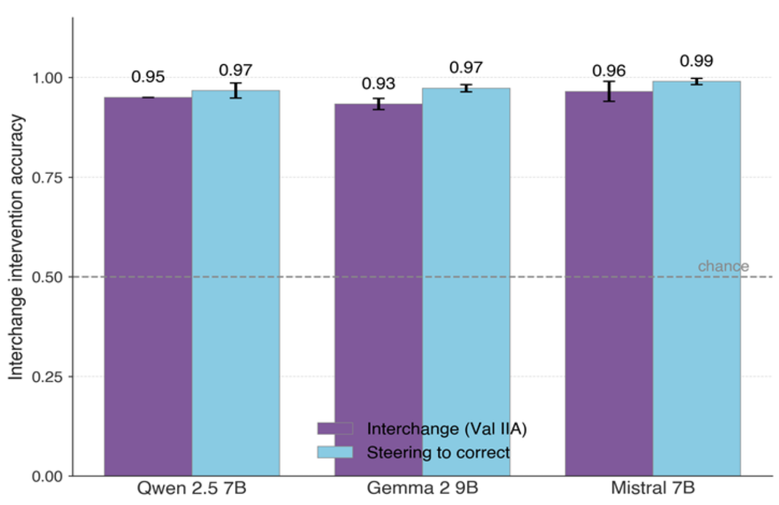
Causal sufficiency does not imply concept-finding

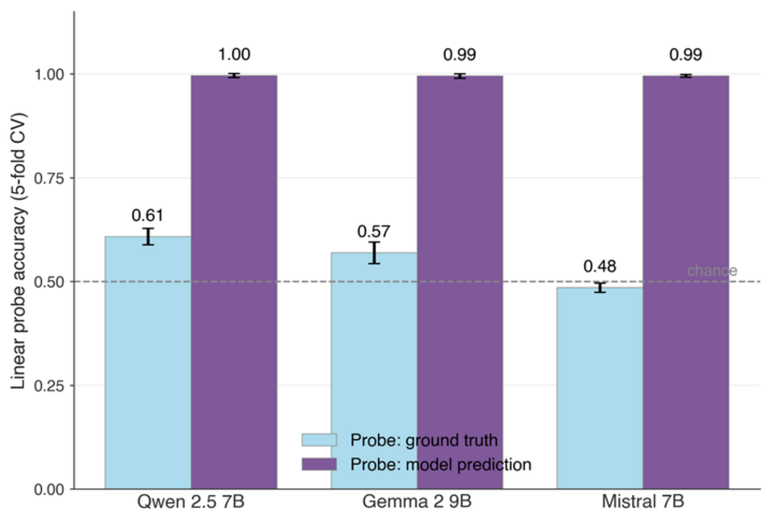
Causal Sufficiency Without Semantic Alignment: How Causal Subspaces Can Masquerade as Semantic Concepts

The question: Mechanistic interpretability treats a causally sufficient subspace as evidence that a model encodes an expected human-interpretable concept. We use a method called Distributed Alignment Search (DAS) for finding causally sufficient subspaces, and ask what DAS finds when accuracy is low (acc. 50–57%). We focus on arithmetic verification ("23 + 47 = 70. Is this true?") across three biased instruction-tuned LLMs.

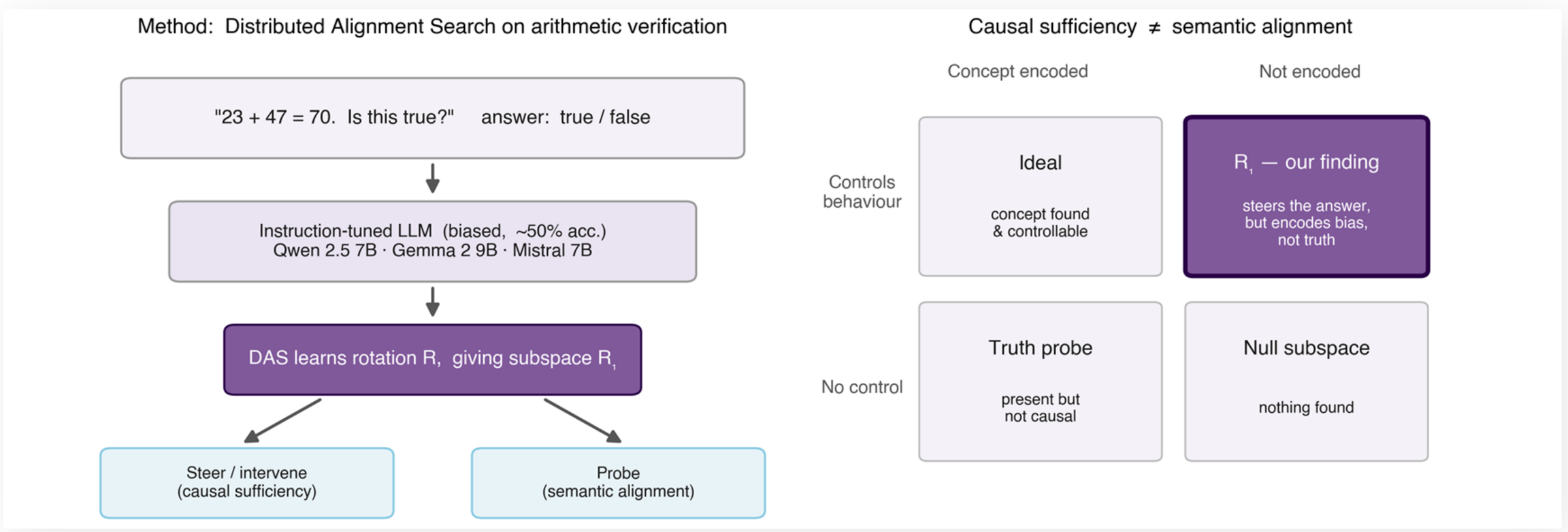
Result 1: The subspace controls model prediction. Patching the subspace achieves 0.93–0.97 interchange accuracy and steers model errors to correct at 97–99%. The subspace even transfers across formats (e.g. twenty-three plus forty-seven equals seventy), and *looks like* a concept of arithmetic truth.



Result 2: The subspace does not encode arithmetic truth. A linear probe reads the subspace activations and achieves ~100% accuracy on the prediction labels, and only 48–61% on the ground truth labels. The subspace encodes model prediction, not arithmetic truth.



Setup



DAS finds a subspace that is causally sufficient but semantically misaligned.

Takeaway & limitation: Causal sufficiency does not entail that a subspace encodes the human concept it appears to capture.

Open question: *how accurate must a model be before a causally sufficient subspace can be claimed as a human-aligned concept?*

PAPER

Mette Friis Andersen, Giovanni Cinà, Sandro Pezzelle

Mech Interp Workshop · ICML 2026