First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models

UNIVERSITY OF CAMBRIDGE

NEURAL INFORMATION PROCESSING SYSTEMS

Theoretical Linguistics Constrains Hypothesis-Driven Causal Abstraction in Mechanistic Interpretability

Suchir Salhan & Konstantinos Voudouris

sas245@cam.ac.uk

How can we develop more hypothesis-driven scientific methodologies for linguistic interpretability (doing Linguistics with Language Models)? How could this be relevant for the broader CogInterp and MechInterp community?

Mechanistic Interpretability is brittle!

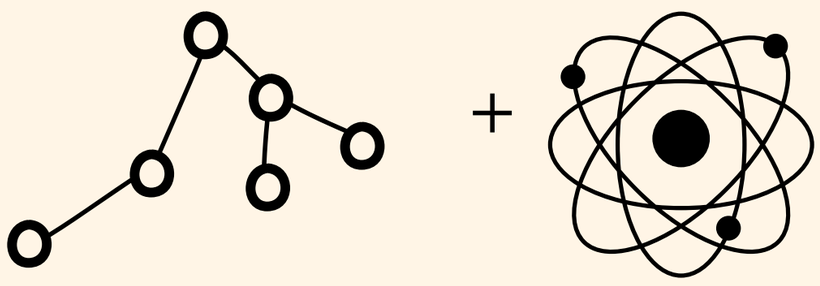
Causal explanations of language model behaviour by mapping hidden activations to hypothesised variables. BUT **alignment maps—functions linking representations to causal factors—are almost unconstrained** (see e.g., Sutter et al 2025).

MI results are hard to reproduce, difficult to generalise, and often unfalsifiable.

Theoretical Linguistics

Linguistic theory provides structured, symbolic templates (phonological rules, morphological processes, syntactic dependencies).

Problem: Without principled constraints, interpretability risks devolving into a combinatorial search over arbitrary mappings, with little guidance for which structures are scientifically meaning.



How can we go from ad-hoc circuit hunting to a scientifically rigorous program of causal abstraction?

How can we generate linguistically constrained causal templates? By lesioning or stimulating the model's internal features involved linguistic rules, we can test whether interventions yield the predicted counterfactual forms—showing that the model encodes not just surface patterns, but the hidden symbolic rules (or rule orderings) themselves.

E.g,. Nasal Assimilation and Opacity in Turkish

Turn on/off the feature that encodes assimilation, or [+back] vowel harmony feature.

	Input	Output	Intervention	Prediction/Counterfactual	
Nasal Assimilation	/sen+de/	[sende]	Lesion	[senge]	Nothing
	/sen+ge/	[seŋge]	Stimulate	[seŋde]	Nothing
Opacity (Harmony)	/kitap+da/	[kitapta]	Lesion		
	/kitap+ta/	[kitapta]	Stimulate		

Stimulating the hidden harmony feature should make a specific internal variable change if the model genuinely encodes harmony; if harmony is not causally represented, nothing in the hidden states will change.

Explanatory Adequacy in CogInterp – how and why do rules arise in a network?

CogInterp (hypothesis-driven, constrained interpretability) could allow us to treat DNNs as experimental model organisms.

By constraining alignment maps to restrictive causal templates, causal abstraction could straightforwardly become a falsifiable framework for modelling the emergence of symbolic representations in language models, if we:

- Restrict alignment maps to subspaces consistent with descriptive linguistic rules
- Treat linguistic generalisations as hypotheses about the internal causal structure of models
- Evaluate causal claims via necessity, sufficiency, and intervention tests
- Quantify representational locality with a localization complexity metric