A Property Induction Framework for Neural Language Models

Kanishka Misra,[1] Julia Taylor Rayz,[1] and Allyson Ettinger[2]
[1]Department of Computer and Information Technology, Purdue University, IN, USA
[2]Department of Linguistics, University of Chicago, IL, USA
kmisra@purdue.edu, jtaylor1@purdue.edu, aettinger@uchicago.edu

Abstract

To what extent can experience from language contribute to our conceptual knowledge? Computational explorations of this question have shed light on the ability of powerful neural language models (LMs)—informed solely through text input—to encode and elicit information about concepts and properties. To extend this line of research, we present a framework that uses neural-network language models (LMs) to perform property induction—a task in which humans generalize novel property knowledge (*has sesamoid bones*) from one or more concepts (*robins*) to others (*sparrows, canaries*). Patterns of property induction observed in humans have shed considerable light on the nature and organization of human conceptual knowledge. Inspired by this insight, we use our framework to explore the property inductions of LMs, and find that they show an inductive preference to generalize novel properties on the basis of category membership, suggesting the presence of a taxonomic bias in their representations.

Keywords: property induction; language models; semantic cognition; generalization; conceptual knowledge

Introduction

There has recently been a growing interest in exploring the limits and potential of language as an environment for learning conceptual knowledge (Elman, 2004; Lupyan & Lewis, 2019)—knowledge that encompasses mental representations of everyday objects/events, and their properties and relations, that together inform our intuitive understanding of the world (Murphy, 2002; Machery, 2009). Computational explorations of this claim often study the extent to which models that learn semantic representations through text alone can capture conceptual knowledge (Lucy & Gauthier, 2017; Forbes et al., 2019; Da & Kasai, 2019; Bhatia & Richie, 2021).

A hallmark feature of the conceptual knowledge acquired by humans is its capacity to facilitate inductive generalizations: inferences that go beyond available data to project novel information about concepts and properties (Osherson et al., 1990; Chater et al., 2011; Hayes & Heit, 2018). For example, our knowledge of taxonomic specificity is reflected when we generalize a novel property of a concept (e.g., *robins have T9 hormones*) more strongly to taxonomically close concepts (*sparrows have T9 hormones*) than to more taxonomically distant concepts (*tigers have T9 hormones*). Inductive generalizations about novel properties (also called *property induction*) therefore provide a context within which we can explore the nature of agents' understanding of conceptual knowledge. In this paper, we develop an analysis framework that uses neural network-based language models (LMs, henceforth) to perform property induction and use this framework to study concept representation in these models. Our framework consists of two stages. In the first stage, we train LMs to evaluate the truth of sentences expressing property knowledge (e.g., *a cat has fur* → True, *a table has fur* → False). In the second stage, we use these property-judgment models to test how the representations from the underlying LMs drive inductive generalization of novel properties—e.g., *has feps, can dax,* etc.

Each stage of our framework sheds light on different aspects of the conceptual knowledge captured by LMs. Using the first stage, we test the extent to which LMs support judgement of whether a property applies to a concept, even when that property has not been seen in task-specific fine-tuning. We find that LMs perform substantially above chance, consistent with the conclusion that they are able to rely on generalizable property knowledge to assess truth of concept-property associations. In the second stage, we use this property judgment framework to study how knowledge representation in the base LMs drives inductive generalization with respect to entirely novel properties. We focus specifically on whether models' inductive preferences indicate reliance on taxonomic information, by testing whether models prefer to generalize within rather than outside of taxonomic categories. To do this, we teach our property-judgment models novel property information such as *robins can dax* via standard backpropagation methods and then test the extent to which they prefer generalizing this novel property to other birds (e.g. *sparrows can dax*) more strongly than to non-birds (e.g. *zebras can dax*). We find that models indeed show a preference for projecting new property knowledge on the basis of taxonomic category membership, suggesting that the models have acquired and represented taxonomic features on which they rely to project novel information.

Our LM-based account of property induction contributes to the field in four primary ways. On the basis of the goals of the task, our framework focuses on reasoning where conclusions do not deductively follow from the premise, unlike the goals of the more commonly-used task of natural language inference (Bowman et al., 2015), and it therefore allows for testing of human-like inferences that are seldom studied in LMs (cf. Bhagavatula et al., 2020). Next, as we show below, our framework opens a new window into exploring how large neural network models of language generalize beyond their training experience, complementing inquiries of models' inductive bias with respect to syntactic structure (McCoy, Frank, & Linzen, 2020) and "universal linguistic constraints" (McCoy, Grant, et al., 2020). Additionally, this work advances research aiming to diagnose the nature and extent of conceptual knowledge in LMs (Da & Kasai, 2019; Forbes et al., 2019; Weir et al., 2020; Bhatia & Richie, 2021) by additionally focusing on how knowledge present in LM representations drives the generalizations they make. Finally, at a high level, our framework contributes to a range of works that have applied connectionist models to the problem of property induction (see Sloman, 1993; Rogers & McClelland, 2004; Saxe et al., 2019).

Testing Property Induction with Arguments

Property induction is often studied in humans through the use of arguments, represented in the following premise-conclusion format, as popularized by Osherson et al. (1990):

$$\frac{\text{Robins have sesamoid bones.}}{\text{All birds have sesamoid bones.}} \tag{i}$$

Argument (i) is read as *"Robins have sesamoid bones. Therefore, all birds have sesamoid bones."* The subject of the premise sentence (*robin*) is referred to as the premise concept (similarly, if there are multiple premises, we have a set of premise concepts), while that of the conclusion is called the conclusion concept. Representing induction stimuli as arguments allows one to use the notion of "argument strength," which quantifies the degree to which a human subject's belief in the premise statements strengthens their belief in the conclusion (Osherson et al., 1990). In many cases, researchers control the type of novel properties provided to participants by using *blank* properties—properties that are synthetically created and are therefore unknown to participants, maximizing the chances that they will use their knowledge of the relations between the premise and conclusion concepts to make generalizations (Rips, 1975; Osherson et al., 1990; Murphy, 2002). In our property induction experiments, we simulate blank properties by using nonce words to synthetically construct novel properties—e.g., *can dax*, *is vorpal*, etc and use them to explore knowledge of conceptual relations in LMs.

The Framework

Computationally, property induction can be viewed as making conditional probability estimates about the conclusion (c), given some premise (π): $p(c \mid \pi)$. We interpret this measure in our framework as a probability that a novel property is applied to a conclusion concept, by a model whose representations reflect the premise information. This interpretation leads to two desiderata that our framework aims to satisfy: (1) the ability to make judgments about the association of properties to concepts, and (2) the ability to accept new property knowledge and then be queried to assess generalization of this new property knowledge to additional concepts. To satisfy (1), we fine-tune existing pre-trained LMs to classify as true or false sentences that associate properties to concepts—i.e., make property judgments. Doing so enables the LMs to estimate the probability that a property applies to a concept, as $p(\texttt{True} \mid \textit{“concept has property”}, \phi)$, where ϕ stands for the parameters of a given LM. We use this approach rather than estimating sequence probabilities—which are relatively more straightforward to compute using LMs—in order to avoid surface-level confounds as observed in similar work by Misra et al. (2021). Next, to satisfy (2), we operationalize induction as the behavior of these LMs (now fine-tuned to make property-judgments) after further adaptation to new properties using standard backpropagation (Rumelhart et al., 1986). The motivation to use backpropagation to perform property induction is simple—it allows the integration of new information in the model by directly updating its representations, which encode knowledge used to inform how the model generalizes. Under this operationalization, we first adapt our LM to reflect the premise information (e.g. $\pi =$ *a robin can dax*) and then use the updated parameters of the model (ϕ') to estimate the probability of a conclusion ($c =$ *a sparrow can dax*) as: $p(\texttt{True} \mid c, \phi')$. A similar operationalization of induction was used by Rogers & McClelland (2004), who reported inductive inferences made by their PDP model of semantic cognition by updating its weights to reflect novel information, which was provided after several steps of training on general conceptual knowledge derived from a toy dataset of concepts and properties. Similar methods have also been used by van Schijndel & Linzen (2018) and Kim & Smolensky (2021) to characterize the adaptation of grammatical knowledge in LMs. We now explain the two stages of our property induction framework in greater detail:

Stage 1: Eliciting Property Judgments using LMs

In the first stage, we constrain LMs to explicitly rely on property knowledge by distinguishing correct (*cat has whiskers*) and incorrect associations (*sparrow has whiskers*) between properties and concepts. We do this by fine-tuning LMs to classify sentences that express concept-property associations to be true or false. Importantly, we fine-tune models in a way that keeps the evaluation sets disjoint in terms of properties—i.e., the model is trained to assess the properties *has feathers, has a tail* and then tested on a distinct set of properties: *can fly, has a beak*. Therefore, in order to succeed on this task (i.e., minimize loss on a disjoint evaluation set), a model must rely on property knowledge encoded in its representations, to enable judgments about properties never seen during fine-tuning. In experiments that follow, we verify the extent to which the models are indeed able to draw on generalized property knowledge in order to succeed in this task. Importantly, this stage assumes the presence of a repository of concepts ($\mathcal{C}$) and associated properties ($\mathcal{P}$) as data for training and testing the model. We create sentences that express property knowledge by pairing properties from $\mathcal{P}$ to concepts from $\mathcal{C}$. We then fine-tune the LM to classify these sentences as true or false. At the end of this stage, we have a trained model (with parameters ϕ) that takes as input a sentence s and produces a probability score $p(\texttt{True} \mid s, \phi)$ corresponding to the degree of truth of s as internalized by the LM. Figure 1A illustrates the property judgment stage.

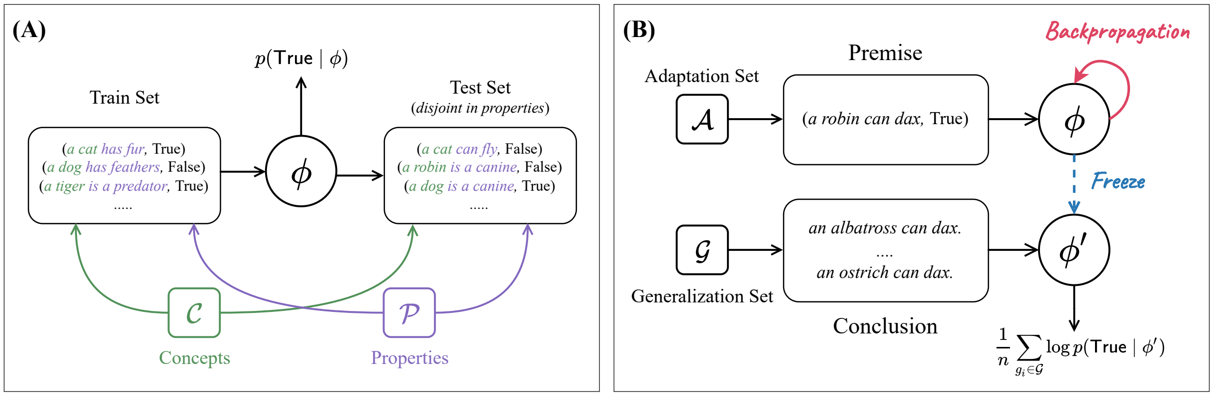

Figure 1: **(A)** Property Judgment Stage describing the training of the property judgment model (with parameters ϕ) to make judgments of truth on sentences expressing concept-property assertions. Sentences created using the concept ($\mathcal{C}$) and property ($\mathcal{P}$) data collected by Devereux et al. (2014); **(B)** Depiction of the Induction Stage, in this case, for testing the generalization of the novel property *can dax* from robin to all birds. Here, $\mathcal{A} = \{\text{ROBIN}\}$, $\mathcal{G} = \{\text{ALBATROSS}, ..., \text{OSTRICH}\}$.

Stage 2: Induction as Adaptation to New Knowledge

In this stage (see Figure 1B), we use the fine-tuned model from the previous stage to perform property induction, which we operationalize as the behavior of the model after adaptation to new property knowledge via backpropagation.

A property induction trial involves (1) a set of premise concepts (which we denote as the adaptation set $\mathcal{A} \subset \mathcal{C}$); (2) a set of conclusion concepts (denoted as the generalization set $\mathcal{G} \subset \mathcal{C}$); and (3) a novel property being generalized from the premise to the conclusion. We construct sentences that associate the novel property to the concepts in $\mathcal{A}$ and $\mathcal{G}$, yielding the premise and conclusion stimuli, respectively (see Figure 1B). To perform property induction, we first adapt the model's parameters ϕ to the premise sentences by using standard backpropagation, yielding an updated state of the model, ϕ', that correctly attributes the concepts in $\mathcal{A}$ with the novel property. We then freeze ϕ' and query the model with the conclusion sentences to obtain the (log) probability of generalizing (or "projecting") the novel property to the concepts in $\mathcal{G}$. We refer to this measure as the "generalization score" (G)—i.e., the strength of projecting the novel property to a set of one or more concepts in the generalization set:

$$\text{G} = \frac{1}{n}\sum_{c_i \in \mathcal{G}} \log p(\text{True} \mid \text{“}c_i \textit{ has property X}\text{”}, \phi') \qquad (1)$$

The model parameters are reset to their original state (ϕ) after this step in order to perform subsequent trials.

We now use components of this framework in two experiments—one for each stage in the framework.

Investigating LMs on Property Judgments

Our first experiment focuses on the first stage of the proposed induction framework. Here, we fine-tune pre-trained LMs to evaluate the truth of sentences attributing properties to concepts—i.e., we want our models to map the sentence *a cat has fur* to True and *a cat can fly* to False. We use an existing semantic property norm dataset to construct our sentences and split them into disjoint evaluation sets, where the properties we test the model on are strictly different from those the model sees during fine-tuning. Therefore, a model must learn to rely on its 'prior' (pre-trained) property knowledge in combination with task specific information it picks up during fine-tuning in order to succeed on this task.

Ground-truth Property Knowledge Data To construct sentences that express property knowledge, we rely on a property-norm dataset collected by the Cambridge Centre for Speech, Language, and the Brain (CSLB; Devereux et al., 2014). The CSLB dataset was collected by asking 123 human participants to elicit properties for a set of 638 concepts, and this dataset has been used in several studies focused on investigating conceptual knowledge in word representations learned by computational models of text (e.g., Lucy & Gauthier, 2017; Da & Kasai, 2019; Bhatia & Richie, 2021). Importantly, property-norm datasets such as CSLB only consist of properties that are applicable for a given concept and do not contain negative property-concept associations. As a result, the works that have used these datasets sample concepts for which a particular property was not elicited and take them as negative instances for that property (e.g., TABLE, CHAIR, SHIRT are negative instances for the property *can breathe*), which can then be used in a standard machine-learning setting to evaluate a given representation-learning model.

Upon careful inspection of the CSLB dataset, we found that the above practice may unintentionally introduce incorrect or inconsistent data. Datasets such as CSLB are collected through human elicitation of properties for a given concept, so it is possible for inconsistencies to arise. One way that this may happen is if some participants choose not to include properties that are obvious for the presented concept (e.g., *breathing* in case of living organisms), while other participants do, resulting in an imbalance that can be left unaccounted for. We found that this was indeed the case: e.g., the property *has a mouth* was only elicited for 6 animal concepts (out of 152), so all other animals in the dataset would have been added to the negative search space for this property during sampling, thereby propagating incorrect and incomplete data. This indicates a potential pitfall of directly using property-norm datasets to investigate semantic representations—and suggests that prior evaluations and analyses (Lucy & Gauthier, 2017; Da & Kasai, 2019; Bhatia & Richie, 2021) may have falsely rewarded or penalized models in such cases. Owing to space constraints, we provide our detailed method and protocol to mitigate this problem in the supplemental materials. The revised dataset that we produce consists of a set of 521 concepts, corresponding to 23 different taxonomic categories (as annotated by the original authors of the CSLB dataset) and 3,735 properties, with 23,107 ground-truth property-concept pairs which we used in our experiment.

For each of our 3,735 properties—associated with k different concepts—we sample k additional concepts that are maximally similar to the k concepts associated with that property, and take these to be negative samples. For instance, for the concept ZEBRA, we want to use HORSE for a negative sample rather than a more distant concept such as TABLE. By doing this, we make the property judgment tasks more difficult, increasing the chances that the models we obtain from this stage focus on finer-grained conceptual/property knowledge as opposed to coarser-grained lexical similarity. For selecting similar concepts we take the *Wu-Palmer similarity* as our similarity function (Wu & Palmer, 1994), which we compute over the subset of the WordNet taxonomy (Miller, 1995) that contains the senses of the 521 concepts considered in our experiments. We then follow the method outlined by Bhatia & Richie (2021) to convert our 46,214 property-concept pairs (23,107 × 2) into natural language sentences, which we then use as inputs to our models. We split these sentences (paired with their respective labels) into training, validation, and testing sets (80/10/10 split), such that the testing and validation sets are only composed of properties that have never been encountered during training (note that properties between training and validation sets are also disjoint). We do this to avoid data leaks, and to ensure that we evaluate models on their capacity to learn property judgment as opposed to memorization of the particular words and properties in the training set. We make our negative sample generation algorithm and the resulting dataset of property-knowledge sentences available in our supplementary materials.

Tested LMs While our framework can be applied to any neural language model, we present results from fine-tuning three pre-trained LM families, based on the precedent of using these models in standard sentence classification tasks (Wang et al., 2018): BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020). All three models use the transformer architecture (Vaswani et al., 2017), and are trained to perform masked language modeling: the task of predicting masked words in context in a cloze-task setup, where models have access to context words to the left and right of the masked word. We report results using the largest models in each of the three families—BERT-large, RoBERTa-large, and ALBERT-xxl—since these variants had the best performance in our preliminary experiments (on a separate validation set). We fine-tune each of the three models on the property knowledge data by minimizing their binary cross-entropy loss on the training set using the AdamW optimizer (Loshchilov & Hutter, 2018). We tune the hyperparameters of the LMs on the validation set, and evaluate the three adapted models on the test set using F1 scores.

Table 1: Performance (F1 score) of the fine-tuned LMs on the test set of the property judgment task. Chance F1 is 0.66.

Model	**Params**	**Test F1**
ALBERT-xxl	206M	0.79
BERT-large	345M	0.78
RoBERTa-large	355M	0.79

Results Table 1 shows the performance of the three models in our property judgment experiments. We find that all three models show similarly high performance on the test set (0.78-0.79), suggesting strong capacities of all three models to assess the application of properties to concepts. Notably, the ALBERT-xxl model shows the same performance as BERT-large and RoBERTa-large despite having ≈130M fewer parameters, suggesting that this property knowledge can be encoded in smaller models with more efficient use of parameters. Furthermore, all three models perform significantly above chance ($p < .001$, FDR corrected).

Investigating Taxonomic Generalizations in LMs using Property Induction

Taxonomic relations between concepts have an important role in studies of human inductive reasoning. Early evidence from Gelman & Markman (1986) indicated a strong preference of children and adults, when making generalizations about new and unfamiliar properties, to do so based on the structure of biological taxonomies and category membership. Building on this, Osherson et al. (1990) documented 13 separate taxonomic phenomena that influenced inductions made by humans. Inspired by these works, we demonstrate how our property induction framework can be used to test whether a similar taxonomic bias is reflected in the LMs used to train the above property judgment models. For instance, if a model is provided with a new property—e.g., *can fep*—that is associated with the concept CAT, to what extent do its representational biases cause it to prefer generalizing or projecting this property to other mammals rather than to fish?

Data We restrict our analysis to the animal-kingdom subset of the concepts in our modified property-norm data, corresponding to a total of 152 animal concepts. We first select the top six categories within this subset: MAMMAL (52), BIRD (36), INSECT (18), FISH (14), MOLLUSK (8), and REPTILE (7). Each instance in this experiment involves one of the six aforementioned categories (of size m) from which we sample n concepts to create the adaptation set, and use the remaining $m - n$ to create the “Within-category” generalization set. We then create two separate “Outside-category” generalization sets. First, we sample the top $m - n$ animal concepts, on the basis of their average cosine similarity with the concepts in the adaptation set (using the representations of the embedding layer of the given model), and take this to be the “Outside$_{similar}$” generalization set. We use this similarity-based sampling technique in order to increase our confidence that observed differences can be attributed to category differences and not to general co-occurrence properties as learned by the models. This choice makes the Outside$_{similar}$ set model-dependent. We then complement this with an equal sized “Outside$_{random}$” generalization set which is model-independent and is composed of concepts randomly selected from the set of animal concepts (excluding those that belong to the main category used for adaptation). We repeat this sampling process 10 times for each of $n = 1, \dots, 5$ adaptation concepts, and 8 novel properties: verb phrases created using nonce words (*can dax, can fep, has blickets, has feps, is a tove, is a wug, is mimsy, is vorpal*). In total, we have 2,400 adaptation trials per model, each involving 3 generalization sets: Within, Outside$_{similar}$, and Outside$_{random}$ to test the property induction behavior of our models.

Method In each trial, we pass the adaptation set to the models and let them minimize their loss (starting from the same optimizer state obtained at the end of the property judgment training phase) until they reach perfect accuracy. Then we compute G for each of our three generalization sets as shown in eq. (1), for each model. Figure 2 shows the average G (over all properties) as a function of the size of the adaptation set.

Results and Analysis We expect models with a preference for category-based generalization to have greater average G value for the “Within” set than for either of the “Outside” sets. From Figure 2, we see that all three models consistently show this pattern—for all models, the average G was significantly greater for “Within” generalization as compared to both “Outside” generalization sets ($p < .001$, according to a Games-Howell test conducted following a Welch’s ANOVA). This suggests that these models show a preference for generalizing newly-learned properties of a concept to other members of that concept’s superordinate category. We also observe that the average generalization score in both categories increases with an increase in the number of adaptation concepts. Notably, this is also robustly observed in humans (characterized as the *premise monotonicity effect* by Osherson et al.)—however, we do not focus on this effect, as it is relatively expected that machine learning models will be more confident in their predictions as the number of samples provided to them increases.

Although the properties provided to the models in our induction experiment are ones that they have never seen during property-judgment training, one may wonder to what extent the models’ inductive behavior can be explained based on taxonomically similar concepts simply being more likely to share properties within the property-judgment training stage—this could call into question how much these generalization patterns tell us about the underlying concept knowledge in the LMs. Under the connectionist perspective of property induction (Sloman, 1993; Rogers & McClelland, 2004), the strength of generalization (of a novel property) to a concept is proportional to the overlap in properties between the premise (adaptation set) and the conclusion (generalization set). We can reasonably expect this to translate to the models that we use here, especially since they are trained to predict the presence and absence of properties. To test the connection between LMs’ generalization behavior and the overlap in training data properties, we first calculate property overlaps between each adaptation/generalization set pair as the ratio of the intersection and union of the ground-truth properties associated with the concepts within the sets (i.e., the jaccard similarity). We then fit a linear mixed-effects model to predict G using the `lme4` (Bates et al., 2015) and `lmerTest` (Kuznetsova et al., 2017) packages in R, for each LM. Our final model included the number of adaptation concepts (`n`), as well as the property overlap (`overlap`) and cosine similarity (`sim`) between the adaptation and generalization sets along with their interaction as fixed effects; and also included random intercepts for the novel property and the trial. Model Specification: `G ∼ n + overlap * sim + (1|property) + (1|trial)`. For all three LMs, we find a positive main effect of the property overlap,[1] suggesting that G was significantly greater for generalization sets whose concepts had greater training data property overlap with those in the adaptation set.

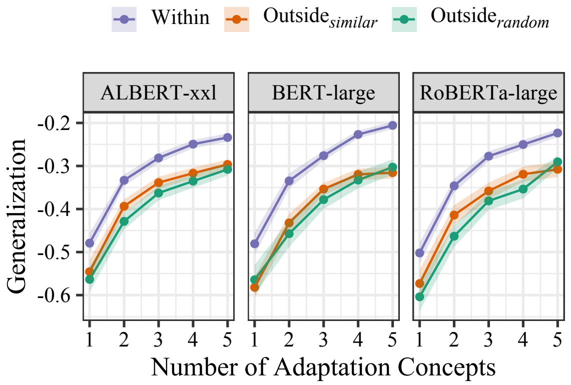

Figure 2: Results from the taxonomic generalization experiment showing generalization scores (G) of the three property-judgment models for ‘Within’ and both the ‘Outside’ generalization sets across different number of adaptation concepts.

While we have established that the models make generalizations that are consistent with the training set statistics, there exist cases where property overlap is in direct conflict with taxonomic category membership. For instance, dolphins share many salient properties with fish and yet are classified as mammals. Motivated by this observation, we ran another

[1] along with that of number of concepts (`n`) as well as the model’s cosine similarity (`sim`), $p < .001$ in all cases, approximated using Satterthwaite’s method; see Suppl. Materials for full results.

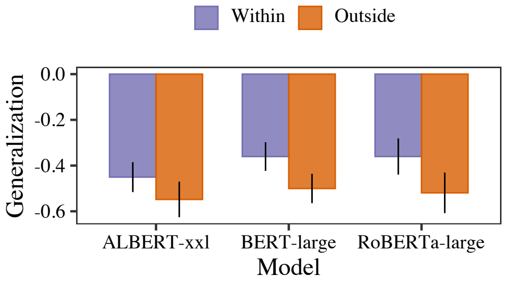

Figure 3: Generalization scores of the models in cases where the Outside category had greater property overlap than the Within category. $N = 48$ trials for each model.

experiment involving only the cases where generalizations based on category membership conflicted with those based on property-overlap. We identified 6 concepts that had greater property overlap with concepts belonging to a different category relative to their own superordinate category: (DOLPHIN, WHALE, TURTLE, SLUG, SNAIL, HIPPO). For each of these 6 concepts, we compare generalization of our previously used 8 novel properties to concepts in the same taxonomic category (Within) vs. concepts in the category with which the concept had greater property overlap (Outside), thereby teasing apart the effect of property-overlap from that of true taxonomic membership. Figure 3 shows results of this experiment. We observe for each model that the inductive preference for the "Within" generalization set was significantly greater than that for the "Outside" generalization set ($p < .001$ in all cases using a paired t-test, FDR corrected). This indicates that while the overall generalization behavior of the models is predicted by training data statistics, the models are robust in showing a taxonomic bias even when this relationship does not hold.

General Discussion and Conclusion

The empirical success of neural network-powered language models (LMs)—especially on high-level semantic tasks—has lent further support to the study of language as a source of semantic knowledge (Elman, 2004; Lupyan & Lewis, 2019). The goal of this paper was to contribute to this line of inquiry by understanding the ways in which LMs generalize novel information about concepts and properties (*a lion can fep*) beyond their training experience. To this end, we developed a framework that used LMs to perform *property induction*—a paradigm through which cognitive scientists have studied how humans use their conceptual repertoire to project novel information about concepts and properties in systematic ways (Rips, 1975; Osherson et al., 1990; Hayes & Heit, 2018). By simulating a similar process in LMs, our framework can yield insights about the inductive preferences that are guided by the LMs' representations and shed light on the nature of the models' conceptual knowledge.

As a motivating case study, we used our property induction framework to study the extent to which LM representations show a preference to project novel properties on the basis of category membership. To this end, we adapted three LMs—fine-tuned to predict the truth of sentences expressing property knowledge—to inputs associating a novel property with one or more concepts. We then compared the models' projection of the novel property between (1) a set of concepts with the same superordinate category as the concept(s) associated with the property, and (2) a pair of concept-sets that were outside of that superordinate category. In a majority of cases, the LMs preferred to project the new property to concepts of the same category, suggesting the influence of taxonomic bias. We hypothesized that some of models' taxonomic category preference could be due to high property overlap between concepts of the same category in property-judgment training—but while these property overlaps were statistically predictive of how models projected novel properties, the preference to generalize to concepts within the taxonomic category persisted even when effects of property-overlap and category-membership were teased apart.

Our results indicate that when LMs—fine-tuned to assess property knowledge—deploy knowledge about novel properties, they are guided in part by representational taxonomic biases beyond simple property-overlap relevant during fine-tuning. While we cannot say precisely what the source of this taxonomic bias is within these models, a simple explanation would be that this bias reflects the nature of the conceptual knowledge that these LMs learn and encode during pre-training. That is, in learning semantic representations of words by predicting them in context, models may have picked up on latent taxonomic knowledge, to which they then show sensitivity when projecting novel property information. This is consistent with existing works that diagnose conceptual knowledge in LMs, finding them to display strong performance in predicting taxonomic category membership (Da & Kasai, 2019; Bhatia & Richie, 2021). Through our results, we learn that this knowledge can additionally be implicitly activated, and in fact guides how new property information is generalized by LMs.

What other phenomena guide the inductive generalizations that LMs make about concepts and properties? Our framework provides a flexible mechanism to simulate and test a broad range of phenomena observed in the human property-induction literature (see Kemp & Jern, 2014; Hayes & Heit, 2018), and to shed light on the extent to which LMs' inductive preferences are consistent with those observed in humans. A potential direction includes testing for a more general class of inductive phenomena that are guided by "intuitive theories" (Murphy, 1993) that provide the context on the basis of which different types of novel properties are projected differently (Carey, 1985; Kemp & Tenenbaum, 2009). For instance, biological information may be projected across a taxonomy (ROBIN and SWAN), whereas behavioral information may be projected on the basis of specific shared properties (HAWK and TIGER). Applying our framework on these phenomena can provide insight into the context-specific flexibility of LM representations, and at a high-level, the kinds of domain knowledge that can be acquired through text-exposure.

Acknowledgements The authors thank the anonymous reviewers and the meta-reviewer for their feedback. K.M. is grateful to the CompLing lab (A.E.'s group) at the University of Chicago for discussion on previous iterations of the ideas presented here. The experiments reported in this paper were partially run on the Gilbreth cluster at Purdue University's Rosen Center for Advanced Computing (two NVIDIA V100 GPUs with 32GB VRAM), and partially run on Hemanth Devarapalli's computational platform consisting of a single NVIDIA 3090 GPU with 24GB VRAM.

Reproducibility We make our code and data available at https://github.com/kanishkamisra/lm-induction. This repository also consists of a supplementary materials document, `supplementary.pdf`, which contains details about our data filtering pipeline, the negative sampling generation algorithm and full set of results from our analyses.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., ... Choi, Y. (2020). Abductive Commonsense Reasoning. In *International Conference on Learning Representations.*

Bhatia, S., & Richie, R. (2021). Transformer networks of human concept knowledge. *Psychological Review*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642).

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2011). Inductive logic and empirical psychology. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the History of Logic* (Vol. 10, pp. 553–624). Elsevier.

Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing* (pp. 1–12). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-6001 doi: 10.18653/v1/D19-6001

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423 doi: 10.18653/v1/N19-1423

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in cognitive sciences*, *8*(7), 301–306.

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do Neural Language Representations Learn Physical Commonsense? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.*

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3), e1459.

Kemp, C., & Jern, A. (2014). A Taxonomy of Inductive Problems. *Psychonomic Bulletin & Review*, *21*(1), 23–46.

Kemp, C., & Tenenbaum, J. B. (2009). Structured Statistical Models of Inductive Reasoning. *Psychological Review*, *116*(1), 20.

Kim, N., & Smolensky, P. (2021). Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 467–470). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.scil-1.59

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I., & Hutter, F. (2018). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations.*

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the first workshop on language grounding for robotics* (pp. 76–85). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-2810 doi: 10.18653/v1/W17-2810

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, *8*, 125–140.

McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., & Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Misra, K., Ettinger, A., & Rayz, J. (2021). Do language models learn typicality judgments from text? In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.*

Murphy, G. L. (1993). Theories and concept formation.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based Induction. *Psychological Review*, *97*(2), 185.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 665–681.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Sloman, S. A. (1993). Feature-based induction. *Cognitive psychology*, *25*(2), 231–280.

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4704–4710).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations.*

Weir, N., Poliak, A., & Van Durme, B. (2020). Probing neural language models for human tacit assumptions. In *CogSci 2020* (pp. 377–383). Cognitive Science Society.

Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. In *32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133–138). Las Cruces, New Mexico, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P94-1019` doi: 10.3115/981732.981751