Competence-based Analysis of Language Models

Adam Davies, Jize Jiang, ChengXiang Zhai
Department of Computer Science, University of Illinois Urbana-Champaign

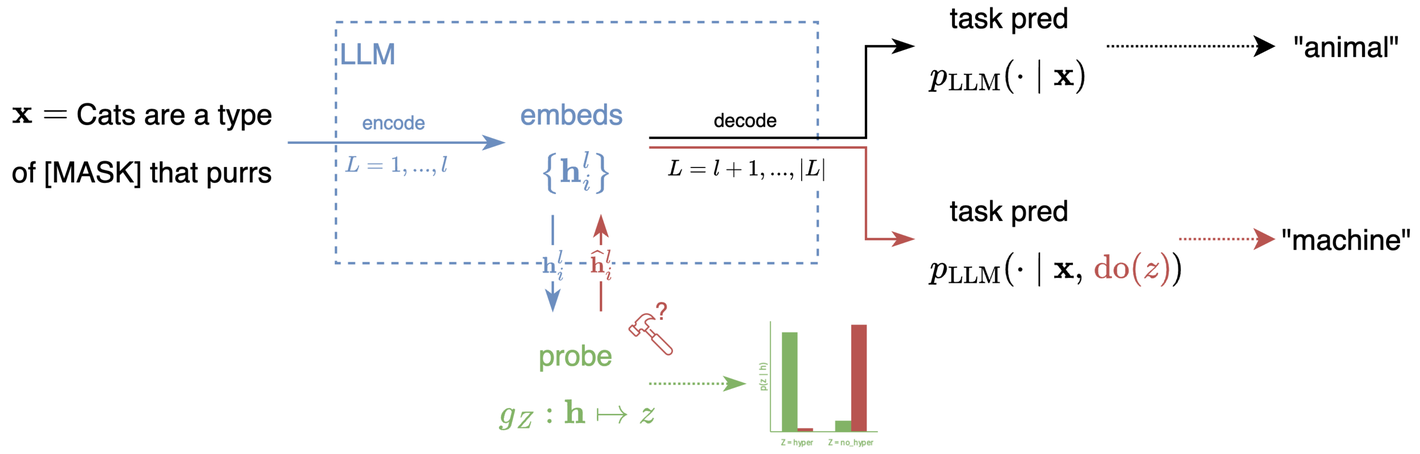
What properties does LLM use to perform a task?

Causal → robust
Spurious → brittle

CAUSAL PROBING

1. **Train probe** to predict property
2. **Intervene on probe** to modify representation
3. **Analyze impact** on model behavior

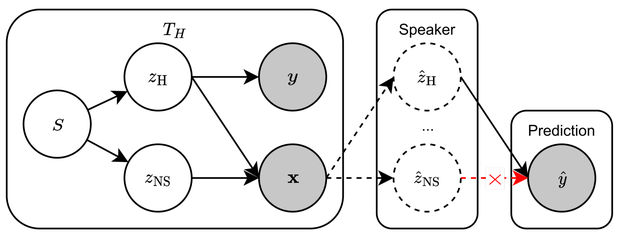
CALM: use causal probing to measure **competence**
- *Change predictions* when modifying *task-causal* properties
- *Stays the same* when modifying *spurious* properties



CALM FRAMEWORK

Causal Task Structure

- LLM M
- Task $\mathcal{T} \sim P(\mathcal{X}, \mathcal{Y})$
- Set of properties $\boldsymbol{Z} = \{Z_j\}$ taking values $\boldsymbol{z} \in \{z_k\}$ for a given input $\boldsymbol{x}$
  - Decompose $\boldsymbol{Z} = \boldsymbol{Z_c} \cup \boldsymbol{Z_e}$ for *causal* vs *environmental* properties
  - Dependency $M(\boldsymbol{X}|do(Z_j)) \neq M(\boldsymbol{X})$ is *spurious* if $Z_j \in \boldsymbol{Z_e}$



- Structural causal model $\mathcal{G}_\mathcal{T}$
  - Nodes $\boldsymbol{Z} \cup \{\mathcal{Y}\}$ for set of properties $\boldsymbol{Z} = \{Z_j\}$
  - Edges denote causal dependencies
  - Decompose $\boldsymbol{Z} = \boldsymbol{Z_c} \cup \boldsymbol{Z_e}$ for *causal* (path to $\mathcal{Y}$) vs *environmental* (spurious, no path)

Measuring Competence

Define *competence* of M w.r.t. $\mathcal{T}$ as alignment with $\mathcal{G}_\mathcal{T}$

$$\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}, \boldsymbol{z}\sim\text{val}(\boldsymbol{Z})}\, S(M(\boldsymbol{x}|\text{do}(\boldsymbol{z})), \mathcal{G}_\mathcal{T}(\boldsymbol{x}|\text{do}(\boldsymbol{z})))$$

- Reformulation of Interchange Intervention Accuracy:
  - IIA uses *interchange interventions:* extract representation of property $Z = z$ from source $\boldsymbol{x_s}$ to patch into target $\boldsymbol{x_t}$
- CALM uses *causal probing interventions* instead:
  - Operate at *concept-level*
  - No need to "borrow" representations from other inputs $\boldsymbol{x_s}$
  - Can study unseen combinations of $\boldsymbol{Z}$ (required for simulating OOD)

EXPERIMENTS

Dataset

LAMA ConceptNet: 14 *lexical inference tasks* for masked-language models
- *Hypernym prediction* (IsA)*:* "cats are a type of [MASK] that purrs"
- Also includes relations like PartOf, HasProperty, etc.

Competence Approximation

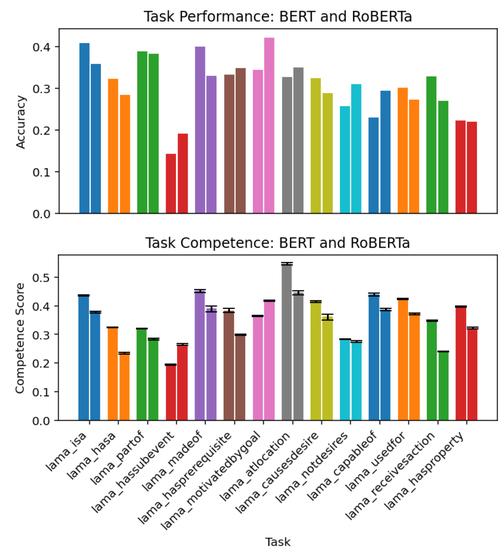
Approximate $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ via $E = (\boldsymbol{Z}, \mathcal{G}_\mathcal{T}, S)$
- $\boldsymbol{Z}$ is the set of relations corresponding to each of the 14 lexical inference tasks
- $\mathcal{G}_\mathcal{T}$ is SCM with a single edge (as determined by the task $\mathcal{T}$; other relations are $\boldsymbol{Z_e}$)
- S is the overlap between top-k predictions before and after intervention

Experiments

Implement interventions using **GBIs**
- Probe is MLP over final embedding layer
- Attack probe using FGSM and PGD (constrain collateral damage via ϵ)

Measure average approximated $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ of **BERT** and **RoBERTa** on each task
- Averaged over 10 experimental runs (randomly re-initialize probes)

RESULTS



BERT (left bars) and RoBERTa (right bars)

- Both models are partially competent across tasks
  - Always higher than random baseline (0.0714)
- Competence is predictive of relative task performance (Spearman's $\rho = 0.508, p = 0.064$)
- Explains earlier findings re: hypernym prediction
  - Great performance with engineered prompts, but fail under small changes to prompts
  - *Explanation:* intermediate competence means models rely on both task-causal and spurious lexical properties

GRADIENT-BASED INTERVENTIONS

Prior work in causal probing has used *INLP:*
- *Nullifies* representation of property
- Assumes *linear representation*

For our experimental setting, we need *nonlinear counterfactual* interventions
- *Nonlinear* required for *relational* properties
- *Counterfactual* required for measuring *competence*

Gradient-Based Interventions

Use *gradient-based adversarial attacks* against probes g to *minimize probe loss* wrt counterfactual target $y' \neq y$

E.g., FGSM: $\boldsymbol{h}' = \boldsymbol{h} + \epsilon\ \text{sgn}(\nabla_{\boldsymbol{h}} \mathcal{L}(g, x, y'))$
- *Flexible:* can use any differentiable probe
- *Controllable:* can modulate perturbation magnitude ϵ

PAPER

ILLINOIS