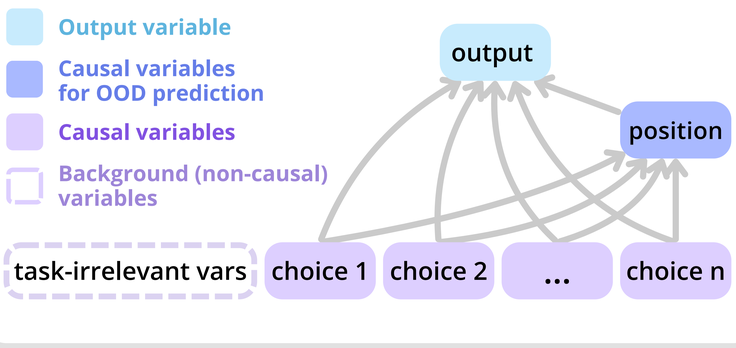
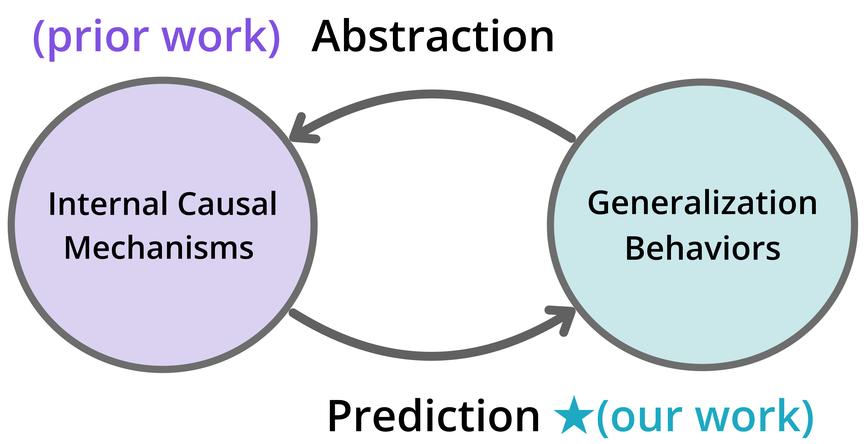
Internal Causal Mechanisms Robustly Predict Language Model Out-of-Distribution Behaviors

Jing Huang, Junyi Tao*, Thomas Icard, Diyi Yang, Christopher Potts*

Interp Finding: Causal Mechanisms of MCQA





Task: Predict OOD Behaviors on MMLU

Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.

ID Scenario	***OOD Scenario***
A. 0	Alpha. 0
B. 4	Bravo. 4
C. 2	Charlie. 2
D. 6	Delta. 6
Answer: B.	**Answer: *Delta*.**

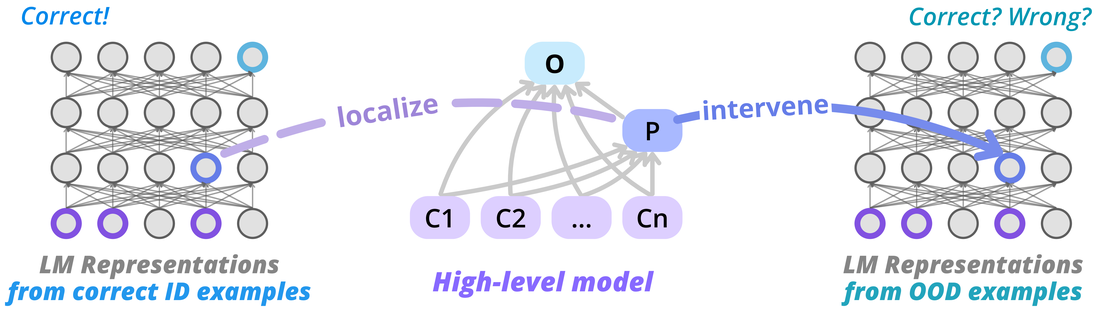
Methods: Abstraction → Prediction

The model solves a task successfully → it likely implements a **systematic solution, i.e. a causal mechanism**

The model implements **the same causal mechanism** on an OOD example → it likely predicts the example **correctly**

Abstract the high-level causal model from ID examples the model correctly solves

Predict the output correctness by checking the implementation of key causal variables

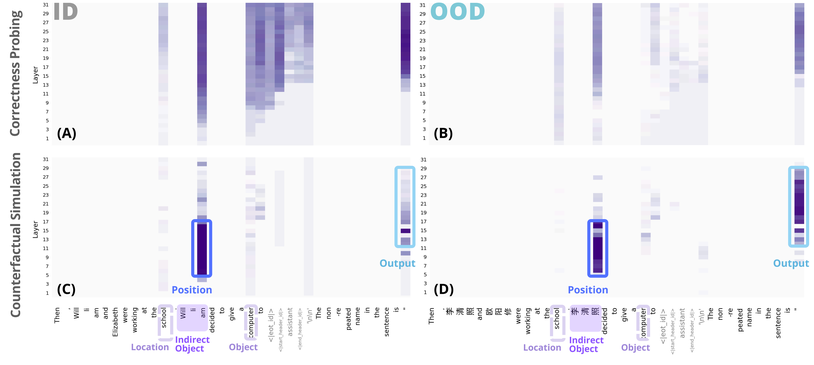


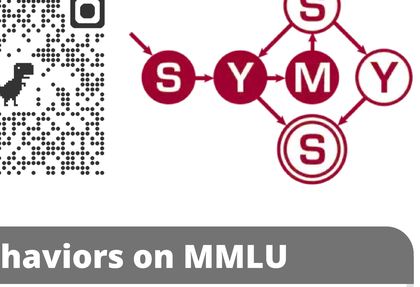
Measure the extent to which an abstraction exists via **interchange intervention accuracy**

Experiment Results

The **most robust features** for correctness prediction are those that play a **causal** role in the model's behavior.

ID and OOD Probing and Intervention Results





Interchange Intervention accuracy reliably predicts model output **correctness**.