Signal in the Noise: Polysemantic Interference Transfers and Predicts Cross-Model Influence

Bofan Gong*, Shiyang Lai*, James Evans, Dawn Song

Berkeley UNIVERSITY OF CALIFORNIA · THE UNIVERSITY OF CHICAGO

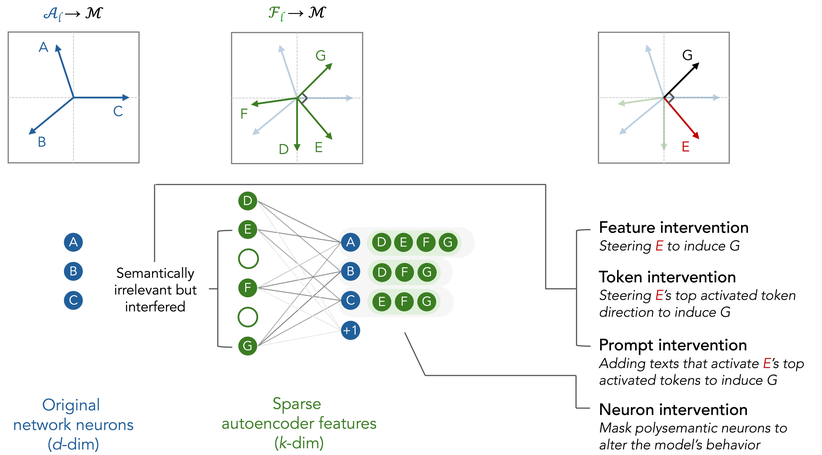
KNOWLEDGE LAB · ICLR

Main question: How do counterintuitive polysemantic structures in large language models (LLMs) causally influence model behavior?

Sub-questions:

(1) To what extent do different models share similar polysemantic structures?

(2) What are the implications of these shared counterintuitive structures for AI interpretability and safety?



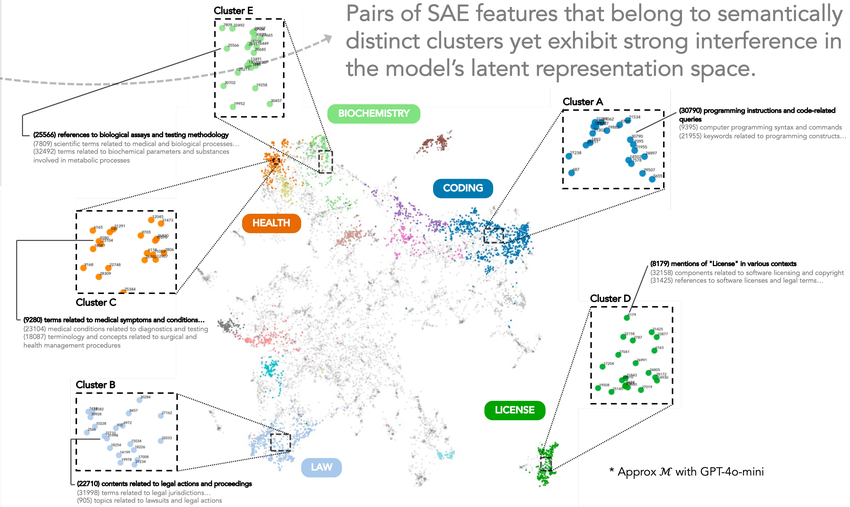
Example:

After trying the new recipe, my brother absolutely __.

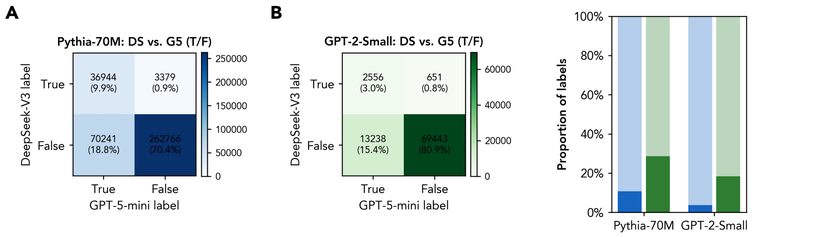
Steer along the "Beethoven" SAE direction

After trying the new recipe, my brother absolutely __.

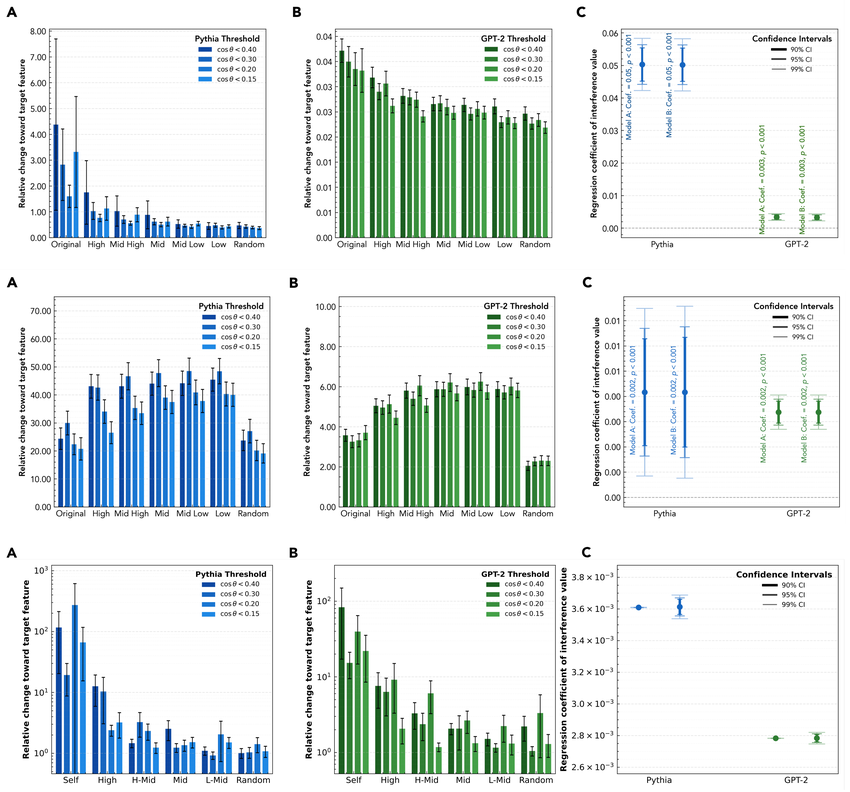
Next token	
fell	Gemma-2-9B
like	Llama-3.1-70B
suffers	Gemma-2-9B
suffers	Llama-3.1-70B



Implication for AI for Science: in most cases, when presented with such counterintuitive yet highly interfering feature pairs, models fail to recognize the underlying association. This suggests that many meaningful structures learned by LLMs remain implicit and cannot be readily verbalized.



Implication for AI Safety: we evidence that counterintuitive feature interference can be systematically leveraged for covert manipulation of model behavior via SAE-based direction steering, gradient-based steering, and prompt injection.



Superposition in large language models is NOT random. Polysemantic features transfer across architectures and training regimes, reflecting stable *yet-not-fully-intelligible* semantic associations that systematically shape model behavior.