LLMs Develop Abstract Representations of Concepts But Don't Rely on Them?

Causality ≠ Invariance: Function and Concept Vectors in LLMs

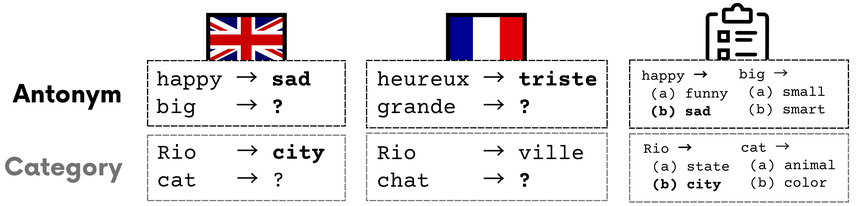
Gustaw Opiełka Hannes Rosenbusch Claire Stevenson

UvA

Motivation

Cognitive science states that **abstract representations** allow for **flexible reasoning in people.** Do they exist in LLMs?

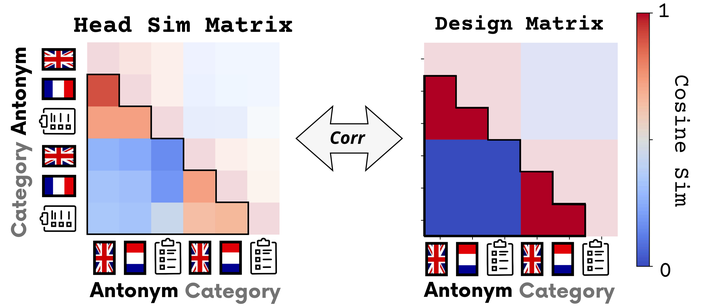
An **abstract representation** of a concept should be **invariant to surface-form changes** (e.g., different languages, open-ended vs. multiple choice prompts).



Methodology

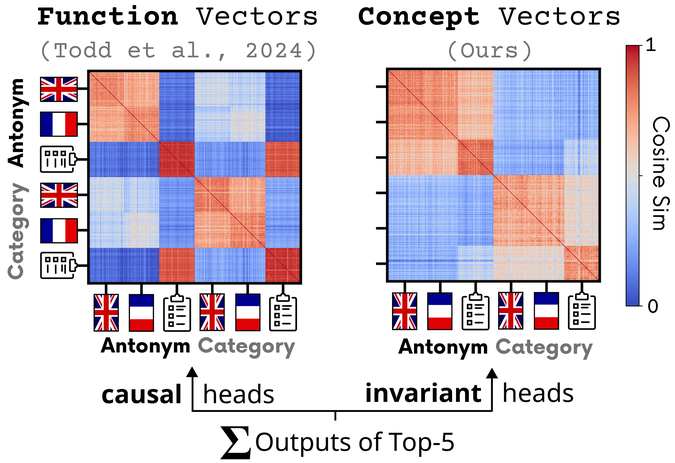
Activation Patching uncovers attention heads that **causally** drive in-context-learning (ICL) performance:

Representational Similarity Analysis (RSA) uncovers attention heads that **invariantly** represent concepts:

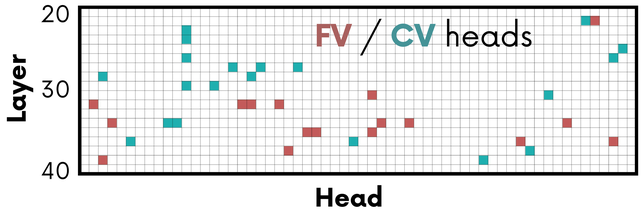


Takeaways

1. Abstract representations exist in LLMs (CVs) but they are not the ones driving ICL performance (FVs)



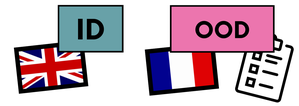
2. Abstract representations are mechanistically separate from causal representations



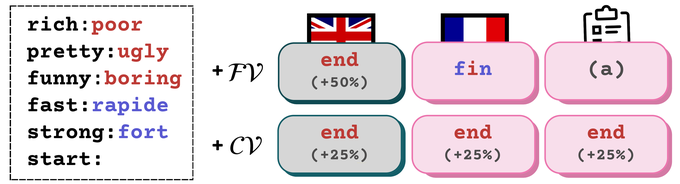
FVs and CVs are in **similar layers** but are composed of **different attention heads.**

3. FVs steer better in-distribution, CVs steer more consistently out-of-distribution

1. extract **antonym** CVs/FVs from ID or OOD prompts



2. steer the model to respond with an **antonym** not **ENG-to-FR**



ICLR