# Cognitively Inspired Interpretability in Large Neural Networks

*Organizers*
**Anna Leshinskaya (anna.leshinskaya@gmail.com)**
Department of Cognitive Sciences, University of California, Irvine, CA, USA
**Taylor Webb (taylor.w.webb@gmail.com)**
Microsoft Research, New York, NY USA 10012

*Invited speakers (in speaking order)*
**Ellie Pavlick (ellie_pavlick@brown.edu)**
Department of Computer Science, Brown University, Providence, RI, USA
**Jiahai Feng (fjiahai@berkeley.edu)**
Department of Computer Science, University of California, Berkeley, CA, USA
**Gustaw Opiełka (g.j.opielka@uva.nl), Claire Stevenson (C.E.Stevenson@uva.nl)**
Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands
**Idan A. Blank (iblank@psych.ucla.edu)**
Department of Psychology, University of California, Los Angeles, CA, USA

**Keywords:** Large Language Models (LLMs); Mechanistic Interpretability; Computational Cognitive Science

## Introduction

Large Language Models (LLMs) and Vision Language Models (VLMs) have become a dominant force in artificial intelligence and have already made a major impact on the cognitive sciences (Binz & Schulz, 2023; Webb, Holyoak & Lu, 2023; McGrath et al., 2024; Choudhury et al., 2025), but debate persists concerning the extent to which they possess emergent cognitive capacities (Denning et al., 2023; Stevenson et al., 2024). Investigation of these systems at the level of behavioral outputs has led to conflicting findings, and the question of how these outputs are generated (at a mechanistic or algorithmic level) remains open. Yet, the abilities they do exhibit behaviorally offer an unprecedented opportunity to answer longstanding questions about how neural networks could, even in principle, achieve abilities that are thought to require structured representations—such as syntactic, combinatorial, and variable-binding operations. In this symposium, we highlight a recent body of work that addresses this gap in understanding by investigating the *internal mechanisms* that support cognitive processing in LLMs and other large-scale neural networks (Feng & Steinhardt, 2023; Lepori et al., 2024; McCoy & Leshinskaya, 2024). The symposium brings together researchers with backgrounds in both computer science and psychology, exploring ways in which mechanistic interpretability research and cognitive science can mutually inform one another.

Ellie Pavlick will begin the symposium talks with a discussion of the ways in which LLMs can be useful as cognitive models, with a particular emphasis on the ways in which they differ from previous generations of connectionist models, and the potential for mechanistic interpretability research to contribute new ideas to cognitive science. Jiahai Feng will then present results that illustrate how LLMs perform dynamic variable-binding, a question of direct relevance to the longstanding debate between connectionist and symbolic approaches. Gustaw Opiełka and Claire Stevenson will then present evidence for the presence of emergent relational representations in LLMs, and highlight the limitations of these representations in more abstract settings. Finally, Idan Blank will present a series of studies that ask whether LLMs employ human-like mechanisms for syntactic processing and language understanding.

## Not-Your-Mother's-Connectionism: LLMs as Cognitive Models (Ellie Pavlick)

Recent advances in AI have led to large neural network models that exhibit human-like behavior across a range of language and reasoning tasks. This (re-)opens important theoretical questions about the nature of the structure that is required to support such behaviors, leading to debates reminiscent of long-running arguments that pit neural network models against explicitly structured symbolic models of the mind. In this talk, I will describe a series of experiments which highlight the ways in which LLMs today appear importantly different from the connectionist systems that inspired these debates originally. I will argue for a more nuanced stance which does not assume neural networks to be diametrically opposed to traditional models of the mind, but still acknowledges the potential of LLMs to teach us something fundamentally new about the structures that govern language and cognition in humans.

## How do language models bind entities in context? (Jiahai Feng)

To correctly use in-context information, language models (LMs) must bind entities with their attributes. For example, given an input "Laura is a physicist. Greg is a nurse.", language models must correctly bind the two people with their respective jobs. We analyze internal LM activations and identify the binding ID mechanism: a representational



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

strategy that marks bound tokens with abstract binding ID vectors. We show, with causal interventions, that every sufficiently large model from the Pythia and LLaMA families uses binding IDs. We further develop a method for decoding binding IDs from internal activations, so that we can extract logical propositions that encode the relations expressed by the input context (e.g. "WorksAs(Laura, physicist)" and "WorksAs(Greg, nurse)"). Despite being trained only on simple templated contexts, our method generalizes to contexts rewritten as short stories and translated to Spanish, suggesting that LMs often utilize simple and robust representations.

## **Relation Vectors Support Relational Reasoning in LLMs** (Gustaw Opiełka and Claire Stevenson)

Analogical reasoning relies on structured relational representations, yet it remains unclear how well LLMs develop such abstractions. Using representational similarity analyses, we find that LLMs linearly encode certain relational concepts—such as antonymy and translation—when performing in-context learning tasks. These representations are primarily carried out by a set of attention heads in early-to-mid layers and remain invariant across low-level input variations, including different languages and task formats (open-ended vs. multiple-choice), while causally influencing model behavior. Notably, these attention heads function as feature detectors independent of the final model output—meaning that even when the model produces an incorrect answer, the correct relational concept may still be internally represented. However, we find that these heads do not linearly encode more abstract relations, such as predecessorship or successorship, which may contribute to LLMs' generalization difficulties in more complex reasoning tasks. In ongoing work, we explore how analyzing the internal representations of abstract concepts can provide insights into LLMs' challenges with more abstract analogical reasoning tasks. Specifically, we investigate conceptual representations in tasks such as letter-string analogies and the Abstraction and Reasoning Corpus (ARC).

## **Do LLMs Comprehend the Meaning of Language?** (Idan Blank)

The ability of LLMs to "understand" language is widely debated. To inform these discussions, we adapt experimental methods from human psycholinguistics to test whether LLMs exhibit signatures of human-like comprehension. One study characterizes the processing interface between meaning and syntax. We test whether any syntactic computations within LLMs are "impenetrable" to semantics—which would render them strikingly different from humans. We study the best a-priori candidates for impenetrability: attention heads specialized for individual syntactic relationships (e.g., verb-object dependencies). We find that even these heads are influenced by semantic plausibility, mirroring human language processing. Another study tests whether the internal workings of LLMs capture thematic roles ("who did what to whom"). We find that LLMs trained only on word prediction learn what thematic roles are (in some attention heads), but do not show human-like integration of this information into overall sentence representations. A final study extends to multimodal LLMs, investigating their pragmatic abilities. We find sensitivity to visual constraints on referring expressions, and explore potential mechanisms for this capacity. These studies reveal both similarities and gaps between LLMs and humans, breaking comprehension into theoretically-informed constructs and promoting a nuanced view of how, and in what sense, LLMs understand language.

## References


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

Choudhury, M., Elyoseph, Z., Fast, N. J., Ong, D. C., Nsoesie, E. O., & Pavlick, E. (2025). The promise and pitfalls of generative AI. *Nature Reviews Psychology*, 1-6.

Denning, J., Guo, X. H., Snefjella, B., & Blank, I. A. (2023). Do Large language Models know who did what to whom?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Feng, J., & Steinhardt, J. (2023). How do language models bind entities in context? In *12th International Conference on Learning Representations*.

Lepori, M. A., Tartaglini, A. R., Vong, W. K., Serre, T., Lake, B. M., & Pavlick, E. (2024). Beyond the Doors of Perception: Vision Transformers Represent Relations Between Objects. In *38th Conference on Neural Information Processing Systems*.

McCoy, M. & Leshinskaya, A. (2024). Relational composition during attribute retrieval in GPT is not purely linear. *Compositional Learning: Perspectives, Methods, and Paths Forward workshop at NeurIPS (38).*

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2024). How Can Deep Neural Networks Inform Theory in Psychological Science?. *Current Directions in Psychological Science, 33*(5), 325-333.

Stevenson, C.E., Pafford, A., van der Maas, H.L., & Mitchell, M. (2024). Can Large Language Models generalize analogy solving like people can?. *arXiv preprint arXiv*:2411.02348.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour, 7*(9), 1526-1541.