A Property Induction Framework for Neural Language Models

Kanishka Misra[1], Julia Taylor Rayz[1], Allyson Ettinger[2]
[1]Purdue University [2]University of Chicago
Correspondence: `kmisra@purdue.edu`

Paper:

PURDUE UNIVERSITY

THE UNIVERSITY OF CHICAGO

Overview and Motivation

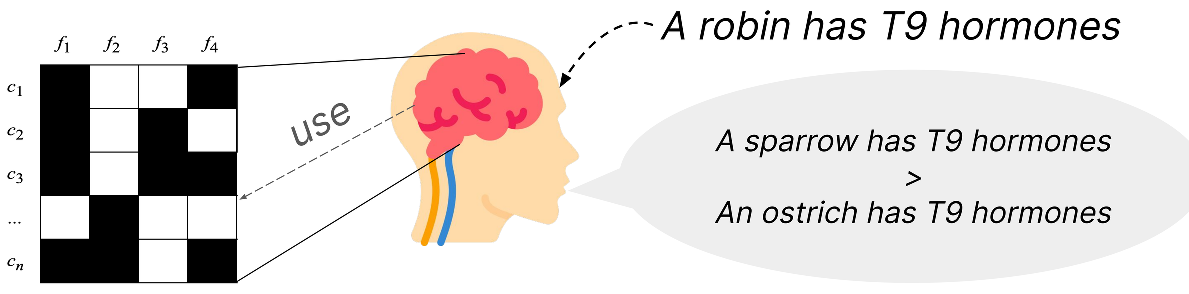
Overall Question: To what extent do models that only rely on language experience learn about everyday concepts and their properties?

Approach: Study the synthetic semantic knowledge of language models by investigating how they perform **property induction.**

Motivation: Property-inductions made by humans have provided context within which cognitive scientists have explored the nature and organization of human conceptual knowledge.

Property Induction

- Inferences that go beyond available data to project novel information about concepts and properties (Osherson et al., 1990; Hayes and Heit; 2018)
- Provide interesting insight into the inductive preferences of humans, in reasoning about concept and property knowledge



Main Idea: *Use Property Induction as a tool to to study how knowledge representation in language models drives inductive generalization with respect to entirely novel properties*

Contributions

In terms of... Goals:

- Different from reasoning that is required for "natural language inference" (Bowman et al., 2015) which is deductive in its formulation.

In terms of... Methodology:

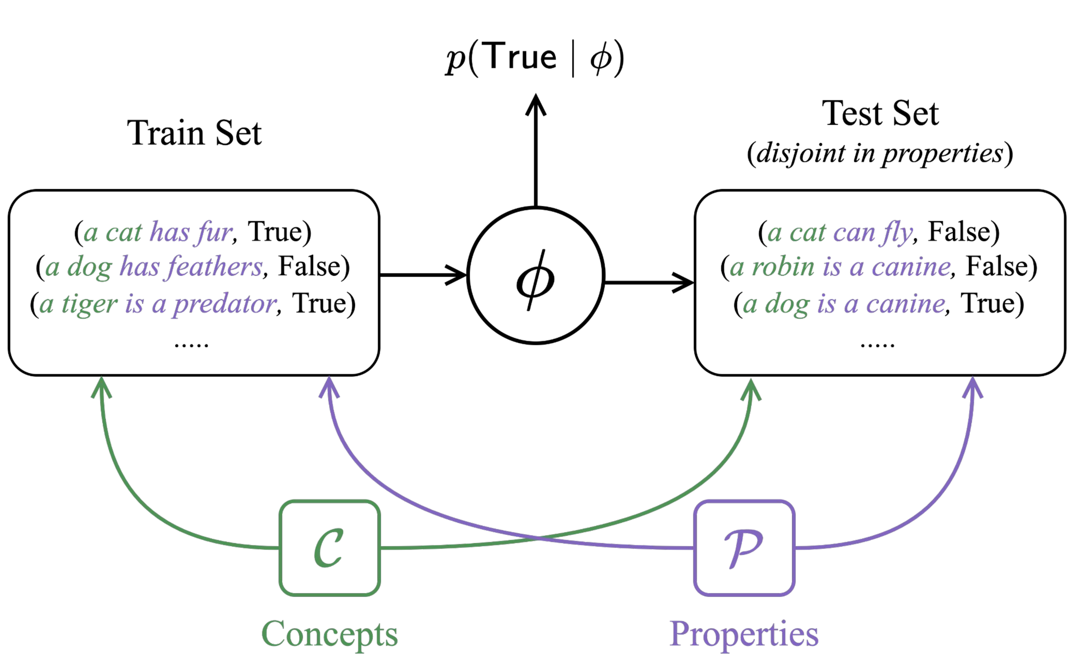
- New paradigm to study generalizations in LMs beyond what they have observed in training.
- Extends line of work on property induction in neural networks (Sloman, 1993; Rogers and McClelland, 2004; Saxe et al., 2019, Misra et al., 2021).

In terms of... Findings:

- When fine-tuned on conceptual knowledge, **LMs acquire a taxonomic preference in generalizing novel property information**, that cannot be explained by simple training data statistics.

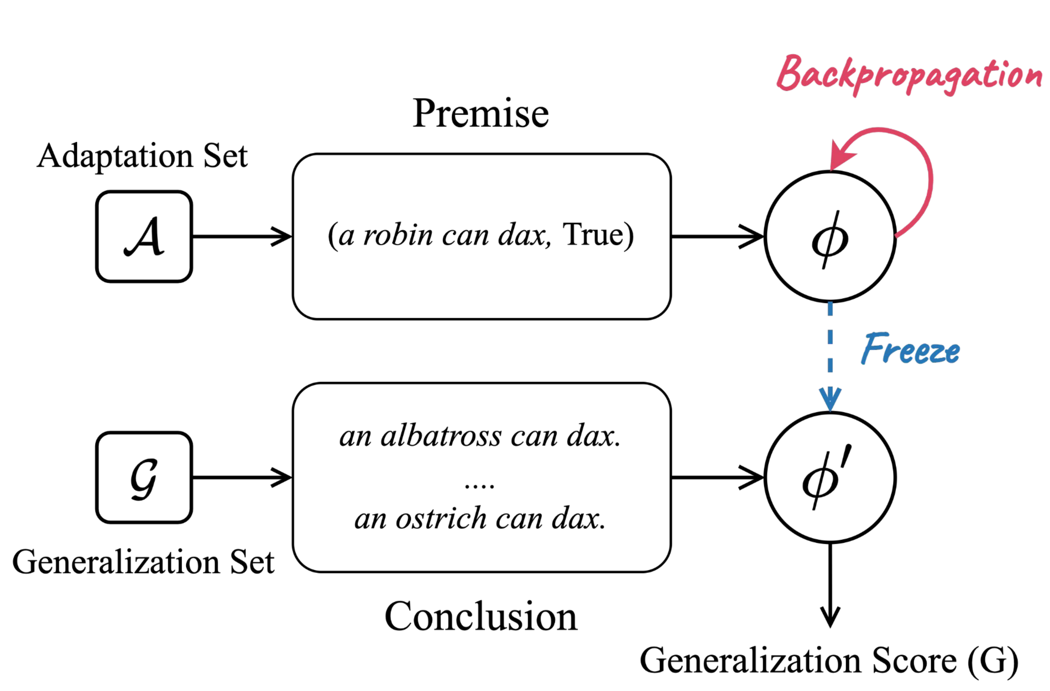
The Framework

Stage 1: Eliciting Property Judgments from LMs



- Equip existing language models with binary judgments of concept-property associations (Bhatia and Ritchie, 2021).
  - *A robin can fly* → **True**
  - *A cat can fly* → **False**
- **Setup:** LM fine-tuned to perform binary classification with disjoint set of properties between train and test sets.
- See exp. 1 for results.

Stage 2: Property Induction as Adaptation



$$G = \frac{1}{n}\sum_{c_i \in \mathcal{G}} \log p(\text{True} \mid \text{“}c_i \textit{ can dax.''}, \phi')$$

Operationalization of induction: Behavior of LM (from stage 1) after it has been further adapted to novel property information.

Property-induction trial:

1. Adapt LM to reflect novel property information for a few known concepts (adaptation set; e.g., *a robin can dax*), and freeze.
2. Query adapted LM to assess generalization of novel property to other concepts (generalization set; e.g., *a canary/giraffe can dax*).
3. Reset LM for next trial.

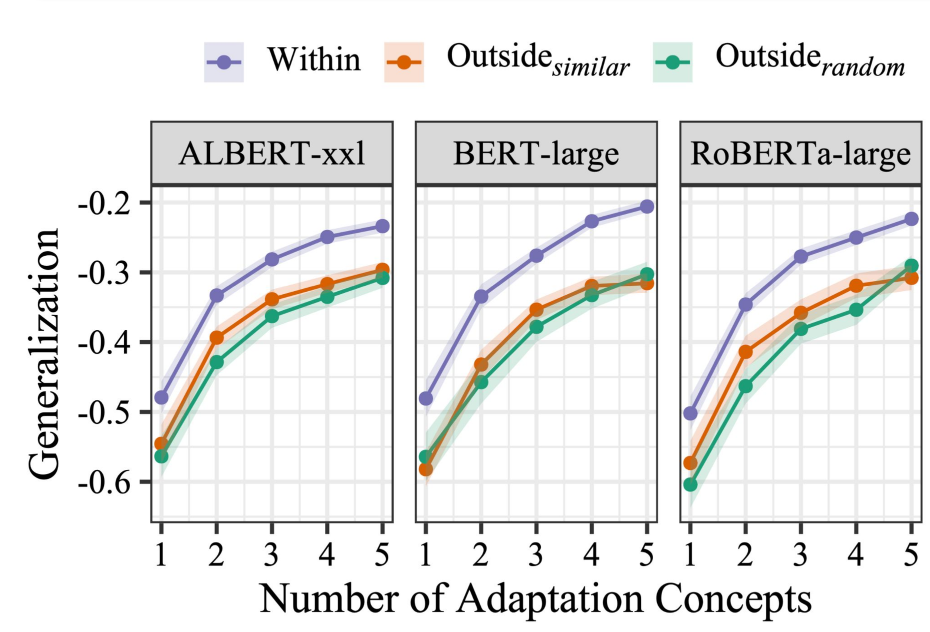
Exp. 1: Property Judgments and LMs

- Models (ALBERT-xxl; BERT-large; RoBERTa-large) fine-tuned on sentences formed by linking concepts to properties - sourced from the CSLB dataset (Devereaux et al., 2014).
- 521 concepts & 3735 properties, corresponding to 46,214 true and false sentences (equal distribution)
- Models show similarly high performance on the test set (0.78 - 0.79).

Model	F1
ALBERT-xxl	0.79
BERT-large	0.78
RoBERTa-large	0.79

Table 1: F1 scores on the test set. Chance = 0.66

Exp. 2: Taxonomic Generalization in LMs



Compare generalization of a novel property (e.g., *can dax*) based on category membership (N = 2400).

Adaptation: *A crow has blickets, True*

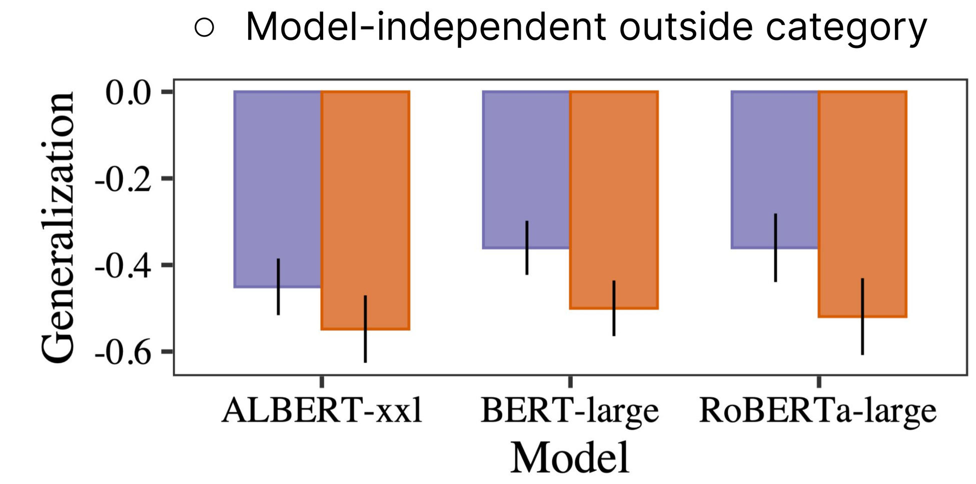
Generalization:

- **Within:** *A <bird> has blickets.*
  - Within-category
- **Outside$_{Similar}$:** *A bat has blickets.*
  - Model-dependent outside category
- **Outside$_{Random}$:** *A table has blickets.*
  - Model-independent outside category

Sub-experiment: Inductive generalizations from concepts that share more properties with superordinate categories ***different*** than their own (N = 48).

Adaptation: *A dolphin can dax, True*

- **Within:** *A <mammal> can dax.*
- **Outside:** *A fish can dax.*



Takeaway: *Models prefer to generalize new properties to concepts that are in the same taxonomic category (Within) as opposed to those that are not (Outside).*

Summary

- Language Models show strong capacities to assess the association of properties to concepts when expressed in natural language form.
- Generalization of novel properties to known concepts in LMs is--at least in part--guided by category membership, indicating the presence of a taxonomic bias.
- **Hypothesis:** Some of models' taxonomic preference could be due to high property overlap between concepts of the same category observed in training (Exp 1).
  - Findings persisted even when property overlap and category membership were teased apart (see sub-experiment)!

Plans for Future Work

Language Models and their evaluation

- Characterize other qualitative reasoning behavior in LMs, inspired from observations in property induction literature.
- Create "Inductive Reasoning" challenge sets that target specific forms of reasoning involving concepts and properties.

Cognitive Modelling?

- Compare against human behavioral results in property induction literature:
  - **Fine-grained taxonomic phenomena** (Osherson et al., 1990)
  - **Theory-based property induction** (Kemp and Tenenbaum, 2009)

Thanks!

- **CompLing Lab** at UChicago for discussions on earlier iterations of the work.
- **Purdue RCAC** and **Hemanth Devarapalli** for compute assistance.
- **Anonymous Reviewers** for their helpful feedback.
- **Charles Kemp, Tom McCoy, Najoung Kim,** and **Keith Ransom** for encouragement and support!
- **Purdue Graduate School** for a Travel award!