meaning over extended interactions. LLMs are increasingly deployed in such interactions, yet LLM evaluations have largely focused on single-turn, constrained-choice settings that miss these dynamics entirely. I will share two lines of work addressing this gap. First, I'll describe RefBank, an open repository of repeated reference game data capturing how humans form conventions over repeated interaction, a resource that we are developing as a multi-turn benchmark to test whether LLMs exhibit human-like audience design. Second, I'll show how moving from constrained to open-ended production reveals systematic deviations in LLM pragmatic competence, with models over-relying on negative politeness strategies even in positive contexts. In both cases, I'll argue that cognitive science offers both the appropriate theoretical frameworks and the measurement tools needed to evaluate LLMs as adaptive communicators.

Evaluating Physics and Psychology in LLMs and VLMs

Tomer D. Ullman

There are multiple algorithms that can get you from a given input (e.g., "what is 5×5") to a given output (e.g., "25"). A great deal of work in cognitive science involves figuring out the underlying algorithms that get people from a certain input to a certain output, and focusing solely on performance and accuracy does not tell you what algorithm got you from input to output. Experiments in cognitive science often involve controls and checks to rule out competing hypotheses, but also rely on certain hidden assumptions about ways in which people obviously are not doing something (e.g., obviously people did not happen to memorize ahead of time the answer to 32819×28229). Current models play havoc with our intuitions as cognitive scientists regarding what is obviously the wrong way of solving a task, and requires a re-assessment of tasks that we take for granted. In this talk, I will consider certain 'obvious' cases involving physical/visual reasoning about images, illusions, and theory-of-mind, and consider small manipulations to those tasks that lead to wonky results. I will emphasize the need to move beyond benchmarked to 'cordoned' generalization tasks.

Speaker Introduction

The symposium will be moderated by Shuhao Fu, who is currently a Postdoctoral Fellow at the Santa Fe Institute working on combining cognitive science and machine learning to study how AI can achieve human-like reasoning.

- **Michael C. Frank** is the Benjamin Scott Crocker Professor of Human Biology at Stanford University and Director of the Symbolic Systems Program. His research investigates language learning and cognitive development in children, often using large-scale behavioral and multimodal datasets.
- **Raphaël Millière** is an Associate Professor of Philosophy at the University of Oxford. His research focuses on understanding modern artificial neural networks through theoretical analysis, behavioral evaluation, and interpretability methods.
- **Robert Hawkins** is an Assistant Professor of Linguistics at Stanford University. He studies language use as a form of social interaction, focusing on how people establish common ground, adapt to their interlocutors, and develop communicative conventions over repeated interaction.
- **Tomer D. Ullman** is the Morris Kahn Associate Professor of Psychology at Harvard University. His work combines computational modeling and behavioral experiments to study human common-sense reasoning, including intuitive physics, theory of mind, and the learning of abstract causal and conceptual structures.

References

Beger, C., Yi, R., Fu, S., Moskvichev, A., Tsai, S. W., Rajamanickam, S., & Mitchell, M. (2025). Do ai models perform human-like abstract reasoning across modalities? *arXiv preprint arXiv:2510.02125*.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., & Cohan, A. (2024). Investigating data contamination in modern benchmarks for large language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8706–8719.

Deveci, İ. E., & Ataman, D. (2025). The ouroboros of benchmarking: Reasoning evaluation in an era of saturation. *arXiv preprint arXiv:2511.01365*.

Du, M., He, F., Zou, N., Tao, D., & Hu, X. (2023). Shortcut learning of large language models in natural language understanding. *Commun. ACM*, *67*(1), 110–120.

Errica, F., Siracusano, G., Sanvito, D., & Bifulco, R. (2025). What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Mitchell, M. (2023). How do we know how smart ai systems are? *Science*, *381*(6654), eadj5957.