# ON SEMANTIC COGNITION, INDUCTIVE GENERALIZATION, AND LANGUAGE MODELS

by

**Kanishka Misra**

**A Dissertation**

*Submitted to the Faculty of Purdue University*

*In Partial Fulfillment of the Requirements for the degree of*

**Doctor of Philosophy**

Polytechnic Institute
West Lafayette, Indiana
December 2023

# THE PURDUE UNIVERSITY GRADUATE SCHOOL
# STATEMENT OF COMMITTEE APPROVAL

**Dr. Julia Taylor Rayz, Chair**

Department of Computer and Information Technology

**Dr. John Springer**

Department of Computer and Information Technology

**Dr. Jin Kocsis**

Department of Computer and Information Technology

**Dr. Victor Raskin**

Department of English

**Dr. Allyson Ettinger**

Departments of Linguistics and Computer Science, University of Chicago

**Approved by:**

Dr. Stephen J. Elliott

*To Kruti and Laxminarayan, for their unwavering love and sacrifice, and for providing me with my conceptual knowledge.*

# ACKNOWLEDGMENTS

*Wow, it really does take a village to raise a PhD!*

– Me

As is often customary, I would like to start by thanking my wonderful advisor, Julia Rayz. Julia and I have been working together since I was an undergraduate—I got into research because she convinced me—and I am very glad that she can see my PhD through to completion. She has the rare ability to come up with research questions and proposals that bring together ideas in semantics, computer science, and cognitive science, from thin air. Working with Julia has been intellectually very satisfying—in fact it was she who introduced me to Osherson et al. (1990), years before I actually decided to pursue inductive reasoning for my dissertation. Julia has put a great deal of faith in me right from the start, and has allowed me to take many, many risks. She introduced me to Natural Language Processing and Cognitive Science research, and did so with utmost patience and support. Allyson Ettinger's role has been equally critical to my scientific training. I got incredibly lucky when Allyson decided to reply to my email all those years ago, and that I got a chance to collaborate with her so closely ever since. She is a wonderful listener and is *always* able to provide highly effective and thoughtful feedback. Writing papers with Allyson has completely transformed me as a scientific communicator—her extraordinary ability to explain the most complex of concepts in the simplest of ways is something I can only hope to live up to. I cannot thank Julia and Allyson enough for their faith and support, but I do hope I can pay it forward some day! I also thank other members of my committee: Victor Raskin, John Springer, and Jin Kocsis, for their support, intellectually stimulating questions, and all sorts of bureaucracy favors (special thanks to John Springer for this). I was fortunate enough to take Victor's seminars, and learn all about the history and nature of semantics and semantic theory, all while munching on some fine Belgian dark chocolate—thank you, Victor!

I am very grateful to Felicia Anderson, Jennifer Cutter, Cindy Salazar, and Kari Ludwig, for helping with the administrative and bureaucratic hurdles over the years—thank you all very much for your hard work, and for putting up with me, especially during travel reimbursement season!

I am indebted to Hemanth Devarapalli, who is one of my closest friends at Purdue. I am not sure if I can thank him enough, but I am grateful for the Lesser and Greater-arc machines, the Library, the magnificent greyhound-mobile, the invaluable memes, and for always being there for me during the good, bad, and the chaotic-neutral times over years.

Foundations of the research presented in this dissertation were laid down during 2020–2022, an unprecedented time of great global panic, despair, and loneliness. I was lucky to share living space/be neighbors with some of the best people during these years: Akancha Pandey, Ananya Sheth, Ankit Manerikar, Janav Udani, Kaushal Jain, Krishna Pillutla, Parul Verma, and Srishti Kush. I thank you all for the exquisite food (accidental pasta-balls and marinara sauce come to mind), wonderful music, and the chaotic (but priceless) memories. Thanks also go to Lina Aboulmouna and Buddy for dropping in time to time during these years and being excellent company!

Najoung Kim has been a great collaborator and an amazing friend but even beyond that she is a fantastic person with very nuanced takes on almost everything I discuss with her. I thank her for her priceless sense of humor, for engaging in useful back and forth about animal models and machine cognition (see §1.1), and for her honest practical advice that always keeps me grounded ("something something *fleeting* sense of glory"). It is sad that I've only known her for a year, but *it is what it is*.

I am lucky to have found a wonderful academic home in Allyson Ettinger's CompLing lab at UChicago, a community I have been a part of since close to its inception, and one I consider very close to my heart. Thanks so much to Sanghee Kim, Mourad Heddaya, Freda Shi, Yan Cong, Chih-chan Tien, Aswathy Ajith, Jiaxuan Li, Lalchand Pandia, Chenghao Yang, Lucas Fegen, Irene Tang, Lang Yu, Brian Kwon, Stephanie Reyes, Andy Yang, and Sam Huang—I am so proud to call you all my colleagues. Special thanks to Sanghee and Mourad for hosting me when I visited in December 2022 (best hosts ever)! I am also thankful for the AKRaNLU lab space at HEAV 108—the place where I started my research career—for providing me with a productive working environment, and for facilitating a wide variety of research discussions over the years. I also often worked from the Vienna and Greyhouse Cafes in West Lafayette—thank you and your delicious coffee for the past 9 years!

I have been fortunate to have worked with many collaborators during my PhD: Aaron Mueller, Adina Williams, Cicero Nogueira dos Santos, Haoyue "Freda" Shi, Hemanth Devarapalli, Geetanjali Bihani, Jon Gauthier, Keren Fuentes, Koustuv Sinha, Siamak Shakeri, Tatiana Ringenberg, Najoung Kim, Roger Levy, and Xinyun Chen—thank you, all! Many thanks to Tal Linzen and Brenden Lake and their groups (CAP lab and H&ML lab) for hosting me at NYU when I was in New York for my internship—I learned a lot from the many discussions we had during my time there, and a good deal of the work presented in ch. 4 is a result of these chats. During this time I was also hosted by Tom McCoy at Princeton—thanks so much for showing me around and for indulging me about word-puzzles, connectionism, abstraction, and generalization. I also thank Ev Fedorenko, Jacob Andreas, Roger Levy, and their groups (EVLab, LINGO lab, and Computational Psycholinguistics Lab) for hosting me at MIT for invited talks—I would like to single out Anna Ivanova and Jon Gauthier and thank them for conspiring to bring me there, and also Niharika Jhingan for helping me with the travel logistics! The community at the intersection between CogSci and NLP/Computational Linguistics is very dear to me, and even though it has grown quite a bit over the years, I want to thank the folks who made it a welcoming environment for me when I was just starting out: Claire Bergey, Forrest Davis, Grusha Prasad, Tom McCoy, Paul Smolensky, Sathvik Nair, Abhilasha Kumar, Philip Resnik, Robert Hawkins, Marty van Schijndel, and Adele Goldberg. I was also lucky enough to get a chance to organize a discussion group on Neural Networks and Cognition at CogSci 2022 with Jay McClelland, Judith Fan, and Felix Binder—thanks for being great co-organizers and I hope we get to do that again some day! I am glad to be in touch with teachers and friends from my high school—thanks to Madhvi Saxena for introducing me to computer science and to Kanchan Ghose Chaudhari for always being supportive! Thanks to my friends—Vasudev Vohra, Arjun Roy, Gautam Wadhwa, Anshul Mathur, Nishank Saini, Vaishnavi Singh, Suhani Tandon, and Prachi Shukla, for all the wonderful memories over the years (14 years now!) Many thanks to friends I made during my undergraduate degree—Lakshmi Achutanand, Mikaylah Gross, Sanjna Vinze, Ashley Jerue, Aditya Gupta, Vishnu Tadimeti, Ridhi Adyanthaya, Joseph Zedan (and the Zedan family)!

Last, but surely not the least, this dissertation has been a testament to the unwavering love, support, and sacrifice of my parents, Laxminarayan and Kruti Misra, and the rest of the Shah and Misra families. I love you all very much!

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

Our ability to understand language and perform reasoning crucially relies on a robust system of semantic cognition (G. L. Murphy, 2002; Rogers & McClelland, 2004; Rips et al., 2012; Lake & Murphy, 2021): processes that allow us to learn, update, and produce inferences about everyday concepts (e.g., CAT, CHAIR), properties (e.g., *has fur, can be sat on*), categories (e.g., MAMMALS, FURNITURE), and relations (e.g., `is-a`, `taller-than`). Meanwhile, recent progress in the field of natural language processing (NLP) has led to the development of language models (LMs): sophisticated neural networks that are trained to predict words in context (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), and as a result build representations that encode the knowledge present in the statistics of their training environment. These models have achieved impressive levels of performance on a range of tasks that require sophisticated semantic knowledge (e.g. question answering and natural language inference), often even reaching human parity. To what extent do LMs capture the nuances of human conceptual knowledge and reasoning? Centering around this broad question, this dissertation uses core ideas in human semantic cognition as guiding principles and lays down the groundwork to establish effective evaluation and improvement of conceptual understanding in LMs. In particular, I build on prior work that focuses on characterizing what semantic knowledge is made available in the behavior and representations of LMs, and extend it by additionally proposing tests that focus on functional consequences of acquiring basic semantic knowledge.

I primarily focus on *inductive generalization* (Hayes & Heit, 2018)—the unique ability of humans to rely on acquired conceptual knowledge to project or generalize novel information—as a context within which we can analyze LMs' encoding of conceptual knowledge. I do this, since the literature surrounding inductive generalization contains a variety of empirical regularities that map to specific conceptual abstractions and shed light on how humans store, organize and use conceptual knowledge. Before explicitly analyzing LMs for these empirical regularities, I test them on two other contexts, which also feature the role of inductive generalization. First I test the extent to which LMs demonstrate typicality effects—a robust finding in human categorization literature where certain members of a category are

considered to be more central to the category than are others. Specifically, I test the behavior 19 different LMs on two contexts where typicality effects modulate human behavior: 1) verification of sentences expressing taxonomic category membership, and 2) projecting novel properties from individual category members to the entire category. In both tests, LMs achieved positive but modest correlations with human typicality ratings, suggesting that they can to a non-trivial extent capture subtle differences between category members. Next, I propose a new benchmark to test the robustness of LMs in attributing properties to everyday concepts, and in making inductive leaps to endow properties to novel concepts. On testing 31 different LMs for these capacities, I find that while they can correctly attribute properties to everyday concepts and even predict the properties of novel concepts in simple settings, they struggle to do so robustly. Combined with the analyses of typicality effects, these results suggest that the ability of LMs to demonstrate impressive conceptual knowledge and reasoning behavior can be explained by their sensitivities to shallow predictive cues. When these cues are carefully controlled for, LMs show critical failures in demonstrating robust conceptual understanding. Finally, I develop a framework that can allow us to characterize the extent to which the distributed representations learned by LMs can encode principles and abstractions that characterize inductive behavior of humans. This framework operationalizes inductive generalization as the behavior of an LM after its representations have been partially exposed (via gradient-based learning) to novel conceptual information. To simulate this behavior, the framework uses LMs that are endowed with human-elicited property knowledge, by training them to evaluate the truth of sentences attributing properties to concepts. I apply this framework to test four different LMs on 13 different inductive phenomena documented for humans (Osherson et al., 1990; Heit & Rubinstein, 1994). Results from these analyses suggest that building representations from word distributions can successfully allow the encoding of many abstract principles that can guide inductive behavior in the models—principles such as sensitivity to conceptual similarity, hierarchical organization of categories, reasoning about category coverage, and sample size. At the same time, the tested models also systematically failed at demonstrating certain phenomena, showcasing their inability to demonstrate pragmatic reasoning, preference to rely on shallow statistical cues, and lack of context sensitivity with respect to high-level intuitive theories.

# 1. INTRODUCTION

Much of our intuitive understanding of the world relies on mental representations (called *concepts*), which store knowledge about everyday objects or events – knowledge involving the class of entities they pick out (called categories), the *properties* that they possess, and their inter-*relations* (Margolis, Laurence, et al., 1999; G. L. Murphy, 2002; Machery, 2009). Processes that underlie how such knowledge is acquired, stored, updated, and used during everyday behavior constitute *human semantic cognition* (McClelland & Rogers, 2003; Rogers & McClelland, 2004; Kemp & Jern, 2014).

Meanwhile, unprecedented progress in the field of natural language processing (NLP), driven primarily by the re-emergence of connectionism (McClelland & Rumelhart, 1986)—now rebranded as Deep Learning (LeCun et al., 2015)—as well as advances in compute hardware have resulted in the development of powerful neural network-based language models (Devlin et al., 2019; Radford et al., 2018; Brown et al., 2020), that are trained to predict words in context (also known as "language modeling"). As a result of this training, these language models (LMs) learn dense vector representations of word forms that reflect the statistical information encoded in their training data—information such as what contexts a given word has occurred in, its statistical dependence on other words, etc. These models are in particular modern day implementations of the *distributional hypothesis* (Harris, 1954; Firth, 1957), a usage based perspective of word meanings that suggests the meaning of a word to be determined by the context it occurs in. Often times, models trained with the language modeling objective are used to perform a number of specialized downstream NLP tasks such as question answering (Rajpurkar et al., 2016), sentiment analysis (Socher et al., 2013), etc., and for this reason, the initial process of their training is often also referred to as "pre-training," and the models so trained are called "pre-trained language models" (PLMs).[1] The general impressive performance of LMs on tasks requiring access to sophisticated semantic knowledge raises a non-trivial possibility that progress in neural network models of language processing can perhaps also translate to progress in modeling how humans might extract meaning from language input.

[1] ↑I will be using 'PLMs' and 'LMs' interchangeably in this dissertation. Therefore, unless stated otherwise, an LM would denote a language model that is pre-trained.

This rather ambitious and optimistic perspective has been the subject of much debate. In particular, Bender and Koller (2020) define natural language meaning to be a mapping from *form* to some type of *communicative intent*, and under this definition, suggest that it is impossible for any model to learn meaning only from form. Lake and Murphy (2021) take a slightly less pessimistic viewpoint and do not discard the notion that LMs or their descendants can encode meaning, but propose that the bar should be more stringent, and that current benchmarks do not cut it. They then lay down multiple non-exhaustive desiderata that computational models must satisfy in order to serve as models of psychological semantics: that it should should demonstrate robust compositionality, encode knowledge about entities goals and motivations, and flexibly update and deploy their beliefs in making inferences. Importantly, they ground their recommendations in the conceptual basis of word meaning (G. L. Murphy, 2002; ch. 11), where words are mapped to broader conceptual structures that allow the intelligent system (largely, the human mind) to plan, communicate, follow instructions, and make inferences in everyday life. Piantadosi and Hill (2022) and Pavlick (2022, 2023) interpret language models to be compatible with conceptual role semantics (Harman, 1982), which defines the meaning of a symbol to be a function of the contexts within which it occurs, similar to the distributional hypothesis. Differing from most accounts, they suggest direct referential grounding and communicative intent to not be the *sine qua non* for meaning, and instead advocate for equal emphasis to be placed on the role of symbols in inference. Importantly, LMs' high dimensional, distributed representations allow them to encode complex relationships between internal states that are present/elicited by the input, giving them a non-trivial chance to capture conceptual structure internally (Pavlick, 2023). Therefore, there is a genuine need for empirical investigations to understand and evaluate the extent to which LMs (and other similar models) capture meaning by analyzing their internal conceptual structure. Pavlick (2023) in particular complements these philosophical discussions by shedding light on a few recent works that indeed make empirical advances towards this goal (Abdou et al., 2021; Patel & Pavlick, 2022; Lovering & Pavlick, 2022).

In this dissertation, I take semantic cognition as a lens to analyze language models ability to demonstrate *conceptual understanding*, which I define to be the collection of behaviors that are compatible with the successful mapping of words to broader conceptual structures that support reasoning and inference. This goal is empirical by definition, as it relies on taking existing observations in studies of human semantic cognition—patterns and behaviors exhibited by humans, along with their theoretical interpretations—as guiding principles to enable effective evaluation and improvement of conceptual understanding in NLP systems. This perspective is also related to many of the aforementioned viewpoints. In particular, like Lake and Murphy (2021), I take the conceptual basis of word meaning (Jackendoff, 1983; G. L. Murphy, 2002; M. L. Murphy, 2010) to be ground truth, where a word's meaning is essentially some mapping to a concept, which is a pointer to a class or category of entities, people, and states. Similarly, in line with Piantadosi and Hill (2022) and Pavlick (2022, 2023), this dissertation takes seriously the treatment of the role of conceptual knowledge and structure in making appropriate inferences and demonstrate reasoning behavior.[2]

Coming back to the goals of this dissertation, I defined conceptual understanding—or the extraction of conceptual meaning—to be compatible with the *mapping* of a word's form to a broader conceptual representation. The nature of this mapping has been constantly contested. One school of thought casts it as a discrete, symbolic mapping between words and concepts (Jackendoff, 1983; G. L. Murphy, 2002; Nirenburg & Raskin, 2004). The other school has put forth the *words-as-cues* view (Elman, 2004; Lupyan & Lewis, 2019), where instead of being mapped symbolically to discrete concepts, words act as *cues* to meaning, and language plays a crucial role in shaping one's semantic knowledge along with other sources such as perception and interaction. Language models can be interpreted as models of the words-as-cues perspective—they learn continuous representations of symbols in context, which can be taken to operate as *cues* that drive downstream behavior of the model. Since the information encoded by LMs is acquired through text data alone, they are models of the words-as-cues view in the strongest sense. Implementing the words-as-cues perspective using neural networks can shed light on debates concerning the nature of the mapping between a word's form to a conceptual representation. In particular, the two schools of thought

[2] ↑as we will see next, an important context within which I evaluate LMs is that of inductive generalization (Hayes & Heit, 2018), which by definition is an instance where humans use their existing conceptual knowledge in making inferences about novel information.

need not necessarily be at odds, since the distributed representations of neural networks (at least those employed by LMs) are composed of low-level sub-symbolic units that "conspire" together to give rise to complex semantic behavior not indifferent to symbolic processing (Smolensky, 1988, 1995). Regardless of the extent to which current or future LMs genuinely approximate symbolic behavior (McCoy, Linzen, et al., 2019), analyzing them for their ability to extract conceptual meaning from the perspective of human semantic cognition can serve as an effective way to test the words-as-cues perspective.

What constitute as candidate behaviors when it comes to the evaluation of conceptual understanding in language models? A considerable amount of work has been done to characterize the semantic knowledge captured by distributional semantic models (Rubinstein et al., 2015; Lucy & Gauthier, 2017; Sommerauer & Fokkens, 2018; Petroni et al., 2019; Weir et al., 2020; Ravichander et al., 2020; N. Li et al., 2021). In particular, most of these works test the extent to which these models can elicit properties of everyday concepts—attributes or predicates that are broadly applicable to the extension of the concept. For example, *can fly* is a property roughly applicable to extensions of SPARROW. Commonsense knowledge about conceptual properties offer a direct and convenient way to go beyond a word's form and access broader unstated knowledge that expected to be a part of its conceptual representation. However, experiments that aim to shed light on conceptual organization and semantic cognition in humans rarely ever only test static knowledge. Instead, they focus on how humans might deploy or use this knowledge in making semantic inferences or generalize to new/unseen items. The rich literature on human inductive reasoning and generalization, spanning more than 40 years, covers many of these behaviors (Feeney & Heit, 2007; Feeney, 2018; Hayes & Heit, 2018). Our knowledge of properties, combined with a myriad of conceptual structure allows us to flexibly make inferences about novel concepts or properties even when they are outside of our perceptual stimulus. For instance, knowing that *penny* is a cat allows us to make the inference that it likely has paws, can make a meowing sound, has a tail, etc. Similarly, when told that robins *have the T9 hormone*, we might readily extend this property to other birds, at least relatively more strongly than when told that zebras *have the T9 hormone*.[3] Numerous studies into human inductive behavior have brought about a

[3] ↑the T9 hormone property is borrowed from Josh Tenenbaum's presentations.

substantial amount of insight into how humans use their background conceptual knowledge in solving a multitude of different kinds of inductive problems (Kemp & Jern, 2014). For instance, if a system predicts that a bird is more likely to share a novel property with other birds than with non-birds suggests that it maintains clear sensitivities to categorical structure. In any case, it is a well known fact that neural networks (and other machine learning models) are prone to relying on spurious cues/correlations to—in many cases—arrive at correct outputs, especially when these spurious cues are not controlled for,[4] making it seem that they possess the genuine ability to mimic more sophisticated semantic processes (Niven & Kao, 2019; McCoy, Pavlick, et al., 2019). In the context of knowledge about conceptual properties, these spurious cues might take the form of simple associations between lexical items of the property and the concept—e.g., a model might successfully predict that birds can fly simply due to the high lexical similarity between 'bird' and 'fly' since they are likely to co-occur during training. This by no means is entirely a negative feature, in fact, we would want models that are semantically competent to learn relatedness between words. However, simply encoding these word/concept associations is different from *also* learning the functional consequences of encoding these word/concept associations (Mahowald et al., 2023)—i.e., demonstrating behavior that is compatible with access and use of conceptual meaning in making inferences and generalizations. If a model is able to predict that robins can fly but cannot reliably make this inference about a novel concept that is a subclass of robins, then it raises doubts about the model's ability to genuinely deploy its 'knowledge,' making it unclear that it has captured conceptual meaning. This is also by no means an overly ambitious ask of distributional semantic models. For instance, Erk (2016) shows in her case study that distributional similarity obtained from models far simpler than today's best language models can be used as a mechanism to reason about properties of concepts—if a model that has no knowledge of the word *alligator* observes it in context, then it can reason about the properties of ALLIGATOR by using its distributional similarity to other concepts such as CROCODILE, etc.

[4] ↑by 'control' I mean the general practice of identifying potential spurious cues than the models could pick up on and devising stimuli that would penalize models that did in fact rely on spurious cues

The above discussion sets up an important desideratum for establishing evidence of conceptual meaning extraction—models should not only capture commonly known semantic properties of concepts, but also show behavior that is compatible with the functional consequences of encoding property knowledge. Therefore, building on considerable amount of work aiming to diagnose conceptual properties in distributional semantic models, I propose in this dissertation further contexts within which we can study language models' understanding of conceptual knowledge and organization. The primary context that this dissertation focuses on will be *inductive generalization* (also known as inductive reasoning, property induction, category-based induction, etc.)—the use of existing knowledge to make inferences about novel situations (Hayes & Heit, 2018), where the reasoner must fundamentally go beyond the available data to make conclusions that are likely but not certain. Typical inductive problems in experimental literature often expose human subjects to partial conceptual information—e.g., specifying a novel property for known concepts, or known properties for a novel concept—and then tasks them to project this information to other known cases—e.g., assigning the novel property to other concepts or predicting other properties for the novel concept. In this manner, researchers have assimilated numerous insight about how humans rely on their background conceptual knowledge to make inductive leaps. A good deal of experiments and analyses in this dissertation borrow from this paradigm and analogously task LMs to project novel conceptual information using stimuli that is controlled to target specific behaviors that are indicative of genuine ability to represent and possess conceptual meaning.

Complementing this important aspect of human semantic cognition, I also study the extent to which LMs demonstrate *typicality effects* (Rosch, 1973, 1975), another important consequence of human conceptual knowledge. Briefly, this phenomena suggests that members of a coherent category (G. L. Murphy & Medin, 1985) fall under a continuum where certain members are better exemplars of the category than are others. For instance, ROBIN and SPARROW are often considered to be typical birds, while PENGUIN and OSTRICH are considered as atypical birds, for most native English speakers in the US (Rosch, 1975). Typicality effects have been ubiquitously observed throughout several experiments concerning categorization/category-based inferences (Rosch, 1973; Rips et al., 1973; Rosch, Simpson, et al., 1976; Kelly et al., 1986; Garrod & Sanford, 1977; Osherson et al., 1990), and as such, overlap in properties are often considered important determinants of what makes a member of a category typical (Wittgenstein, 1953; Rosch & Mervis, 1975). Thus, typicality can be considered as an important candidate context within which we can study conceptual knowledge in LMs. Additionally, I also revisit the question of property knowledge in language models, in order to establish their prerequisite knowledge, which will be important when we test for reasoning that exclusively relies on property knowledge that models have internalized. I do so by devising tests that control for superficial cues and avoid making isolated conclusions, in a manner that is relatively more robust compared to prior works aiming to answer similar questions.[5]

Overall, these enquiries translate to the following research questions, which together advance our knowledge of how well LMs—perhaps the most powerful models of the "words-as-cues to meaning" perspective (Elman, 2004; Lupyan & Lewis, 2019)—can encode and extract conceptual meaning from language input:

**RQ 1:** To what extent do pre-trained language models capture typicality effects in making category inferences?

**RQ 2:** **RQ 2.1:** How robustly do pre-trained language models capture the properties of everyday concepts?

**RQ 2.2:** How robustly do they endow these properties to a novel concept that is a subclass of existing concepts?

**RQ 3:** To what extent do the inductive generalizations of pre-trained language models, driven by their learned representations, align with empirical regularities and phenomena observed in human inductive behavior?

These questions, along with the collection of analysis methods that aim to answer them, and their results, form the main contribution of this dissertation.

[5] ↑I expand in great detail on why conclusions derived from prior works are not robust enough in §2.2

## 1.1 Why study language models?

Why would we want to study language models? The above discussion points to one obvious reason: their close connection to many related viewpoints of how meaning can be extracted/learned from language: distributional hypothesis (Harris, 1954; Firth, 1957), conceptual role semantics (Harman, 1982), and the words as cues to meaning perspective (Elman, 2004; Lupyan & Lewis, 2019). LMs are by far the most powerful implementations of these ideas, and because they are able to demonstrate impressive performance on tasks that clearly require access to sophisticated semantic knowledge, they provide us with an unique opportunity to test the aforementioned viewpoints computationally. A clear advantage of LMs over traditional distributional semantic models is that they can be analyzed for behavior that goes beyond just simply storing word associations and relatedness measures—e.g. they can be tested for the extent to which they can demonstrate complex reasoning behavior *using* the distributional information they have acquired (RQs 2 and 3). This is primarily due to their fundamental ability to handle sequences—as opposed to single words, in the case of older vector space models—which can allow for greater expressivity of inputs and enable us to place words in contexts from which their conceptual role can be established.

The seemingly impressive semantic behavior of language models makes it perhaps tempting to consider them as genuine models of how humans process and understand language. However, I am not considering LMs to be explicit models of human semantic cognition in this dissertation. This is largely because outside the fact that LMs learn from distributions of word forms and process inputs using a series of linear and non-linear units along with some sequential constraints, what computational, algorithmic, or even implementational hypotheses (Marr, 1982) they encode in the context of semantic cognition is rather unknown. As a result, even if we observe behavioral alignment between LMs and humans (which would indeed be an interesting result) it alone does not qualify an LM as a cognitive theory, thereby casting doubt on its explanatory role (McCloskey, 1991; Guest & Martin, 2023).[6] Instead, inspired by N. Kim (2021), I take the perspective of McCloskey (1991), who suggests the

[6] ↑for instance, the alignment in behavior could also be due to multiple realizability where two systems with completely different processing mechanisms can still achieve the same input-output mapping (Quine, 1951; Fodor & Pylyshyn, 1988)

usage of artificial neural networks (ANNs) as animal models,[7] where to the extent that there are shared properties between the human and the animal system, a better understanding of the semantic knowledge and processing in the animal system could offer insights about the human system. Analogous to how animal models provide us with the advantage of performing manipulations such as lesions into specific parts of the brain connectome or histological investigation of tissues—those that are otherwise non-trivial or even unethical to perform with humans—ANNs also provide greater degrees of freedom to delve deeper into the ways in which a system can allow for complex semantic behavior to emerge from low-level units. For instance, one can have full control of the training environment of an ANN and even "switch off" some of its neurons in order to simulate a lesion. The primary advantage of this viewpoint is that once a set of models have been discovered to somewhat replicate the behavior of the human system of interest, they can then be subjected to analyses as described above, allowing one to disentangle relevant factors that are conducive to the target behavior from those that are not. For instance, N. Kim (2021), whose dissertation focuses on compositional generalization (Fodor & Pylyshyn, 1988), argues:

> "[...] the goal is first to explore whether we can train a set of models that generalize compositionally without building in or teaching them explicit rules of symbol manipulation that we assume to underlie those generalizations, and then identify factors or model components that contribute to constraining the learning space."

However, even before getting a better grasp at the level of performing experimental manipulations, one must first lay down the groundwork in actually scrutinizing the network for replicating the behavior of the human system of interest (here, semantic cognition). That is precisely what this dissertation aims to do—propose contexts within which we can analyze the extent to which LMs show behavior that is compatible with conceptual understanding.

[7] ↑see also Toneva (2021), who takes NLP systems to be "model organisms" for language processing in the human brain

## 1.2 Structure of this dissertation

This dissertation consists of three parts, each corresponding to a chapter, and each answering the research questions stated above. Before embarking on these chapters, I first summarize important background information in Chapter 2: distributional semantics (Harris, 1954; Firth, 1957), modern language models based on the transformer architecture (Vaswani et al., 2017), related literature on the study of conceptual knowledge in language models, and important terminology and phenomena observed in studies of human inductive generalization. The main chapters are summarized as follows:

In **Part 1 (Chapter 3)**, I develop analyses that answer RQ1. In particular, I propose two tests that involved language processing, and when applied to humans yielded sensitivities to typicality effects. In the first test, I develop stimuli that express taxonomic knowledge (*a robin is a bird*), and analyze LM-derived measure of their truth verification. In the second test, I use an important result in inductive generalization where humans showed sensitivities to typicality effects, and study LM behavior in extending new property information for individual items to their categories. Applying these tests to 19 different LMs, I found that in both tests the LMs demonstrated non-trivially positive—but also modest—correspondence with humans, showing some promise in encoding relevant conceptual knowledge, but also suggesting that text-based exposure alone is insufficient to acquire typicality knowledge. This chapter is adapted from:

> **Misra, K.**, Ettinger, A., & Rayz, J. (2021). Do language models learn typicality judgments from text?. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, 43, pages 216–222.

**Part 2 (Chapter 4)** focuses on RQs 2.1 and 2.2. Here, I propose COMPS, a collection of minimal pair sentences that jointly tests pre-trained language models on their ability to attribute properties to concepts and their ability to demonstrate property inheritance behavior. Analyses of 29 different LMs on COMPS reveal that they can easily distinguish between concepts on the basis of a property when they are trivially different, but find it relatively difficult when concepts are related on the basis of nuanced knowledge representations. Furthermore, LMs can demonstrate behavior consistent with property inheritance to a great extent, but fail in the presence of distracting information, which decreases the performance of many models, sometimes even below chance. This lack of robustness in demonstrating simple reasoning raises important questions about LMs capacity to make correct inferences even when they appear to possess the prerequisite knowledge. This chapter is adapted from

> **Misra, K.**, Rayz, J., & Ettinger, A. (2023). COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

In **Part 3 (Chapter 5)**, I present a framework that simulates inductive generalizations using the representations of pre-trained language models. This framework operationalizes inductive generalization as the behavior of an LM after its representations have been partially exposed (via gradient-based learning) to novel conceptual information. To simulate this behavior, the framework uses LMs that are endowed with human-elicited property knowledge, by training them to evaluate the truth of sentences attributing properties to concepts. Using this framework, I analyze the extent to which LMs demonstrated 13 different empirical regularities that have been observed in human inductive generalization literature. The findings suggest that while LMs can demonstrate sensitivities to hierarchical structure, conceptual similarity, diversity, and monotonicity, they were unable to capture typicality effects, context specificity, and demonstrate generalization behavior that is compatible with pragmatic reasoning. Importantly I found pre-training (i.e., acquisition of distributional semantic knowledge) to be the prime driver of the successes of the models, and at the same time also insufficient to result in other important behavior indicative of successful conceptual abstraction. This chapter greatly modifies content presented in:

- **Misra, K.** (2022). On semantic cognition, inductive generalization, and language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* 36(11), pages 12894-12895.

- **Misra, K.**, Rayz, J., & Ettinger, A. (2022). A Property Induction Framework for Neural Language Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44, pages 1977–1984.

Finally, in **Chapter** 6 I summarize findings from the three parts, discuss specific and broader limitations of the research program laid out in this work, and sketch proposals for future work to bridge artificial intelligence and cognitive science.

# 2. BACKGROUND

In this chapter, I briefly review important literature spanning computational linguistics, engineering, and cognitive science, which form the theoretical and methodological background for the dissertation.

## 2.1 Distributional Semantics and Language Models

The distributional semantics (DS) approach offers a “usage-based” perspective of word meaning, and states that a word’s meaning can be characterized by the meanings of words that occur in its observed linguistic contexts (Lenci, 2008; Erk, 2016; *i.a.*). The origins of this approach can be traced back to American structuralists (Harris, 1954) as well as British lexicographers (Firth, 1957), who collectively put forth the so-called *distributional hypothesis*. Computationally, DS has been operationalized using vector-space models (VSMs), that—as their name suggests—represent words numerically using continuous vectors of real numbers, learned from (typically large) text corpora. Regardless of the method used to construct such vectors, a hallmark consequence of building VSMs is that vectors of words that are related (e.g., *cat-dog*, or *swim-swam*, etc.) tend to be similar in vector space (Landauer & Dumais, 1997; Mikolov, Sutskever, et al., 2013; *i.a.*). That is, by relying on the distributional properties of a word’s *form*, VSMs tend to approximate and aggregate the myriad of relations that the word takes part in, as reflected in the text corpus. Vector space models of distributional semantics have come a long way—from the earliest boolean vector representation model of Spärck-Jones (1964), to the latent semantic analysis (LSA) model developed by Landauer and Dumais (1997), to the introduction of neural network-based vector-space models (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013; Pennington et al., 2014) to finally the present groundbreaking advances made by language models (e.g. Devlin et al., 2019; Peters et al., 2018; Radford et al., 2018) that continue to scale in their size and performance. With the reemergence of connectionism, now re-branded as Deep Learning (LeCun et al., 2015), neural networks have now become a staple model for learning VSMs. Typically, a neural network based model maps words to randomly initialized dense vectors—known as embeddings—which are then updated (along with the rest of the neural-network’s parameters) to maximize the statistics of a given text corpus. For instance, earlier neural-network based VSMs would be learned by optimizing simple feed-forward networks to either: (1) predict a word given a window of $h$ words around it (Mikolov, Sutskever, et al., 2013); (2) predict the words that occur in a window of $h$ words, given a target word (Mikolov, Chen, et al., 2013); or (3) encode global co-occurrence statistics between words (Pennington et al., 2014). As a result, the input weight matrix of the trained neural network, which projects words onto an $n$-dimensional space, forms the representation for each word, resulting in a vector-space. Building on the success of neural network-based models of distributional semantics, recent work has reconciled the many different ways of learning of word distributions using a more general data-fitting task known as 'language modelling,' the goal of which is to predict a word given either unidirectional or bidirectional context.[1] More formally, language modeling is the task of assigning likelihoods to sequences of words. Borrowing notation from the study of formal languages, let $\mathcal{V}$ be a finite set of words in a language, and $\mathcal{V}^*$ be the infinite set of sequences of the words in $\mathcal{V}$, then a language model (LM) specifies a function $LM : \mathcal{V}^* \rightarrow \mathbb{R}$ that assigns probabilities to sequences in $\mathcal{V}^*$—let $S_i \in \mathcal{V}^*$ be an $n$-length sequence of words from $\mathcal{V}$, $S_i = (w_1, w_2, \dots, w_n)$, the goal of a language model is to estimate:

$$LM(S_i) = p(S_i) = p(w_1, \dots, w_n) \tag{2.1}$$

Using the chain rule of probability, the above equation becomes:

$$LM(S_i) = p(w_1) \times p(w_2 \mid w_1) \times \cdots \times p(w_n \mid w_1, \dots, w_{n-1}), \tag{2.2}$$

That is, the task of language modeling essentially involves treating sentences as sequences of words, and estimating the probabilities of words in context to assess the likelihood of a sentence. In doing so, LMs essentially answer the question about how *statistically plausible* a given sequence of words is in the language, the data of which they are trained on. By definition, there are no architectural constraints placed on LMs, so in principle, any mechanism

[1] ↑unidirectional context refers to the sequence of words that appear either to the left or right of the predicted word (left in English), e.g., *'I was reading a ___'* Bidirectional contexts on the other hand involve words present in both directions of the predicted word, e.g., *'I was ___ a book.'*

that operates over sequences can be used to estimate $p(S_i)$. However, almost all LMs today are trained using the transformer architecture (Vaswani et al., 2017), the development of which has seen unforeseen breakthroughs in NLP and AI. As such, transformer based LMs are perhaps the strongest computational models of distributional semantics at the time of writing, and are therefore the main subjects of study in this dissertation. In what follows, I describe the practice of language modeling using transformers in more detail.

### 2.1.1 Language Modeling using Transformers

Transformers (Vaswani et al., 2017) are sequence models built on the concept of *attention* (Bahdanau et al., 2014). In the context of sequence modelling, attention is a mechanism that constructs representations of a given time-step by informing it by a weighted combination of the representations that represent the time-steps on which the current time-step is conditioned. For example, in a left-to-right sequence model, the $t^{th}$ word is predicted by the words that precede it $(w_1, \dots, w_{t-1})$, and therefore the hidden state representation of $w_t$ is constructed using the hidden states of the previous words, each of which is weighted by its interaction with the input at the current time-step (usually additive or dot-product based). This interaction is also referred to as the process of "attention," and since the weights add up to 1, they are collectively called the "attention distribution." Attention is an useful mechanism in sequence-learning as it enables long-range interactions between representations, thereby facilitating better encoding of long-distance dependency into the model. It has seen several successes in application to LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014) architectures, as well as their bi-directional counterparts. With the transformer architecture, Vaswani et al. (2017) take a drastic step and train a model that does away with recurrence and is instead completely reliant on attention. While the original transformer architecture was an encoder-decoder network trained to perform machine-translation, it can also be used as a language model – e.g., GPT2 (Radford et al., 2019) is a decoder-only transformer LM, while BERT (Devlin et al., 2019) performs a version of language modeling using the encoder alone, while more recent works use the entire encoder-decoder setup to perform language modeling. The transformer architecture revolves around two main ideas that have now become mainstay in much of sequence-modeling within NLP:

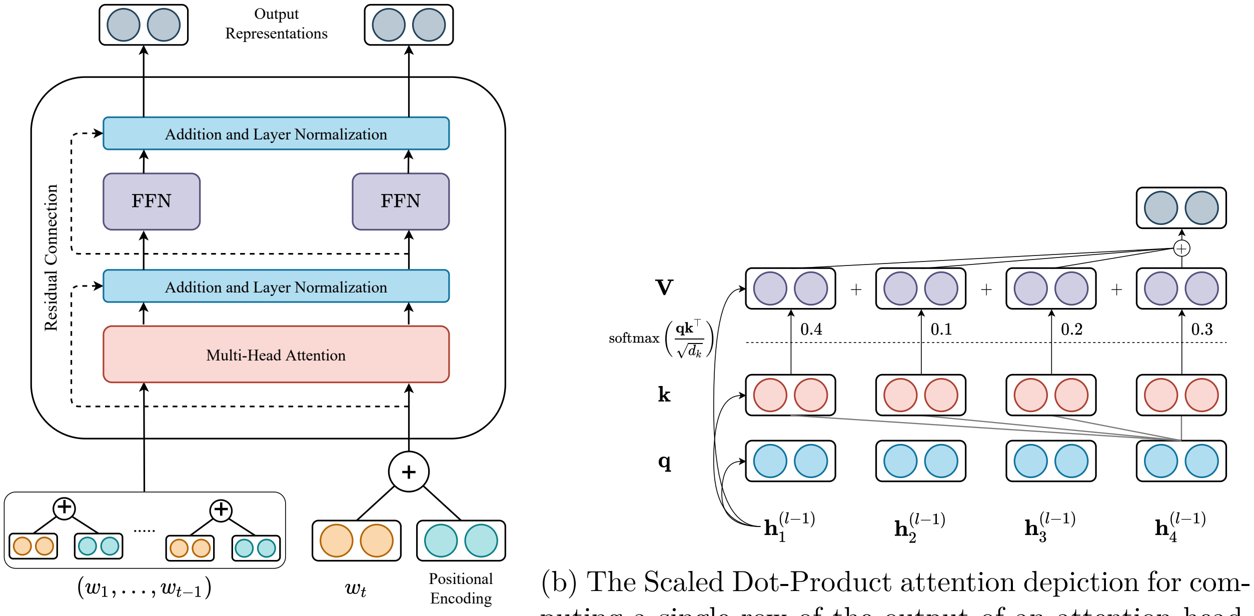
### Multi-headed Self-Attention

Self-attention is a component within transformers that relates different positions of the sequence and composes them together in order to form the representation of a specific position, allowing the model to learn long-range dependencies at every layer. For a given input consisting of $n$ tokens, self-attention is described as an interaction between a token's query representation to a set of $n$ key-value representation pairs. Specifically, each token's input representation ($\mathbb{R}^{1\times d}$) is projected to three individual vectors: (1) queries that represent the focal word being operated on, packed together into a matrix $\mathbf{Q} \in \mathbb{R}^{n\times d_k}$; (2) keys that represent the word that the query is being related to, packed together into a matrix $\mathbf{K} \in \mathbb{R}^{n\times d_k}$; and (3) values that correspond to words that the keys refer to, packed together into a matrix $\mathbf{V} \in \mathbb{R}^{n\times d_v}$. The attention vector of a single token, then, is a weighted sum of the value vectors, where the weight is calculated as a softmax transformation of the scaled dot-product between the queries and the keys (for this reason, it is also referred to as "scaled dot-product attention"). Mathematically, the output representations (packed as a matrix, the row vectors of which are representations for each token) formed by a single computation of self-attention are computed as follows:

$$\textsc{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.3}$$

A standard transformer architecture often employs multiple such attention components, each of which is referred to as an attention "head." Multiple such attention heads are concatenated and then linearly projected to produce the final output ($\mathbb{R}^{n\times d}$) of the multi-headed attention module in a transformer layer:

$$\begin{aligned} \textsc{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= [\textsc{head}_1; \ldots; \textsc{head}_h]\mathbf{W}^o, \\ \text{where } \textsc{head}_i &= \textsc{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \end{aligned} \tag{2.4}$$



(a) A transformer layer. Adapted and modified from Misra (2020). Original architecture by Vaswani et al. (2017).

(b) The Scaled Dot-Product attention depiction for computing a single row of the output of an attention head. Similar computation is performed for all input tokens to make up the entire head which is then concatenated with the other $h - 1$ heads (see Equation (2.4)).

**Figure 2.1.** Components of the Transformer Architecture (Vaswani et al., 2017).

Note that the above equations involve a slight abuse of notation, i.e., here $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$ (different from eq. (2.3) which describes general computation), $\mathbf{W}_i^Q \in \mathbb{R}^{d \times d_k}, \mathbf{W}_i^K \in \mathbb{R}^{d \times d_k}, \mathbf{W}_i^V \in \mathbb{R}^{d \times d_v}$, and $\mathbf{W}_i^o \in \mathbb{R}^{hd_v \times d}$. Note that in left-to-right models the dot-products in the attention computation for all words to the right of a given input word are set to $-\infty$ so that the prediction of a given word is always conditioned on words to its left. This will be different for other variations of the language modelling task, as we will see in the next section. Figure 2.1b depicts the attention operation for a single row in the output of an attention head.

### Positional Encodings

Instead of using recurrence, transformers incorporate sequential information by using positional encodings – embeddings that exclusively have a 1-1 mapping with the position of a given token (i.e., the position encoding of any word in the second position is always the same). These embeddings are of the same dimension as the word embedding: $d$, so that the two can be summed in order to infuse the word's distributional information with information about its position. Specifically in Vaswani et al. (2017),

$$\mathbf{p}_i = \begin{cases} \sin(i/10000^{2dim/d}) & \text{if } dim \text{ is even} \\ \cos(i/10000^{2dim/d}) & \text{if } dim \text{ is odd} \end{cases} \tag{2.5}$$

$$\mathbf{h}_i^0 = e(w_i) + \mathbf{p}_i, \tag{2.6}$$

where $\mathbf{h}_i^0$ represents the input representation (of the $i^{th}$ word) to the first transformer layer, $e(w_i)$ represents the embedding of the $i^{th}$ word, and $\mathbf{p}_i$ is its positional embedding, and $dim$ is the dimension of a given $\mathbf{p}$ vector.

Using multi-headed attention and positional encodings, the computations of a single transformer block can be summarized by the following equations:

$$\mathbf{h}_i^0 = e(w_i) + \mathbf{p}_i \tag{2.7}$$

$$\tilde{\mathbf{x}}^\ell = \textsc{MultiHead}(\mathbf{h}^{\ell-1}, \mathbf{h}^{\ell-1}, \mathbf{h}^{\ell-1}) \tag{2.8}$$

$$\mathbf{x}^\ell = \textsc{LayerNorm}(\tilde{\mathbf{x}}^\ell + \mathbf{h}^{\ell-1}) \tag{2.9}$$

$$\tilde{\mathbf{h}}^\ell = \text{FFN}(\mathbf{x}^\ell) \tag{2.10}$$

$$\mathbf{h}^\ell = \textsc{LayerNorm}(\tilde{\mathbf{h}}^\ell + \mathbf{x}^\ell), \tag{2.11}$$

where LayerNorm stands for the layer normalization operation (Ba et al., 2016), and FFN is a standard two-layer feed-forward network with an arbitrary hidden dimension. To perform language modeling—predicting words in context—the output of the last layer at each step is projected over all the tokens in the model's vocabulary, resulting in unnormalized logits which can then be converted into probabilities. Figure 2.1a depicts a single transformer layer.

The advent of efficient hardware and compute combined with engineering progress in better LM architectures—part of which is summarized above—has brought about a standard paradigm in the NLP field that proposes to learn "general" language representations by training LMs on large corpora, thereby encoding rich distributional information in their representations which can be adapted for specific NLP applications such as question answering (Rajpurkar et al., 2016), sentiment analysis (Socher et al., 2013), etc. The language modeling objective (of maximizing the probability of a given piece of text) in particular is a convenient way for training the architectures discussed above since it has the unique property of leveraging the statistical environment of text itself for supervision, circumventing the need to collected labelled data, and allowing models to scale both in terms of expressivity and training data. By being trained on large corpora using the language modeling objective, models are able to learn highly sophisticated distributional information about words and their contexts and represent them in their underlying parameters. In doing so, the representations of these models provides a useful prior that allows them to achieve impressive performance when *fine-tuned* on engineering applications (Howard & Ruder, 2018; Peters et al., 2018; Devlin et al., 2019). As a result, such models are often also called "pre-trained" language models. The pre-train then fine-tune paradigm is ubiquitous in the current state of NLP research. A number of different pre-training schemes and language modeling objectives have been proposed

### Autoregressive Language Modeling

The most straightforward way of performing language modeling is to predict words in a unidirectional manner,[2] where the probability of a word ($w_t$) is estimated as:

$$p_\theta(w_t \mid w_1, \ldots, w_{t-1}), \tag{2.12}$$

[2] ↑this was in fact the only "type" of language modeling that existed, until 2018, when Devlin et al. released the BERT pre-print. Note that language modeling as a task has been around much longer than the pre-training paradigm. In fact, its development can be traced back to Shannon (1948).

where $\theta$ represents the parameters of the model. This method is especially convenient since it allows generating natural language, by means of conditional sampling. Examples of models that are trained using this objective include the GPT family of models (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020) and Transformer-XL (Dai et al., 2019), both of which use the decode portion of the transformer architecture and mask the attention values for words to the right of a given predicted word.

### Masked Language Modeling

Diverging from the traditional language modeling task, models trained using the masked language modeling objective, introduced in the BERT model (Devlin et al., 2019). Instead of predicting words by conditioning on words to their left (unidirectionally), masked language modeling specifies an objective where words are predicted by conditioning on all words to their left *and* right, i.e., it allows for bidirectional modeling of context. This setup closely resembles *Cloze*-tasks (Taylor, 1953). In general masked language models operate using a special 'mask' token, which specifies the position of the word to be predicted. For instance, BERT (Devlin et al., 2019) uses the following representation of an input during training.[3]

[CLS] The quick brown [MASK] was jumping over the lazy dog. [SEP]

Mathematically, masked language models estimate the following measure:

$$p_\theta([\textsc{mask}] = w_t \mid W_{\setminus t}), \tag{2.13}$$

where $W_{\setminus t}$ denotes the set of words in the input that exclude the $t^{th}$ position. In addition to the masked language modeling objective, BERT was also trained on the 'Next Sentence Prediction' task—where it predicts if in a pair of sentences, the second sentence sequentially follows the first one. This was primarily motivated by the fact that several NLP tasks require the modeling of the relationship between two pieces of text, such as natural language

[3] ↑in addition to the [MASK] token, BERT also uses two other special tokens, [CLS] as the token to represent the entire sentence, during fine-tuning, and [SEP], to denote sentence boundaries.

inference (Bowman et al., 2015), and reading comprehension (Rajpurkar et al., 2016). In addition to BERT, several other models have been proposed: RoBERTa (does away with the next sentence prediction task, and optimizes the BERT training process; Liu et al., 2019), ALBERT (replaces the next sentence prediction task with a sentence ordering objective, and reduces overall parameter count by sharing parameters across layers; Lan et al., 2020), etc.

## 2.2 Conceptual Knowledge in Minds and Machines

The goal of this dissertation is to characterize the extent to which language models, perhaps the most powerful implementations of distributional semantics, can demonstrate conceptual understanding—the collection of behaviors that are consistent with successful mapping of words to broader conceptual structures that go beyond what is provided to them in their input. Versions of this goal have been pursued numerous times in computational linguistics and natural language processing. Here, I review important methodological contributions related to this overall goal made in the past few years.

Past enquiries into semantic and conceptual knowledge encoded in distributional models have primarily focused on testing the extent to which the models encode 'properties' of noun concepts—symbolic expressions (often represented using natural language—e.g., "*can fly*") that correspond to attributes or features of the concept. For instance, it would be desirable for the representation of CAT to encode information that is consistent with properties such as *has whiskers, has paws, is a mammal, is an animal, has a tail*, etc. Such investigations are especially useful because 1) distributional semantic models often learn dense representations for words that are otherwise uninterpretable, so having a mapping from these representations to a space of symbolically meaningful representations can improve our understanding of the aspects of meaning that distributional semantic models are able to capture; and 2) properties provide a natural pathway to make contact with broader grounded conceptual representations that go beyond the word's form. Erk (2016) describes an interesting case study for how distributional evidence can support noisy and probabilistic property inferences about concepts denoted by words. Erk uses the example of an *alligator* and sketches out experiments to make probabilistic inferences about its properties by comparing its vector representation against that of *crocodile* and *trout*, whose properties are assumed to be known. She concludes that because distributional data keeps track of the various predicates for words as well as their selectional restrictions, they may provide an approximate and uncertain account of property knowledge. Contemporary works that have been proposed since then fall primarily under two distinct paradigms, focusing on different aspects of distributional semantic models (here restricted to sequence models such as LMs): (1) representational probing (Ettinger et al., 2016; Belinkov et al., 2017; Adi et al., 2017; Conneau et al., 2018; Hupkes et al., 2018), which investigates the properties encoded in the vector space of the model; and (2) behavioral evaluation, which leverages the fundamental task of language modeling, where models are deemed to encode meaning sensitive capacities if they demonstrate predictive behavior that is consistent with correct semantic knowledge. In what follows I describe both these paradigms, discuss their advantages and shortcomings, and then propose means to go beyond simple property knowledge, leading into *inductive generalization*, which stands out as the distinct contribution of this thesis to the study of semantic knowledge of language models.

The premise of probing methods is to answer the question: *"Can the representational space of a given model make a particular property readily accessible to a minimal classifier?"* The experimental procedure of the probing paradigm resembles that of a typical classification problem. First, the researcher identifies the property of interest, e.g., `has wings`. Next, the researcher collects samples (in this case, words) that may or may not possess the property, e.g., *robin, crow, cheetah, table*, etc., and splits them into disjoint experimental sets (train, validation, test). Then, a 'minimal' classifier referred to as a *probe*[4] is trained on the training split, taking as input the words' meaning representations corresponding to the model under investigation, and classifying them accordingly (true vs. false with respect to the concept possessing the property in this case). Finally its performance is evaluated on the disjoint test set to form conclusions about how well the property's information is reflected in the input word representation. This allows a straightforward means to compare the quality of representations learned by the model—by simply swapping input representations provided to

[4] ↑usually a multi-layer perceptron or logistic regression classifier. In some cases also called as a *diagnostic classifier* (Adi et al., 2017; Hupkes et al., 2018).

the probe (Tenney et al., 2019). Probing has been extremely popular in the interpretability sub-field of NLP, having been used to understand the extent to which language models and other sequence encoders make available in their representations grammatical features (Hewitt & Manning, 2019; Klafka & Ettinger, 2020), function word information (**empty citation**), phrasal composition (Yu & Ettinger, 2020), entity states (B. Z. Li et al., 2021), etc.

From the perspective of conceptual knowledge, probing classifiers have also been used to decode the extent to which vector space models of distributional semantics make properties of everyday concepts available (Rubinstein et al., 2015; Lucy & Gauthier, 2017; Sommerauer & Fokkens, 2018), and this has since been extended to also analyze language model representations (N. Li et al., 2021; Derby et al., 2021). In this case, a separate probe (usually a multi-layer perceptron or SVM classifier) is trained for every individual property, and the probe's test set performance is used to conclude about how well the specific property's information is encoded within the model's representations. While probing has revealed quite a bit about what sequence mdoels such as LMs are able to encode and make available, its use as a method for diagnosing conceptual properties from vector representations of LMs can suffer from multiple limitations—ones that are beyond existing methodological debates surrounding the paradigm at large (Ravichander et al., 2021; Belinkov, 2022), which I summarize in what follows.

First, probing classifiers test for property knowledge in an implicit manner—i.e., it is the probe that detects the presence and absence of the target property from a vector representation. This could be problematic since (1) many properties are often perfectly correlated with each other and therefore would reward/penalize the model equally regardless of the amount of variability in their genuine knowledge of the different correlated properties (i.e., a model could perfectly capture the property of being a mammal but poorly capture the property of which animals lay eggs, which in reality is almost perfectly correlated with the former, and yet the probe might assign the same performance to the model); and (2) the probe might just pick up on something spurious in the vector space that happens to be correlated with the property in the training data (of the probe), which again would misrepresent the actual knowledge present in the model. Next, probing requires supervision, and therefore unfortunately restricts the space of properties that can be tested for, which is an issue since many properties are often idiosyncratic and core defining a concept (e.g., zebras having striped patterns on their bodies). Finally, LMs' encoding of conceptual knowledge is likely to be distributed across neurons/units that span multiple layers and may require all of these units to be jointly activated (hold some value) in order to make the property knowledge available. As a result, using probes on isolated, intermediate states may not genuinely test for semantic knowledge elicited by the input concept, leading to unfaithful conclusions. Note that this issue is expected to be encountered largely for language model-based instantiations of distributional semantics, since classical models are often single-layer and therefore only have one single representation.

Diverging from the probing paradigm, which focuses on individual representations at a time, approaches that fall under behavioral evaluation treat the entire model as one large black-box (Alishahi et al., 2019) and design tests that target specific linguistic behavior. These behavioral methods primarily rely on the fundamental ability of language models to perform word prediction in context, and as such design "incomplete" stimuli that are used as inputs to the model of interest, the predictions of which are then analyzed to summarize its performance. For instance, consider the problem of subject-verb agreement, where the goal is to test the extent to which models' predictions for a verb-expecting input context (see (1)) agree with that of the subject (*penguins*).

(1) The penguins by the iceberg ___

Testing LMs on multiple such stimuli can thereby reveal the extent to which they encode simple hierarchical structure of English language sentences. Similar to the above example, behavioral evaluations (Warstadt et al., 2020) might often also include pairs of minimally differing input, one of which violates the rules of the given language, and measure the extent to which models' sequence probabilities for acceptable inputs such as (2a) is greater than that for paired unacceptable inputs such as (2b).

(2) a. The robins next to the tree **are** dancing.

    b. The robin next to the tree **is** dancing.

Regardless of the particular method used, there have been numerous accounts of behavioral testing that have brought about plenty of insight into the linguistic capacities of language models. These range from a plethora of syntactic analyses (Gulordava et al., 2018; Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020), to tests for negation and semantic roles (Kassner & Schütze, 2020; Ettinger, 2020), to a range of pragmatic tests (Ettinger, 2020; Davis & Van Schijndel, 2020; Pandia et al., 2021; Parrish et al., 2021; Schuster & Linzen, 2022; S. J. Kim et al., 2022), etc. A modern reinterpretation of this paradigm has resulted in the now ubiquitous *prompting* methods that allow the querying of natural language instructions and/or few-shot examples of ground-truth input-output pairs to models such as GPT3 (Brown et al., 2020), allowing them to demonstrate complex reasoning behavior (Wei, Wang, et al., 2022).

Behavioral methods seem to alleviate many of the concerns that I raised in the context of the probing paradigm. First, behavioral testing allows us to explicitly query for the precise knowledge we are interested in—i.e., we can simply query the model to complete the input "*a zebra has* ___" with the appropriate knowledge (*stripes*). Second, since behavioral testing is completely unsupervised, it places no restrictions on the types of properties/conceptual knowledge being tested, unlike probing. Finally, since behavioral tests utilize the entire model, knowledge that is distributed across multiple different can readily be activated without any constraints, and contribute to semantic behavior.

As a result, many recent works have adopted the behavioral paradigm in order to test models for conceptual and property knowledge. Notably, this was pioneered by the "Language Models as Knowledge Bases" work by Petroni et al. (2019). Since then, researchers have proposed a range of behavioral tests and stimuli that directly query for semantic knowledge of models (Weir et al., 2020; Shwartz & Choi, 2020; Ravichander et al., 2020; Hanna & Mareek, 2021; Abdou et al., 2021; i.a.). Importantly, these tests largely follow the original single-context examples, where models are queried using cloze test examples (Taylor, 1953) such as:

(3) a. a ___ can fly.

b. penguins are ___.

c. chair is a type of ___.

While these are certainly non-trivial improvements over the probing paradigm, behavioral testing of language models for conceptual knowledge—at least in the way done so far—has its own drawbacks. Nearly every behavioral analysis of LMs that tests for conceptual knowledge does so by using isolated input contexts such those as in (3). Here, models are evaluated positively if they successfully assign the target completion non-trivially positive probability. However, without any other constraints, these models can get away with predicting, for e.g., "*sparrow*" for (3a) while also assigning greater probability to "*dog*" than to other concepts that can fly. Conclusions derived using such isolated examples, then, cannot be considered as fully robust. Relatedly, such isolated cloze contexts often involve queries that have multiple correct answers. That is, the relations that are being tested as part of the property phrase are often many-to-many—e.g., in (3b), the ___ can be filled with anything ranging from *birds*, *animals*, *vertebrate* to even subjective evaluations such as *cute*.[5] This can further bias the conclusions one can draw, since they will be critically dependent on what the researcher has set to be the correct answer.

Despite their clear limitations and trade-offs, both the probing and behavioral paradigms have brought about important insights with respect to the knowledge that pre-trained language models are able to capture—both in their representations as well as their behavior. High level conclusions from both lines of work suggest—almost convergently—that distributional semantic models can capture some properties substantially better than others, and more often than not, these properties are taxonomic in nature—e.g., `is a bird`, `is an animal`, `is a type of furniture`, etc (Rubinstein et al., 2015; Weir et al., 2020; Hanna & Mareek, 2021). Properties often expected to be absent from text-only models, such as those requiring a visual or sensory perception (`is yellow in color`, `is heavy`, etc.) were often captured poorly. A potential explanation for the strong performance on taxonomic properties could be that hyponyms and hypernyms are often highly likely to co-occur in the same contexts, and are often also substitutable, and that this information can reasonably be captured by models that exclusively encode distributional statistics (Erk, 2016). These

[5] ↑though this is likely to be universal.

results are useful as they provide us with evidence for the kinds of knowledge that is clearly captured by distributional semantic models, as well as areas where they might clearly be lacking, thereby paving ways to build more robust systems.

The above line of research has shed considerable light on the abilities of LMs and distributional semantic models to capture static conceptual knowledge—associations between concepts and their properties that LMs acquire from their training. This leaves open the question of how this knowledge guides the models' ability to plan, reason, and communicate information about the world, which is an important functional property of human semantic cognition (Carey, 1985; Rogers & McClelland, 2004; G. L. Murphy, 2002; Rips et al., 2012). I argue that knowledge of conceptual properties only scratches the surface of deeper investigations into conceptual representations and meaning. A model that can successfully encode and extract conceptual meaning should demonstrate behavior that goes beyond just capturing static properties of concepts. That is, if a model is able to show evidence for *storage* of conceptual knowledge but cannot show behavior that is consistent with this knowledge (i.e., performing reasoning), then this raises clear doubts into the extent to which the model genuinely possess knowledge about concepts it has encountered. Are these models simply memorizing data or can they productively demonstrate useful generalizations? This question is ever so relevant especially in modern times where it might often be intractable or even impossible to have full characterization of a language model's training data,[6] which leaves open the alternative explanation that models are simply producing output that they might have memorized from their training data, as opposed to showing generalization (though see McCoy et al., 2023). **I therefore suggest that capturing properties may certainly be necessary to show evidence of conceptual meaning, but it is not sufficient.** This argument is analogous to that of many cognitive scientists about the state of research in categorization, a fundamental cognitive ability – while knowing a category is fundamental to human semantic cognition, its primary purpose is to make inferences about novel items (Smith & Medin, 1981; G. L. Murphy, 2002; Rips et al., 2012; G. L. Murphy, 2016). Therefore, testing for functional consequences of static conceptual information, such as reasoning, can provide us with contexts that can advance our understanding of the extent to which

[6] ↑e.g., as of June 14, 2023, OpenAI has still not disclosed what data GPT4 was trained on.

distributional learning of word representations supports positive extraction and encoding of conceptual meaning.

While a majority of the dissertation will primarily focus on reasoning behavior enabled by conceptual knowledge, one important context within which we can study functional consequences of capturing property information is that of *typicality effects* (Rosch, 1973, 1975). In particular, the ability to encode properties of various concepts allows one to compare concepts on the basis of how many properties they end up sharing. This, combined with the hierarchical organization gives rise to graded measure of how well an object/item represents its entire category. That is, category exemplars are considered to by typical members when they have high family resemblance with the members of the category (Rosch & Mervis, 1975), occur more frequently as instances of the category (Barsalou, 1985), and 'fit the primary goal of the category' (Barsalou, 1985). For instance, in many western, English-speaking cultures, robins are considered to be typical birds, while penguins are not. The existence of typicality effects has been ubiquitous throughout numerous experiments dealing with categories (Rips et al., 1973; Rosch, 1973; Rosch, Simpson, et al., 1976; Kelly et al., 1986; Garrod & Sanford, 1977; Osherson et al., 1990), and as such is an important property of human conceptual organization/categorization. The next chapter delves deeper into the extent to which language models can demonstrate typicality effects.

Going beyond typicality effects, perhaps an even more important and comprehensive context that can shed light on deeper conceptual meaning extraction and use is *inductive generalization* (Osherson et al., 1990; Hayes & Heit, 2018), which allows humans to make inferences on being partially exposed to novel information. For instance, encountering or reading about *buddy* the dog allows one to make likely—but uncertain—inferences such as he has paws, he can bark, has a tail, etc. Similarly, on learning that polar bears and penguins share a novel property that one has never encountered before allows them to make inferences about the property that go beyond available data—that perhaps other animals that live in cold conditions also share that property. In fact, there are several different types of inductive problems that humans encounter and solve on a daily basis (Kemp & Jern, 2014). Importantly, researchers have assimilated multiple different phenomena that humans readily show while making inductive inferences, each of which requires high level abstractions over conceptual structures (Rips, 1975; Osherson et al., 1990; Heit & Rubinstein, 1994; Medin et al., 2003; Kemp & Tenenbaum, 2009; Kemp, 2011; Hayes & Heit, 2018). Therefore, one important contribution of this dissertation is to propose analyses and tests for language models that borrow from the principles of human inductive behavior.

In what follows, I zoom further into the topic of inductive generalization and discuss ways in which it has revealed how humans deploy their acquired conceptual knowledge.

## 2.3 Inductive Generalization

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). Such inductive generalizations are also referred to as *property induction* (Kemp & Jern, 2014). 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*). Inferences such as these are often interpreted as a form of reasoning where a conclusion does not deductively follow from a premise (Kemp & Jern, 2014; Feeney & Heit, 2007), and instead the reasoner makes inductive leaps in order to process and generalize newly encountered information, by relying on its existing background knowledge of concepts and properties.

Humans reason inductively on a daily basis—when learning meanings of new words (Heibeck & Markman, 1987), predicting the trajectory of objects during motion (Spelke, 1990), etc., and therefore the topic of induction is pervasive throughout the study of human cognition. The importance of studying inductive reasoning is best highlighted by Tenenbaum et al. (2007), who describe inductive generalization ("inference") as:

> *..a fundamental part of everyday life, and for cognitive scientists, a fundamental phenomenon of human learning and reasoning in need of a computational explanation.*

The general principles of induction can be applied to a myriad of different problems in the study of cognition. However, in this dissertation I consider a subset of the universe of inductive problems (Kemp & Jern, 2014) that deal with knowledge of concepts and properties, often collectively called **category-based induction** (Osherson et al., 1990), **property induction** (Gelman & Markman, 1986), or **inductive generalization** (Saxe et al., 2019). I will refer to these as inductive generalization throughout this dissertation.

Patterns of property induction observed in humans have shed a great deal of light on the nature and organization of semantic cognition. For instance, early evidence from Gelman and 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. Similarly, a number of researchers (see Hayes et al., 2010; Hayes & Heit, 2018; for a review) have assimilated various regularities that have been robustly observed to occur in humans' projection of novel information—ranging from taxonomic phenomena such as *typicality, diversity, monotonicity,* etc. (Osherson et al., 1990) to sensitivity towards *salient property knowledge* (Heit & Rubinstein, 1994; Medin et al., 2003).

Based on the above experimental evidence, many researchers posit that property inductions made by humans are governed by domain-specific knowledge (Carey, 1985; G. L. Murphy & Medin, 1985; G. L. Murphy, 1993) that is activated by the novel stimuli, which provides the basis of generalization. For instance, when given the novel information "*dolphins have blickets*," an agent that activates knowledge about biological taxonomies is likely to project it to other mammals, while an agent that activates behavioral knowledge is more likely to project it to fishes instead. Due to its emphasis on domain-specificity, this perspective is therefore likely to challenge computations that arise from domain-general learning systems (Rogers & McClelland, 2008). This is relevant because in a way, LMs can be interpreted to be domain-general—they are generally trained on a diverse set of data[7] that span multiple domains, lack any explicit semantic/structural inductive bias, and their learned representations generally perform better when they are specifically adapted for a task (e.g.,

[7] ↑for instance, BERT is trained on Wikipedia articles and texts in novels.

Devlin et al., 2019; Gururangan et al., 2020). Human-like property induction can therefore be potentially challenging for LMs, and therefore warrants thorough exploration.

On the basis of goals, inductive generalizations such as the ones described above are fundamentally different from the task of 'Natural Language Inference' (Bowman et al., 2015; Williams et al., 2018; NLI in short), also called 'Recognizing Textual Entailment' (Dagan et al., 2005; Dagan et al., 2013; RTE in short). In brief, NLI is the task of predicting the logical relationship between two sentences (premise and hypothesis) – either the premise *entails* the hypothesis, or it *contradicts* it.[8] There have been numerous iterations of the NLI task over the years, each involving a dataset of sentence pairs along with the ground truth label, typically collected by crowdsourcing. While early versions of the task instructed workers to infer whether the hypothesis is "likely true" given a premise (Dagan et al., 2005), present-day NLI tasks have used phrases like "definitely true" in their instructions (Bowman et al., 2015). Despite the differences in the instructions, the data in these tasks often resemble deductive inferences in which the relation between a premise and a hypothesis is *entailment* iff. the hypothesis is true when the premise is true. This is contradictory to the kinds of reasoning demonstrated during property induction, which necessarily focuses on likely but uncertain inferences, given a premise. Therefore, an LM-based account of property induction allows the study of human-like generalization (Linzen, 2020) that is yet to be studied in NLP systems.

### 2.3.1 Terminology

Inductive generalization in humans is often studied experimentally through the use of inductive arguments, which are often represented in the following premise-conclusion setup, as popularized by Osherson et al. (1990):

$$\frac{\text{Robins have sesamoid bones.}}{\text{Canaries have sesamoid bones.}} \tag{2}$$

[8] ↑in some cases, there is also a third type of relation, labelled as *neutral*. In such cases, the exact relationship between the premise and the hypothesis cannot be discerned.

Argument (2) is read as: *"Robins have sesamoid bones. Therefore, canaries have sesamoid bones."* In this notation, one or more premise(s) are written above the line, followed by a single conclusion, written under the line. Arguments with conclusions involving two or more concepts will be written separately. The subject of the premise sentence (Robin, in this case) 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 inductive generalization in such a manner allows us to use the construct of "argument strength," which in most cases quantifies the degree to which a human subject's belief in the premise statements strengthens their belief in the conclusion (Osherson et al., 1990). This then lets us compare different arguments in controlled experiments to shed light on regularities and phenomena typically observed in how humans reason inductively.

Inductive arguments such as argument (2) generally take two forms, based on the taxonomic status of the premise and the conclusion concepts. When the conclusion concept is strictly a taxonomic parent of the premise concept (at arbitrary levels of a specified taxonomy), we have a *general* argument:

$$\frac{\text{Tigers use the Dihedron neurotransmitter.}}{\text{Felines use the Dihedron neurotransmitter.}} \tag{3}$$

When this is not the case, we have a *specific* argument:

$$\frac{\text{Tigers use the Dihedron neurotransmitter.}}{\text{Whales use the Dihedron neurotransmitter.}} \tag{4}$$

Sometimes, an argument may contain a premise concept that is not included (taxonomically) by the conclusion concept. In such a case, we have a *mixed* argument:[9]

$$\begin{array}{l} \text{Tigers use the Dihedron neurotransmitter.} \\ \text{Ostriches use the Dihedron neurotransmitter.} \\ \hline \text{All mammals use the Dihedron neurotransmitter.} \end{array} \tag{5}$$

An important component of property induction is the type of property being used in arguments. In many cases, researchers aim to control the type of novel properties provided to participants, by using *blank* properties – that are true but assumed to be “expert knowledge” and are therefore unknown to participants, or created synthetically. The primary reason behind this is to force participants to use their knowledge of the relations between the premise and conclusion concepts/categories to make generalizations (G. L. Murphy, 2002).

In some experiments and analyses, we will use arguments that deal with multiple properties of the same—usually, novel—concept. These instances are often referred to as object generalization in the taxonomy of inductive problems proposed by Kemp and Jern (2014):

$$\begin{array}{l} \text{A wug has no legs.} \\ \text{A wug can swim.} \\ \hline \text{A wug has fins.} \end{array} \tag{6}$$

A special case of object generalization, known as category-to-feature inference (Estes, 1994) or property inheritance (Sloman, 1998; G. L. Murphy, 2002), involves inferences that are made for an entity whose higher level category is known/given, allowing recursive access to a wealth of knowledge via the hierarchical structure of concepts. For instance, knowing implicitly that cats have paws allows us to make inferences that endow the property to *cookie*, a novel cat that we might encounter (see argument (7)). This kind of reasoning has been ubiquitously studied by many classical theories of semantic memory (see Smith &

[9] ↑note that ostriches are not mammals.

Estes, 1978; for a review), and plays a generally important role in the study of knowledge representation and concept organization.

$$\frac{\text{cookie is a cat.}}{\text{cookie has paws.}} \tag{7}$$

### 2.3.2 Touchstone Phenomena in Human Inductive Generalization

Research in human inductive generalization typically focuses on the relative argument strength of inductive arguments as judged by human participants in an experimental setting. Such experiments often provide participants with pairs of inductive arguments and task them to select the argument the conclusions of which appeared to be more believable given the premise (Rips, 1975; Osherson et al., 1990). Alternatively, participants could also directly be asked to provide a scalar value that indicates the extent to which the conclusion is likely, given the premise (Heit & Rubinstein, 1994). By tracking patterns over multiple inductive trials, researchers have characterized a number of phenomena that typically deal with people's knowledge about concepts and their properties, and drive induction in humans. In what follows, I summarize hallmark phenomena in human induction, assimilated by cognitive scientists over the years.

#### Phenomena based on Taxonomic Similarity

We begin with induction phenomena based on the *hierarchical* organization of concepts. Here, generalization of newly-observed properties are made across *taxonomies*. A taxonomy can be defined as a tree-like data structure, whose nodes represent concepts (or categories), and edges represent a special relation called '`is_a`,' which generally refers to a subsumption relation – e.g., a DOG `is_a` CANINE suggests that the concept of CANINE *contains* the concept of DOG. The nodes of a taxonomy at the same level often form a particular type or 'kind', for instance CANINE and FELINE are often at the same level (under CARNIVORE), and at each level the concepts get progressively more specific than the ones "above" them. At the very top lies the most general node (e.g., ENTITY in case of WordNet). If in a taxonomy,

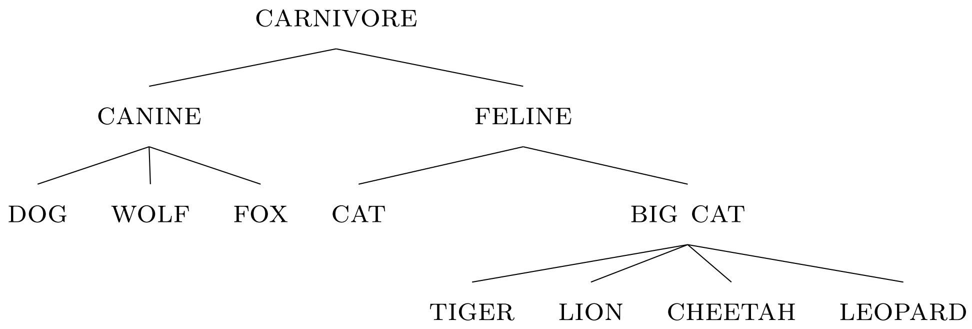


**Figure 2.2.** An example of an incomplete taxonomy of a few CARNIVORE concepts, taken from WordNet (Miller, 1995).

concept A is at a higher level than concept B, then we say that A is the *superordinate* of B, and B is the *subordinate* of A. Superordinates and subordinates are also called *hypernyms* and *hyponyms* in linguistics (Jurafsky & Martin, 2020).

Taxonomies have a privileged status in the study of concepts and semantic cognition in general – from inquiries into the nature of semantic memory (Collins & Quillian, 1969) to studies on graded, typicality effects (Rosch, 1975) and the 'basic-level' (Rosch, Mervis, et al., 1976; Rosch & Lloyd, 1978), to the word learning research by Xu and Tenenbaum (2007), *inter alia*. In the context of property induction, Gelman and Markman (1986) were one of the first ones to observe the role of taxonomic relations emerge in patterns of induction in children and adults. In their experiments, Gelman and Markman first displayed a picture of a flamingo to their subjects and informed them that it had a "right aortic arch," then displayed a picture of a bat and told them that it had a "left aortic arch." The subjects were then shown a picture of a blackbird (that visually resembled the bat more than it did the flamingo), and were asked whether it had a right or a left aortic arch. The subjects attributed the "right aortic arch" to the blackbird, indicating a *taxonomic bias* in their inductive judgments. Building on research by Gelman and Markman (1986), Osherson et al. (1990) use the notion of *similarity* to describe their observations in a landmark paper that documented 13 separate induction phenomena related to taxonomies. While Osherson et al. do not explicitly define what they mean by similarity – a likely candidate could be property overlap (Tversky, 1977),[10] especially in the context of this dissertation. All 13 phenomena compiled by Osherson et al. (1990) deal with taxonomic relations between concepts involved in the premise and conclusion portion of inductive arguments. For instance, sometimes the phenomena would involve the generalization from specific categories to a general category that subsumes them (i.e., projections of a novel property from DOG to the entire CANINE category), while sometimes it may involve the generalization between concepts at similar levels (for instance, generalization of a new property from LION and MICE to WOLF). The following are the various phenomena reported by Osherson et al. (1990) in their work:

**Premise Typicality** The phenomenon of *typicality* suggests that certain members of a category are more representative to it, than are others (Rosch, 1975). For instance, APPLE may be considered to be a more typical fruit than TOMATO. Like in many other cognitive phenomena involving concepts and categories, typicality also modulates property induction. Novel properties are more likely to be generalized to the rest of the category when linked with typical—as opposed to atypical—concepts. That is, argument (8) is considered to be stronger than argument (9).

$$\frac{\text{Robins have sesamoid bones.}}{\text{All birds have sesamoid bones.}} \tag{8}$$

$$\frac{\text{Penguins have sesamoid bones.}}{\text{All birds have sesamoid bones.}} \tag{9}$$

**Premise Diversity** The diversity of premise concepts also tends to influence inductive judgments, i.e., premises with similar concepts are less likely to generalize to the entire category (that subsumes them) as compared to those that contain different ones (different, but still subsumed by the conclusion category). That is, argument (10) is considered to be stronger than argument (11).

[10] ↑this could correspond to the jaccard similarity between the binary feature representations of two categories. Mathematically, $sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

$$\begin{array}{l}\text{Lions have the T9 hormone.}\\ \text{Mice have the T9 hormone.}\\ \hline \text{All animals have the T9 hormone.}\end{array}\tag{10}$$

$$\begin{array}{l}\text{Lions have the T9 hormone.}\\ \text{Tigers have the T9 hormone.}\\ \hline \text{All animals have the T9 hormone.}\end{array}\tag{11}$$

This phenomena is also applicable to specific arguments, i.e., the pattern holds even if 'all animals' in the conclusion of the above arguments is replaced by 'giraffe.'

**Conclusion Specificity** The homogeneity of the conclusion concept with respect to the premise concepts affects inductive generalization of novel properties. For the same set of premise concepts (e.g., ROBIN, SPARROW), inductive arguments with more specific conclusion concepts (e.g., BIRD) are stronger than those with more general conclusion concepts (e.g., ANIMAL). This implies that more evidence is typically needed to support a conclusion whose concept has greater coverage relative to the premise concepts. For example, argument (12) is stronger than argument (13):

$$\begin{array}{l}\text{Robins require Vitamin K.}\\ \text{Sparrows require Vitamin K}\\ \hline \text{All birds require Vitamin K.}\end{array}\tag{12}$$

$$\begin{array}{l}\text{Robins require Vitamin K.}\\ \text{Sparrows require Vitamin K}\\ \hline \text{All animals require Vitamin K.}\end{array}\tag{13}$$

**Premise Monotonicity** The inclusivity of the premise concepts promotes inductive generalization. That is, arguments with greater number of premise concepts are generally stronger than those with fewer premise concepts. For example, argument (14) is stronger than argument (15):

$$\begin{array}{l} \text{Foxes use Vitamin K to produce clotting agents in their blood.} \\ \text{Pigs use Vitamin K to produce clotting agents in their blood.} \\ \text{Wolves use Vitamin K to produce clotting agents in their blood.} \\ \hline \text{All mammals use Vitamin K to produce clotting agents in their blood.} \end{array} \tag{14}$$

$$\begin{array}{l} \text{Foxes use Vitamin K to produce clotting agents in their blood.} \\ \text{Pigs use Vitamin K to produce clotting agents in their blood.} \\ \hline \text{All mammals use Vitamin K to produce clotting agents in their blood.} \end{array} \tag{15}$$

This phenomena is also applicable to specific arguments, i.e., the pattern holds even if 'all mammals' in the conclusion of the above arguments is replaced by 'gorillas.'

**Non-monotonic Effects** Inductive arguments can be made weaker by the addition of a premise concept that converts them into a mixed argument. Argument (16) is stronger than argument (17).

$$\begin{array}{l} \text{Crows secrete uric acid.} \\ \hline \text{All birds secrete uric acid.} \end{array} \tag{16}$$

$$\begin{array}{l} \text{Crows secrete uric acid.} \\ \text{Gorillas secret uric acid.} \\ \hline \text{All birds secrete uric acid.} \end{array} \tag{17}$$

This phenomena is also applicable to specific arguments, i.e., the pattern holds even if 'all birds' in the conclusion of the above arguments is replaced by 'Peacocks.'

**Premise-Conclusion Asymmetry** This phenomenon was first observed by Rips (1975). Single premise arguments are often not symmetric, i.e., the strength of an argument with a premise $P$ and a conclusion $C$ is different than the argument that has $C$ as its premise, and $P$ as its conclusion.

**Inclusion Fallacy** A general argument can sometimes be made weaker by converting it into a specific argument. In such cases, experiment subjects can sometimes rate an argument to be stronger when its conclusion concept represents a whole category than when it represents a subset of that category. Argument (18) is rated to be stronger than argument (19).

$$\frac{\text{Robins have an ulnar artery.}}{\text{All birds have an ulnar artery.}} \tag{18}$$

$$\frac{\text{Robins have an ulnar artery.}}{\text{Ostriches have an ulnar artery.}} \tag{19}$$

This phenomenon is so named because if a given concept having a particular property makes people think its superordinate also has that property to an extent, then all subordinate concepts should have the property at least as much. Since this is not the case, we have a *fallacy*. This is similar to the conjunction fallacy (Tversky & Kahneman, 1983).

### Phenomena based on Salient Background Knowledge

Even though taxonomies have a privileged status in the study of semantic cognition, the full range of inductive phenomena goes well beyond the hierarchical organization of concepts. Indeed, a majority of works following that of Osherson et al. (1990) have experimentally confirmed the existence of inductive behavior in humans that cannot be explained by taxonomic similarity, and instead show patterns that are in direct contrast to the ones discussed previously (see Hayes et al., 2010; Hayes & Heit, 2018; for a review). In such cases, inductive generalization is observed to occur between concepts not on the basis of taxonomic relations, but rather on the basis of salient property knowledge that is shared between them. For instance, Heit and Rubinstein (1994) found inductive generalization of behavioral properties such as “swimming in a zig-zag pattern” to be strongly generalized from sea-mammals such as dolphins to fish as compared to other mammals such as giraffes and tigers. Such observations highlight the flexibility of inductive inferences, that is free from the shackles of any one single *theory* of domain knowledge (G. L. Murphy, 1993; G. L. Murphy & Medin, 1985), e.g. taxonomic relations between concepts. In such a viewpoint, property induction can be viewed as a process that involves the identification of the appropriate domain knowledge, controlled by the context of the premise concepts and the property being generalized, and then using the features implied by the domain knowledge to carry out the generalization (G. L. Murphy, 2002). Even if we are to allow *similarity* to be the driver of induction, as suggested by Osherson et al. (1990), it is likely to be updated dynamically, possibly by considering only the features of the domain knowledge activated by the premise during induction.

The context specified by the premise in an inductive argument can vary in a myriad of different ways. One possible way could be when the property being generalized is not truly unfamiliar, or *blank* (Heit & Rubinstein, 1994). Recall that the original experimental paradigm to study inductive reasoning used blank properties—properties that are likely to be unfamiliar to the subjects of the experiments—in the premise (Rips, 1975; Osherson et al., 1990). Heit and Rubinstein (1994) experimentally found that taxonomically motivated similarity may not always account for induction when the property being generalized is non-blank. Other ways in which the context of the premise may be modulated is when the premise concepts share an implicit property that dictates the kind of induction that is being made (e.g. polar bears and penguins, which activate the domain of "living in cold climate"), as discussed by Medin et al. (2003). Note that the aforementioned possibilities of premise context modulation is certainly not exhaustive, but crucial enough to warrant substantial discussion. I further expand on them in what follows.

**Context Sensitivity** When the property in a given inductive argument are not *blank* and is instead familiar to the reasoner, its strength is determined by the domain knowledge/context relevant to the property. For instance, Heit and Rubinstein (1994) found anatomical properties such as "*have an ulnar artery*" to generalize taxonomically – i.e., argument (20) is stronger than argument (21) since sparrows and hawks are both birds and are likely to share anatomical characteristics.

$$\begin{array}{l}\text{Sparrows have a ulnar artery.}\\ \hline \text{Hawks have a ulnar artery.}\end{array} \tag{20}$$

$$\frac{\text{Tigers have a ulnar artery.}}{\text{Hawks have a ulnar artery.}} \tag{21}$$

However, when the property in the inductive argument was behavioral ("*study their food before attacking*"), then the generalization pattern reverses (as compared to arguments (20) and (21) – i.e., argument (22) is stronger than argument (23) since tigers and hawks are both predators, and hence their similarity along this axis trumps their taxonomic relation.

$$\frac{\text{Tigers study their food before attacking.}}{\text{Hawks study their food before attacking.}} \tag{22}$$

$$\frac{\text{Sparrows study their food before attacking.}}{\text{Hawks study their food before attacking.}} \tag{23}$$

**Nondiversity by Property Reinforcement** Even when the properties *are* blank, there exist cases where inductive phenomena covered in §2.3.2 can be reversed. For instance (Medin et al., 2003) found that the Premise Diversity phenomenon is reversed when concepts in the premise have a salient feature. In such a case, a general argument can be weakened even in the presence of diverse premise concepts when the conclusion concept did not share the implicit salient feature that was reinforced by the premise. Argument (25) is weaker than argument (24) since it has the implicit property of "*living in cold habitats*" that is only applicable to a specific subset of the ANIMAL category.

$$\begin{array}{l} \text{Polar bears have Property } X \\ \text{Antelopes have Property } X \\ \hline \text{All animals have Property } X \end{array} \tag{24}$$

$$\begin{array}{l} \text{Polar bears have Property } X \\ \text{Penguins have Property } X \\ \hline \text{All animals have Property } X \end{array} \tag{25}$$

**Nonmonotonicity by Property Reinforcement** Similar to the Nondiversity by Property Reinforcement phenomenon, the presence of salient features also overrides the Premise Monotonicity phenomenon (Medin et al., 2003). Consider arguments (26) and (27). The premise of argument (27) reinforces the property of "*being a bear*" – that "Property $X$" is likely only a property of bears, this is less likely in the case of argument (26), where the reinforcement is weak. As a result, argument (26) is stronger than argument (27).

$$\frac{\text{Brown bears have Property } X}{\text{Buffalo have Property } X} \tag{26}$$

$$\frac{\begin{array}{l}\text{Brown bears have Property } X\\ \text{Grizzly bears have Property } X\\ \text{Polar bears have Property } X\end{array}}{\text{Buffalo have Property } X} \tag{27}$$

# 3. INVESTIGATING TYPICALITY EFFECTS IN LANGUAGE MODELS

This chapter adapts work from the following publication:

> Misra, K., Ettinger, A., & Rayz, J. (2021). Do language models learn typicality judgments from text?. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43).

It has been reproduced here with slight alterations.

## 3.1 Introduction

Perhaps one of the most important findings in the study of human categorical knowledge is the phenomenon of *typicality*, the observation that certain members of a category are considered to be more representative of the category than are others (G. L. Murphy, 2002). As observed in the pioneering work of Rosch (1975), native English speakers tend to rate *robins* and *canaries* as more typical birds than *penguins* and *emus*, *chairs* and *sofas* as more typical furniture than *clocks* and *vases*, etc. Apart from modulating categorization experiments, typicality reliably affects people's *use* of categories. In particular, typicality differences in stimuli strongly predict response times in taxonomic sentence verification tasks (Rips et al., 1973; Rosch, 1973) and category production (Rosch, Simpson, et al., 1976) – human subjects take longer to verify sentences expressing atypical category-membership knowledge as compared to typical ones, and similarly are more likely to produce typical examples of categories as compared to atypical ones. In the context of concept learning, typical items facilitate faster concept acquisition than do atypical items (Rosch, Simpson, et al., 1976). Typicality also prominently affects category-based inductive reasoning judgments (Rips, 1975; Osherson et al., 1990):[1] that is, subjects more readily extend new information (*has T9 hormones*) about typical (ROBIN)—as opposed to atypical (PENGUIN)—items to other members of their superordinate category (BIRD). In summary, typicality is a salient and impactful phenomenon in the study of human category knowledge.

[1] ↑an important class of inductive generalization problems solved by humans

In this chapter, we propose tests—inspired by the aforementioned experimental studies—that target the question of typicality in Language Models (LMs). Why should we investigate typicality effects in LMs? I provide three reasons. First, as established in ch. 1 and ch. 2, LMs are perhaps the most powerful implementations of a perspective where words do not directly map onto symbolic concepts but are instead cues to conceptual meaning (Elman, 2004; Lupyan & Lewis, 2019)—i.e., words directly affect 'mental states', which leads to semantic behavior. Therefore, by learning representations for words from their usage, LMs encode a possible way in which words can affect internal states of a system (Pavlick, 2022, 2023), and are therefore possible accounts of how everyday concepts are used (as words). Since typicality crucially affects people's behaviors and judgments for stimuli involving everyday concepts and categories, any genuine account of concept or category-use *must* ideally also exhibit typicality effects. Second, LMs encode many desirable prerequisite properties of graded structure by construction – they are composed of continuous vector spaces and predict probability distributions over tokens, both of which are graded by definition. Investigations involving typicality can therefore serve as a means by which we can relate the graded structure in LMs with human conceptual structures. Finally, there are a number of empirical investigations of LMs (and other distributional semantic models) that suggest them to be adept at predicting category membership of everyday concepts as well as their commonly known properties (Lucy & Gauthier, 2017; Petroni et al., 2019; Weir et al., 2020; Ettinger, 2020)[2] – factors that are at the center of what makes an item a typical (or atypical) member of a category (described by Rosch and Mervis (1975) as *family resemblance*). By analyzing them for for typicality effects, we can further cast light on the extent to which LMs go *beyond* simply capturing static categorical and property knowledge for concepts (which can potentially be explained by simpler word association effects or even memorization from their pre-training data), and encode deeper nuances of conceptual structure.

In order to investigate LMs for typicality effects, we adopt two prior experimental paradigms where humans showed clear sensitivities to typicality in processing of textual stimuli. First, we build on prior work analyzing conceptual and categorical knowledge in

[2] ↑though the question of just how robustly they capture category and property knowledge will be addressed in the next chapter.

LMs, and test whether typicality effects modulate LM judgments of taxonomic sentence verification (*"a robin is a bird"*) as they do in humans (Rips et al., 1973; Rosch, 1973). Complementing this simple and direct test of taxonomic category membership, we add a layer of complexity, and investigate the manifestation of typicality effects in LMs on the basis of how they extend new information about items (*"robins can dax"*) to all members of a category (*"all birds can dax"*), inspired by tests targeting psychological strength of inductive arguments (Rips, 1975; Osherson et al., 1990). Though the human experiments that inspire these tests do not *explicitly* target typicality as a phenomenon, typicality effects still robustly modulate human behavior on them. Hence, we examine whether LMs show comparable typicality effects on stimuli similar to those used in the above experiments.

## 3.2 Methodology

### 3.2.1 Models Investigated

We conduct our analyses on pre-trained LMs based on the transformer architecture (Vaswani et al., 2017). Our choice of LMs is motivated by recent evidence that shows qualitative alignment of category knowledge (*"a robin is a <u>bird</u>"*, *"a <u>bear</u> has fur, has claws."*) in pre-trained LMs (Ettinger, 2020; Weir et al., 2020). Although we focus on a particular type of pre-trained LMs (transformers) in this work, the tests we propose can be applied to any LM. We investigate two broad classes of transformer-based pre-trained LMs: **(1) Autoregressive LMs**, trained autoregressively (left to right) to predict one word at a time, when conditioned on exclusively the left context; and **(2) Masked LMs**, that access context of the word to be predicted bidirectionally, e.g., the models are optimized to predict correct completions (*airplane* or *bird*) to sentences such as *"the* [MASK] *flew away,"* where [MASK] represents the hidden word. We apply our tests on GPT (Radford et al., 2018) and GPT2 (Radford et al., 2019) as our Autoregressive LMs, and ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), BERT, (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as our Masked LMs. In addition, we use compressed versions of the above models (Sanh et al., 2019): distilGPT2, distilBERT-base, and distilRoBERTa-base. All transformer-based pre-trained

**Table 3.1.** Number of subordinate concepts ($N$) per superordinate category (Rosch, 1975).

| Superordinate | $N$ | Superordinate | $N$ |
|---|---|---|---|
| FURNITURE | 60 | VEGETABLE | 56 |
| TOOL | 60 | CLOTHING | 55 |
| TOY | 60 | BIRD | 54 |
| WEAPON | 60 | FRUIT | 51 |
| SPORT | 59 | VEHICLE | 50 |

LMs were accessed using the `minicons`[3] library, a wrapper around the `transformers` library (Wolf et al., 2020).

### 3.2.2 Data and Stimuli

#### Item typicality data

For both experiments, we use as our primary source the list of 565 item-typicality ratings compiled by Rosch (1975) across 10 different categories. In the original human experiments, 209 native speakers of English were tasked to rate the "goodness of example" for various items of each given category, on a scale of 1 (most typical) to 7 (least typical). The statistics of the items and categories is presented in Table 3.1. It should be noted that the experiments we base our tests on involve sensitivities to typicality measured using different quantities (response times and raw typicality ratings), but make none or only a small subset of results available. Therefore, we use the Rosch (1975) ratings as the common "ground-truth" typicality ratings for our experiments.

#### Stimuli Setup

Because the models we investigate are sentence processors, and because all of our tests involve propositions about items and categories expressed as sentences, we rely on using sentence stimuli in our experiments. Every stimulus consists of two components: (1) condition, which

[3] ↑https://github.com/kanishkamisra/minicons

**Table 3.2.** Examples of stimuli used in our experiments. Our measures take the form: $\log p(\text{predicted} \mid \text{condition})$

| Experiment | Stimulus |
|---|---|
| Taxonomic Sentence Verification | $\underbrace{\text{A robin is a}}_{\text{condition}}\ \underbrace{\text{bird}}_{\text{predicted}}$ . |
| Category-based Induction | $\underbrace{\text{Saws can dax.}}_{\text{condition}}\underbrace{\text{All tools can dax.}}_{\text{predicted}}$ |

is a noun phrase/sentence consisting of the item (*robins, sparrows, eagles, etc.*); and (2) predicted material, which consists of the super-ordinate category (*bird*). The exact linguistic format in which it appears depends on the experiments — we use single words as the predicted material in our Taxonomic Sentence Verification experiment while for our Category-based induction experiment we use an entire sentence as our predicted material. In evaluating typicality measurements of various items for a given category, the predicted material remains constant, while the condition changes depending on the item. Table 3.2 shows examples of stimuli we use in each of our experiments.

### 3.2.3 Measures

Following precedent set by previous work evaluating conceptual knowledge in pre-trained LMs, we use the models' probability estimates as our main variable of interest. Specifically, we focus on the log-probability of the word or statement represented in the predicted material, given the condition:

$$\log p_{LM}(\textit{predicted} \mid \textit{condition}), \tag{3.1}$$

that is, we are measuring the effect on the probability of the predicted part (held constant for a given category) due to the item mentioned in the condition. Our reason for separating the item from the predicted material is two-fold: (1) it avoids skewed measurements due to the choice of determiner (*a* vs *an*) that precedes the item in the condition (a model might assign higher value to $p(\textit{ostrich} \mid \textit{an})$ simply due to a component that is sensitive to determiner prefixes), or when the model does not include the item word in its vocabulary,[4] and (2) it aids in factoring out the role played by the frequency of the item in the condition – the model can prefer an item over the other simply due to its frequency in the training corpus. While it is straightforward to compute our conditional probability measure for Autoregressive LMs by using the chain-rule, we rely on recent work by A. Wang and Cho (2019) and Salazar et al. (2020) to approximate sequence log-probabilities in Masked LMs by summing the conditional log-probabilities of all words in the stimuli.

## 3.3 Experiments and Analyses

We use two experimental paradigms where humans showed clear sensitivities to typicality effects—(1) Taxonomic Sentence Verification; and (2) Category-based Induction. In what follows, we first describe the typicality-related phenomenon observed in each of these experiments, and then describe how the experiment and phenomenon can be linked to LM-derived measures. We then discuss our precise experiment, and finally report and analyse its results.

### 3.3.1 Taxonomic Sentence Verification

#### Description of Phenomenon

Typicality effects in the sentence verification paradigm were introduced by Rips et al. (1973) and Rosch (1973). Subjects were tasked with verifying the truth of sentences expressing taxonomic propositions, such as *"An $X$ is a $Y$"*—where $X$ and $Y$ are the item and category, respectively. The subjects consistently responded faster to verifying the truth of propositions where $X$ was a typical member of $Y$ than when it was an atypical one.

#### Linking Phenomenon to LMs

We draw on the aforementioned findings and investigate whether typicality is able to account for difference in the word probabilities to complete taxonomic sentences by our tested LMs. Linking our hypothesis to the original experiment requires a simplifying assumption that

[4] ↑E.g., RoBERTa and GPT2 segment the word *ostrich* into *ostr* and *ich*, and during estimation, the probability of *ich* given that it is preceded by *ostr* is anomalously high, skewing the overall sequence probability.

an LM's sequence log-probability is proportional to its plausibility for a sequence. That is, we assume and expect a semantically sound LM to show overall high probability scores for semantically plausible propositions, which in this case, are simple taxonomic propositions[5]. Therefore, LMs that are more sensitive to typicality effects should show greater magnitudes for the measure $\log p_{LM}(Y \mid \textit{An } X \textit{ is a})$ when $X$ is a more typical member of $Y$.

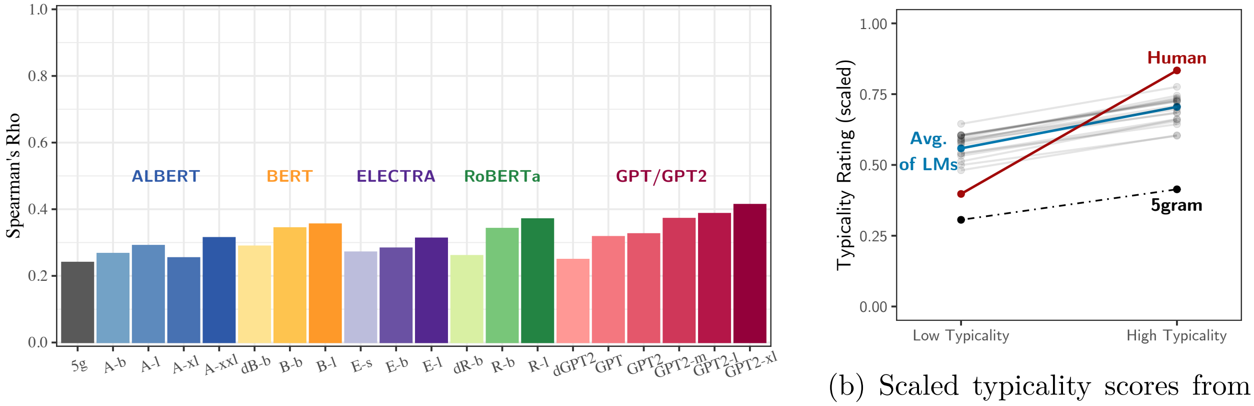
### Experiment

We follow Rips et al. (1973) and Rosch (1973) and construct sentences expressing taxonomic propositions using items from the Rosch (1975) data, i.e., *"An X is a Y,"* amounting to 565 unique propositions. We test for typicality effects by measuring the Spearman correlation ($\rho$) of the sequence log-probability $\log p_{LM}(Y \mid \textit{An } X \textit{ is a})$ with the human typicality ratings for items, as collected by Rosch (1975). This correlation measure reflects the extent to which the predictive estimates of an LM reflect typicality information—or information that underlies it—to assess taxonomic verification in sentences. Additionally, we perform a median split on the Rosch (1975) ratings by the items' typicality ratings, per category, leaving us with two sets of typical and atypical ratings. We then compute the average log-probabilities assigned to items in each set and compare them to the average ratings elicited by humans. All scores in this analysis are re-scaled to be between 0 and 1.

### Results

Figure 3.1 show results from our correlation and typicality-effect comparisons. Non-trivially positive but modest correlations between LM log-probabilities and human typicality ratings ($\rho \in [0.24,\ 0.41]$, $p < .001$) suggest that LMs' judgments of taxonomic propositions are moderately reflective of typicality effects. Though all LMs assign greater probability scores to category items with high—as compared to low—typicality (see Figure 3.1b), they are consistently less extreme as compared to humans ($p < .001$ across all models). Correlation of 5-gram LM log-probabilities, though weakest in magnitude, is highly competitive with

[5] ↑However, we acknowledge that this might not always be the case. For instance, LMs are largely insensitive to negation and semantic role-reversals (Ettinger, 2020).



(a) Spearman correlation ($\rho$) measured between LM log-probabilities on the taxonomic sentence verification stimuli and typicality ratings from Rosch (1975). Models from the same family are arranged in an increasing order of total number of parameters.

(b) Scaled typicality scores from LMs (log-probabilities on the taxonomic sentence verification stimuli) and Humans (raw ratings) between low and high typicality category members.

**Figure 3.1.** Results from the Taxonomic Sentence Verification experiment.

certain smaller yet highly expressive LMs (ALBERT-b, ALBERT-xl, distilGPT2, and distilRoBERTa). This suggests that a substantial portion of the observed correspondence between model and human typicality judgments can be attributed to fairly simpler sequential statistical effects in word prediction (e.g. memorizing n-grams). Interestingly, with the exception of the ALBERT family of Masked LMs, models with greater number of parameters tend to show greater correspondence with humans on taxonomic judgments ($\rho = 0.82, p < .001$), suggesting that the information needed to distinguish typical vs. atypical category members during taxonomic attribution requires greater model expressivity.

### 3.3.2 Category-based Induction

#### Description of Phenomenon

As mentioned in §2.3.2, typicality plays a salient role in modulating human behavior on inductive reasoning problems involving categories (Rips, 1975; Osherson et al., 1990; i.e., category-based induction). In particular, typicality effects were one of the 13 phenomena documented by Osherson et al. (1990) – for single-premise general arguments[6], human sub-

[6] ↑see §2.3.1 for what general arguments are.

jects were more likely to believe in the conclusion when the category mentioned in the premise was more typical member of that in the conclusion. That is, the argument strength of argument (28) was found to be greater than that of argument (29), since ROBIN is a more typical BIRD than is PENGUIN.

$$\frac{\text{Robins have sesamoid bones.}}{\text{All birds have sesamoid bones.}} \tag{28}$$

$$\frac{\text{Penguins have sesamoid bones.}}{\text{All birds have sesamoid bones.}} \tag{29}$$

### Linking Phenomenon to LMs

The Osherson et al. (1990) study explicitly targets the degree to which uncertain statements such as "all birds have T9 hormones" are judged in light of new information about a subordinate category such as *robins*. Analogously, we are interested in assessing whether sophisticated LMs show similar behavior in assigning probabilities to conclusions when conditioned on premises whose categories vary based on their typicality. If LMs show sensitivity to the typicality of items in this setting, i.e., their log-probability is greater for conclusions with typical versus atypical premise, then we take this as preliminary evidence that typicality—or the factors that underlie it—modulates inductive inference in LMs. We formulate an approximation of inductive argument strength ($AS$) in an LM as the probability it assigns to the conclusion when conditioned on a given premise. For instance $AS(\textit{robin}, \textit{bird})$ for the property *"has T9 hormones"* is given by:

$$\log p_{LM}(\textit{“All birds have T9 hormones.”} \mid \textit{“Robins have T9 hormones.”})$$

The premise and conclusion sentences naturally fit within our stimulus setup discussed earlier — the premise sentence is the condition, and the conclusion sentence the predicted material.

### Experiment

For our items and categories, we again use data from Rosch (1975). Since Osherson et al. do not make all of their blank predicates available, we construct synthetic properties using novel (or pseudo) words such as *dax*, *wugs*, *feps*, *vorpal*, etc., as objects of common verbs such as *has, can, is a, etc.* This allows us to conform to the blank predicate condition applied by Rips (1975) and Osherson et al. (1990) since these words are either outside of the model's effective vocabulary,[7] or are usually rare in the corpora that the models are trained on. We create between 15 to 30 properties[8] for all items in each category, resulting in a total of 12,180 premise-conclusion pairs across 10 categories. An example of the stimuli we use for our category-based induction task is shown in Table 3.2. We calculate the *AS* metric for each premise-conclusion pair with each of our tested LMs.

Conditioning our LMs as we do here has two potential confounds: **(1) Premise Order Sensitivity (POS)**: A model might estimate high probabilities for words in the conclusion simply because it is relying on lexical cues in its premise (Misra et al., 2020), instead of processing the premise *compositionally* and making inferences about items possessing a property. To study this confound, we compute the LMs' average probability for the conclusion sentence when prefixed by a shuffled version of the premise (10 times, with random seeds). We then calculate POS as the difference between this measure (which we call $AS_{\textit{shuffled}}$) and the original $AS$ score, for each item:

$$\text{POS}_i = AS_i - AS_{\textit{shuffled},i} \tag{3.2}$$

It would be desirable to have POS values that are greater than 0, signifying that the LM is indeed sensitive to the correct word-order structure of the premise. Figure 3.2a displays

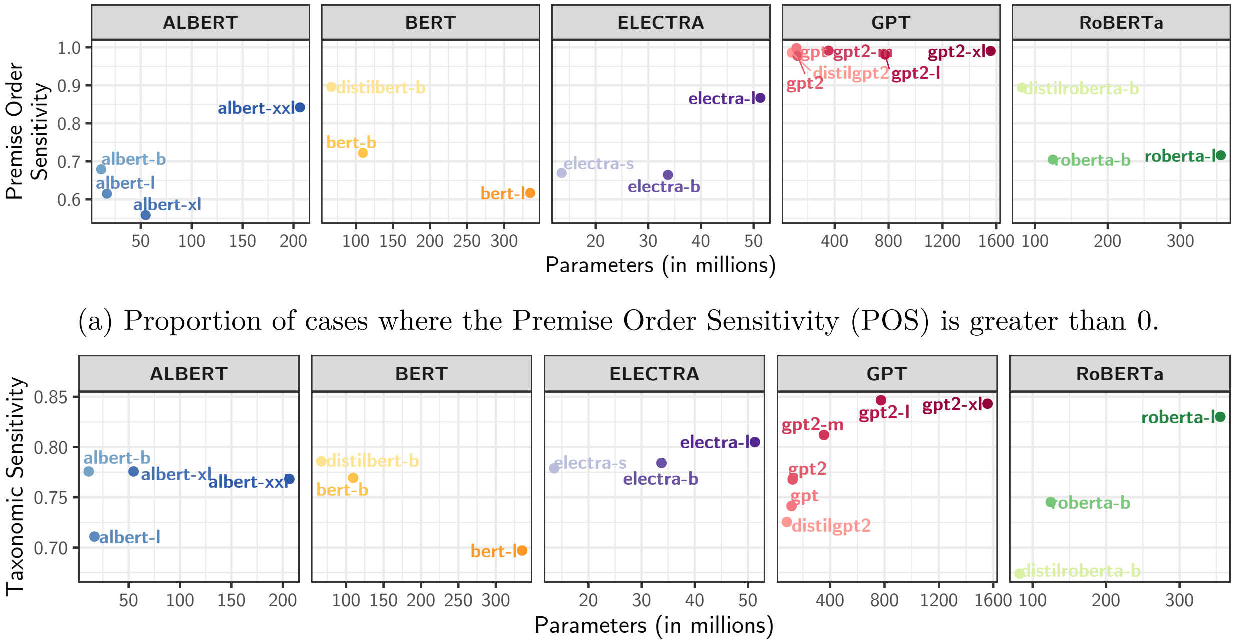
[7] ↑Due to their tokenization mechanism, the LMs we study are always able to encode any text through 'word pieces' instead of relying on <unk> tokens. So even if these words appear to be outside of the models' effective vocabulary, they will always be tokenized.

[8] ↑The choice of properties depends largely on the class of word the items belong to, such that syntactic constraints are met. For instance, if *dax* is a verb, it would be ungrammatical to have "can dax" as a property of sports, which can be better paired with properties such as "involve" and "require". The entire unique set of synthetic properties and our construction method is made available in our supplementary materials.

the proportion of cases (out of 12,180) where POS was greater than 0, against the models' total number of trainable parameters. We observe that only Incremental LMs show nearly complete sensitivity to the word-order of the premise. Masked LMs on the other hand do show sensitivity in a majority of cases, but are still far below incremental LMs, suggesting that this confound greatly affects their results, and is likely not to affect Incremental LMs. No particular effect of number of parameters was observed. **(2) Taxonomic Sensitivity (TS)**: LMs might tend to repeat the property phrase mentioned in the predicted material with high probability when prefixed by a sentence containing it, i.e., repeating *"can dax"* in the conclusion when already conditioned on the same phrase in the premise (Holtzman et al., 2019), confounding the degree to which the conclusion is generated using the taxonomic relationship between the premise and the conclusion categories. In order to study this tendency, we compute the LMs' probabilities for conclusions consisting of a different category with the exact same property as the original (for instance, *"All fruits are slithy"* given *"Sofas are slithy"*). We call this measure $AS_{\mathit{flipped}}$. Just like in our POS calculation, our TS measure for each item is calculated as the difference between the original measure and the flipped measure:

$$\text{TS}_i = AS_i - AS_{\mathit{flipped},i} \tag{3.3}$$

As is the case with POS, it is desirable to have a strongly positive value for sensitivity towards this confound (i.e., TS $> 0$). Figure 3.2b shows the proportion of cases (out of 12,180 for each model) where TS was greater than 0, against the models' total number of trainable parameters. From Figure 3.2b, there is a non-trivial proportion of cases where models produce greater $AS$ scores for the conclusion when the premise concept had no taxonomic relation to the conclusion concept as compared to when it did, suggesting that in many cases models might not be processing based on the taxonomic relation between premise and conclusion and may simply be assigning high probabilities to the property phrase because it was listed in the preceding context – it is common-knowledge that LMs are highly conducive to repeating text already seen in the input context (Holtzman et al., 2019).



(a) Proportion of cases where the Premise Order Sensitivity (POS) is greater than 0.

(b) Proportion of cases where the Taxonomic Sensitivity (TS) is greater than 0.

**Figure 3.2.** Confounds from the Category-based Induction Experiment.

We find that a substantial amount of variance in our original $AS$ scores is in fact captured by both these confounds (overall $r^2 = 0.43$, $\beta_{\text{TS}} = 0.68$, $\beta_{\text{POS}} = -0.04, p < .0001$ in both cases). We regress these relationships out from our $AS$ scores by first fitting a multiple regression model to predict $AS$ using our confounds, and then subtracting its linear relationship with TS and POS as follows:

$$\begin{aligned}
AS &= \beta_0 + \beta_1\text{TS} + \beta_2\text{POS} + \epsilon \\
AS' &= AS - \beta_1\text{TS} - \beta_2\text{POS} \\
&= \beta_0 + \epsilon \qquad \text{(Adjusted } AS\text{)}
\end{aligned}$$

Using the adjusted $AS$ scores in each LM, we compute the score of generating the conclusion (scaled between 0 and 1) for each category, item, and synthetic property, and average them to get the model's overall score for extending new information about an item to its category. We compute the Spearman's correlation ($\rho$) between our normalized adjusted $AS$

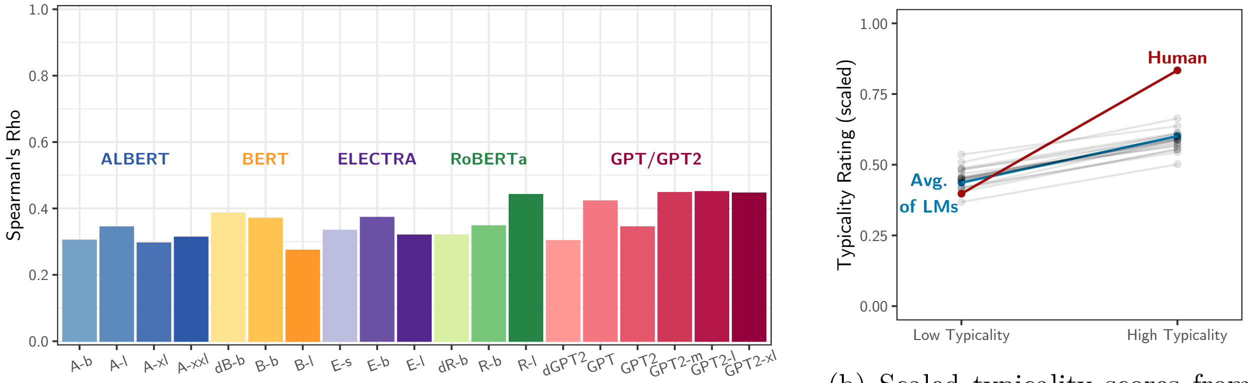
scores and the human typicality ratings from Rosch (1975), and compare average $AS$ scores (across all blank properties that we used in this experiment) and human typicality ratings assigned to low and high-typicality items.

### Results

Figure 3.3 summarizes results from our category-based induction experiments. When LMs extend information about an item to its category, they are moderately but positively influenced by its typicality ($\rho \in [0.27, 0.45]$, $p < .001$ everywhere). This influence is above and beyond their usual predilection towards repeating sequences and being lexically sensitive to items present in the premise (Misra et al., 2020). In terms of the training objectives of the models, Autoregressive LMs show stronger correspondence with human ratings as compared to Masked LMs of comparable size, suggesting that they are slightly more sensitive to the typicality of the premise item in generating the conclusion. We notice almost no effect of model size (in terms of parameters) on the results, suggesting that the factors that underlie typicality effects in property induction are likely independent of the number of parameters of an LM. This contradicts the set of results reported in the previous experiment (§3.3.1), where typicality-sensitive attribution of items to their superordinate categories was generally found to be improved by scaling up the overall expressiveness of the model.

## 3.4 General Discussion

Extensive research in the field of cognitive science has highlighted the prevalent role played by typicality in studies of categories—that certain items (*chair*) are considered to be better representatives of a category (*furniture*) than others (*vase*). Motivated by recent evidence showing pre-trained LMs to capture patterns exhibiting conceptual and categorical knowledge, we presented two experiments targeting sensitivities to typicality in LMs. The first experiment targets typicality directly, in its role played in associating items to their taxonomic categories (*"football is a <u>sport</u>"*). Our second experiment complements this by instead assessing the extent to which the "knowledge" of category typicality is used to extend information about items (*"football involves blicking"*) to their respective categories (*"all*



(a) Spearman correlation ($\rho$) measured between average *AS* scores and human typicality ratings compiled by Rosch (1975). Models from the same family are arranged in an increasing order of total number of parameters.

(b) Scaled typicality scores from LMs (as represented by their *AS* values) and Humans (raw ratings) between low and high typicality category members.

**Figure 3.3.** Results from the Category-based Induction experiment.

*sports involve blicking"*). We investigate typicality effects in LMs by evaluating their log-probabilities in response to stimuli as measures of (1) taxonomic verification and (2) inductive argument strength (when conditioned on a premise). For each test, we made the simplifying assumption that the likelihood assigned by the LM to the sentence stimuli corresponds to the variables of interest—strength of category membership in the first experiment, and argument strength in the second. Overall, the pre-trained LMs showed positive but modest correlations with human typicality ratings in both experiments, and were, on average, far less extreme in distinguishing between typical and atypical items than humans. We also observed that a considerable amount of sensitivity to typicality effects can be attributed to the mechanisms available to simpler LMs (5-gram), relative to the sophisticated pre-trained LMs that we studied here, suggesting that the representational mechanisms in most models that are optimized to reflect the statistics in training corpora only account for a minimal gain over correspondence that is afforded by simpler sequential statistics. Results on pre-trained LMs suggests that the statistical associations that inform their word probabilities are modestly sensitive to human-elicited typicality ratings in (1) attributing items to their category members, as well as (2) making complex inductive inferences about categories when conditioned on new information about the items. While our taxonomic sentence verification experiments showed typicality correspondence to increase with model size, this was not the case in our induction experiments, suggesting that extending new information about items to their categories in a manner that is positively modulated by typicality effects does not scale with an increase in parameters. We leave fine-grained exploration of specific language modelling factors affecting typicality correspondence for future work.

LMs are trained by exclusively relying on distributional evidence to inform their word predictions. In our experiments, we find that while the aforementioned word prediction capacities show qualitatively similar patterns of associating concepts with human-produced property norms (Weir et al., 2020), they show weak agreement with the typicality effects that are robustly elicited in humans (G. L. Murphy, 2002; and references therein). This suggests that solely relying on text is insufficient for exhibiting quantitatively similar categorical knowledge to that in humans, and highlights the limitations of using word-prediction capacities from state-of-the-art pre-trained LMs as mechanisms to model semantic cognition. This is in line with work in knowledge acquisition through text, which suggests large textual corpora to lack real world grounding, in that these corpora represent language use but distort general knowledge about the world (Gordon & Van Durme, 2013; Shwartz & Choi, 2020). Even though text data contain encyclopedic knowledge, they miss out on the more perceptual or semi-perceptual features that can be learned through visual input, and that have been found to better align with human ratings of typicality, albeit on non-taxonomic categories (Lake et al., 2015). Another line of work supporting the lack of typicality signal in textual corpora is that of Bergey et al. (2020). These authors analyze parent-child interactions using models that are similar to—but less-sophisticated than—pre-trained LMs, and find them to negatively align with typicality ratings on adjective-noun compounds. The authors conclude from their findings that much of what children hear (corresponding to language use by the parent) is atypical, as opposed to typical information about noun concepts (specifically with respect to the adjectives that modify them). While our results also shed light on the difficulty of acquiring knowledge about typical members of categories, they do suggest the presence of some typicality effects, by contrast to the findings of Bergey et al. (2020)—raising the possibility that associations in text that impact typicality of adjective-noun compounds could be independent of, or even run in opposition to, those that impact taxonomic categories. At the same time, considering that we do see non-zero correspondence with human typicality ratings, our results also suggest that textual corpora are not fully devoid of associations that may align with empirical phenomena underlying typicality effects. Taking this as inspiration, future work on modeling of typicality through text will likely require models to correct for the distorted frequency of atypical items mentioned in text, and potentially also include features informed from a more grounded source of knowledge. One promising way of doing so could be to let LMs and their representations adapt to text data that explicitly specify semantic knowledge about everyday concepts and categories (Bhatia & Richie, 2021)—e.g., in the form of statements such as `a robin has wings` or `dolphins are mammals`. Explicitly encoding semantic features into LMs could possibly make them compliant with feature-based hypotheses of typicality (Rosch & Mervis, 1975; Rosch, Simpson, et al., 1976) and inductive reasoning (Sloman, 1993), and better facilitate research into other key facets of semantic cognition (Rogers & McClelland, 2004) in computational models that learn through text.

While the aforementioned proposals to improve LM correspondence with typicality and graded category structure involve additional training/fine-tuning or switching the class of models, recent work has extended the body of work presented in this chapter by proposing alternate measure derivations from the same types of models studied here. In particular, Renner et al. (2023) used one of the 19 models analyzed in this work (BERT-base) and combined measures that were derived from the representations of the model (Chronis & Erk, 2020) along with the relative information content between a category vs. the category member as tracked by the WordNet taxonomy (Miller, 1995), to estimate typicality values.[9] This resulted in improved correspondence to human typicality ratings by almost 34% (Spearman correlation of 0.396 using our method vs. 0.528 using their method, for BERT-base). While this approach objectively improves the correlation with human typicality judgments, they have little to no bearing on the main research questions posed in this dissertation. This is primarily because the main driver of improvement is the addition of the WordNet based similarity metric. That is, the correspondence obtained using *only* BERT's internal representations were either slightly worse, or only marginally better than the word-prediction based measure proposed in this work, depending on the ground-truth data used. Therefore, there

[9] ↑this work was published at the time of writing of this dissertation.

is little to be said about how distributional statistics could contribute to broader conceptual abilities beyond what our results already suggest. Nevertheless, one lesson learned from the improvements brought about by Renner et al. (2023) is that perhaps explicit integration of hierarchical conceptual organization within the LMs (Bouraoui et al., 2020) could result in more robust conceptual representations and better alignment with typicality effects. At the same time, this would also require testing on multiple different contexts where typicality modulates behavior—for instance, also testing on inductive generalization problems like we have, in this chapter.

# 4. A TEST FOR ROBUST PROPERTY KNOWLEDGE AND ITS INHERITANCE

This chapter adapts work from the following publication:

> Misra, K., Rayz, J., & Ettinger, A. (2023). COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

It has been reproduced here with slight alterations, and is in accordance with the Copyright and Licensing guidelines stated on the ACL Anthology website, where the publication resides.[1]

## 4.1 Introduction

The ability to learn, update and deploy ones knowledge about concepts (ROBIN, CHAIR) and their properties (*can fly*, *can be sat on*), observed during everyday experience is fundamental to human semantic cognition (G. L. Murphy, 2002; Rogers & McClelland, 2004; Rips et al., 2012). Knowledge of a concept's properties, combined with the ability to infer the `IsA` relation (Sloman, 1998; M. L. Murphy, 2003) leads to an important behavior known as *property inheritance* (Quillian, 1967; Smith & Estes, 1978; G. L. Murphy, 2002), where subordinates of a concept inherit its properties. For instance, one is likely to infer that an entity called *luna* can meow, has a tail, is a mammal, etc., even if *all* they know is that it is a cat. Importantly, property inheritance can also be viewed as an instance of *inductive generalization* (see §2.3) where the agent is provided a novel object (or concept) with a category label, and must infer what properties it possesses (Kemp & Jern, 2014).

As discussed in Ch. 1 and Ch. 2, the close connection between a word's meaning and its conceptual representation makes the aforementioned abilities crucial to conceptual understanding (G. L. Murphy, 2002; Lake & Murphy, 2021; as also established in ch. 2), making it

[1] ↑see `https://aclanthology.org/faq/copyright/`

critical for computational models of language processing to also exhibit behavior consistent with these capacities. Indeed, modern pre-trained language models (PLMs; Devlin et al., 2019; Brown et al., 2020; etc.) have made impressive empirical strides in eliciting general knowledge about real world concepts and entities (Petroni et al., 2019; Weir et al., 2020; *i.a.*), as well as in demonstrating isomorphism with real world abstractions like direction and color (Abdou et al., 2021; Patel & Pavlick, 2022), often times without even having been explicitly trained to do so. At the same time, their ability to robustly demonstrate such capacities has recently been called to question, owing to failures due to reporting bias (Gordon & Van Durme, 2013; Shwartz & Choi, 2020), lack of consistency (Elazar et al., 2021; Ravichander et al., 2020), and sensitivity to lexical cues (Kassner & Schütze, 2020; Misra et al., 2020; Pandia & Ettinger, 2021).

In this chapter, we cast further light on PLMs' ability to robustly demonstrate knowledge about concepts and their properties. To this end, we introduce Conceptual Minimal Pair Sentences (COMPS), a collection of English minimal pair sentences, where each pair attributes a property (*can fly*) to two noun concepts: one which actually possesses the property (ROBIN), and one which does not (PENGUIN). Following standard practice in the minimal pairs evaluation paradigm (Warstadt et al., 2020; etc.), we test whether PLMs prefer sentence stimuli expressing correct property knowledge over those expressing incorrect ones. COMPS can be decomposed into three subsets, each containing stimuli that progressively isolate deeper understanding of the task of attributing properties to concepts, by adding controls for more superficial heuristics. Our first subset—COMPS-BASE—measures the extent to which PLMs attribute properties to the right concepts, while varying the similarity of the positive (ROBIN) and the negative concepts (PENGUIN [high] vs. TABLE [low]). This controls for the possibility that models are relying on coarse-grained concept distinctions. For instance, in this setup a model should prefer (1a) over both versions of (1b).

(1) a. A **robin** can fly.

b. *A (**penguin/table**) can fly.

Next, drawing on the phenomenon of property inheritance, the COMPS-WUGS set introduces a novel concept, WUG, expressed as the subordinate of the positive and negative concepts from a subset of the COMPS-BASE set, and tests the extent to which PLMs successfully attribute it the given property when it is associated with the positive concept. This increases the complexity of the reasoning task, as well as the distance between the associated concept (ROBIN) and property (*can fly*). These manipulations help to control for memorization of the literal phrases being tested, forcing models to judge properties for a novel concept that inherits the property from a known concept. In this task, given that a model successfully prefers (1a) over (1b), it should also prefer (2a) over (2b):

(2) a. A wug is a **robin**. Therefore, a wug can fly.
    b. *A wug is a **penguin**. Therefore, a wug can fly.

The final subset—COMPS-WUGS-DIST, combines the aforementioned controls by using negative concepts as distracting content and inserting them into the COMPS-WUGS stimuli. Specifically, we transform the stimuli of COMPS-WUGS by creating two subordinates for every minimal pair; one for the positive concept (ROBIN, subordinate: WUG) and the other for the negative concept (PENGUIN, subordinate: DAX), which acts as a distractor. This way, we control for the possibility that models may be relying on simple word associations between content words—of which there are only two in the prior tests—by introducing additional, irrelevant but contentful words into the context. Here, we consider models to be correct if they prefer (3a) over (3b), given that they prefer (1a) over (1b):

(3) a. A **wug** is a robin. A **dax** is a penguin. Therefore, a **wug** can fly.
    b. *A **wug** is a robin. A **dax** is a penguin. Therefore, a **dax** can fly.

Together, the three sets of stimuli tease apart more superficial predictive behaviors, such as contextual word associations, from more robust reasoning behaviors based on understanding of concept properties. While we can expect superficial predictive strategies to be brittle in the face of shallow perturbations and irrelevant distractions, robust property knowledge and reasoning behaviors should not.

## 4.2 Motivation for using minimal pairs

Why should we use minimal pairs to test PLMs for property knowledge and/or its inheritance? I present an argument by contrasting and relating it to alternative test formats and paradigms. In particular, prior work in exploring property knowledge in PLMs has adopted two different paradigms: one which uses probing classifiers to test if the applicability of a property can be decoded from the representations of LMs (Forbes et al., 2019; Da & Kasai, 2019; Derby et al., 2021); and the other which uses cloze-testing, in which LMs are tasked to fill in the blank in prompts that describe specific properties/factual knowledge about the world (Petroni et al., 2019; Weir et al., 2020). As argued in §2.2 both these approaches—though insightful—have key limitations for evaluating property knowledge, and that minimal pair testing overcomes these limitations to a beneficial extent.

Briefly, the probing setup does not allow the testing of property knowledge in a precise manner. Specifically, several properties are often perfectly correlated in datasets such as the one we use here (see §4.3.1). For example, the property of being an animal and being able to breathe and grow, etc., are all perfectly correlated with one another. Even if the models true knowledge of these properties is highly variable, probing its representations for them could yield the exact same result, leading to conclusions that overestimate the model's capacity for some properties, while underestimating for others. Furthermore, many properties are often idiosyncratic (*has striped patterns on its body*) and are therefore sparsely distributed in the world, making them non-trivial to be learned by a simple probing classifier. Similarly, standard cloze-testing of PLMs (Petroni et al., 2019; Weir et al., 2020; Jiang et al., 2021) also faces multiple limitations. First, it does not allow for testing of multi-word expressions, as by definition, it involves prediction of a single word/token. Second, it does not yield faithful conclusions about one-to-many or many-to-many relations: e.g. the cloze prompts "Ravens can ___." and "___ can fly." do not have a single correct answer. This makes our conclusions about models knowledge contingent on choice of one correct completion over the other.

The minimal pair evaluation paradigm combines useful properties of both the aforementioned paradigms and by doing so, presents solutions to their limitations. In particular, it uses the concept of 'negative samples' from the probing paradigm, while explicitly representing property knowledge by generalizing the cloze-testing method to multi-word expressions—i.e., by focusing on entire sentences. This allows for a straightforward way to assess correctness: the choice between multiple correct completions is transformed into one between correct and incorrect, at the cost of having several different instances (pairs) for testing knowledge of the same property. Additionally, the minimal pair paradigm allows us also to shed light on how the nature of negative samples affects model behavior, which has been missing in approaches using probing and cloze-testing. Finally, the usage of minimal pairs is a well-established practice in the literature, having been widely used in works that analyze syntactic knowledge of PLMs (Marvin & Linzen, 2018; Futrell et al., 2019; Warstadt et al., 2020), as well as the potential social biases they might encode (Nangia et al., 2020; e.g.). COMPS complements this body of work by introducing minimal-pair testing to the study of conceptual knowledge in PLMs.

## 4.3 Dataset Design and Generation

### 4.3.1 Ground-truth Property Knowledge data

For our ground-truth property knowledge resource, we use a subset of the CSLB property norms collected by Devereux et al. (2014), which was further extended by Misra et al. (2022). The original dataset was constructed by asking 123 human participants to generate properties for 638 everyday concepts. Contemporary work has used this dataset by taking as positive instances all concepts for which a property was generated, while taking the rest as negative instances (Lucy & Gauthier, 2017; Da & Kasai, 2019; etc.) for each property. While this dataset has been popularly used in related literature, Misra et al. (2022; see also ch. 5) recently discovered striking gaps in coverage among the properties included in the dataset.[2] For example, the property *can breathe* was only generated for 6 out of 152 animal concepts, despite being applicable for all of them—as a result, contemporary work can be expected to have wrongfully penalized models that attributed this property to animals that could indeed breathe, and similarly for other properties. To remedy this issue, Misra et al. (2022) manually extended CSLB's coverage for 521 concepts and 3,643 properties. We refer

[2] ↑See also Sommerauer and Fokkens (2018) and Sommerauer (2022), who also discuss this limitation.

to this extended CSLB dataset as XCSLB, and we use it as our source for ground-truth property knowledge.

### 4.3.2 Choice of Negative Samples

We rely on a diverse set of knowledge representation sources to construct negative samples for COMPS. Each source has a unique representational structure which gives rise to different pairwise similarity metrics, on the basis of which we pick out negative samples for each property:

#### Taxonomy

To sample from taxonomies, we consider a hierarchical organization of our concepts, by taking a subset of WordNet (Miller, 1995) consisting of our 521 concepts. We use the `wup` similarity (Wu & Palmer, 1994) as our choice of taxonomic similarity.

#### Property Norms

We use the XCSLB dataset and organize it as a matrix whose rows indicate concepts and columns indicate properties that are either present (indicated as 1) or absent (indicated as 0) for each concept. As our similarity measure, we consider the jaccard similarity between the row vectors of concepts. This reflects the overlap in properties between concepts, and is prevalent in studies utilizing conceptual similarity in cognitive science (Tversky, 1977; Sloman, 1993; etc.).

#### Co-occurrence

We use the co-occurrence between concept words as an unstructured knowledge representation. For quantifying similarity, we use the cosine similarity of the GloVe vectors (Pennington et al., 2014) of our concept words.

### Sampling Strategy

Each property ($p_i$) in our dataset splits the set of concepts into two: a set of concepts that possess the property ($Q_{p_i}$), and a set of concepts that do not ($\neg Q_{p_i}$). We sample $\min(|Q_{p_i}|, 10)$—i.e., at most 10—concepts from $Q_{p_i}$ and take them to be our positive set. Then for each concept in the positive set, we sample from $\neg Q_{p_i}$ the concept that is most similar (depending on the source) to the positive concept and take it as a negative concept for the property. We additionally include a negative concept that is randomly sampled from $\neg Q_{p_i}$, leaving out the concepts sampled on the basis of the three previously described knowledge sources. Examples of the four types of negative samples for the concept ZEBRA and the property *has striped patterns* are shown in Table 4.1.

**Table 4.1.** Negatively sampled concepts selected on the basis of various knowledge representational mechanisms, where the property is *has striped patterns*, and the positive concept is ZEBRA.

| **Knowledge Rep.** | **Negative Concept** | **Similarity** |
|---|---|---|
| Taxonomy | HORSE | 0.88 |
| Property Norms | DEER | 0.63 |
| Co-occurrence | GIRAFFE | 0.75 |
| Random | BAT | - |

### 4.3.3 Minimal Pair Construction

Following our negative sample generation process, we end up with total of 49,324 pairs of positive and negative concepts that span across 3,643 properties (14 pairs per property, on average). Every property is associated with a property phrase—a verb phrase which expresses the property in English, as provided in XCSLB. Using these materials, we construct our three datasets of minimal pair sentence stimuli, examples of which are shown in Figure 4.1.

### COMPS-BASE

The COMPS-BASE dataset contains minimal pair sentences that follow the template:

"`[DET] [CONCEPT] [property-phrase]`."

where `[DET]` is an optional determiner, and `[CONCEPT]` is the noun concept. Applying this template to our generated pairs results in 49,324 instances. See Figure 4.1a for an example.

### COMPS-WUGS

We test property inheritance in PLMs using only the animal kingdom subset of COMPS-BASE (152 concepts, 944 properties, and 13,888 pairs), keeping the same negative samples. We convert the original minimal pair sentences in COMPS-BASE, in which the positive concept is an animal, into pairs of two-sentence stimuli by first introducing a new concept (WUG) to be the subordinate of the concepts in the original minimal pair. We then express its property inheritance in a separate sentence. Our two sentence stimuli follow the template:

"A wug is a `[CONCEPT]`. Therefore, a wug `[property-phrase]`."

Although we use *wug* as our running example for the subordinate concept, we use four different nonsense words {*wug*, *dax*, *blicket*, *fep*} equal numbers of times, to avoid making spurious conclusions based on a single nonsense word.[3] Introducing an intervening novel concept allows us to robustly control for simple word-level associations between concepts and properties that models might have picked up during training. Figure 4.1a shows an example.

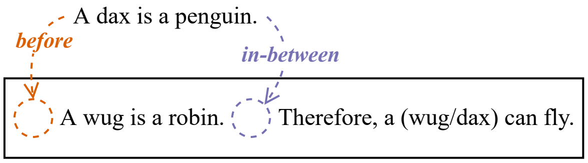
### COMPS-WUGS-DIST

To add distracting information, we follow Pandia and Ettinger (2021) and convert the COMPS-WUGS stimuli by associating a different subordinate concept (DAX) with the negative concept (`[NEG-CONCEPT]`), and inserting it **before** or **in-between** the sentence containing

[3] ↑As we describe in §4.5, we also tried a different set of nonce words, to address concerns about possible impacts of using nonce words from existing literature (e.g., *wug*).

**Property:** *can fly*
**Positive:** ROBIN
**Negative:** PENGUIN
**Subordinate:** WUG
**COMPS-BASE:** A (**robin/penguin**) can fly.
**COMPS-WUGS:** A wug is a (**robin/penguin**). Therefore, a wug can fly.

(a) COMPS-BASE and COMPS-WUGS instances.



(b) Distraction scheme for stimuli in COMPS-WUGS-DIST, where the distractor is inserted either **before** or **in between** each COMPS-WUGS stimulus.

**Figure 4.1.** Examples of materials used in our experiments. In this example, ROBIN is the positive concept.

the positive concept and its subordinate, separately. This results in two subsets (**before** and **in-between**) of three-sentence minimal pair stimuli, which differ in the subordinate to which the property is attributed. We use the following template to create our stimuli:

"A **wug** is a `[CONCEPT]`. A **dax** is a `[NEG-CONCEPT]`. Therefore, a (**wug**/**dax**) `[property-phrase]`."

That is, we have stimuli that resemble COMPS-WUGS but instead deal with a pair of competing subordinate concepts in context.[4] See Figure 4.1b for an example.

## 4.4 Methodology

### 4.4.1 Models Investigated

We investigate property knowledge and property inheritance capacities of 22 different PLMs, belonging to six different families. We evaluate four widely used masked language modeling (MLM) families: (1) ALBERT (Lan et al., 2020), (2) BERT (Devlin et al., 2019), (3) ELECTRA (Clark et al., 2020), and (4) RoBERTa (Liu et al., 2019); as well as two autoregressive language modeling families: (1) GPT2 (Radford et al., 2019), and (2) the GPT-Neo (Black et al., 2021) and GPT-J models (B. Wang & Komatsuzaki, 2021) from EleutherAI. We also use distilled versions of BERT-base, RoBERTa-base, and GPT2, trained using the method described by Sanh et al. (2019). Table B.1 shows each models parameters, vocabulary

[4] ↑We again choose from our list of four nonsense words (*wug*, *dax*, *blicket*, and *fep*), which amounts to 12 unique ordered pairs, after accounting for counterbalancing.

**Table 4.2.** An example of matched stimuli across different COMPS subsets, as well as conditional log-probabilities elicited by GPT-J. Here, the property of interest is *has hooves*, the positive concept is HORSE, and the negative concept is DOG. The negative concept in this case was sampled using the co-occurrence knowledge representation method (see §4.3.2). Emboldened words indicate items that are different in the minimal pair. Refer to §4.4.2 for discussion on how 'Score' is computed.

| COMPS subset | Stimulus | Score |
|---|---|---|
| BASE | A **horse** has hooves. | -3.829 |
| | A **dog** has hooves. | -4.963 |
| WUGS | A fep is a **horse**. Therefore, a fep has hooves. | -2.153 |
| | A fep is a **dog**. Therefore, a fep has hooves. | -3.392 |
| WUGS-DIST (**before**) | **A wug is a dog.** A fep is a horse. Therefore, a **fep** has hooves. | -2.919 |
| | **A wug is a dog.** A fep is a horse. Therefore, a **wug** has hooves. | -2.895 |
| WUGS-DIST (**in-between**) | A fep is a horse. **A wug is a dog.** Therefore, a **fep** has hooves. | -3.616 |
| | A fep is a horse. **A wug is a dog.** Therefore, a **wug** has hooves. | -3.092 |

size, and training corpora. All models were accessed using `minicons` (Misra, 2022),[5] a python library that serves as a wrapper around Huggingface's `transformers` (Wolf et al., 2020), and provides a unified mechanism for eliciting log-probabilities in batch-wise manner for any autoregressive or masked LM that is accessible through the huggingface hub, or is trained using the transformers library. Outside of these models, we also investigated 7 different GPT-3/3.5 models, in separate experiments with a smaller subset of the COMPS stimuli (see §4.5.4).

### 4.4.2 Measuring Performance

To evaluate models on COMPS, we compare their log-probabilities for the property phrase—conditioned on contexts (to the left) containing the positive and negative noun concepts. That is, we hold the property phrase constant, and compare across minimally differing

[5] ↑`https://github.com/kanishkamisra/minicons`

conditions to evaluate the probability with which a property is attributed to each concept. For example, we score stimuli in COMPS-BASE, e.g., "*A dog can bark.*" as:

$$\log p(\text{can bark.} \mid \text{A dog}),$$

its corresponding stimulus in COMPS-WUGS, "*A wug is a dog. Therefore, a wug can bark.*" as:

$$\log p(\text{can bark.} \mid \text{A wug is a dog. Therefore, a wug}),$$

and similarly—assuming CAT as the negative concept—the corresponding stimuli in our COMPS-WUGS-DIST subset, "*A wug is a dog. A dax is a cat. Therefore, a wug can bark.*" as:[6]

$$\log p(\text{can bark.} \mid \text{A wug is a dog. A dax is a cat. Therefore, a wug}).$$

This approach to eliciting conditional LM judgments is equivalent to the "scoring by premise" method (Holtzman et al., 2021), which has been shown to result in stable comparisons across items. Additionally, this also takes into account the potential noise due to frequency effects or tokenization differences (Misra et al., 2021). Estimating these conditional log-probabilities using auto-regressive PLMs can be directly computed in a left-to-right manner. For MLMs, we use their conditional pseudo-loglikelihoods (Salazar et al., 2020) as a proxy for conditional log-probabilities.

Based on this simple method of eliciting relative acceptability measures from PLMs, we evaluate a model's accuracy on all COMPS stimuli as the percentage of times its log-probability for a property is greater when conditioned on the context that attributes the property to the positive—as opposed to the negative—concept. Since all cases are forced-choice tasks between two instances, chance performance is set to 50%. Table 4.2 shows examples of all COMPS stimuli and GPT-J's conditional log-probabilities for them.

[6] ↑Here we show an example where the distractor is added **in-between** the context specifying the positive concept, and the queried property knowledge.

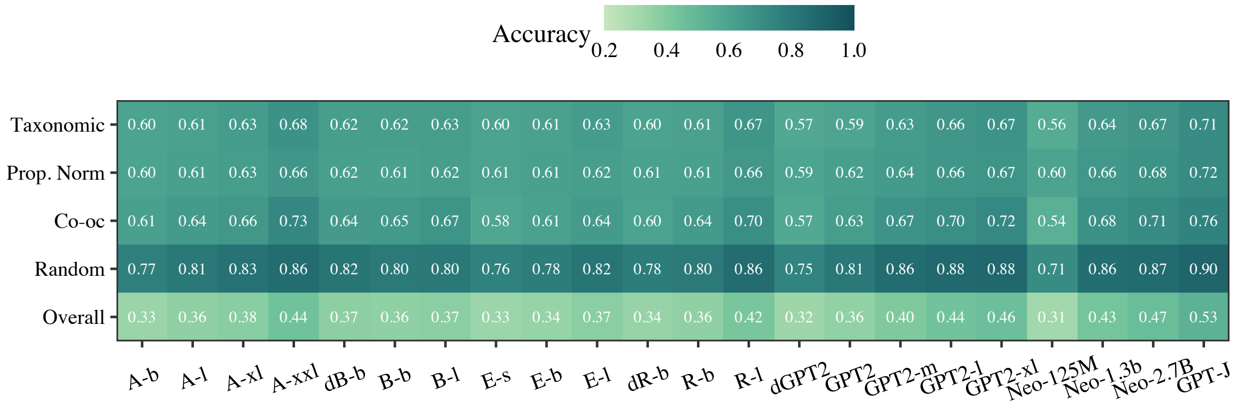


**Figure 4.2.** Accuracies of PLMs on COMPS-BASE under various negative sampling schemes. Chance performance for all rows is 50%, except for 'Overall,' where it is 6.25%. Refer to Table B.1 for unabbreviated model names.

## 4.5 Experiments and Analyses

### 4.5.1 Base property knowledge of PLMs and their sensitivity to similarity effects

We begin by evaluating the 22 PLMs on COMPS-BASE. Here we focus on the extent to which models robustly associate properties to the correct concepts across stimuli with varying kinds of similarity between the positive and negative concepts. We report accuracies of the 22 PLMs on COMPS-BASE across the four different negative sampling schemes that we specified in §4.3.2. We additionally report a more stringent accuracy measure that we refer to as 'Overall accuracy,' which is calculated for every property and its positive concept, as the percentage of times a model correctly attributes the property to the positive concept in **all four types of negative sampling schemes**. Chance performance for only the 'Overall' case is then 6.25% ($0.5^4 \times 100$). Figure 4.2 shows these results.

From Figure 4.2, we see that models strongly distinguish between positive and negative concepts in cases where they are dramatically different—i.e., where negative concepts were sampled randomly (e.g., BEAR [positive] vs BOTTLE [negative] for the property *can breathe*). However performance drops substantially when there are subtler differences between the two concepts—e.g, the concepts WALRUS (positive) and SHARK (negative) for the property *is a mammal*. For instance, the best performing model in any similarity-based negative sampling scheme (GPT-J, 76%, ‘Co-oc’) only slightly outperforms the worst model in the random negative sampling scheme (Neo-125M, 71%). The performance of PLMs is not substantially different across the three similarity-based negative sampling schemes, suggesting that the dynamics of model sensitivity in attributing properties to concepts are largely harmonized across various types of similarities. As a result of models’ insensitivity in presence of similar negative concepts, the overall accuracies are very modest in value, with the overall accuracy of the best performing model (GPT-J) being only 53%. This overall performance is, however, significantly above chance (6.25%).

### 4.5.2 How does performance on COMPS-BASE vary by property type?

Devereux et al. (2014) have categorized the properties that we use in our experiments to lie in 5 different categories: (1) **Taxonomic**, e.g., *is a mammal*, *is a vehicle*, etc.; (2) **Functional**, e.g., *can keep the body warm*, *is used to hit nails*, etc.; (3) **Encyclopedic**, e.g., *uses electricity*, *is warm blooded*, etc.; (4) **Visual Perceptual**, e.g., *has webbed feet*, *has thick fur*, etc.; and (5) **Other Perceptual**, e.g., *makes grunting sounds* and *is sharp*, etc. We report results of the 22 PLMs on the COMPS-BASE stimuli across the five different property types, in Figure 4.3.

From Figure 4.3, we observe that PLMs are substantially stronger in eliciting taxonomic properties of concepts as compared to other types, with highest overall accuracy being 71%, as compared to 48% on encyclopedic properties, 50% on visual perceptual properties, 57% on functional properties, and 43% on non-visual perceptual properties. Recall that chance accuracy for the ‘Overall’ scenario is just 6.25%, so these scores are fairly high. This corroborates evidence from previous work in analyzing property knowledge of distributional semantic models as well as LM representations to lack perceptual knowledge (Lucy & Gauthier, 2017; Da & Kasai, 2019; Rubinstein et al., 2015; Weir et al., 2020), likely due to reporting bias (Gordon & Van Durme, 2013; Shwartz & Choi, 2020). However, different to most of these works, the gap between performance on perceptual properties and non-perceptual properties is small. We conjecture that this could be primarily due to the extension of the CSLB by Misra et al. (2022), which lead to an increase in coverage of property knowledge for several properties. For instance, the property *has teeth* was mentioned only for 45 out of 67 potential concepts, having been left out for concepts such as CALF,[7] BUFFALO, KANGAROO, etc. So it could be the case that previous research has underestimated the extent to which property knowledge is encoded by PLMs and other distributional semantic models of language.

### 4.5.3 Property Inheritance in PLMs

Having established the base property knowledge of PLMs, we now investigate the extent to which they can show behavior that is consistent with reasoning required to handle property inheritance. We first investigate their performance on COMPS-WUGS, created using the subset of COMPS-BASE containing only animal concepts (see §4.3.3 for stimulus construction). Table 4.3 shows average accuracies obtained by PLMs on our property inheritance stimuli, and compares them to average accuracies on COMPS-BASE—aggregating across all negative sampling schemes. Recall that the stimuli in COMPS-WUGS present a more challenging property attribution task than in COMPS-BASE, by not only controlling for coarse-grained similarity effects, but also introducing an intervening novel concept that is expected to inherit the properties of the positive concept. By measuring attribution of properties more indirectly, these stimuli increase the complexity of the reasoning and control for memorization of the literal phrase initially tested with COMPS-BASE.

Table 4.3 shows the average accuracy of the PLMs on each subset of COMPS. Despite the increase in complexity, we see that PLMs actually show slightly stronger performance on COMPS-WUGS (68.9%) than on COMPS-BASE (67.1%). This means that there are instances in which models prefer the property in the positive context over the negative context (4a > 4b), but show the opposite behavior in COMPS-BASE (4d > 4c).

(4) a. A wug is a **robin**. Therefore, a wug can fly.
b. A wug is a **penguin**. Therefore, a wug can fly.
c. A **robin** can fly.
d. A **penguin** can fly.

[7] ↑the young one of a cow, and not the muscles in the vertebrate body

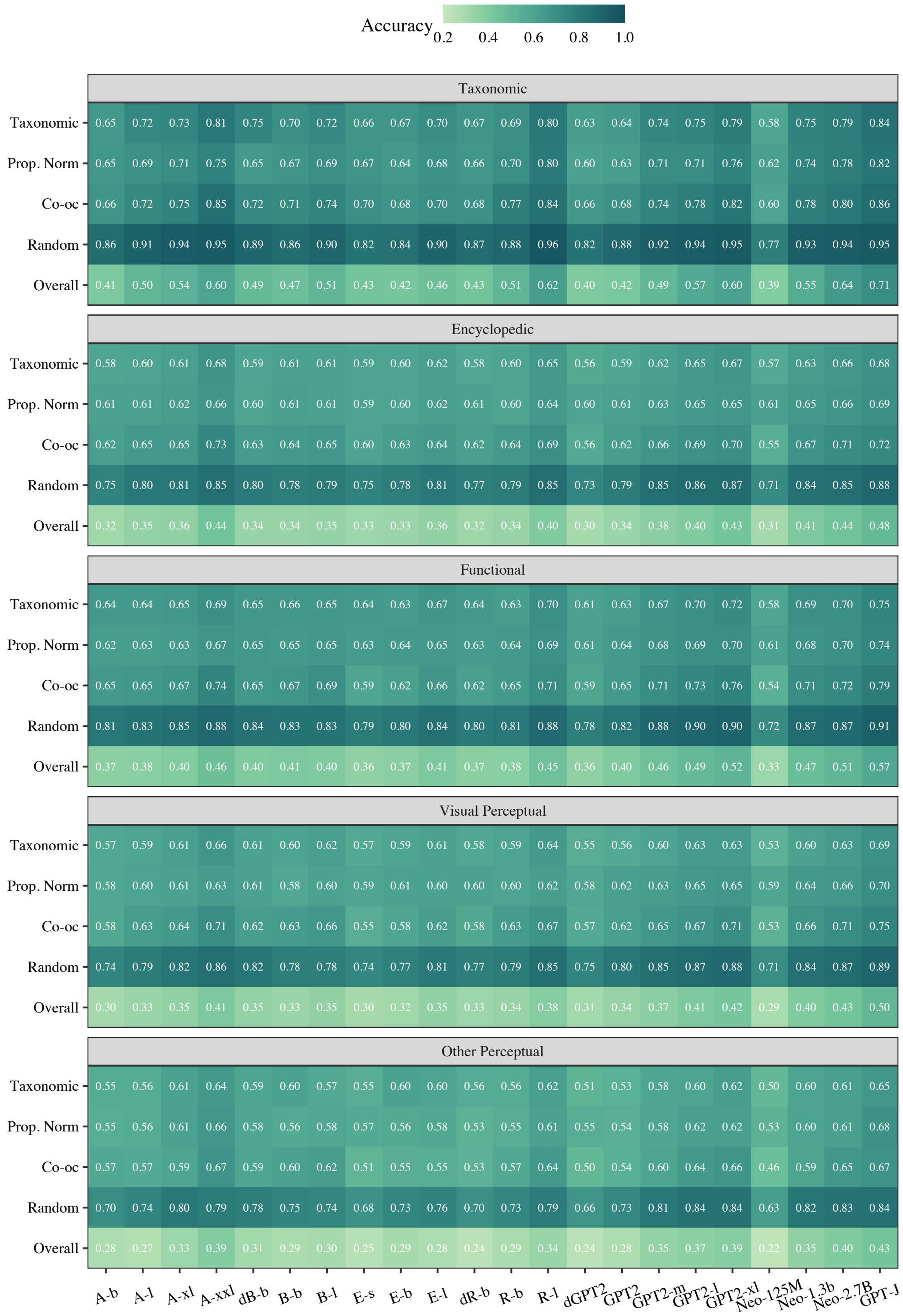


**Figure 4.3.** COMPS-BASE performance across five property types annotated in CSLB (Devereux et al., 2014).

**Table 4.3.** Average accuracy (and standard error of the mean) of PLMs ($N = 22$) on each of our COMPS subsets. Chance performance is 50% throughout.

| **COMPS subset** | **Size** | **Acc.** |
|---|---|---|
| BASE | 49.3K | $68.4_{1.7}$ |
| BASE (*animal kingdom only*) | 13.8K | $67.1_{2.0}$ |
| WUGS | 13.8K | $68.9_{2.3}$ |
| WUGS-DIST (**before**) | 13.8K | $59.2_{3.9}$ |
| WUGS-DIST (**in-between**) | 13.8K | $47.2_{4.5}$ |

This pattern of performance could lead to spurious conclusions that models are successfully executing property inheritance, when in fact they show a lack of the pre-requisite property knowledge based on their failure on COMPS-BASE. We will discuss these inconsistencies in more detail below. Overall, however, the relatively strong performance on COMPS-WUGS suggests that models are largely unaffected when we control for simple memorization of tested phrases—e.g., *robin can fly*—by linking known concepts to properties through an intervening subordinate concept (*wug*). This suggests that models are not relying on simple memorization, but does not control for the possibility that simple association between content words (*robin* and *fly*) could be responsible for seemingly desirable behavior—for this we turn to COMPS-WUGS-DIST. The COMPS-WUGS-DIST test assesses whether models retain strong property attribution performance when content words in the context are not all relevant for the property prediction. The stimuli thus include irrelevant distractor concepts and their subordinates—which, in a robust model, should not affect attribution of the property to the correct concept (see §4.3.3 for stimulus construction).

From Table 4.3, the average accuracies of PLMs on both subsets of COMPS-WUGS-DIST (**before** and **in-between**) indicate that overall, models now show clear degradation in property inheritance performance as a result of the distracting information. Specifically, the PLMs' performance drops by 9.7 points on instances when the distracting information is added **before** the relevant context and queried property, and by 21.7 points on instances where it is added **in-between** the two, relative to the undistracted property inheritance stimuli (COMPS-WUGS). Notably, the latter drop in performance brings models level with

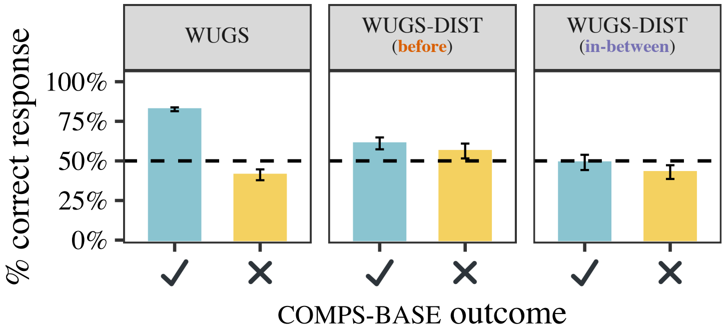


**Figure 4.4.** Distribution of model performance on COMPS-WUGS and COMPS-WUGS-DIST (both subsets) across possible outcomes (correct = ✓, incorrect = ✗) of the models on corresponding minimal pairs in COMPS-BASE. Error bars indicate 95% CI, while dashed line indicates chance performance (50%).

chance accuracy (we fail to reject the null hypothesis that avg. accuracy of models is 50%; $p = .62$, Wilcoxon signed rank exact text), highlighting a pronounced lack of robustness in PLMs' capacity to attribute properties to the correct concepts in their input context.

### Accounting for spurious performance

The COMPS-WUGS results above raise the concern that models are often showing spurious performance: accurately demonstrating property inheritance behavior without actually possessing the right property knowledge. To shed more light on this potential issue, we plot the distribution of model accuracies on our property inheritance stimuli (COMPS-WUGS and COMPS-WUGS-DIST) divided based on their outcomes on the corresponding stimuli in COMPS-BASE. Figure 4.4 shows these distributions. In COMPS-WUGS and both subsets of COMPS-WUGS-DIST, models show this spurious correct behavior on 41.3%, 55.6%, and 42.8% of instances in which they produce incorrect judgments on the corresponding COMPS-BASE stimuli (yellow bars in Figure 4.4). This non-trivial proportion of cases with spurious performance further reinforces the idea that PLMs' successful predictions on these tests are likely relying on heuristics rather than robust inferences about property knowledge. We can remove the effects of these spurious instances by filtering to items in which models give the

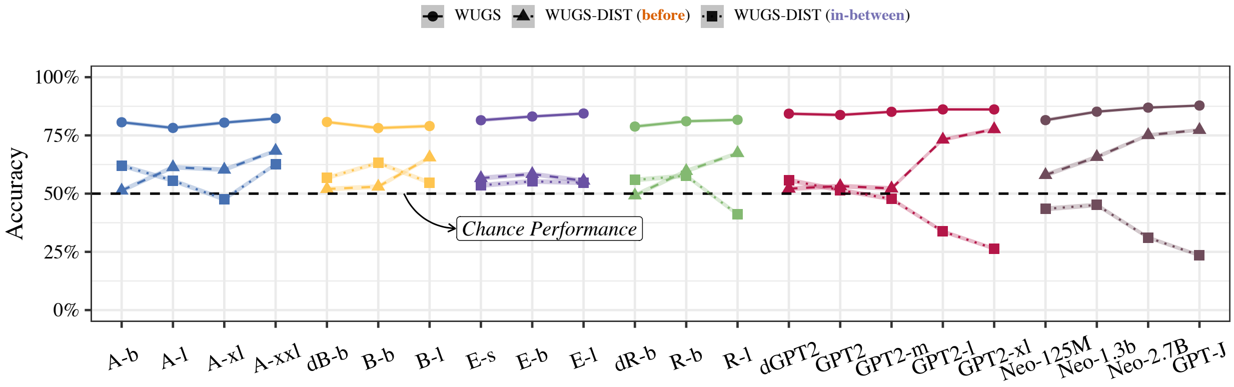


**Figure 4.5.** Accuracies of individual models (grouped by family, in increasing order based on number of parameters) on COMPS-WUGS and COMPS-WUGS-DIST. Black dashed line indicates chance performance (50%). Refer to Table B.1 for unabbreviated model names. Error bands indicate 95% Bootstrap CIs.

correct answer on COMPS-BASE (blue bars in Figure 4.4)—though we see that the overall conclusions remain the same after this filtering.

### On the pronounced effect of proximity in autoregressive PLMs

Our previous discussion summarized the aggregate property inheritance behavior of the 22 PLMs we considered—we now zoom in for a model-wise analysis. Figure 4.5 shows models' relative accuracies on COMPS-WUGS and COMPS-WUGS-DIST, filtering to items with correct COMPS-BASE performance, as in the blue bars of Figure 4.4. Consistent with our overall findings, we observe distracting content to substantially degrade model performance across the board.[8] A particularly noteworthy pattern is that the degradation in autoregressive PLM families—GPT2 and EleutherAI—shows a stark sensitivity to *proximity effects*. While these classes of model seem to suffer less when distracting content is added **before** the context containing the positive concept (thus placing the distraction farther from the queried property), they show substantially worse performance when the opposite is the case (i.e., when distraction is added **in-between**, and is therefore closer to the queried property). This

[8] ↑See also Pandia and Ettinger (2021) for a similar degradation of performance on cloze-tasks involving factual retrieval.

degradation due to proximity of the distracting content becomes **catastrophically worse as models grow larger in the number of pre-trained parameters**—in fact bringing their performance down to as much as 26.2 points **below chance** (in GPT-J, which has 6B parameters). While MLMs also show similar levels of degraded performance in presence of distraction, they do not seem to show any systematic sensitivity to proximity effects, likely due to their bidirectional nature.

### 4.5.4 Testing GPT-3/3.5

Recent work in scaling PLMs to hundred billion parameters has led to models such as GPT-3 (Brown et al., 2020), which are significantly larger than the largest model tested in the results discussed above (i.e., GPT-J, with 6B parameters). Testing them on the entire set of COMPS stimuli (49K + 3 × 13.8K pairs of sentences) is prohibitively expensive since they are only accessible through paid APIs. Nonetheless, we sampled a small set of COMPS stimuli—which we term as miniCOMPS—in order to get a glimpse of how well substantially larger PLMs elicit property knowledge and demonstrate reasoning behavior compatible with property inheritance. Specifically, we created miniCOMPS by sampling 1200 minimal pairs from each of our original COMPS subsets (matched in terms of real world concepts and properties across the subsets), such that all pairs of nonce words in the resulting miniCOMPS-WUGS-DIST end up being sampled equal number of times (100 times each).

#### Models

As test subjects, we chose four GPT-3 models (Brown et al., 2020): `ada`, `babbage`, `curie`, `davinci`, with the last one being the largest (at 175B parameters), and an additional fifth `davinci`-based model called `text-davinci-001`, which fine-tunes `davinci` on human-written demonstrations. We also test the recently proposed GPT-3.5 models, `text-davinci-002` and `text-davinci-003`, which improve over `davinci` by additionally fine-tuning it on code *and* human-written demonstrations (Ouyang et al., 2022).[9] All these models are autoregressive

[9] ↑These models are also known as InstructGPT, as discussed in `https://platform.openai.com/docs/model-index-for-researchers`.

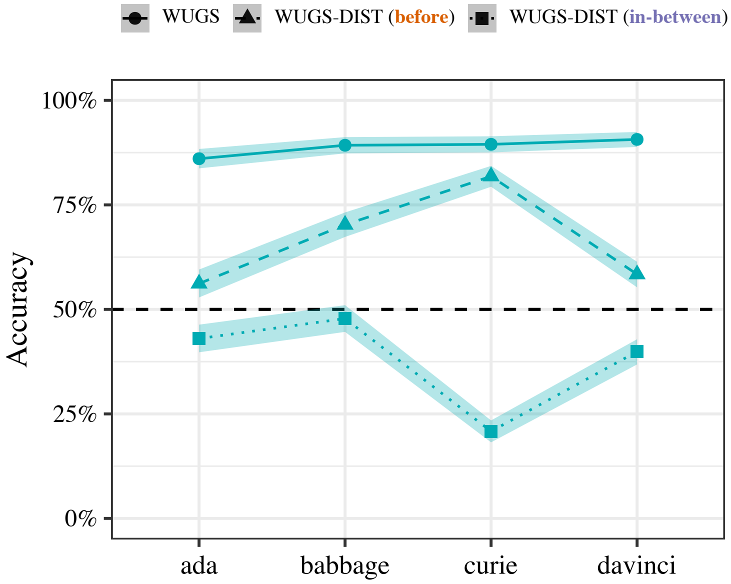


**Figure 4.6.** Accuracies of GPT-3 models (arranged in increasing order of the number of trained parameters) on miniCOMPS-WUGS and miniCOMPS-WUGS-DIST. Black dashed line indicates chance performance (50%). Error bands indicate 95% Bootstrap CIs.

in nature, so we use the same scoring and evaluation method as described in §4.4.2. Since the four original GPT-3 models (`ada`, `babbage`, `curie`, `davinci`) are trained using the same LM objective on the same corpora, we analyze them separately from `text-davinci-001`, `text-davinci-002`, and `text-davinci-003`, which we only compare to `davinci`. We do this to remain consistent with the way we displayed results in §4.5—ordering models based on their number of trained parameters—and also because models in the `text-davinci-XXX` series use the same underlying `davinci` model augmented with additional training mechanisms (e.g., reinforcement learning and fine-tuning on human-feedback) and data (e.g., code) instead of increasing its size, to our knowledge. The total cost of the experiments was $14.74 (as of February 3, 2023).

### Results

Figure 4.6 shows the performance of the four GPT-3 models on miniCOMPS-WUGS and miniCOMPS-WUGS-DIST, while Figure 4.7 compares GPT-3 `davinci` to its code and human-

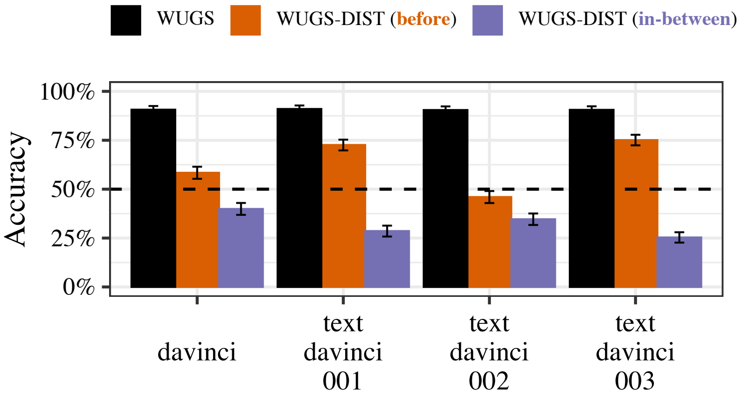


**Figure 4.7.** Accuracies of `davinci` models (GPT-3 and GPT-3.5) on miniCOMPS-WUGS and miniCOMPS-WUGS-DIST. Black dashed line indicates chance performance (50%). Error bars indicate 95% Bootstrap CIs. `davinci` and `text-davinci-001` are GPT-3 (Brown et al., 2020) models, while `text-davinci-002` and `text-davinci-003` are GPT-3.5 models.

feedback adapted counterparts. From Figure 4.6, we see robustness issues to persist even for GPT-3 models, similar to our main results. Models perform remarkably well in the absence of distraction (i.e., on miniCOMPS-WUGS), but struggle in its presence, especially when it is closer to the queried property. In particular, performance on miniCOMPS-WUGS-DIST (**before**) increases with an increase in parameters until the largest model (`davinci`), where the performance drops closer to chance. On miniCOMPS-WUGS-DIST (**in-between**), all models perform catastrophically worse than chance. This noteworthy pattern of proximity-based degradation in performance mimics the results shown in Figure 4.5, though we do not see a systematic decline in performance with an increase in parameters as observed in the GPT2 and EleutherAI models—with the 175B parameter model (`davinci`) demonstrating an increase in performance over the relatively smaller `curie` model.

While the above results demonstrate that simply scaling autoregressive PLMs is unlikely to overcome the lack of robustness against distracting content, we now test whether augmenting these large PLMs by additionally training on code (GPT-3.5 models) and aligning them with human-provided demonstrations (`text-davinci-001` and both GPT-3.5 models) could lead to any improvements. For instance, training on code could provide training signals to PLMs that encourage entity tracking (N. Kim & Schuster, 2023), which could potentially enable them, in our case, to resolve which subordinate concept (e.g., *wug* vs. *dax*) the target property is more likely to be associated with. Similarly, aligning with human-written demonstrations could potentially improve their truthfulness, which in our case, could lead to them to prefer correct property assignments. However, from Figure 4.7, we see no noteworthy improvements demonstrated by these augmented models. All augmented models achieved similar accuracies on COMPS-WUGS as the `davinci` model (within 90.5% and 91%), suggesting that their augmentations preserved the general associations between the lexical items that denote everyday concepts and properties. On stimuli containing distraction (i.e., both subsets of COMPS-WUGS-DIST), either the models performed systematically worse as compared to `davinci` (with `text-davinci-002` showing below-chance performance on both subsets), or they showed mixed results, where an improvement on COMPS-WUGS-DIST (**before**) was accompanied by a decline on COMPS-WUGS-DIST (**in-between**).

Together, these results suggest that neither an increase in scale nor additional training methods such as alignment with human instructions/feedback or training on code prevents models from being distracted in associating properties to novel subordinate concepts introduced in the input context. In fact, the catastrophic effects of proximity-based distraction persists even for the most recent state of the art GPT-3/3.5 models.

### 4.5.5 Choice of nonce words

Nonce words constitute an important design decision for our stimuli—we followed precedents in language acquisition research (Berko, 1958; Gopnik & Sobel, 2000; *i.a.*) and used previously existing nonce words (such as *wug* and *blicket*) to represent novel concepts in context. While these are expected to be novel for humans, they may appear in pre-training corpora on which PLMs are usually trained.[10] This raises a potential concern that PLMs could already be biased toward certain properties for these words (e.g., *wug* is commonly depicted as a bird), and may struggle to associate them with different properties.[11] To explore this empirically, we conducted experiments with alternative nonce words (generated synthetically,

[10] ↑e.g., *wug* appears in wikipedia: `https://en.wikipedia.org/wiki/Jean_Berko_Gleason` (accessed on Jan 23)

[11] ↑We thank Reviewers 1 and 3 of the paper (Misra et al., 2023), Najoung Kim, and Kyle Mahowald for raising this concern.

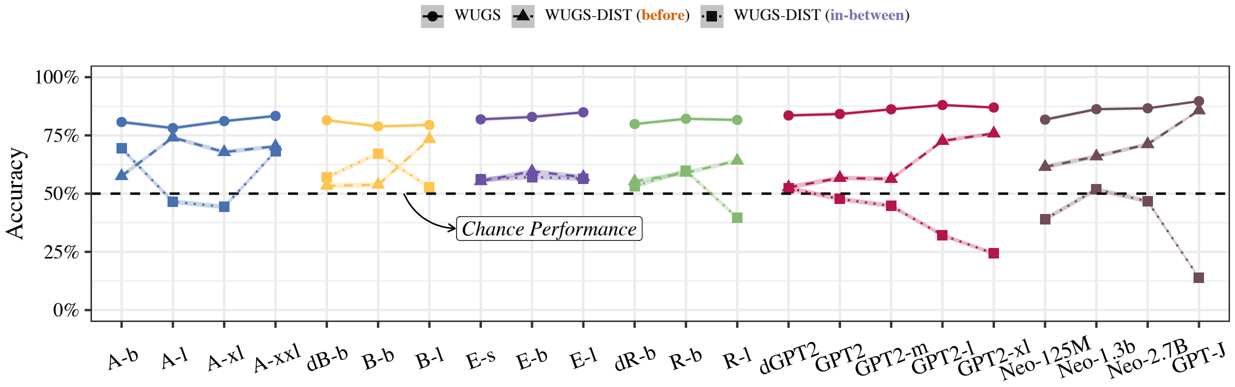


**Figure 4.8.** Accuracies of individual models (grouped by family, in increasing order based on number of parameters) on COMPS-WUGS and COMPS-WUGS-DIST with **synthetically constructed nonce words**. Black dashed line indicates chance performance (50%). Refer to Table B.1 for unabbreviated model names.

similar to N. Kim et al. (2022)). Specifically, we constructed novel character sequences—each assigned as a replacement for our original four nonce words—of lengths ranging from 4-8 by sampling in an alternate fashion from consonants (odd positions) and vowels (even positions).[12] A replication of Figure 4.5 using the stimuli with these newly sampled nonce words is shown in Figure 4.8.

On comparing figures 4.5 and 4.8, we observe largely similar patterns of results on stimuli containing nonce words constructed using randomly sampled characters. That is, models generally performed well on COMPS-WUGS, while they struggled on COMPS-WUGS-DIST. There were some exceptions: (1) GPT-Neo 1.3B and 2.7B showed improvements (relative to the original stimuli) in cases where distraction is added closer to the queried property (i.e., **in-between**), though they still hover around chance performance, and additionally the performance of GPT-J, like in the original results is still substantially below chance; and (2) there were non-trivial improvements demonstrated by ALBERT models (large and xl) on the **before** subset of COMPS-WUGS-DIST, and BERT-large on the **in-between** subset of COMPS-WUGS-DIST.

[12] ↑the resulting set of words is: {*ruhisin, kifosa, rosibif, lepuvu*}, still amounting to 12 unique ordered pairs in the COMPS-WUGS-DIST stimuli.

**Table 4.4.** Average agreement (× 100) in PLMs' preference on stimuli containing original and synthetically constructed nonce words.

| **Stimuli** | **Avg. Agreement** |
|---|---|
| COMPS-WUGS | $93.1_{0.8}$ |
| COMPS-WUGS-DIST (**before**) | $73.9_{4.6}$ |
| COMPS-WUGS-DIST (**in-between**) | $73.0_{4.6}$ |
| Overall | $80.0_{2.9}$ |

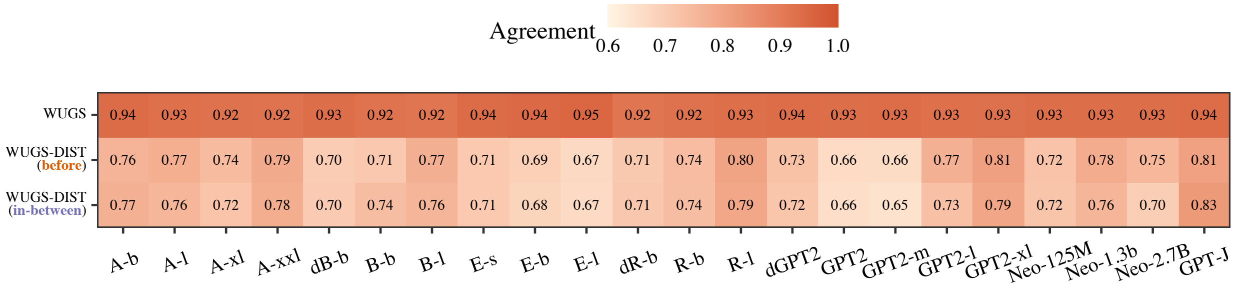


**Figure 4.9.** Proportion of cases (in COMPS-WUGS and both subsets of COMPS-WUGS-DIST) **where each listed model's preference on the original stimuli matches that in stimuli with synthetically constructed nonce words**, measured as the 'Agreement'. An agreement of 1.0 suggests that a given model's preferences are perfectly matched across both sets of stimuli.

To precisely quantify the difference between the two sets of results, we measured the agreement between the predictions of the PLMs for both sets of stimuli, taken as the proportion of minimal pairs in which the models' relative preference agree. Figure 4.9 shows individual model agreement on COMPS-WUGS and COMPS-WUGS-DIST, while Table 4.4 shows agreement percentages averaged across all models. From these results we observe models to show greater robustness to the variability introduced by the choice of nonce words in stimuli with one novel concept (COMPS-WUGS) than in stimuli with multiple novel concepts (COMPS-WUGS-DIST). Despite this discrepancy, there is generally a high average agreement (80%) between a given model's set of decisions on stimuli with original and alternative nonce words.

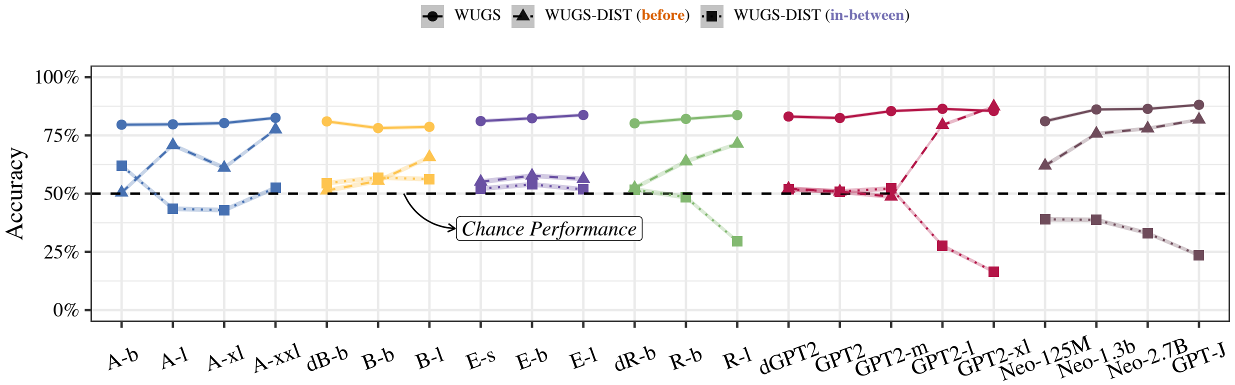


**Figure 4.10.** Accuracies of individual models (grouped by family, in increasing order based on number of parameters) on COMPS-WUGS and COMPS-WUGS-DIST with **alternate framing of novel taxonomic information**. Black dashed line indicates chance performance (50%). Refer to Table B.1 for unabbreviated model names.

### 4.5.6 Framing of novel taxonomic information

Another relevant stimulus design decision is the phrasing for introducing novel concepts in context. While we used *"A wug is a* `[CONCEPT]`*"* for our main experiments, we additionally tested with an alternate framing: *"A wug is a type of* `[CONCEPT]`*."*, where *wug* indicates the novel concept. In all cases, we simply alter the template, keeping everything else constant, including the choice of nonce words.

Figure 4.10 shows accuracies of the models on stimuli with this alternate phrasing, while Figure 4.11 and Table 4.5 show individual and averaged overall agreement between models' preference on original and the alternatively-phrased stimuli, respectively. The agreement percentages between models' preferences are quite high (average agreement being 90%)—in fact even greater than the agreement observed as a result of altering the nonce words (Table 4.4), further cementing the robustness of our results.

## 4.6 General Discussion

The overall goal of this chapter (and COMPS) was to shed light on the extent to which PLMs can robustly (1) attribute to real world concepts (e.g., HORSE, WHALE) their properties (e.g.,

**Table 4.5.** Average agreement (× 100) in PLMs' preference on stimuli containing original (*A wug is a* `[CONCEPT]`.) and alternate framing of novel taxonomic information (*A wug is a type of* `[CONCEPT]`.).

| **Stimuli** | **Avg. Agreement** |
|---|---|
| COMPS-WUGS | $93.4_{1.2}$ |
| COMPS-WUGS-DIST (**before**) | $88.5_{3.1}$ |
| COMPS-WUGS-DIST (**in-between**) | $88.0_{3.6}$ |
| Overall | $90.0_{2.6}$ |

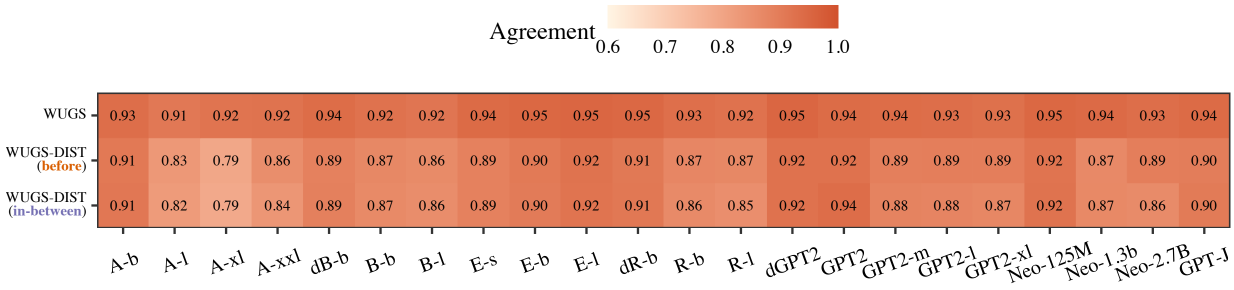


**Figure 4.11.** Proportion of cases (in COMPS-WUGS and both subsets of COMPS-WUGS-DIST) **where each listed model's preference on the original stimuli matches that in stimuli with alternate framing of novel taxonomic information**, measured as the 'Agreement'. An agreement of 1.0 suggests that a given model's preferences are perfectly matched across both sets of stimuli.

*is a mammal*); and (2) demonstrate behavior consistent with *property inheritance*: a reasoning process in which concepts are endowed with the properties of their superordinates (Smith & Estes, 1978; Sloman, 1998; G. L. Murphy, 2002). Testing PLMs for these abilities allows us to ask key questions about how they encode and transfer their background world knowledge (i.e., their inductive generalizations). To target these capabilities more precisely, and mitigate potential inflation of performance by superficial heuristics such as coarse-grained similarity, memorization, and word association, we propose incrementally increasing levels of controls in constructing our minimal pair stimuli, progressively making the task of attributing properties to concepts more challenging.

Findings from our initial experiment on COMPS-BASE established that the basic capacity of models to attribute properties to everyday concepts is largely coarse grained. PLMs were more successful in making correct property attributions when the candidate concepts were radically different, and struggled when the concepts shared semantic relations or had high co-occurrence. On testing for 'property inheritance' behavior (via COMPS-WUGS), PLMs initially appeared to demonstrate reasonable success, but they also showed spurious behavior in achieving correct performance on a non-trivial number of instances for which they did not succeed in the prerequisite base condition. Furthermore, this performance declined substantially in the presence of distracting information (COMPS-WUGS-DIST), providing further evidence that what property knowledge and reasoning we appear to see in these PLMs is more reliant on superficial heuristics than on ideal reasoning behavior. Of particular note is our finding of *catastrophic distraction* in large autoregressive PLMs, whose sensitivity to proximity effects brings their overall performance well under chance, especially when scaled up to billions of parameters, demonstrating a case of *inverse-scaling* (McKenzie et al., 2023).

Contemporary work has highlighted the promise of PLMs on high-level tasks requiring—among other things—access to proper relational knowledge between concepts (see Petroni et al., 2019; Safavi & Koutra, 2021; Piantadosi & Hill, 2022). By drawing on the concept of property inheritance, our experiments target reasoning ability based on perhaps the most well-established of relations—the taxonomic or the `IsA` relation (M. L. Murphy, 2003). Recent work has also alluded to the proficiency of PLMs in capturing taxonomic information about everyday objects and entities (Weir et al., 2020; Chen et al., 2021; Hanna & Mareek, 2021), though perhaps not systematically, (see Ravichander et al. (2020)). Findings from our controlled experiments suggest that PLMs' approximation of the consequences of the taxonomic relation is at best noisy, in light of clear failures especially in presence of similarity-governed competition. We conclude from our analyses that instead of robustly extracting relational information and reasoning about properties of concepts, it is likely that the PLMs tested here are optimized to prefer superficial cues in making word predictions, leading to mistakes and inaccuracies in presence of irrelevant and distracting information. Since robust natural language understanding will be critically reliant on understanding of property knowledge and implications of property transfer, these findings strongly motivate the adoption of rigorous assessment methods when it comes to evaluating for conceptual knowledge and its functional consequences in PLMs.

# 5. A FRAMEWORK FOR SIMULATING INDUCTIVE GENERALIZATION IN LANGUAGE MODELS

This chapter adapts work from the following publications:

1. Misra, K. (2022). On semantic cognition, inductive generalization, and language models. In *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(11), 12894-12895.

2. Misra, K., Rayz, J., & Ettinger, A. (2022). A Property Induction Framework for Neural Language Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*.

They have been reproduced here with significant alterations.

## 5.1 Introduction

One of the hallmark features of human semantic cognition is its capacity to facilitate inferences that go beyond available data to project novel information about concepts and properties, also known as *inductive generalization* or *property induction* (Osherson et al., 1990; Feeney & Heit, 2007; Chater et al., 2011; Kemp & Jern, 2014; Hayes & Heit, 2018). Empirical studies of inductive generalization in adults and children have assimilated a variety of touchstone inductive phenomena, thereby revealing aspects of conceptual knowledge that influence how humans deploy novel information. Analogously, the principles of inductive generalization can provide us with further context within which we can study the nature of LMs understanding of conceptual knowledge and organization (as also argued in §2.3. The previous two chapters have indeed featured—albeit in a very narrow sense—concept of inductive generalization as a means to study PLMs' conceptual knowledge, operationalizing it using the models' natural ability to perform word prediction.[1] In this chapter, we scrutinize this aforementioned approach and propose an alternate analysis framework that uses

[1] ↑Ch. 3 uses a single phenomenon, typicality, which is known to modulate human inductive behavior; while ch. 4 uses a very specific type of inductive generalization (object generalization or property inheritance) to study the extent to which PLMs robustly attribute properties to concepts.

PLMs to simulate inductive generalizations about everyday concepts and their properties. In particular, this framework makes more direct contact with PLMs' representational space, in that the integration of novel information (premises in an inductive argument) constitutes genuine updates to the models' learned representations, which will have a causal effect on the models' responses to queries about how that information will apply in other settings. This enables us to investigate whether the models' learned representations place novel information in their pre-constructed space whereby they can make desirable inductive inferences, which further allows us to questions about how the representations of the PLMs drive generalizations, and what kinds of abstractions they encode. Additionally, this allows us to offer complementary insight similar to other works that have followed similar principles of inductive generalization as we pursue here, though they do so to shed light on the models' capacity to make syntactic and grammatical generalization (van Schijndel & Linzen, 2018; Prasad et al., 2019; N. Kim & Smolensky, 2021), as opposed to semantic phenomena.

I use this framework to specifically simulate hallmark phenomena observed in human inductive literature (Osherson et al., 1990; Heit & Rubinstein, 1994) which have collectively shed light on how the human mind stores, organizes, and uses acquired conceptual knowledge to make inductive leaps given some novel information. For instance, the phenomenon of *conclusion specificity* sheds light on sensitivities to the taxonomic organization of concepts—humans were more likely to project a novel property from robins to all birds (more specific, lower level node in a taxonomy) than to all animals (less specific, higher level node). Similarly, the phenomenon of *premise diversity* highlights how humans reason about category coverage: they are more likely to project a property to all birds when two diverse birds (robins and ostriches)—likely to cover the category of birds better—share a property than when two highly similar birds (robins and sparrows) share the property. Therefore, testing if LMs can systematically show behavior that is compatible with these phenomena can allow us to form conclusions about the extent to which their representations encode the abstractions (such as hierarchical category membership relation) and principles (category coverage) that are evidently observed in at least one system that has mastered *conceptual understanding*: the human mind. In the analyses that follow, we will be comparing results from models with and without pre-trained representations, which will allow us to conclude about the precise benefits that pre-training might provide to models when it comes to conceptual knowledge and reasoning.

### 5.1.1 Desiderata

Recall from §2.3 that inductive generalization is often studied in humans through the use of arguments, represented in the following premise-conclusion format, as popularized by Osherson et al. (1990):

$$\frac{\texttt{concept A} \text{ has } \texttt{property X}.}{\texttt{concept B} \text{ has } \texttt{property X}.} \tag{30}$$

Argument (30) is read as *"A has property X. Therefore, B has property X."* The subject of the premise sentence (`concept A`) 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). Computationally, argument strength has been viewed as an estimate of the conditional probability about a conclusion, given some premise (Kemp & Tenenbaum, 2003; Rogers & McClelland, 2004; Shafto, Kemp, Baraff, Coley, et al., 2005; Tenenbaum et al., 2007; Kemp & Tenenbaum, 2009):

$$p(\texttt{conclusion} \mid \texttt{premise}) \tag{5.1}$$

This measure can be interpreted as the model's estimate of the probability that the proposition expressed by the conclusion is true, given that the premise is true. This is compatible with multiple different types of inductive generalizations—ones that involve generalization of a novel property across known concepts, or those that focus on generalizations involving novel concepts. For instance, in case of a novel property being generalized, the measure would be interpreted as the probability that a property is shared between the conclusion and the premise concepts. Similarly, in the case of a novel concept being introduced, it would be interpreted as the probability that the novel concept has one or more known properties (conclusion), given that it has other known properties (premise) This interpretation leads to two desiderata that the proposed framework aims to satisfy:

① The ability to make judgments about the association of concepts to properties, expressed in natural language—i.e., provide an estimate for eq. (5.1).

② The ability to accept new conceptual knowledge and then be queried to assess generalization or projection of this new conceptual knowledge to additional instances (concepts/properties).

### 5.1.2 Scrutinizing the word prediction account of inductive generalization

One candidate method to perform inductive generalization with language models borrows directly from the 'LMs as Knowledge bases' paradigm (Petroni et al., 2019), by using word prediction-based judgments of property knowledge. Probability estimation is natural to LMs – they can readily assign likelihoods to any given piece of text, and can therefore also make relative probability comparisons between statements expressing property knowledge. For instance, assigning probability scores to statements such as "*a cat has fur*" or "*a dog can fly*." In this case, a competent LM should ideally assign greater probability to the former as compared to the latter (see ch. 4 for exactly how robust models are in assigning properties to concepts). Extending this method to inductive generalization–or any task dealing with conceptual knowledge—would require one to use an LM's probability for a sequence as the sequence's semantic plausibility, like in many previous works (Davison et al., 2019; Porada et al., 2021). Operating under this very assumption, the previous two chapters—and their associated publications—used a word-prediction based method of performing inductive generalization, with the one proposed in Misra et al. (2021; and also ch. 3) being the *first* to investigate inductive generalization in PLMs. That is, for the argument:

$$\frac{\text{Tigers have T9 Hormones.}}{\text{Cheetahs have T9 Hormones.}}, \tag{31}$$

a word prediction based account of inductive generalization would use the following measure as a given language model's approximation of the argument's strength:

$$p(\textit{Cheetahs have T9 hormones.} \mid \textit{Tigers have T9 hormones.}) \tag{5.2}$$

At first glance, the word prediction-based account holds up quite well to both the desiderata. Using sequence probability measures of the conclusion sentence given the premise is certainly a straightforward—and perhaps obvious in the context of PLMs—method to estimate eq. (5.1). It appears to satisfy desideratum ①, under the assumption that the sequence probability of a sentence is an estimate of its statistical plausibility given the PLM's training data and learned parameters. Similarly, since the contents of the premise sentence(s) are present in the same input context as the conclusion, they are always available to the transformer model (the primary subjects of this investigation) while it is queried with the conclusion sentence, thereby satisfying desideratum ②. However, a closer look raises principled concerns with this approach. The task of word prediction by itself is quite unconstrained when it comes to reasoning problems. In the absence of precise contextual information, models might assign a large probability density to a number of acceptable continuations given a prefix (Holtzman et al., 2021). While minimal pairs control for this issue—especially in the way we utilized in chapter 4—they have no real effect on the actual log-probabilities that the models produce. Furthermore, while having the premise in the same context as the conclusion is a necessity (under desideratum ②), it also makes it more likely for models to leveraging surface level cues such as coarse grained lexical similarities (Misra et al., 2020). For example, a model might assign greater probability to "bird" merely due to the presence of "robin" in the input context. This is especially seen in ch. 4, where models would assign greater probability to tokens of the property phrase (*can fly*) even for concepts that cannot actually fly (PENGUIN), due to the presence of distractor concepts (ROBIN). Finally, eq. (5.2)—the measure of argument strength as extracted from LMs—is more an estimate of the extent to which a conclusion *sequentially* follows the premise in free-text, different from the interpretation of eq. (5.1) which aims to quantify the extent to which a conclusion is *true* given that a premise is *true*. In summary, using word prediction alone without any further pressures to focus on the fundamentally pre-requisite task of assessing the extent to which properties apply to concepts divorces the experimental context of inductive generalization from the LM. Note that this criticism is not a criticism of word prediction as an objective for training language models. Instead, it scrutinizes the interpretation of word-prediction as a means to achieve inductive reasoning without the appropriate experimental context. Therefore, these limitations do not preclude the fact that the PLMs might not be engaging in deeper world-modeling to arrive at their next word distributions—this is an interesting and active area of research (see for instance B. Z. Li et al., 2021; N. Kim & Schuster, 2023)—but due to the concerns raised above, may not paint the full picture of the representational competence of PLMs which this chapter (and the framework proposed herein) aims to characterize.

## 5.2 The framework

A solution to the above problems is presented by the influential work of Rogers and McClelland (2004), who present a connectionist model of semantic cognition. Specifically, Rogers and McClelland (2004) developed a feed-forward neural-network that is trained to predict a range of properties for a given input concept (1 if the property exists for a concept, and 0 otherwise) – treated as the model's *experience*. For instance, for the concept ROBIN the model has to predict 1 for the properties *is a bird, can fly, can grow, can sing,* etc. and 0 for the properties *is a tree, can swim,* etc. Rogers and McClelland (2004) relate a wide range of empirical phenomena in semantic cognition to the patterns observed during and after training of this model. Immediately relevant to this dissertation is their experiments using the feed-forward network to make inductive inferences exactly like the kind discussed in **??** – by projecting a new property (e.g., *can queem*), that they make up, from a single concept to all other concepts in their toy dataset. To operationalize induction, the authors use the same method as they used for optimizing the model's parameters to fit the data: *predictive error driven learning* (Rumelhart et al., 1986), more commonly known as *backpropagation*. Specifically, the authors first train the network using the initial set of properties for each concept, where the made-up property is always set to 0. Then, at the $e^{th}$ epoch, for a single concept, they let the model also predict 1 for the made-up property and backpropagate the error (which now also includes a component for the new property). They then query the model's predictions for linking the new property to all other concepts, completing one round of an inductive inference.

In the context of an LM-based inductive generalization framework, two insights can be gained from Rogers and McClelland (2004), which can help alleviate the issues pointed out in §5.1.2. First, we can take a positive step towards constraining the model's parameters to exclusively use property knowledge by further training the already pre-trained LM on a task similar to that used by Rogers and McClelland (2004). Note that the kind of inductive generalization that this dissertation is concerned with is first and foremost a task of judging how likely are one or more properties attributable to a concept (or a set of concepts). Therefore, by 'amplifying' the aspects of knowledge needed to successfully predict that a ROBIN *is a bird, can fly, can grow, can sing,* etc., we are better suited to assess the capacity of the LM to perform inductive generalizations. Under this viewpoint, the measure $p(\text{conclusion} \mid \text{premise})$—now characterized as the probability that the conclusion is true, given that the premise is true—provides relatively more insight about the conceptual knowledge of the model as opposed to its word-prediction-based analogue. This is especially because in order to succeed at making judgments about concept-property associations, a competent model will *have* to rely on property knowledge – i.e., learn inductive preferences that favor accurate conceptual knowledge over/in addition to word-sequences. Second and more importantly, backpropagation has desirable properties to simulate induction. Assuming that we have a model that is trained to predict a concept's property knowledge, we can integrate new property information (i.e., the premise in an inductive argument) by simply backpropagating the error made by the model on the new information. Doing so allows the integration of new information in the model as updates to its representations, which encode knowledge used to inform how the model generalizes. Another desirable consequence of casting induction as backpropagation is that we can now have an explicitly measurable form of "acceptance" of the new information (desideratum ②). That is, we can quantify the extent to which the model's representation encode the novel information by measuring its loss after each round of backpropagation, which also allows us to control the rounds of backpropagation for a given experiment. This unlocks opportunities to analyze the model at a deeper level – e.g. measuring how *fast* different kinds of information are integrated, or how likely is new information generalized compared to the number of backpropagation steps needed, etc., which were otherwise not possible in the earlier account.

Taking into account the above insights, I propose a framework that simulates inductive generalization in language models. In particular, I interpret the measure specified in eq. (5.1) in the proposed framework as the probability that the information in conclusion is true, by a model whose representations *reflect* the premise information. To satisfy desideratum ①, the framework fine-tunes LMs to classify as true or false sentences that associate concepts to properties—i.e., make property judgments, which will automatically allow the models to estimate the probability that a property applies to a concept—as p(concept has property = True). To satisfy desideratum ②, the framework operationalizes inductive generalization as:

> *the behavior of a language model—constrained to make judgments about how likely a property can be attributed to a concept—after being adapted to novel conceptual information using predictive-error driven learning, or backpropagation.*

These decisions map to two separate stages in the framework, both of which I describe in further detail below:

### 5.2.1 Stage 1: Eliciting property judgments from LMs

The goal of the first stage is to construct a model that can make judgments about property knowledge expressed as natural language sentences. In particular, this model should distinguish correct concept-property associations (*cat has whiskers*) from those that are incorrect (*sparrow has whiskers*). Based on the model's objectives, it is similar to the network described by Rogers and McClelland (2004), but instead of using simple, localist representations as input, it uses dense distributed representations that are learned during pre-training across several inter-connected layers and give rise to semantic behavior.[2] We construct such

[2] ↑though importantly the very first layer (embedding), the input to the model is a localist one-hot representation of each word in the sentence.

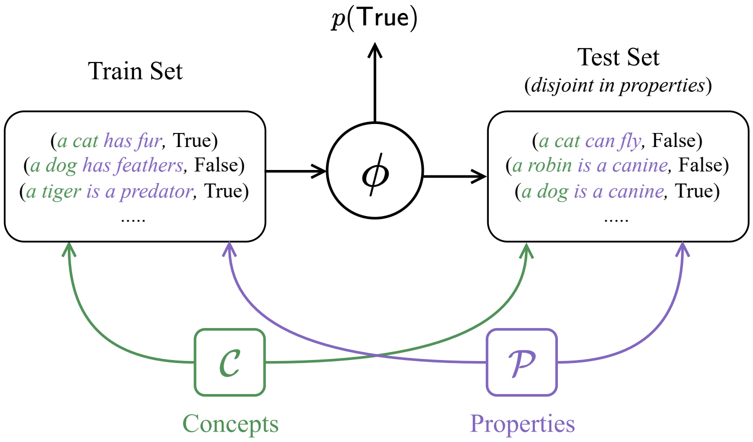


**Figure 5.1.** Depiction of the property judgment stage. Concepts and properties are sampled to create experimental splits consisting of sentences and labels corresponding to whether the assertion expressed by the sentence is true or false. The model (with parameters $\phi(.)$) takes in as input these sentences and returns the probability of the True label. Crucially, the model is evaluated on a set of property judgments that are disjoint from the training set (in this case, in terms of properties).

a model by fine-tuning existing LMs to classify sentences that express concept-property associations to be true or false.[3] 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

[3] ↑using a binary classification set up might be considered too stringent, however to the best of my knowledge, there are no existing resources that can support fine-grained graded judgments at a large enough scale.

of this stage, we have a trained model (with parameters $\phi$) that takes as input a sentence $s$ and produces a probability score $p(\textsf{True} \mid s, \phi)$ corresponding to the degree of truth of $s$ as internalized by the LM. For example, the model would—hypothetically—produce as output $p(\textsf{True} \mid \textit{a cat can fly}, \phi) = 0.04$, or $p(\textsf{True} \mid \textit{a cat has fur}, \phi) = 0.96$. Figure 5.1 describes the property judgment stage.

### 5.2.2 Stage 2: Inductive generalization as updates to LM representational space

In this stage (see Figure 5.2), we use the fine-tuned model from the previous stage to perform inductive generalization, 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 5.2).

To perform property induction, we first adapt the model's parameters $\phi$ to the premise sentences by using standard backpropagation, yielding an updated state of the model, $\phi'$, that correctly attributes the concepts in $\mathcal{A}$ with the novel property. We then freeze $\phi'$ 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:

$$\mathrm{G} = \frac{1}{n} \sum_{c_i \in \mathcal{G}} \log p(\textsf{True} \mid \textit{``}c_i \textit{ has property X''}, \phi') \tag{5.3}$$

Before every round of induction, the model parameters are reset to their original state ($\phi$), i.e., the state acquired as a result of the property judgment stage. Additionally unless otherwise stated, the model is backpropagated until it correctly predicts the label(s) of the premise sentence(s).

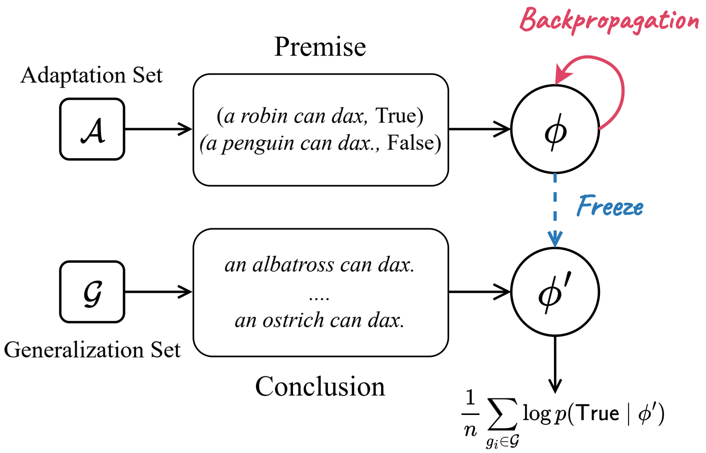


**Figure 5.2.** Depiction of the Induction Stage, in this case, for testing the generalization of the property *can dax*. Here, the model is provided with information that *a robin can dax* is true, but *a penguin can dax* is not. The model has to then deploy this knowledge across all other birds. $\mathcal{A} = \{\text{ROBIN}, \text{PENGUIN}\}$, $\mathcal{G} = \{\text{ALBATROSS}, \ldots, \text{OSTRICH}\}$.

Consider the following example (also depicted in figure 5.2): where a novel property (can dax) is applied to ROBIN but not to PENGUIN, and the goal of the experiment is to test how it is projected to other birds (the set $B = \{\text{CROW}, \ldots, \text{OSTRICH}\}$), the components of the inductive generalization step are as follows:

$$\begin{aligned}
\mathcal{N} &= \{\textit{can dax}\} \\
\mathcal{A} &= \{\text{ROBIN},\ \text{PENGUIN}\} \\
\mathcal{G} &= B - \mathcal{A} \\
\text{Premise} &= \{(\textit{a robin can dax.}, \mathsf{True}), \\
&\qquad (\textit{a penguin can dax.}, \mathsf{False})\} \\
\text{Conclusion} &= \{\textit{a crow can dax.}, \ldots, \\
&\qquad \textit{an ostrich can dax.}\}
\end{aligned}$$

### 5.2.3 Alternate Formulations

While the above description of the framework is compatible with any type of neural network that can process sentences, it is not restrictive to binary classification settings. That is, a similar operationalization of inductive generalization can be construed using models that perform word prediction but are in fact constrained to reflect property knowledge. This way, one can keep the flexibility of word-prediction without violating the construct validity issues raised in §5.1.2. Such a construal can be possible with models trained in the same way as the COMET model (Bosselut et al., 2019; Hwang et al., 2021), which uses autoregressive models such as GPT2 (Radford et al., 2019) to predict the tokens that make up the property phrase (*can fly*) given a concept (*robin*). This is done by reformulating the basic language modeling objective to one that penalizes incorrect generation of properties given a concept.

## 5.3 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.

### 5.3.1 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, prior 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 3,643 properties, with 30,076 unique ground-truth property-concept pairs which we used in our experiment.

For each of our 3,643 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 and Richie (2021) to convert our 60,152 property-concept pairs (30,076 × 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 appendix C.

### 5.3.2 Tested PLMs

While our framework can be applied to any PLM, we present results from fine-tuning four pre-trained LM families, based on the precedent of using these models in standard sentence classification tasks (A. Wang et al., 2018; A. Wang et al., 2019): BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), DeBERTa v3 (He, Liu, et al., 2021; He, Gao, et al., 2021). All four 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 on BERT-large, RoBERTa-large, ALBERT-v2-xxl, and DeBERTa v3-large, each of which uses approximately the same amount of trainable parameters. We fine-tune each of the PLMs 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 hyper-parameters of the LMs on the validation

**Table 5.1.** Performance of the fine-tuned PLMs on the (1) test set of the property judgment task (F1), and (2) a subset of COMPS-BASE with properties that do not occur in the models' training set (accuracy reported per negative sample type). Chance stands for a hypothetical model that is equivalent to a coin flip.

| Model | Test F1 | COMPS-BASE Accuracy | | | |
|---|---|---|---|---|---|
| | | Taxonomic | Prop. Norms | Co-occ. | Random |
| Chance | 0.66 | 0.50 | 0.50 | 0.50 | 0.50 |
| BERT-large | 0.75 | $0.75_{0.01}$ | $0.78_{0.01}$ | $0.74_{0.01}$ | $0.90_{0.009}$ |
| ALBERT-xxl | 0.74 | $0.80_{0.01}$ | $0.80_{0.01}$ | $0.81_{0.01}$ | $0.94_{0.007}$ |
| RoBERTa-large | 0.76 | $0.77_{0.01}$ | $0.80_{0.01}$ | $0.77_{0.01}$ | $0.91_{0.009}$ |
| DeBERTa-large | **0.77** | $\mathbf{0.86}_{0.01}$ | $\mathbf{0.86}_{0.01}$ | $\mathbf{0.88}_{0.01}$ | $\mathbf{0.96}_{0.006}$ |

set,[4] and evaluate the four adapted models on the test set using F1 scores. Following ch. 4, we also evaluate on the subset of the COMPS-BASE stimuli that do not appear in the models' training set ($N = 14,437$). That is, we use the models' ability to estimate $p(\textsf{True})$ for the minimal pair sentences in COMPS-BASE and take their accuracies to an additional measure that sheds light on their ability to associate properties to concepts in a generalizable manner.

### 5.3.3 Results and Analysis

Table 5.1 shows the performance of the fine-tuned PLMs on the test set as well as on the subset of COMPS-BASE disjoint from the models' training data, further decomposed based on the negative sample type. Note that chance performance for the test set (F1) is 0.66, and that on each decomposed subset of COMPS-BASE is 0.50. From table 5.1, we see that all models perform well above chance across all evaluation settings ($p < .001$, FDR corrected), suggesting strong capacities of all four models to assess the application of properties to concepts. DeBERTa-large in particular stands out as the best model out of the four, consistently achieving the highest F1 score/Accuracy in all settings, highlighting its distinctive robustness and mastery of the property judgment task. Similar to the trend observed in ch. 4, the

[4] ↑we searched the best learning rate from the list: {1e-04, 3e-04, 5e-04, ... 9e-07}

models performed best on COMPS-BASE stimuli where the negative sample was randomly chosen (and was likely to be trivially different from the positive concept, e.g., ROBIN and CHAIR for the property *can fly.*) as compared to those where the negative samples were semantically similar to the positive concept. At the same time, there was a considerable reduction in the gap between the performance achieved on stimuli with semantically similar negative concepts and those with randomly chosen negative concepts. This suggests that perhaps one way to overcome lack of robustness in assigning properties to concepts could be to adapt models specifically for the task, where instead of assigning probabilities over large vocabulary spaces, models only have to map sequences to two values (the True or False labels).

## 5.4 Inductive Generalization Experiments

Using models trained using the property judgment task, I now describe a series of experiments that simulate well known empirical regularities observed in human inductive generalization literature. Specifically, I first test the models' ability to demonstrate key phenomena observed by the pioneering work of Osherson et al. (1990), which exclusively focuses on inductive behavior across taxonomically organized categories (e.g., ANIMAL, MAMMAL, BIRD, etc). Importantly these phenomena usually assume (and design) the nature of the property to be largely ambiguous and novel, and instead focus on systematic patterns observed as a result of manipulating the categories/concepts present in these arguments. Following the Osherson et al. experiments, I then turn to cases where conceptual relations *and* the nature of the property being generalized both end up mattering, where human inductive behavior is different, depending on the type of property, thus showcasing their (humans') meta-sensitivity to different conceptual structures. For the phenomena we consider here, we either construct synthetic stimuli using the guiding principles of the given phenomenon or use existing stimuli for cases where authors have made them available.

While our main subjects in these experiments are the pre-trained LMs fine-tuned to perform property judgment, we additionally also use models with the same architecture but those that are randomly initialized before being trained to perform property judgment.

Results from these comparisons will shed light on the extent to which pre-trained features result in the patterns we observe. As a sanity check, all randomly initialized models are at chance on the property judgment task (see table 5.1, and have overfit their training set. In all cases, models are adapted using the same learning rate that was being used just before their best performance on the development set of the property judgment task.

### 5.4.1 Empirical Regularities from Osherson et al. (1990)

Our first set of experiments focus on the empirical regularities first discovered and documented by the pioneering work of Osherson et al. (1990). These regularities (also discussed in depth in §2.3.2) primarily focus on principles of hierarchical structure and organization of concepts and categories. Osherson et al. (1990) do not make their stimuli available and instead describe in-depth the principles behind each phenomenon. For instance, in the premise diversity phenomenon the premise concepts/categories of stronger arguments should be *different* from each other, relative to those in weaker arguments. We rely on these principles to construct synthetic stimuli that adhere to every phenomenon of interest. We only focus on eleven out of the thirteen stimuli, and leave out *(premise-conclusion identity)* and *(premise-conclusion inclusion)* since they do not involve genuine abstractions—the former case is a sanity check, while the latter is deductive rather than inductive. Below I describe a general framework that allows the sampling of an arbitrary amount of stimuli that encode the Osherson phenomena.

To construct our synthetic stimuli, we define a world $\mathcal{W} = (\mathcal{C}, \mathcal{P}, \mathcal{T}, \mathcal{M})$ where $\mathcal{C}$ is the set of animal concepts, $\mathcal{P}$ is the set of properties, $\mathcal{T}$ indicates a taxonomy—a hierarchically organized structure with all items of $\mathcal{C}$ at the lowest level. Finally, $\mathcal{M}$ is a matrix whose row indices indicate concepts and column indices indicate properties, and whose values are either 1, if the concept has a property, or 0, if it does not. As a result, the rows of matrix are binary vector representations of concepts. In terms of notation, we say that the first concept (in our case, ALLIGATOR) has the 22nd property (*basks in the sun*) if $\mathcal{M}_{1,22} = 1$. Representing concepts and properties using binary vectors allows us to define similarities between concepts—by comparing their respective vectors. One popular similarity metric between binary vectors is the jaccard similarity, which we use to operationalize the notion of property overlap—i.e., the extent to which two concepts share known properties. Mathematically, it is defined as:

$$sim(c_1, c_2) = \frac{\sum_{i=1}^{|\mathcal{P}|} \min(\mathcal{M}_{1,i}, \mathcal{M}_{2,i})}{\sum_{i=1}^{|\mathcal{P}|} \max(\mathcal{M}_{1,i}, \mathcal{M}_{2,i})} \tag{5.4}$$

Equation (5.4) will play an important role in sampling concepts for the adaptation and generalization sets, since many of the phenomena covered by Osherson et al. (1990) deal with conceptual similarities between the premise and conclusion. Relatedly, property overlap constitutes an important factor in a prominent connectionist model of human inductive generalization (Sloman, 1993). While $\mathcal{W}$ can be constructed using any set of concepts and properties, we specifically use the data described by the training set of our property judgment models, and restrict ourselves to the animal kingdom, the largest well-defined taxonomy within our collection of training concepts. By doing so, we ensure that the models' apparent "knowledge" (set of outputs corresponding to concept-property association) aligns with the properties encoded by the sampled stimuli. Overall, the instance of $\mathcal{W}$ we construct for the experiments reported in this dissertation comprises of 152 concepts and 726 properties,[5] which results in the matrix $\mathcal{M}$ to be of size 152 × 726. For our taxonomy $\mathcal{T}$, we extract the subtree of WordNet (Miller, 1995) that entirely subsumes our concepts, represented at its leaf nodes, similar to our design in the original property judgment experiments.

We generate stimuli for a given phenomenon using our instance of $\mathcal{W}$ as follows: First we sample adaptation and generalization concepts according to the constraints of the phenomenon (described below), resulting in strong and weak adaptation-generalization sets. Next, we pair them with multiple different novel properties belonging to the set: {*can dax, can fep, has dax, has blickets, is a tove, is a wug, is mimsy, is vorpal*}. We then follow the method described in §5.2.2 and perform multiple trials of inductive generalization using models obtained from the property judgment task (see §5.2.1). To measure the models' inductive behavior, we analyse their average generalization score (see eq. (5.3)) over all eight

[5] ↑out of 991 total animal-related properties, distributed across training, test, and validation sets.

**Table 5.2.** Empirical regularities from Osherson et al. (1990) that LMs are tested for and their respective number of unique generated adaptation-generalization pairs. Symbols in parentheses indicate type of argument: 'G' suggests general, 'S' suggests specific, and for cases with both 'G' and 'S', we have two unique sets of stimuli, one for each type.

| **Phenomenon** | $N$ | **Phenomenon** | $N$ |
|---|---|---|---|
| Premise Conclusion Similarity (S) | 240 | Premise Monotonicity (G, S) | 180 |
| Premise Typicality (G) | 120 | Premise Nonmonotonicity (G, S) | 150 |
| Premise Diversity (G,S) | 342 | Inclusion Fallacy | 240 |
| Conclusion Specificity (G) | 150 | Premise Conclusion Asymmetry | 2000 |

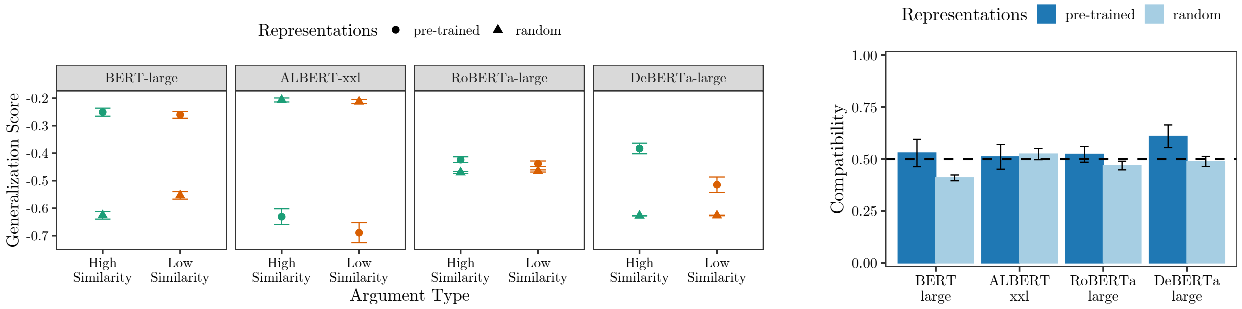
novel properties,[6] for the strong and weak arguments. We additionally measure the ***compatibility*** of models with respect to the behavioral pattern associated with the phenomenon: i.e., calculate the percentage of time a model shows greater generalization scores for stronger versus weaker arguments. These analyses will result in clear observations about whether or not models on average are able to capture the phenomena. For instance, it would be desirable for models to show greater generalization scores for diverse premises than for highly similar ones.

Below, I describe the generation process for the eleven different Osherson phenomena, as well as our results from the aforementioned analyses. Unless otherwise specified, the adaptation set concepts are sampled from the following higher level categories: {ANIMAL, MAMMAL, BIRD, REPTILE, INSECT, FISH}, in all cases. Table 5.2 shows the list of Osherson et al. (1990) phenomena we simulate in this experiment.

### Premise Conclusion Similarity

**Generation:** To create the adaptation set, we sample 10 pairs of items from our 6 categories, and for each concept pair, we sample 2 concepts from their nearest neighbors (as computed using eq. 5.4) as the 'high similarity' generalization sets, and similarly 2 concepts from the set of most dissimilar concepts as the 'low similarity' generalization sets. Doing

[6] ↑this differs from standard practice where human participants/models are often analyzed for a single property at a time.



(a)

(b)

**Figure 5.3.** Premise-Conclusion Similarity Results. **(a):** Average generalization scores (along with 95% CI) obtained by the models on High and Low similarity arguments. **(b):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations.

so results in a total of 240 unique adaptation-generalization pairs. Combined with our 8 different novel properties, this results in 1,920 unique inductive arguments.

**Results:** Figure 5.3 shows the average generalization scores of the models for arguments with high and low similarity between the two adaptation concepts and the generalization concept (figure 5.3a), as well as the compatibility of the models with the phenomena (figure 5.3b). We observe that only the DeBERTa-large model that is fine-tuned on the property judgement task starting from pre-trained representations is able to demonstrate the premise-conclusion similarity phenomenon (compatibilty value of ≈0.60). All other models fail to show this behavior, regardless of the nature of their initial representations (pre-trained vs. random). Note that this phenomenon includes two adaptation concepts in the premise throughout, in order to mimick the conditions specified by Osherson et al. (1990). Therefore, in order for a model to demonstrate this phenomenon, it must implicitly consider the *joint* similarity of the premise concepts and the conclusion concept (approximating a process similar to a logical conjunction). Since this may not necessarily be trivial for models to perform, especially via backpropagation, we turn to a simpler setting of having only a single adaptation concept in the premise, and measure the extent to which training-data similarity (computed using eq. (5.4)) modulates models' inductive behavior. To this end, we compute

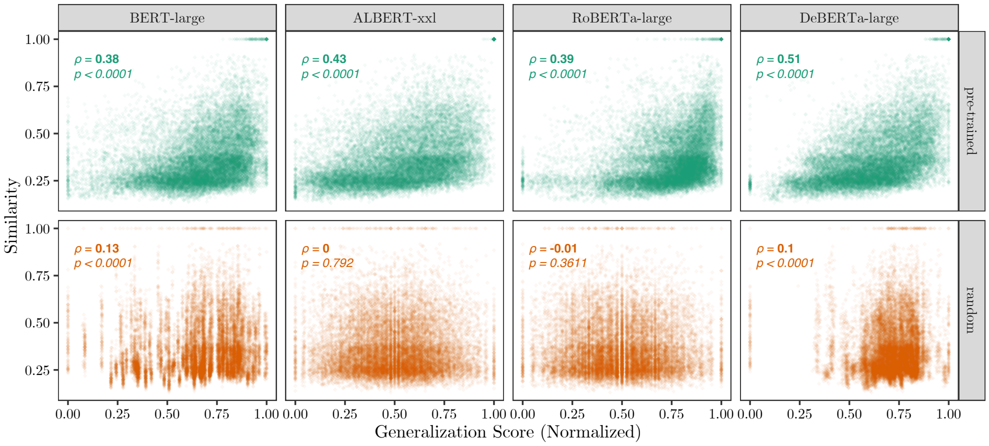
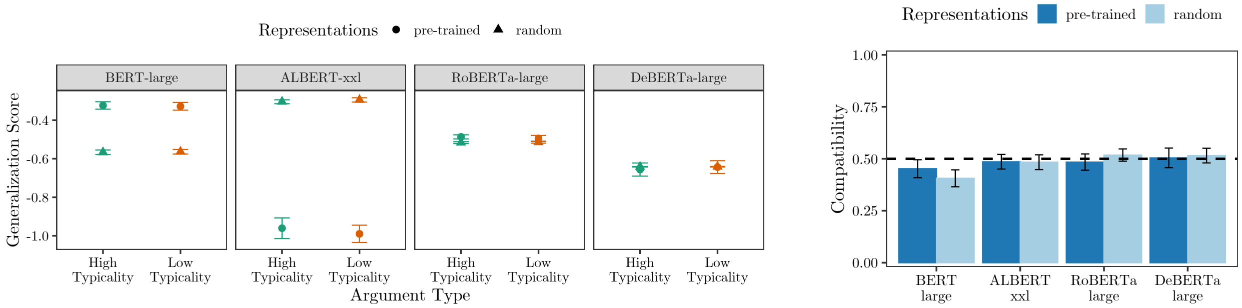


**Figure 5.4.** Scatter plots showing Normalized Generalization Scores and Concept Similarities obtained from models' training data (see eq. (5.3)), along with the Spearman correlation ($\rho$) between them, across all models.

the generalization scores (normalized to be between 0 and 1) for every pair of concepts in our world $\mathcal{W}$, and measure their correlation against pairwise similarities between the concepts. Figure 5.4 shows these results for all our models. From this figure, we see that all models that start with pre-trained representations show non-trivially positive correlation scores ($\rho \in [0.38, 0.51]$, $p < .0001$ everywhere, FDR corrected), with DeBERTa-large showing the strongest value (0.51). Models that start from randomly initialized representations show correlation values that are either lower than their pre-trained counterparts (BERT-large and DeBERTa-large), or are not significant (ALBERT-xxl and RoBERTa-large). Overall, this suggests that pre-trained representations can in principle be adapted to demonstrate inductive behavior that is compatible with simple conceptual similarities between premise and conclusions. However, they may not necessarily engage in behavior that resembles a simple composition over these similarities (i.e., when the premise contains more than one adaptation concepts).

### Premise Typicality

**Generation:** We use all members of every category with at least 10 members as our adaptation concepts—i.e., members of MAMMAL (52), BIRD (36), INSECT (18), and FISH



(a)

(b)

**Figure 5.5.** Premise Typicality Results. **(a):** Average generalization scores (along with 95% CI) obtained by the models on High and Low similarity arguments. **(b):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations.

(14)—and use these categories as the generalization set. This results in a total of 120 unique adaptation-generalization pairs, giving us 960 different arguments, all of which are general in nature (i.e., generalization from ROBIN to BIRD). We then compute typicality measures for the adaptation concepts with respect to their category by considering their low-dimensional representations of their binary concept-property vectors (from $\mathcal{M}$), as computed using a singular value decomposition (see §C.3). This method was inspired by the work of Saxe et al. (2019), who used a similar formulation to mathematically define how a simple linear connectionist network may be able to encode typicality. We perform a median split of the typicality measures, giving us sets of high and low typicality members for each category. We then randomly pair members from these sets, giving us a total of 60 strong and weak adaptation-generalization pairs, and 480 different arguments.

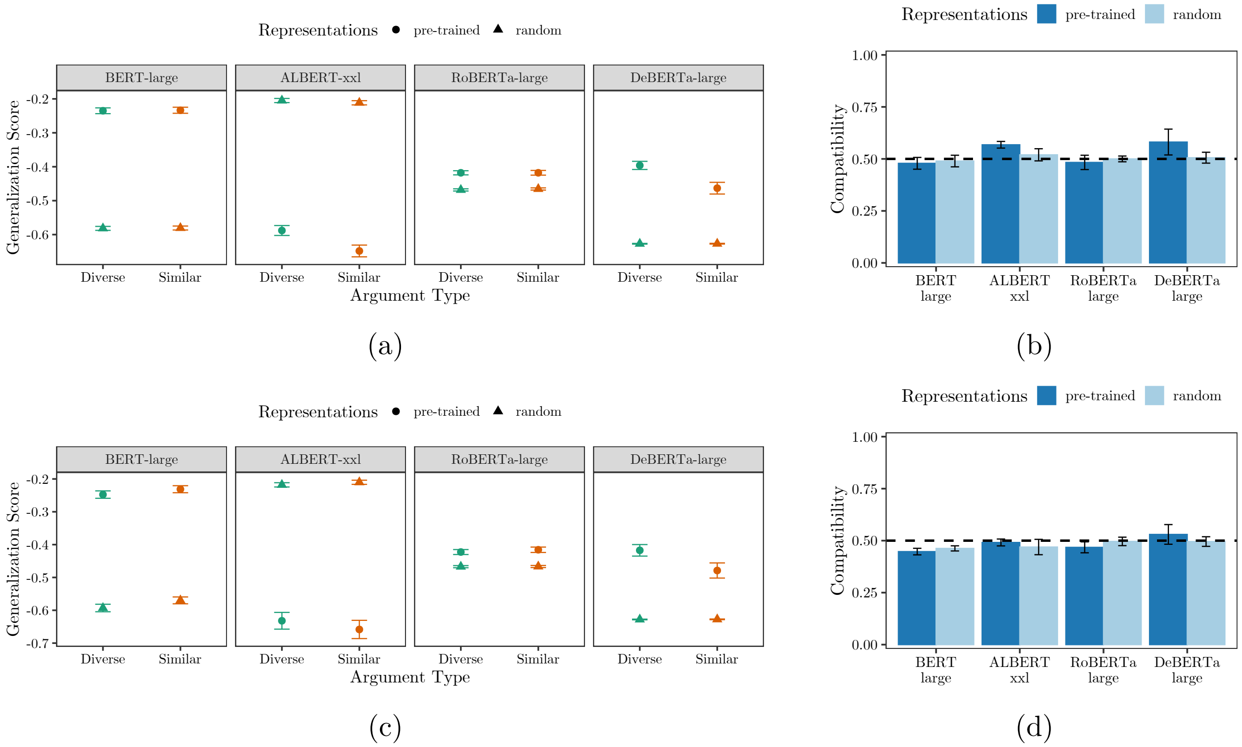
**Results:** Figure 5.5 shows the average generalization scores of the models for arguments that have high and low typicality adaptation concepts (figure 5.5a), as well as the compatibility of the models with the phenomena (figure 5.5b). From these figures, we see that none of the models are able to demonstrate this phenomenon, with all of their compatibility hovering at chance. That is, the models are either indifferent to the typicality of the adaptation concept when generalizing its properties to rest of the members of the target category, or in some cases show the opposite behavior. That is, even if their representations encode individual-concept similarities and distinctions, it does not necessarily translate to graded measures of "category-ness" as operationalized by the method described in §C.3.

### Premise Diversity

**Generation:** We start with populating our adaptation set by first sampling from our 6 categories $\min(10, |C|)$ concepts, where $|C|$ refers to the size of the category (number of concepts within it). We formulate diverse premises as ones where the concepts are maximally different (e.g., ROBIN and PENGUIN). So, in order to create diverse (strong) and similar (weak) premises, we sample three concepts each from the nearest and most dissimilar neighbors. This creates at most 30 diverse pairs and 30 similar pairs for each superordinate category, and in total we end up with 342 unique adaptation-generalization pairs,[7] giving us 2,736 different arguments. We repeat this process to create stimuli for specific arguments, by replacing superordinate category members in the generalization set to just one randomly sampled category member.

**Results:** Figure 5.6 shows the average generalization scores of the models for general (figure 5.6a) and specific (figure 5.6c) arguments with diverse and similar adaptation concepts, as well as the compatibility of the models with the phenomena (General: figure 5.6b, Specific: figure 5.6d). For general arguments, we observe that only ALBERT-xxl and DeBERTa-large (both with pre-trained representations) demonstrate the desired phenomenon—these models are more likely to project a property to other members of a category when a diverse set of concepts share the property. For specific arguments however, only DeBERTa-large (again with pre-trained representations) is able to demonstrate behavior that is compatible with the phenomenon above chance. In all cases, models trained starting from random representations fail to demonstrate a premise diversity effect.

[7] ↑some categories such as REPTILE had fewer than 10 concepts.



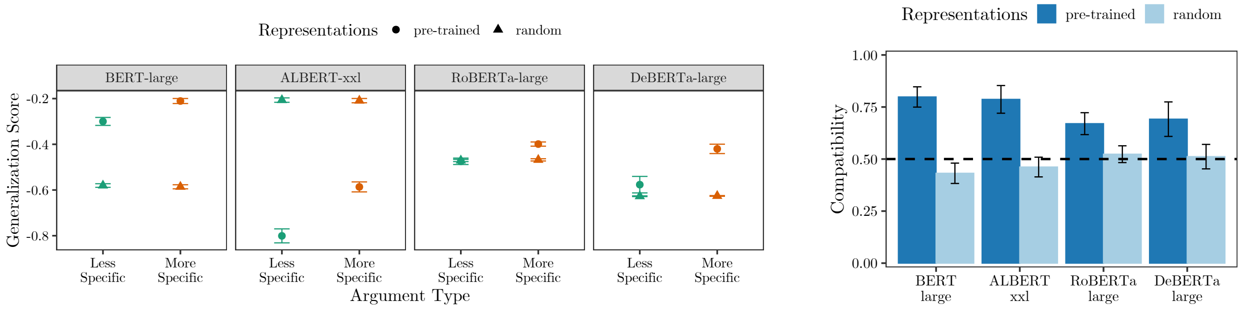
(a) (b)

(c) (d)

**Figure 5.6.** Premise Diversity Results. **(a) and (c):** Average generalization scores (along with 95% CI) obtained by the models on general (a) and specific (c) arguments containing diverse and similar premises. **(b) and (d):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations on general (b) and specific (d) arguments.

### Conclusion Specificity

**Generation:** Here, we only consider 5 categories (BIRD, MAMMAL, FISH, INSECT, REPTILE), and leave out ANIMAL to serve as the general, less specific, category. We then compare generalization of a property from the premise to ANIMAL against that from the same premise to all members of either of the 5 categories. To create our adaptation set, we sample pairs of 15 concepts from each of our 5 sets of superordinate categories. For each pair, we create two generalization sets – one including only the members of the specific superordinate category, and one including all members of the ANIMAL category. This give us 150 unique adaptation-generalization pairs, and therefore 1,200 arguments.



(a) (b)

**Figure 5.7.** Conclusion Specificity Results. **(a):** Average generalization scores (along with 95% CI) obtained by the models on arguments with general and specific conclusions. **(b):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations.

**Results:** In order to successfully demonstrate this effect, a model must make sufficient distinctions between specific and general categories—i.e., be sensitive to the hierarchical organization of concepts. Figure 5.7 depicts the average generalization scores of models for arguments with the same premise (e.g., ROBIN) to general (ANIMAL) and specific (BIRD) categories (figure 5.7a), as well as the percentage of cases where generalization to specific categories was stronger than to general ones–i.e., the models' compatibility (figure 5.7b). Note that the general category perfectly encompasses each of the specific ones. From figure 5.7, we observe that all models that started with pre-trained representations robustly showed this effect, and that their compatibility was substantially above chance. Models with random initial representations were either below chance or just above it (ALBERT-xxl and DeBERTa-large).

### Premise Monotonicity

**Generation:** We create adaptation sets by sampling 1, 2, and 3 concepts from each high level category. For general arguments, we include all members of the high level category, while for specific arguments we randomly sample one member. We perform this sampling process 10 times for each of our 6 high level categories, and end up with 180 unique adaptation-

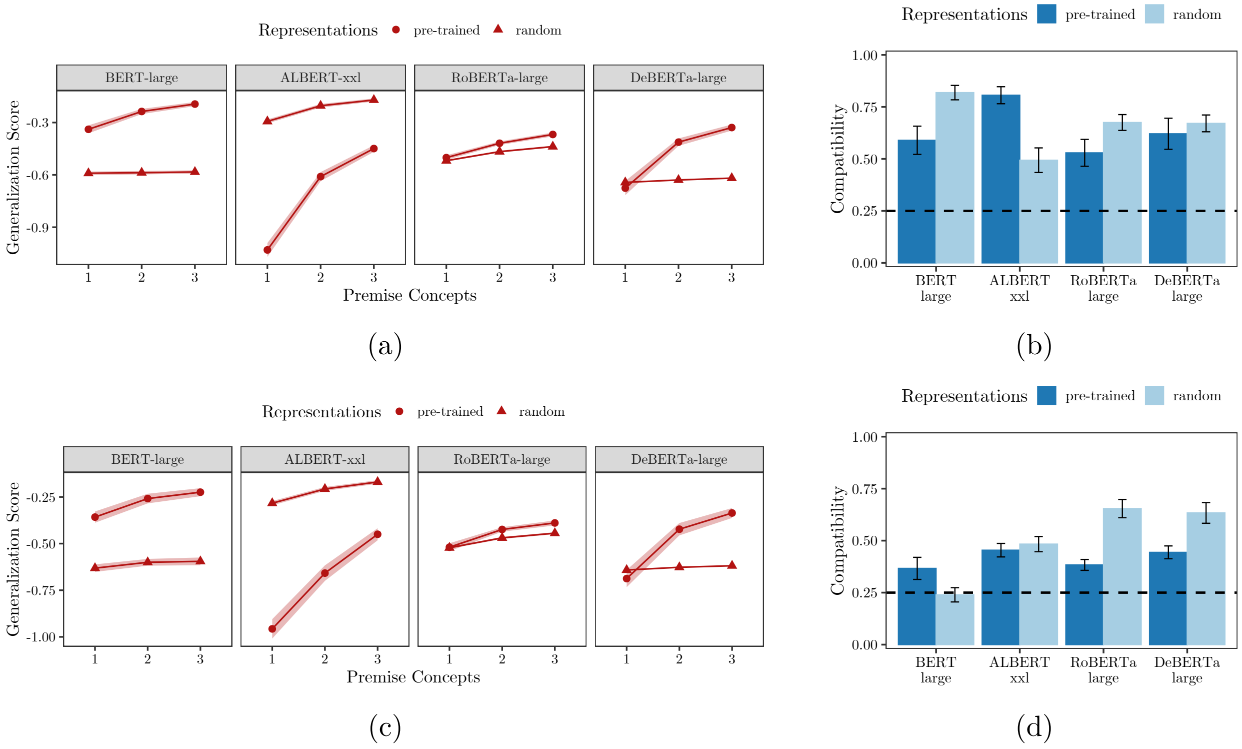


**Figure 5.8.** Premise Monotonicity Results. **(a) and (c):** Average generalization scores (along with 95% CI) obtained by the models on general (a) and specific (c) arguments containing 1, 2, and 3 premises. **(b) and (d):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations on general (b) and specific (d) arguments.

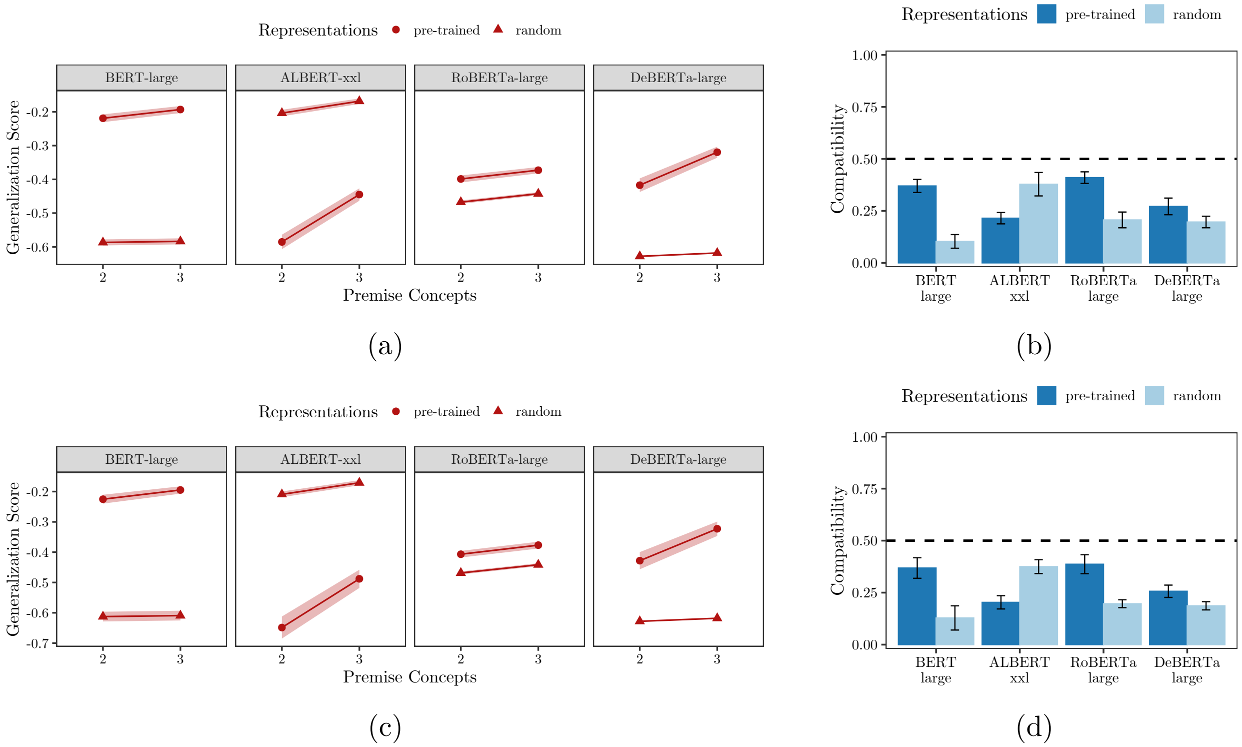
generalization pairs for each argument type (specific and general).

**Results:** Figure 5.8 shows the average generalization scores of the models for general (figure 5.8a) and specific (figure 5.8a) arguments with increasing number of adaptation concepts, as well as the compatibility of the models with the phenomena (General: figure 5.8b, Specific: figure 5.8d). We see that almost all models (with the exception of BERT-large with randomly initialized representations) successfully the premise monotonicity effect—i.e., their projection of a property to other concepts is proportional to the amount of evidence about the property provided in the premise. This result is relatively unsurprising, since we can reasonably expect data-driven models to be more confident in their predictions as the number of samples provided to them increases, and so even models trained from randomly initialized representations are able to demonstrate this effect.

### Premise Nonmonotonicity

**Generation:** To create the adaptation set we first sample two concepts from our five lower level categories (BIRD, MAMMAL, FISH, INSECT, REPTILE), giving us our strong argument premise concepts. We then separately sample one concept from outside of each category and pair it with the two previously sampled concepts to get the weak argument premise concepts. In all cases, we measure generalization to the members of the lower category. Repeating this process 15 times for each of our five categories gives us 150 total adaptation-generalization pairs.

**Results:** Figure 5.9 shows the average generalization scores of the models for general (figure 5.9a) and specific (figure 5.9a) arguments that are either mixed (3 adaptation concepts, weaker) or perfectly general (2 adaptation concepts, stronger), as well as the compatibility of the models with the phenomena (General: figure 5.9b, Specific: figure 5.9d). We see that all models, regardless of their representations' initial state fail to capture this effect. Recall that in this case, arguments where the generalization concept converts the argument into a mixed argument (i.e., adding PIG to the premise of ROBIN, EAGLE → BIRD) are the weaker ones. Only in such a case, this effect contradicts the effect of premise monotonicity. Here, only the reasoner that can robustly capture the principles of category membership and coverage (how well do the premise concepts cover the category present in the conclusion) can successfully demontrate this effect. As such, we take this test to serve as a possible check for whether models employ genuine abstraction over category boundaries in their monotonicity effect observed above or if they are simply matching observed strings of the property phrase (e.g., *can dax*) to their corresponding labels (True). In the latter case, the models would tend to more strongly predict True with greater number of premises since all of them contain the same property phrase. Since all models seem to show below-chance compatibility with



(a)

(b)

(c)

(d)

**Figure 5.9.** Premise Nonmonotonicity Results. Average generalization scores (along with 95% CI) obtained by the models on general (a) and specific (c) arguments containing 2, and 3 premises. **(b) and (d):** Compatibility scores achieved by models fine-tuned from pre-trained and random initializations on general (b) and specific (d) arguments.

this phenomenon, their behavior is likely more compatible with the "string-label matching" explanation rather than one that is sensitive to category boundaries and coverage.

### Inclusion Fallacy

**Generation:** We reused the adaptation-generalization sets from the premise typicality stimuli, due to the close relation between the two phenomena. In particular, we paired each specific argument (generalization from a lower level concept to another lower level concept) with another argument with a general conclusion (to a higher level category). For example, the argument ROBIN → OSTRICH was paired with ROBIN → BIRD. As a result we end up having 120 specific arguments, compared against 120 general ones, for each of the 8 novel

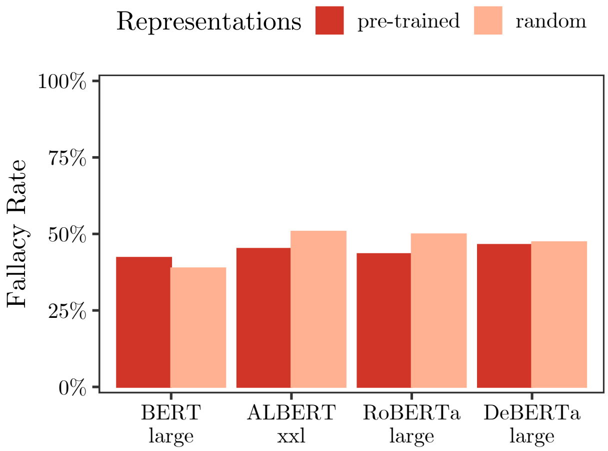


**Figure 5.10.** Percentage of time models (fine-tuned starting from pre-trained and random representations) demonstrate the inclusion fallacy effect.

properties.

**Results:** Figure 5.10 shows the percentage of cases where models demonstrated inclusion fallacy. For this phenomenon, we are only interested if models show the desired behavior or not—that is, we no longer have a notion of "performance" or "ground-truth" like in the previous analyses. Therefore, a model is compatible with this phenomenon just in case it demonstrates behavior similar to inclusion fallacy, where its generalization from a concept to a specific concept within a larger category is lower than that to the entire category. It is also important to note that this behavior is expected, and emerges as a natural property of the way we calculate generalization to a higher level category. That is, since we compute generalization score to say, BIRD, by averaging over the log-probabilities for the individual bird concepts, there are bound to be cases where some of generalization scores are lower than that of the averaged value. From Figure 5.10, we see that this is indeed observed for all the models we considered. While it is no surprise that the models demonstrate this behavior, this analysis can be treated as a sanity check for necessary behavior that a framework for inductive generalization *must* demonstrate.

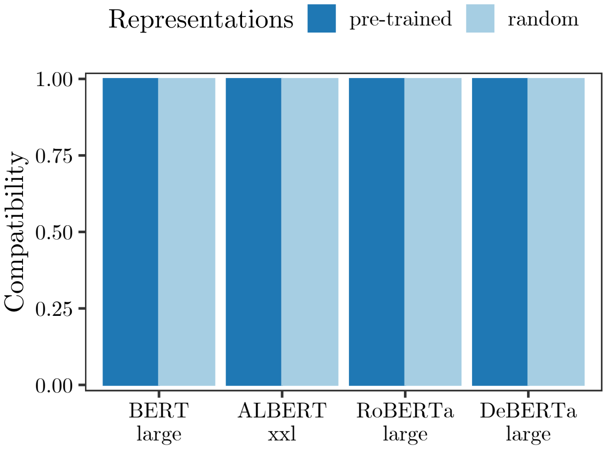


**Figure 5.11.** Compatibility scores achieved by models fine-tuned from pre-trained and random representations for showing premise-conclusion asymmetry effects.

### Premise Conclusion Asymmetry

**Generation:** We randomly sample 1000 adaptation and generalization sets (with each generalization set comprising of single concepts) from all possible pairs from our list of concepts. We then pair them each sampled adaptation-generalization instance with its reversed version—where the generalization concept becomes the adaptation concept, and vice-versa. In total, this gives us 2000 arguments across 8 novel properties.

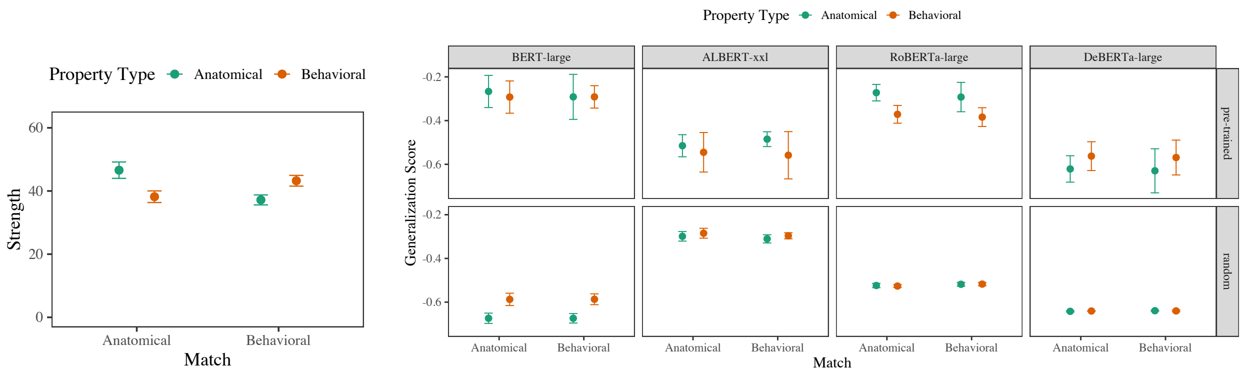
**Results:** For our analyses, we simply check if the generalization score from an adaptation concept to a generalization concept is *different* from the value obtained from its reversed counterpart. We expect this to be satisfied *all* the time in our framework, since it is highly unlikely for models to produce the same probability value for a label with changes to the surface form of the input. Figure 5.11 shows the compatibility of all our models for this phenomenon, and unsurprisingly, we see all models to perfectly satisfy it. Like the inclusion fallacy case, this too can serve as a direct sanity check for inductive generalization frameworks. Importantly models and frameworks that only consider similarity metrics like cosine similarity, to produce inductive generalization scores will be unable to demonstrate this phenomenon, as these metrics are by definition symmetric.

### 5.4.2 Context Sensitive Inductive Generalization

While the previous set of experiments and analyses dealt explicitly with inter-concept/category taxonomic relations by using properties that were essentially "novel" and ambiguous, we now turn to a more complicated inductive phenomenon that removes this assumption: *context specificity* (Gelman & Markman, 1986; Heit & Rubinstein, 1994; Shafto, Kemp, Baraff, Coley, et al., 2005; Kemp & Tenenbaum, 2009). This phenomenon questions the assumption that there is only one type of similarity that modulates induction, and instead proposes inductive behavior to be context-specific. Here, the conceptual content of the property may allude to specific contexts which give rise to different similarities (or other related measures) that guide inductive behavior. For instance, while participants projected anatomical properties (*has an omentum in its body*) along biological taxonomies (BEAR and WHALE), they would often project properties that describe behavior (*moves in a zig-zag motion*) to behaviorally related concepts (TUNA and DOLPHIN), often cutting across taxonomies.

Synthetically constructing stimuli that can test for context specificity in inductive generalization can be non-trivial, as coming up with property descriptions that elicit a particular type of context requires nuanced expertise. Therefore to test the extent to which the our models show context-specific inductive behavior, we rely on stimuli made available by Heit and Rubinstein (1994), one of the first works that discovered this behavior.

Specifically, the Heit and Rubinstein stimuli contain 14 pairs of concepts, 7 of which are related anatomically (and therefore taxonomically), and 7 of which are share specific salient behaviors. Importantly, for every anatomically related pair, there is a behaviorally related pair with the same conclusion concept (giving us minimal pairs). Heit and Rubinstein (1994) test human inductive behavior on 28 different properties, 14 of which are anatomical in nature, and the other 14 are behavioral. Upon asking 41 University of Michigan students to rate how likely the conclusion is given that the premise is true, they found that for anatomically matched concepts, the subjects made stronger generalizations for anatomical properties than for behavioral properties, and vice versa (see figure 5.12a). We borrowed these stimuli as is, and conducted our experiments with our eight property judgment models. Briefly, we computed generalization scores for all individual arguments, and then computed



(a)

(b)

**Figure 5.12.** Context Sensitivity Results. **(a):** Human inductive generalization behavior based on property type and match. **(b):** Average generalization scores (along with 95% CI) produced by the models on arguments with concepts that are either anatomically or behaviorally related and properties that are anatomical or behavioral in nature.

averages separately for arguments for anatomically and behaviorally matched concepts across both types of properties. Figure 5.12b shows these average generalization scores for all our models.

On comparing figure 5.12a and figure 5.12b, we observe that none of the models seem to demonstrate the kinds of contextual sensitivities clearly observed in humans, regardless of the nature of their initial representations (pre-trained vs. random). The models either showed no significant differences in generalizing anatomical vs. behavioral properties to anatomically/behaviorally matched concepts, or showed a marked preference toward one kind of property. For instance, while other models showed the former pattern, BERT-large (random) demonstrated a substantial preference towards behavioral properties, in that it on average was more likely to project behavioral properties across concepts than anatomical ones, regardless of the ways in which the concepts were related. While these results are from a study with a very small sample size (only 7 pairs per property type), it provides preliminary evidence that perhaps models may not reflect the flexibility/meta-reasoning ability required to demonstrate genuine contextual sensitivities during inductive generalization.

The above results suggest that despite the fact that the models we have used to conduct inductive generalization represent words as a function of their context (thus being context-sensitive by definition), they may not systematically distinguish high level contextual differences between properties (anatomical vs behavioral). One explanation could be that perhaps this only happens when models are being adapted—i.e., the models might still be representing non-trivial differences between concepts when paired with anatomical vs. behavioral properties, but not use them in making inductive generalizations. To test this, we tracked the cosine similarities between the vector representations of the premise and conclusion concepts at various layers of the model (without making any updates). Briefly, let $C_g$ be a conclusion concept, with anatomically matched concept $C_a$, and behaviorally matched concept $C_b$. Furthermore, let $P_a$ be an anatomical property and $P_b$ be a behavioral property, and $\texttt{extract}_l(c, p)$ be a function that extracts the representation of a concept $c$ when paired with the property $p$ from a model's $l^{th}$ layer. Then, if models are successfully representing higher level contextual differences in the interpretation of the concepts paired with the properties, we expect the following two behaviors in the similarities between the contextual representations of the concepts at layer $l$:

$$sim(\texttt{extract}_l(C_a, P_a), \texttt{extract}_l(C_g, P_a)) > sim(\texttt{extract}_l(C_b, P_a), \texttt{extract}_l(C_g, P_a)),$$

and

$$sim(\texttt{extract}_l(C_b, P_b), \texttt{extract}_l(C_g, P_b)) > sim(\texttt{extract}_l(C_a, P_b), \texttt{extract}_l(C_g, P_b)),$$

where $sim()$ is the cosine similarity between two vectors. That is, we expect similarity of an adaptation concept with the generalization concept to be high when the property they are paired to match their conceptual similarities (anatomical and behavioral) as compared to when the concept matches contradict the type of the property.

We calculated the proportion of cases where this behavior was observed across all layers, and used it as a version of compatibility, as estimated or made available by the models' representational space. Figure 5.13 shows our results for each layers, and also shows the analogous compatibility scores achieved by models using the inductive generalization frame-

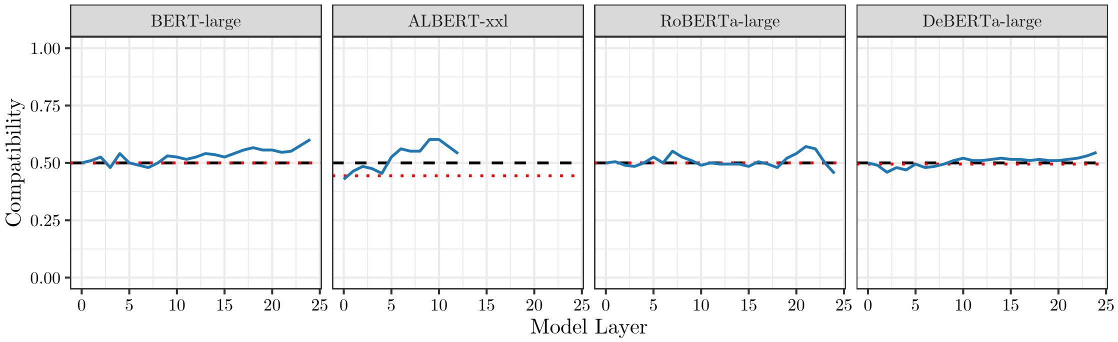


**Figure 5.13.** Compatibility computed based on contextual representations of the pre-trained models (blue) across their intermediate layers. Black dashed line indicates chance performance, while red dotted line indicates compatibility as a result of the inductive generalization experiments.

work. Note that all metrics here are from pre-trained models since models fine-tuned using random-vectors would often produce cosines that only differed in the 5th or 6th significant digit in their decimals, rendering little meaning to inter-representational differences. From Figure 5.13, we see that despite models showing no notion of context sensitivity when made to inductively generalize can still somewhat store high level differences between concept representations when paired with different types of properties. For instance in all three models, the higher layers tend to show non-trivially above chance compatibility, and this is largely observed for BERT, ALBERT, and RoBERTa. Overall, this suggests that in generalizing a novel property to other concepts, models may not fully harness the conceptual information made available at individual layers.

## 5.5 Discussion

Inductive generalization, or the ability to use one's background knowledge to make uncertain inferences about novel situations is an important consequence of acquiring conceptual knowledge, and fundamental to human semantic cognition (G. L. Murphy, 2002; Hayes & Heit, 2018). Humans engage in inductive reasoning about concepts and categories on a daily basis (Hayes & Heit, 2018), and often readily show a variety of systematic patterns that have revealed considerable insight into how they organize, use, and update their conceptual knowledge. Experimental studies into human inductive behavior typically provide human subjects with novel conceptual information and then test the extent to which they project/generalize this information to other known concepts/properties. In this chapter (and dissertation as a whole), I have proposed methods and analyses that use inductive generalization as a context where we can study language models, in order to conclude about how exposure to a word's distributional statistics can allow the learning and representation of conceptual meaning.

While the previous chapters have used inductive generalization as means to probe language models for conceptual knowledge and reasoning, they primarily did so by casting inductive reasoning as an instance of word-prediction in context, and largely focused on limited number of cases (either one phenomena or one type of inductive generalization). This chapter critically analyzes the word-prediction based account of inductive reasoning in language models, raising questions into its construct validity. In particular, using the conditional probability of a conclusion, given a premise, as a measure that arises due to reasoning can be inappropriate in principle, as it is more of a measure of how likely the two sentences are to be in a sequence rather than one derived from a genuine update of semantic knowledge. As a solution, we propose a novel analysis framework, inspired by the influential work of Rogers and McClelland (2004), that makes a more representational contact with the target model in simulating inductive generalizations. In particular, this framework operationalizes inductive generalization as the behavior of an LM after its representations have been partially exposed (via gradient-based updates) to novel conceptual information. To simulate this behavior, the framework uses LMs that are endowed with human-elicited property knowledge, by training them to evaluate the truth of sentences attributing properties to concepts. Using LMs trained exclusively to perform property judgment takes a positive step over the concerns raised in prior work and provides the models with the experimental context that matches that given to humans in cognitive psychology experiments.

Using this framework, we asked two related questions: (1) how do language models' representations drive their generalization behavior? and (2) what kind of conceptual abstractions are supported by language model representations? To study these questions, we primarily focused on 11 different systematic phenomena observed in human inductive reasoning literature (Osherson et al., 1990). In particular, I used the principles described by

Osherson et al. (1990) to generate stimuli that encoded these phenomena, each focusing on particular abstractions that humans might rely on in demonstrating inductive generalization. For example, the phenomena of conclusion specificity crucially tests for sensitivity to the hierarchical structure of category organization (generalization from a bird to other birds is stronger than to all animals).

**Table 5.3.** Summary of the models' performance at demonstrating the phenomena discussed by Osherson et al. (1990). 'PT' represents models that are fine-tuned starting from pre-trained representations, while 'R' represents those that start using random representations.

| **Phenomenon** | **Strong** | **Weak** | **BERT-large** | | **ALBERT-xxl** | | **RoBERTa-large** | | **DeBERTa-large** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | PT | R | PT | R | PT | R | PT | R |
| Premise Conclusion Similarity | ROBIN, SPARROW<br>↓<br>CANARY | ROBIN, SPARROW<br>↓<br>OSTRICH | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| Premise Conclusion Similarity (single) | LION<br>↓<br>CHEETAH | LION<br>↓<br>SKUNK | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| Premise Typicality | ROBIN<br>↓<br>BIRD | PENGUIN<br>↓<br>BIRD | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Premise Diversity (general) | ROBIN, PENGUIN<br>↓<br>BIRD | ROBIN, CANARY<br>↓<br>BIRD | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| Premise Diversity (specific) | LION, DEER<br>↓<br>RAT | LION, CHEETAH<br>↓<br>RAT | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| Conclusion Specificity | CROCODILE, IGUANA<br>↓<br>REPTILE | CROCODILE, IGUANA<br>↓<br>ANIMAL | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| Premise Monotonicity (general) | ROBIN<br>↓<br>BIRD | ROBIN, PENGUIN<br>↓<br>BIRD | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Premise Monotonicity (specific) | TUNA, SALMON<br>↓<br>SHARK | TUNA, SALMON, HERRING<br>↓<br>SHARK | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Premise Nonmonotonicity (general) | CROCODILE, IGUANA<br>↓<br>REPTILE | CROCODILE, IGUANA, EAGLE<br>↓<br>REPTILE | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Premise Nonmonotonicity (specific) | TUNA, SALMON<br>↓<br>SHARK | TUNA, SALMON, GIRAFFE<br>↓<br>SHARK | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Inclusion Fallacy | ROBIN<br>↓<br>BIRD | ROBIN<br>↓<br>OSTRICH | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Premise Conclusion Asymmetry | - | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Total Compatible** | - | - | 6 | 4 | 7 | 4 | 6 | 4 | **9** | 5 |

Using these stimuli, we analyzed four different pre-trained LMs: BERT-large (Devlin et al., 2019), ALBERT-xxl (Lan et al., 2020), RoBERTa-large (Liu et al., 2019), and DeBERTa-large (He, Gao, et al., 2021). We additionally also analyzed these models' randomly initialized counterparts. In all eight cases, the models were fine-tuned to predict the truth of sentences that attributed properties to concepts. Table 5.3 summarizes the results from these experiments, along with examples of strong and weak inductive arguments, as predicted by the principles of each phenomenon.

In general, we observe that our models are able to demonstrate some, but not all, of the Osherson et al. (1990), suggesting clear—but expected—differences with the patterns observed in humans, and by extension, also with the abstractions that they encode/rely on. Additionally, the phenomena for which we observed a general consensus among all the models (regardless of their initial state—pre-trained vs. random) were either expected for data-driven models (Premise Monotoncity), or were a natural consequence based on how our framework and generalization measurement is set up (Inclusion Fallacy and Premise Conclusion Asymmetry), or were cases of complete failure to encode the target phenomena (Premise Typicality and Premise Non-monotonicity). In all other cases (conclusion specificity, premise conclusion similarity, premise diversity), we observed models starting from a pre-trained representational space to show generally greater capability to demonstrate desirable behavior than their randomly initialized counterparts, indicating the clear advantages of learning distributional information through pre-training.

The DeBERTa-large model, with pre-trained initial state, stood out as the best model, showing non-trivial compatibility with 9 out of 12 unique inductive phenomena that we tested.[8] Interestingly, the same DeBERTa model also stood out in our property-judgment experiments, consistently outperforming BERT (6/12), ALBERT (7/12), and RoBERTa (6/12), on not only the unseen test instances of the property judgment task, but also on subsets of the COMPS-BASE stimuli (Misra et al., 2023; see also ch. 4), which were controlled for shallow similarity-based effects via the use of multiple different negative samples. Overall, this suggests that models that can robustly capture properties of everyday concepts can also be expected to show important inductive behavior.

Delving deeper into individual phenomena that were otherwise not predictable to be demonstrated by the models, we find all pre-trained models to successfully demonstrate sen-

[8] ↑general and specific arguments were counted separately, and similarity analyses contains an additional single-argument phenomenon tested post-hoc.

sitivity to hierarchical organization of conceptual structure. That is, models were more likely to project a novel property from a bird such as *robin* to all birds (more specific conclusion) than to all animals (more general conclusion). Since none of the random models were able to show this effect, this result provides strong evidence for pre-training (and by extension distributional learning) to impart an inductive preference to models with respect to hierarchical conceptual structure. Similarly, our pre-trained models were also able to show an effect for similarity between premise and conclusion concepts, but only for single arguments. Here, models were more likely to project a property to a concept that is similar to the adaptation/premise concept. Importantly this was strongest in the case of DeBERTa, again highlighting its relative mastery over conceptual knowledge. However, when considering cases with two adaptation concepts—where models had to combine the similarities of both the adaptation concepts to the conclusion concept—we only observed DeBERTa to show alignment with the phenomena. Similarly, premise diversity is an important phenomena that not only tests for hierarchical sensitivities, but also for precise and nuanced differences between the premise concepts themselves. Demonstrating this phenomenon requires reasoning about *coverage* (Osherson et al., 1990) of a category, where premise items that covered the category better (ROBIN, OSTRICH) would be more likely to generalize a novel property, and those that were too similar (ROBIN, CANARY) to each other would be less so. For general premise diversity arguments, we find ALBERT and DeBERTa show non-trivial compatibility whereas for specific arguments only DeBERTa showed this phenomenon. Overall, this suggests that while basic taxonomic information can be imparted, or made accessible through pre-training, diversity and complex similarity based phenomena might require a more robust encoding of everyday conceptual knowledge (i.e., generalize better outside of the models' training).

Turning to cases where models—at least the pre-trained ones—failed to capture the target phenomenon, we first begin with premise typicality. Every model that we tested using our inductive generalization framework seemed to have failed at demonstrating stronger generalizations for typical as opposed to atypical premises. At first glance this may seem counter-intuitive, since models are clearly sensitive to the taxonomic categories of items, and also generally capture their properties well, suggesting that they have enough information to discern which concepts were more typical/overall more similar to other concepts in the category. One possible explanation for this failure could be that the models never receive training signal for higher level categories. For instance, they are not trained on properties that are salient to being a bird—instead, all properties in the training data are reflected equally. As a result, their knowledge about penguins not being able to fly, something that is core to the concept of PENGUIN, is expected to be encoded in a manner that is similar to that about commonly known birds' ability to fly (which is not core to those birds). This lack of property salience could therefore be a limiting factor in learning graded conceptual structure in a manner similar to collective human beliefs. One potential solution to this issue could be to follow Bhatia and Richie (2021), and repeat the sentences expressing concept-property pairings as many times as in common property norms data sets—i.e., as many times as people had generated a given property for a given concept during data collection (also called production frequency).[9] Since their model is trained in a similar manner, it can easily be analyzed using the framework developed in this chapter, something we plan to do in the future. Another case where our models showed behavior opposite to that observed in humans was that concerning premise non-monotonicity—a phenomenon where a property that is shared by fewer category members is more strongly projected to all members of that category, as compared to where the property is shared by a larger number of concepts that do not all fall under the same category. For instance, BROWN BEAR → MAMMAL is stronger than BROWN BEAR, GRIZZLY BEAR → MAMMAL, since the second argument, despite having more premise concepts, is more compatible with the generalization that suggests that only bears have the property. In order to demonstrate this behavior, a reasoner must genuinely take into consideration these suspicious coincidences, which requires making assumptions about the sampling of the premise (Xu & Tenenbaum, 2007; Ransom et al., 2016). Instead, we see that models, having been updated using gradient descent algorithms seem to stick to their strong preference in matching property-phrase strings to the True label. This paves the way for future work where perhaps gradient-based updates to the representational space

[9] ↑Note that since we manually extended a popular property norms dataset(Devereux et al., 2014; see appendix C) to prevent training on incomplete data, we were not able to faithfully include their production frequency and therefore we exclusively trained the models on presence or absence of property

could be combined with communicative intent (Grice, 1989)—where relevance of the precise relationship between the premise concepts is taken into account, i.e., that it is perhaps intentional for the arguments to contain concepts with semantically salient shared properties (Medin et al., 2003).

Diverging from our analyses of the Osherson et al. (1990) phenomena, we also considered an important finding in the human inductive reasoning literature: *context sensitivity* (Heit & Rubinstein, 1994). Unlike in the case of Osherson et al. (1990), human behavior here is no longer a function of a single, vague type of similarity. Instead, inductive generalization from one concept to the other is dependent on the type of property being projected, and the type of intuitive theory it evokes (G. L. Murphy & Medin, 1985; G. L. Murphy, 1993; Kemp & Tenenbaum, 2009). For instance *moves in a zig-zag motion* is more likely to be shared by concepts such as TUNA and WHALE who may not be hierarchically similar but certainly share many behavioral properties, while *has a liver with two chambers* is likely to be shared by BEAR and WHALE, since both of them are mammals and are likely to share anatomical properties. On testing models on the stimuli made available by Heit and Rubinstein (1994), we found all our models to fail in displaying this nuanced type of behavior, and either did not make any distinctions based on the property type, or showed the opposite behavior. That is, they were not making precise high level distinctions between the property type and the semantic relations shared by concepts, needed to demonstrate this phenomenon. This was a counter-intuitive finding since these models are *well-known* for being context sensitive (Nair et al., 2020; Reif et al., 2019), at least at the word level—since they are able to represent similar usages of words in similar ways (e.g., *bat* as the mammal vs. sporting equipment). To get a better grasp around this issue, we also tracked the behavior of the intermediate representations of the adaptation and generalization concepts as produced by the models and found that a slight amount of high level context sensitivity is indeed present in the representations. This suggests that our inductive framework is unable to leverage this information in making its updates. Regardless, it is important to note that these conclusions are derived from an experiment that only contained 14 different concept pairs, and therefore, might not be conclusive. While it is non-trivial to come up with properties that evoke different types of intuitive theories, future work should ideally strive to build a more comprehensive analyses data that can test for a range of different intuitive theories. Additionally, future work should also perhaps harness the representational information—especially context sensitivity—in making inductive projections.

# 6. CONCLUSION and FUTURE WORK

Our ability to understand language and perform reasoning crucially relies on a robust system of semantic cognition (G. L. Murphy, 2002; Rogers & McClelland, 2004; Lake & Murphy, 2021): processes that allow us to learn, update, and produce inferences about everyday concepts (e.g., CAT, CHAIR), properties (e.g., *has fur, can be sat on*), categories (e.g., MAMMAL, FURNITURE), and relations (e.g., `is-a`, `taller-than`). Taking core ideas in semantic cognition research as guiding principles, this dissertation has laid down the groundwork to enable the evaluation of conceptual knowledge and understanding in pre-trained language models—perhaps the most sophisticated computational models that learn from statistics contained in large text corpora. In particular, I build on prior work that focuses on characterising what semantic knowledge is made available in the behavior and representations of language models, and extend it by additionally proposing tests that focus on functional consequences of acquiring basic semantic knowledge. To do so, I primarily focus on the interesting—and perhaps unique—ability of humans to rely on acquired conceptual knowledge to project and deploy novel information: *inductive generalization* (Osherson et al., 1990; Feeney & Heit, 2007; Kemp & Tenenbaum, 2009; Hayes & Heit, 2018), and use it as a context within which the conceptual understanding of language models can be characterized. Before doing so I additionally analyse the extent to which LMs can reflect *typicality effects* (Rosch, 1975)—graded measures of goodness of examples for everyday categories—and revisit the question of the extent to which LMs can robustly elicit everyday semantic knowledge (*cats have fur*) and deploy it by making inferences about novel concepts (another instance of inductive generalization). Finally, I develop a framework that can allow us to characterize the extent to which the distributed representations learned by LMs can encode principles and abstractions that characterize inductive behavior of humans. I now summarize the findings from these analyses.

## 6.1 Summary of Findings

The first context within which I analyzed conceptual understanding in LMs was that involving typicality effects in making category inferences. Typicality is an important functional consequence of human conceptual acquisition, and has bearing on how the human mind organizes knowledge about category structure (G. L. Murphy, 2002). A long-observed finding is that category-membership inferences are graded—i.e., certain members of a category (BIRD) are more typical (ROBIN, SPARROW) than are others (PENGUIN, OSTRICH), and this is reflected in almost every categorization experiment involving human subjects. Language models not only encode important prerequisite properties required to demonstrate typicality effects by construction—e.g., producing graded responses as a result of distributional learning—they also to a large extent map items to their categories with reasonable accuracy (Ettinger, 2020; Weir et al., 2020). To what extent can this ability of theirs account for *typicality effects*? **(RQ1)** Answering this question can broadly shed light on the extent to which the statistical associations in textual corpora can give rise to human-like conceptual structure.

In Ch. 3, I studied the behavior of 19 different LMs on two tasks in which humans showed clear sensitivities to typicality in processing of textual stimuli that varied only in terms of the concept: 1) verifying the truth of sentences expressing category membership (*a robin/penguin is a bird*) and 2) extending novel properties of objects to their superordinate category (*robins/penguins have ulnar arteries. therefore, all birds have ulnar arteries*)—an instance of inductive generalization. In both tasks, LMs demonstrated positive but moderate correlations (up to 0.42) with gold-standard human typicality ratings. Furthermore, LMs on average, also produced behavior that distinguished between high and low typicality category members in a manner that was qualitatively similar (but also relatively less extreme) to humans. Interestingly, correlation to human preference monotonically increased with the number of parameters in the LM, suggesting that building models of word-prediction that have greater expressivity in their representational space could lead to increased capacities in reflecting patterns of graded category-membership. At the same time, the rather modest magnitudes of the correlations indicate the difficulty of acquiring the specific nuances of human like conceptual organization from distributional signals in text.

Next in ch. 4, I revisit the topic of property knowledge (how well can a system store and retrieve properties of everyday concepts), a common context within which prior work has repeatedly investigated conceptual knowledge in LMs and other distributional semantic models. An important limitation that these prior works suffer from, however, is that they form their conclusions about how well LMs capture property knowledge in an isolated manner. That is, they focus only on the models' ability to predict properties for concepts that do indeed possess the properties, without necessarily considering the extent to which a property is deemed equally (or even more) likely given a concept that does not in fact possess the property. Furthermore, one desideratum developed in this dissertation is that LMs must go beyond just encoding static knowledge that they might as well have picked up from pre-training, and produce behavior that is compatible with this knowledge. One context that allows testing for such a scenario is *property inheritance* (Quillian, 1967; Smith & Estes, 1978; Sloman, 1993), a type of inductive generalization where concepts tend to inherit the properties of their superordinate concept. Demonstrating property inheritance over natural language text requires robust extraction of relational information (*a wug is a sparrow* → `is_a(wug, sparrow)`), implicit recall of commonsense knowledge (*sparrows can fly*), and perform simple inference (*a wug can fly*). Testing LMs for these abilities allows us to answer a key question: how robustly do they encode and transfer conceptual knowledge? **(RQ2.1 and 2.2)**

To this end, and to additionally circumvent the issue of drawing conclusions from isolated property judgments, I developed the COMPS (Conceptual Minimal Pairs) benchmark (Misra et al., 2023). COMPS is a collection of minimal pair sentence stimuli, each of which attributes properties (e.g., *can fly*) to: (i) real world concepts (ROBIN vs. PENGUIN), and (ii) their newly introduced subordinates (WUG), instantiated by supplying a prefix sentence that indicates taxonomic relations (e.g., *a wug is a robin/penguin*). By controlling for undesirable behavior such as sensitivity to rote memorization and distracting lexical associations, COMPS progressively isolates deeper understanding of concept properties, and penalizes models that rely on spurious heuristics. Analyses of 31 state of the art LMs spanning seven different model families, including GPT-3 and GPT-3.5, revealed that while they can correctly attribute properties to real-world concepts and show behavior consistent with property inheritance in simple settings, they struggle to do so robustly, and in particular show failures in presence of irrelevant distracting information. Importantly this failure is especially catastrophic when the distracting information is sequentially more recent—in that the performance of almost all LMs, regardless of expressivity or training data or objective is at or below chance level performance. For instance, on processing the stimuli "*a wug is a robin. a dax is a penguin*," models were more likely to attribute *can fly* to "dax" than to "wug," despite demonstrating desirable behavior on cases without distraction. This is a valuable finding, as it offers a potential explanation of LMs' failures on important semantic processing capacities, as shown by recent work (Ettinger, 2020; Ravichander et al., 2020; Elazar et al., 2021; Pandia & Ettinger, 2021; N. Kim, 2021; Schuster & Linzen, 2022; etc.): *LMs tend to rely on simple predictive cues in context over processing language compositionally, which results in their critical failures on controlled tasks that cannot be solved by simple predictive processing.* That is, LMs' reliance on heuristics that may have enabled better word prediction—lexical cues (Misra et al., 2020; Kassner & Schütze, 2020) and sequentially recent information (this work)—seems to trump robust information processing capacities critical to successful extraction of conceptual knowledge, leading to revealing failures where these heuristics do not lead to desirable behavior (McCoy, Pavlick, et al., 2019). In general, COMPS poses a severe challenge for popular state of the art LMs, many of which achieve below-chance performance, and can potentially serve as an important benchmark to tease apart robust conceptual understanding from shallow processing in language processing systems.

Finally in ch. 5, I turned to a more comprehensive and exclusive analysis of inductive generalization in LMs **(RQ3)**. In particular, I proposed and developed a framework that simulates inductive generalization in LMs by tasking them to project novel information about concepts and properties. This framework moves beyond the pure word-prediction based accounts of "reasoning" in language models, which could suffer from construct validity issues (see §5.1.2), and instead makes a more representational contact with the LMs, similar in spirit to work by Rogers and McClelland (2004). My framework satisfies two important desiderata of a computational model of inductive reasoning: (1) the ability to make judgments about the attribution of properties to concepts (experimental context provided in human experiments), and (2) the ability to accept and reflect new conceptual proposition and then be queried to assess generalization or 'projection' of this new knowledge. Specifically, the framework operationalizes inductive generalization as the behavior of an LM after its representations have been partially exposed (via gradient-based learning) to novel conceptual information. To simulate this behavior, the framework uses LMs that are endowed with human-elicited property knowledge, by training them to evaluate the truth of sentences attributing properties to concepts. For instance, if the inductive problem involved robins having the property *has T9 hormones*, and tested the extent to which canaries also have the property, then the underlying LM would first be adapted to attribute *have T9 hormones* to robins via backpropagation, after which its representations would be frozen and it would then be queried with "*canaries have T9 hormones*". Therefore, the framework makes a direct causal contact with the LMs' representational space—in that the projection of novel information is a causal consequence of its integration within the LMs' representations. This allows us to perform analyses that primarily target the extent to which an LM places novel information within its pre-constructed space, in a manner that is conducive to systematic regularities and abstractions that are observed in human inductive literature.

Using this framework I analysed the extent to which four different language models, typically used for classification tasks, can capture 12 different systematic regularities that humans readily show in deploying novel information, as documented by Osherson et al. (1990). Importantly, I performed these analyses with models that are fine-tuned to predict the truth value of sentences that attribute properties to concepts. Furthermore, I compared the inductive behavior of models that started from pre-trained information—thereby encoding principles of distributional semantics—versus those that start from a randomly initialized space (and are likely to overfit to the property-judgment training set). Results from these analyses highlight the clear benefits of pre-training in endowing models with important sensitivities such as that to the hierarchical organization of concepts, category coverage (via premise diversity), and conceptual similarity. For instance, models were more likely to project a novel property such as *can dax* or *has feps* from a ROBIN to BIRDS than to ANIMALS. However, these benefits were more strongly observed for the DeBERTa model (He, Gao, et al., 2021), which was able to capture 9 out of the 12 tested phenomena. Importantly, the DeBERTa model also consistently and substantially outperformed the other three models (BERT, ALBERT, RoBERTa), in correctly attributing to concepts, properties that were unseen during the fine-tuning task, and did so robustly (as measured using a subset of COMPS). Nevertheless, sensitivity to hierarchical organization was observed for models that started from a pre-trained representational space and was absent for all their counterparts that started from random initialization, providing evidence for a systematic relationship between distributional learning (via pre-training) and taxonomic organization.

Despite this partial observation of desirable inductive behavior, all models, regardless of their starting state (pre-trained vs. random), failed to show two important patterns: typicality and non-monotonicity.[1] First, with respect to *typicality effects* in making inductive generalizations, we find models to make no particular distinction between typical and atypical members, generalizing from both types of premises to all members of the category with nearly the same preference. That is, even if models can successfully encode sensitivity to category membership relations and even project properties based on individual conceptual similarities, they are unable to capture typicality effects. One potential explanation for this observation could be the fact that the properties they are trained on are reflected equally. That is, the models' knowledge about penguins not being able to fly, something that is core to what makes them atypical members of the BIRD category, is given the same weight as all the properties that are shared commonly across all birds. As a result, they have implicit evidence about what properties are salient to being a prototypical bird and those that are not that we are able to account for using the method applied in appendix C. Next, moving to failures of model on the *nonmonotonicity* phenomena we find them to stick to their proclivity of projecting a property more strongly as the number of premise concepts increases (i.e., always showing *premise monotonicity* effects). That is, they are more likely to project a property to all birds when robins, sparrows, and tigers share it as compared to when just robins and sparrows share it. This is opposite to humans who would override their monotonicity behavior and use the suspicious coincidence that the latter argument only had birds in the premise to rate it as the stronger inductive argument. The explanation for this behavior in models is simple—they are being adapted to map sentences to two labels: True or False, and could therefore be susceptible to string-label matching heuristics. During adaptation they observe more evidence for the property *can dax* to be associated with True in the weaker argument (3 times) than in the stronger argument (2 times). They can then use this information and map *can dax*-containing sentences to True with greater log-probability in the weaker argument,

[1] ↑these spanned three different types of stimuli, one for premise typicality and two for non-monotonicity.

which in this case has more number of premises. From our results it seems that this proclivity is likely driving their inability to reason about the concepts themselves, thereby leading to consistent failures. Importantly it is still unclear if the models are using this bias strictly, or there is still genuine scope of pragmatic behavior regarding suspicious coincidence and sampling assumptions (Xu & Tenenbaum, 2007; Ransom et al., 2016) encoded within the representations of these models. One potential way to further test for these cases would be to (1) include more bizarrely different concepts in the premise—e.g., robin, sparrow, screwdriver → bird vs. robin, sparrow → bird; or (2) test with more fine-grained category members (e.g., felines or different types of bears) where there is an even greater effect of suspicious coincidence (Medin et al., 2003).

Outside of the Osherson et al. (1990) phenomena, I additionally analyzed models on the extent to which they showed context sensitive induction—where different types of properties were projected differently. This phenomena arises specifically when the properties are no longer "blank" (as was the case for Osherson et al.), and inductive generalization is instead viewed as a process that first identifies the appropriate domain knowledge or intuitive theory (G. L. Murphy & Medin, 1985; G. L. Murphy, 1993, 2002) evoked by the context of the premise and the property being generalized, and then uses the features implied by the domain knowledge to carry out the generalization. For instance, anatomical properties such as *has a liver with multiple chambers* were more likely to be shared by taxonomically related concepts while a behavioral property such as *moves in a zig-zag motion* was instead shared between concepts that shared one or more properties related to the property being projected, often crossing taxonomic boundaries (Heit & Rubinstein, 1994). Testing models on a small set of stimuli (made available by Heit and Rubinstein (1994)) that targeted this phenomenon lead to further systematic failures where none of the models demonstrated the sensitivity. This suggested that LMs perhaps have yet to properly encode precise distinctions between property type in projecting novel information. However, a follow-up analysis of only the models' representation for concepts to which the novel properties were being attributed suggested some level of sensitivity at least at higher layers of the models, where the contextual similarity of concepts matched that of the property type (showing desirable behavior as much as 60% of time). Overall this demonstrated some evidence for models' to store and represent non-trivial distinctions between different kinds of properties and their interaction with a concept's representation when paired. The divergence between models' inductive behavior versus contextual representation content is expected, since the final output value of the model is a function of all intermediate representations at all layers, not just the ones that demonstrated the desirable sensitivity to property types. Future work could therefore explore using a more sophisticated update mechanism that integrates novel information in a manner that is proportional to the models' representation of context sensitivities at different layers.

## 6.2 Limitations

### Culture and linguistic assumptions in testing for typicality effects

One important limitation of the work presented here is that it specifically tests for conceptual knowledge using English stimuli, and additionally only tests models that were trained on corpora that overwhelmingly contained English. Importantly, the specific type of English that dominates these corpora is likely to be representative of Western, college educated adults. For instance, the GPT and EleutherAI family of models were trained on corpora containing Reddit,[2] the users of which are primarily from the U.S.[3] In sum, to the extent to which LMs indeed pick up on robust conceptual knowledge from the distributional statistics of words they have been trained on, their representations reflect a worldview that is likely to be aligned with the demography of the U.S. This is not an unreasonable guess, since scores of empirical work have shown LMs and other distributional models to pick up on socio-cultural norms, preferences, and harmful stereotypes present in their training data (Bolukbasi et al., 2016; Caliskan et al., 2017; Bender et al., 2021; Arora et al., 2023; Ramezani & Xu, 2023; i.a.). In ch. 3, these LMs were compared for typicality effects against human data that was collected in the 1970s (Rosch, 1975). Typicality effects, the presence of which is observed robustly, can still vary in their relative ordering of categories/concepts as the geographical and linguistic background of the human subjects changes (Schwanenflugel & Rey, 1986). This points to a natural limitation of our work in its failure to account for the gap between

[2] ↑https://reddit.com

[3] ↑According to `Semrush` (Semrush, 2023), 42.19% of reddit users are from the US, as of June 21, 2023

the socio-cultural statistics that would have been present if LMs were trained on data in mid to late 1900s versus those that are trained on more recent data.

### Testing conceptual knowledge and reasoning using a zero-shot setup

Using a zero-shot setup to test PLMs for human-like capacities such as property inheritance, as I have done in this dissertation and its associated papers—specifically Misra et al. (2021) and Misra et al. (2023)—has recently come under scrutiny. In particular, Lampinen (2022) argues that such a setup could be problematic because PLMs are trained to imitate the language produced by countless individuals with different beliefs, cultures, and behaviors. As a result, PLMs are likely to be handicapped in assigning sufficient probability mass to the desired family of continuations, given minimal prompts without any particular task-specific context. Instead, Lampinen (2022) suggests the need for PLMs

> "[...] to be guided into an experiment-appropriate behavioral context, analogously to the way cognitive researchers place humans in an experimental context, and orient them toward the task with instructions and examples."

This criticism is valid, and it is possible that models could overcome their lack of robustness to distraction effects by observing examples of our stimuli in context, though this has largely been shown in PLMs that are significantly larger than the ones I have tested here (Brown et al., 2020; Chowdhery et al., 2022; Wei, Tay, et al., 2022).[4] Indeed, recent work has demonstrated these larger PLMs to achieve strong performance on other types of reasoning—such as those required for solving math problems, reversing sequences, etc.—by priming models to produce additional textual content that represents intermediate reasoning steps and explanations (Nye et al., 2021; Wei, Wang, et al., 2022; Lampinen et al., 2022), in a few-shot setting.[5] While it is non-trivial, and uninformative to test for typicality effects using multiple in-context example, a few-shot version of COMPS stimuli could expose models

[4] ↑though see recent work by Shi et al. (2023), who show distraction effects in such large PLMs in solving arithmetic reasoning problems, even after using sophisticated in-context prompting methods such as Chain-of-Thought (Wei, Wang, et al., 2022), Least-to-Most (Zhou et al., 2022), and Self-Consistency (X. Wang et al., 2022).

[5] ↑See also Sinha et al. (2023), who analyze PLMs comparable in size to those studied in this work in a few-shot minimal-pair setting.

to the possibility of leveraging heuristics that are naturally absent in the zero-shot setup, and therefore such a setup would critically require the design of additional controls, which I leave for future work.

### Assumption of ideal reasoning behavior

Another limitation of our work (here, COMPS) is that it takes ideal and robust property inheritance behavior as the monolithic gold-standard for human cognition, something that recent work has cautioned against (Pavlick & Kwiatkowski, 2019; Dasgupta et al., 2022; Webson et al., 2023). Although we[6] relied on a database of concept-property pairs that were largely generated by human participants, whether or not humans will be robust to the types of distraction that were observed in PLMs is an open question and requires further investigation. However, notably we are not making direct comparisons between models and humans here—we argue that our primary contribution of controlled stimuli that tease apart shallow processing from robust conceptual reasoning in PLMs bears substantial merit that is independent from any comparisons between humans and computational systems. Furthermore, we emphasize that we are setting a reasonable—and to a certain extent, human-independent—desideratum in this work, which is that models should robustly capture ground-truth knowledge about everyday concepts and their properties and reflect this knowledge in their inferences about newly introduced concepts.

### On the limits of binary classification for inductive generalization

The analysis of inductive generalization in language models that I have carried out in this dissertation has primarily been conducted using a setting where the models are adapted to perform property judgments in a binary classification setting. While this design decision was made to ensure that the models' objectives are aligned with the experimental context presented in inductive reasoning experiments, is a bit of a radical change from their natural environment of learning distributions of words in context—instead of producing a output vector as large as their vocabulary, the models are constrained to instead produce a 2-

[6] ↑my co-authors and I

dimensional one (True vs. False). Additionally, this strict true vs. false classification also misses out on the saliency of each individual property where models are unable to encode the fact that having stripes on its body, for example, is salient to a zebra—this is largely an artifact of our manual extension of the original property norms dataset our analyses are based on, as opposed to a limitation of the framework or the binary classification setting (c.f. Bhatia & Richie, 2021). Therefore, even though we see clear benefits of pre-training, the binary classification setting may make it a non-zero possibility for models to not faithfully display their full range of conceptual understanding capacities that they might have actually gained from pre-training. Note that this is a limitation of our analyses and not the proposed framework. As mentioned in §5.2.3, the framework can be readily applied to models that stick to their word-prediction capacities, where instead of predicting the truth of sentences, they are trained to predict the properties given a concept[7] in a manner similar to COMET (Bosselut et al., 2019; Hwang et al., 2021).

## 6.3 Proposals for future work

Building on the the findings of this dissertation, as well as the limitations discussed above, there are two obvious routes to pursue in improving our conclusions about the extent to which computational models of distributional semantics and the "words as cues to meaning" perspective can encode important functional consequences of encoding conceptual knowledge (if at all). The first line of work is to conduct a more comprehensive analysis of typicality effects in LM behavior. Specifically, this would not only involve additional contexts and experimental data where typicality modulated human responses on textual stimuli (Kelly et al., 1986; Garrod & Sanford, 1977), it would also benefit from using LMs whose training corpora can be directly accessed. This is primarily because it will allow us to form hypotheses about the relationship between specific types of co-occurrence patterns and the extent to which they affect typicality effects in LMs. For instance, we could use hearst patterns (Hearst, 1992) to characterize the amount of explicit evidence that is available to LMs with respect to

[7] ↑this may lead to an intended consequence of models not being robust due to the lack of signal about what properties *should not* be predicted for concepts, but this could be mitigated by employing contrastive learning (Khosla et al., 2020) or unlikelihood training (Welleck et al., 2020).

taxonomic category memberships. Additionally, since one important limitation here was our failure to take geographical, linguistic, and socio-cultural factors that determine typicality effects (Schwanenflugel & Rey, 1986), another promising line of work is to perform analyses on separate mono-lingual LMs each trained separately on data from target languages, and rerunning the analysis presented and proposed here. The next line of work would be to extend COMPS to a few-shot scenario, thereby providing additional experimental context, as suggested by Lampinen (2022). However, simply taking random samples of COMPS stimuli as in-context examples still leaves models open to using other shallow heuristics that are orthogonal to the target reasoning behavior (property inheritance or category-feature inference). For instance, consider the following example of a possible conception of few-shot COMPS-WUGS-DIST stimuli:

> `Context:`
>
> ***ABC*** *is a robin. DEF is a penguin. Therefore,* ***ABC*** *can fly.*
>
> ***XYZ*** *is a horse. PQR is a dog. Therefore,* ***XYZ*** *has hooves.*
>
> `Test sample:`
>
> ***KLM*** *is a capybara. GHI is an elephant. Therefore, (****KLM****/GHI) has a trunk.*

An LM here could simply use the heuristic that it is always the first entity (ABC, XYZ) that possesses the target property and assign it greater probability. This would fail in the event the test sample defines a context where the second, and not the first entity is one to whom the property must be endowed. Similarly, another heuristic could be to assign the property to the most recent concept. A promising design decision, then is to create stimuli the few-shot contexts of which are compatible with principles of genuine reasoning as well as shallow heuristics, while the test items are only compatible with genuine reasoning behavior. Additional controls such as separating form vs. meaning, in a manner similar to N. Kim and Schuster (2023), where the in-context examples' mentions of the taxonomic context differs in surface-form to the test example (e.g., *is a* vs. *is a type of*), could further improve our ability to conduct robust tests into the models' reasoning behaviors.

Apart from the above two obvious avenues of future research, an important open problem that the present work highlights is inability of pre-trained models to flexibly adhere to different contexts/structures in making inductive generalizations. In the context of human inductive behavior, these context sensitivities arise from encoding of different "intuitive theories" (Carey, 1985; G. L. Murphy & Medin, 1985; G. L. Murphy, 1993) which govern how a novel property will be generalized. Computationally, these intuitive theories have been operationalized using graph structures (Kemp & Tenenbaum, 2003; Shafto, Kemp, Baraff, Coley, et al., 2005; Kemp & Tenenbaum, 2009), which explicitly define the relations between concepts, and thereby dictate extent to which a novel property would be shared by one or more concepts. However, an analysis of context-sensitive inductive generalization is non-trivial to conduct with real data, since it not only involves generation/collection of real properties that can be said to generalize in according to a particular intuitive theory, but also involves making assumptions of the models pre-training data—it is unclear whether the training data has sufficient signal to map tested properties, or their lexical content to the knowledge of the intended intuitive domain. The lack of information about the models training data can weaken the general conclusions we can make about their context sensitivity. As a solution, I propose to train and evaluate models on data that is generated synthetically, and contains the intended intuitive theory or structure that governs the basis of inductive inferences. Having full control of the experimental data will allow us to more robustly answer questions about the extent to which distributional learning over carefully designed semantic knowledge that explicates the relations between concepts—a form of indirect grounding (Merrill et al., 2021)—can successfully support generalization that is compatible with the structure reflected in the training data. This proposal has two main problems that we must solve: (1) the creation of synthetic datasets that reflect the targeted conceptual structure, and (2) a method to instil the target structure into the models, and measure success.

Analogues to both these problems have recently been addressed using a novel paradigm that aims to combine the principles of structured bayesian models with that of artificial neural networks, with a specific focus on language learning (McCoy & Griffiths, 2023). Specifically, bayesian models can reliably encode structural constraints (with a top-down approach) from very little data (relative to neural networks) but often struggle when it comes to performing inference over natural data (Griffiths et al., 2010). At the same time, neural networks learn structure bottom-up and require data that are several orders of magnitude larger, but are flexible enough to interact with any type of naturalistic data (McClelland et al., 2010). This paradigm combines the two by using meta-learning (Thrun & Pratt, 2012),[8] where models are trained on a series of tasks that bear high level similarities to the main task, after which it can learn the main task easily. Here, structured Bayesian models are constructed to encode the inductive biases and constraints of the target language, and then used to generate data for multiple virtual tasks. Then, a neural network is 'meta-trained' on these tasks to arrive at initialization that would rapidly learn a new target language—i.e., desirably encode its structural constraints.

Using meta-learning—specifically model agnostic meta-learning (MAML; Finn et al., 2017)—can certainly aid in solving problem 2, provided that problem 1 is solved: how should we design targeted Bayesian models that encode the principles of the intuitive theories we want to test for/also encode in our models? To address problem 1, I propose to use the influential work of Kemp and Tenenbaum (2008, 2009), which defines methods to encode complex conceptual structure (tree, ring, chain, arbitrary graph) into prior distributions, which can then straightforwardly guide inductive generalizations. For instance, a prior that encodes properties of a taxonomic structure, is likely to assign greater probability for concepts in the same taxonomic category to share properties. Importantly this prior can act as a generative model, and therefore allows samples to be drawn from it. For instance, we can draw a sample of 1000 different features for our n synthetic concepts, and use this matrix as a training data where a model receives a synthetic concept as input and predicts its known features (represented symbolically using verb-phrases), using a simple language modeling objective, similar to (Bosselut et al., 2019; Hwang et al., 2021). Importantly, we can draw multiple such training data samples as virtual tasks for ANNs (that underlie LMs) to be meta-trained using MAML (Finn et al., 2017). This process can then be repeated for any arbitrary structure, provided that we can easily define a prior that encodes the structure's properties (Kemp & Tenenbaum, 2008, 2009; Lake et al., 2018). Running our framework on models meta-trained using these synthetic data will allow us to conclude if they have successfully encoded the abstractions that align with the implicit structure of the data. Importantly,

[8] ↑a precursor to this work that does not use bayesian models is that by McCoy et al. (2020)

we can do so without relying on the limited amount of data provided by studies of human inductive behavior, since we will be dealing with synthetic data.

# REFERENCES

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? a case study in color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. https://openreview.net/forum?id=BJh6Ztuxl

Alishahi, A., Chrupaa, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, *25*(4), 543–557.

Arora, A., Kaffee, L.-a., & Augenstein, I. (2023). Probing pre-trained language models for cross-cultural differences in values. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 114–130. https://aclanthology.org/2023.c3nlp-1.12

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, *11*(4), 629.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207–219. https://doi.org/10.1162/coli_a_00422

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do neural machine translation models learn about morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–872.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bergey, C., Morris, B. C., & Yurovsky, D. (2020). Children hear more about what is atypical than what is typical. *CogSci 2020*, 501–507.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*(2-3), 150–177.

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

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5297715

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. https://doi.org/10.18653/v1/P19-1470

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2020). Inducing relational knowledge from bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 7456–7463.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Callan, J., Hoy, M., Yoo, C., & Zhao, L. (2009). Clueweb09 data set. https://lemurproject.org/clueweb12/

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* (pp. 553–624). Elsevier.

Chen, C., Lin, K., & Klein, D. (2021). Constructing taxonomies from pretrained language models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4687–4700. https://doi.org/10.18653/v1/2021.naacl-main.373

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 227–244. https://doi.org/10.18653/v1/2020.conll-1.17

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *International Conference on Learning Representations*. https://openreview.net/forum?id=r1xMH1BtvB

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136.

Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 1–12. https://doi.org/10.18653/v1/D19-6001

Dagan, I., Glickman, O., & Magnini, B. (2005). The pascal recognising textual entailment challenge. *Machine Learning Challenges Workshop*, 177–190.

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, *6*(4), 1–220.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Davis, F., & Van Schijndel, M. (2020). Interaction with context during recurrent neural network sentence processing. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2744–2750.

Davison, J., Feldman, J., & Rush, A. (2019). Commonsense knowledge mining from pretrained models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1173–1178. https://doi.org/10.18653/v1/D19-1109

Derby, S., Miller, P., & Devereux, B. (2021). Representation and pre-activation of lexical-semantic knowledge in neural language models. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 211–221. https://doi.org/10.18653/v1/2021.cmcl-1.25

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. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, *9*, 1012–1031. https://doi.org/10.1162/tacl_a_00410

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

Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, *9*(17), 1–63. http://dx.doi.org/10.3765/sp.9.17

Estes, W. K. (1994). *Classification and cognition*. Oxford University Press.

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 134–139.

Feeney, A. (2018). Forty years of progress on category-based inductive reasoning.

Feeney, A., & Heit, E. (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge University Press.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955*. Studies in linguistic analysis.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of NAACL-HLT 2019*, 32–42.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Garrod, S., & Sanford, A. (1977). Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 77–90.

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

Gokaslan, A., & Cohen, V. (2019). Openwebtext corpus.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205–1222.

Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30. https://dl.acm.org/doi/abs/10.1145/2509558.2509563?casa_token=mE3LH0NgZXYAAAAA:jV9pdGKqpOSLdftVM3UudHk0sa9nhH_xUspKq9oeBYEnQ9FK-yDUCenVi9ofiqGHqSL0eNnqVIgKvA

Graff, D., Kong, J., Chen, K., & Maeda, K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*, *4*(1), 34.

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 1–15.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. *Proceedings of NAACL-HLT 2018*, 1195–1205.

Gururangan, S., Marasovi, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

Hanna, M., & Mareek, D. (2021). Analyzing BERT's knowledge of hypernymy via prompting. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 275–282. https://doi.org/10.18653/v1/2021.blackboxnlp-1.20

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*(2), 242–256.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

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

Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley interdisciplinary reviews: Cognitive science*, *1*(2), 278–292.

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. https://aclanthology.org/C92-2082

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child development*, 1021–1034.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 411.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *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)*, 4129–4138.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

Holtzman, A., West, P., Shwartz, V., Choi, Y., & Zettlemoyer, L. (2021). Surface form competition: Why the highest probability answer isn't always right. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7038–7051. https://doi.org/10.18653/v1/2021.emnlp-main.564

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. https://doi.org/10.18653/v1/P18-1031

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725–1744. https://doi.org/10.18653/v1/2020.acl-main.158

Hupkes, D., Veldhoen, S., & Zuidema, W. (2018). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, *61*, 907–926.

Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2021). (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 6384–6392.

Jackendoff, R. (1983). *Semantics and cognition* (Vol. 8). MIT Press.

Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, *9*, 962–977.

Jurafsky, D., & Martin, J. H. (2020). *Speech & Language Processing, 3rd Edition.* https://web.stanford.edu/~jurafsky/slp3/

Kassner, N., & Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7811–7818. https://doi.org/10.18653/v1/2020.acl-main.698

Kelly, M. H., Bock, J. K., & Keil, F. C. (1986). Prototypicality in a linguistic context: Effects on sentence structure. *Journal of memory and language*, *25*(1), 59–74.

Kemp, C. (2011). Inductive reasoning about chimeric creatures. *Advances in neural information processing systems*, *24*, 316–324.

Kemp, C., & Jern, A. (2014). A taxonomy of inductive problems. *Psychonomic bulletin & review*, *21*(1), 23–46.

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, *33*, 18661–18673.

Kim, N. (2021). *Compositional linguistic generalization in artificial neural networks* (Doctoral dissertation). Johns Hopkins University.

Kim, N., Linzen, T., & Smolensky, P. (2022). Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv preprint*. https://arxiv.org/abs/2212.10769

Kim, N., & Schuster, S. (2023). Entity tracking in language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3835–3855. https://doi.org/10.18653/v1/2023.acl-long.213

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

Kim, S. J., Yu, L., & Ettinger, A. (2022). "no, they did not": Dialogue response dynamics in pre-trained language models. *Proceedings of the 29th International Conference on Computational Linguistics*, 863–874. https://aclanthology.org/2022.coling-1.72

Klafka, J., & Ettinger, A. (2020). Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4801–4811. https://doi.org/10.18653/v1/2020.acl-main.434

Lake, B. M., Lawrence, N. D., & Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive science*, *42*, 809–832.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. *CogSci 2015*, 1243–1248.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., & Hill, F. (2022). Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Lampinen, A. K. (2022). Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *arXiv preprint arXiv:2210.15303*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, *20*(1), 1–31.

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143

Li, N., Bouraoui, Z., Camacho-Collados, J., Espinosa-Anke, L., Gu, Q., & Schockaert, S. (2021). Modelling general properties of nouns by selectively averaging contextualised embeddings [Main Track]. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 3850–3856). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/530

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5210–5217. https://doi.org/10.18653/v1/2020.acl-main.465

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

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

Lovering, C., & Pavlick, E. (2022). Unit testing for concepts in neural networks. *Transactions of the Association for Computational Linguistics*, *10*, 1193–1208.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *Proceedings of the First Workshop on Language Grounding for Robotics*, 76–85. https://doi.org/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.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Margolis, E., Laurence, S., et al. (1999). *Concepts: Core readings*. Mit Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348–356.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310–322.

McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science*, *2*(6), 387–395.

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

McCoy, R. T., & Griffiths, T. L. (2023). Modeling rapid language learning by distilling bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701*.

McCoy, R. T., Linzen, T., Dunbar, E., & Smolensky, P. (2019). RNNs implicitly implement tensor-product representations. *International Conference on Learning Representations*. https://openreview.net/forum?id=BJx0sjC5FX

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. https://doi.org/10.18653/v1/P19-1334

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., … Perez, E. (2023). Inverse scaling: When bigger isn't better.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517–532.

Merrill, W., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, *9*, 1047–1060. https://doi.org/10.1162/tacl_a_00412

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

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

Misra, K. (2020). *Exploring lexical sensitivities in word prediction models: A case study on bert* (Master's thesis). Purdue University Graduate School.

Misra, K. (2022). Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's sensitivity to lexical cues using tests from semantic priming. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4625–4635. https://doi.org/10.18653/v1/2020.findings-emnlp.415

Misra, K., Ettinger, A., & Rayz, J. (2021). Do language models learn typicality judgments from text? *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. https://escholarship.org/uc/item/9n77r9mr

Misra, K., Rayz, J., & Ettinger, A. (2023). COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2920–2941. https://aclanthology.org/2023.eacl-main.213

Misra, K., Rayz, J. T., & Ettinger, A. (2022). A property induction framework for neural language models. *Proceedings of the 44th 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.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic bulletin & review*, *23*, 1035–1042.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Murphy, M. L. (2010). *Lexical meaning*. Cambridge University Press.

Nagel, S. (2016). CC-News. http://web.archive.org/save/http://commoncrawl.org/2016/10/newsdataset-available.

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized word embeddings encode aspects of human-like word sense knowledge. *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, 129–141. https://aclanthology.org/2020.cogalex-1.16

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. MIT Press.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664. https://doi.org/10.18653/v1/P19-1459

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

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

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Pandia, L., Cong, Y., & Ettinger, A. (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 367–379. https://doi.org/10.18653/v1/2021.conll-1.29

Pandia, L., & Ettinger, A. (2021). Sorting through the noise: Testing robustness of information processing in pre-trained language models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1583–1596. https://doi.org/10.18653/v1/2021.emnlp-main.119

Parrish, A., Schuster, S., Warstadt, A., Agha, O., Lee, S.-H., Zhao, Z., Bowman, S. R., & Linzen, T. (2021). NOPE: A corpus of naturally-occurring presuppositions in English. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 349–366. https://doi.org/10.18653/v1/2021.conll-1.28

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*. https://openreview.net/forum?id=gJcEM8sxHK

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*, 447–471.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220041.

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, *7*, 677–694.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of EMNLP 2014*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Porada, I., Suleman, K., Trischler, A., & Cheung, J. C. K. (2021). Modeling event plausibility with consistent conceptual abstraction. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1732–1743. https://doi.org/10.18653/v1/2021.naacl-main.138

Prasad, G., van Schijndel, M., & Linzen, T. (2019). Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 66–76.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Quine, W. V. (1951). Main trends in recent philosophy: Two dogmas of empiricism. *The philosophical review*, *60*(1), 20.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. https://doi.org/10.18653/v1/D16-1264

Ramezani, A., & Xu, Y. (2023). Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796.

Ravichander, A., Belinkov, Y., & Hovy, E. (2021). Probing the probing paradigm: Does probing accuracy entail task relevance? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3363–3377. https://aclanthology.org/2021.eacl-main.295

Ravichander, A., Hovy, E., Suleman, K., Trischler, A., & Cheung, J. C. K. (2020). On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, 88–102. https://aclanthology.org/2020.starsem-1.10

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., & Kim, B. (2019). Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, *32*.

Renner, J., Denis, P., Gilleron, R., & Brunellière, A. (2023). Exploring category structure with contextual language models and lexical semantic networks. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2277–2290. https://aclanthology.org/2023.eacl-main.167

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665–681.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, *12*(1), 1–20.

Rips, L. J., Smith, E. E., & Medin, D. L. (2012). *Concepts and categories: Memory, meaning, and metaphysics*. Oxford University Press.

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

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689–714.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language* (pp. 111–144). Elsevier.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rosch, E., & Lloyd, B. B. (1978). Cognition and categorization.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*(4), 491.

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 726–730. https://doi.org/10.3115/v1/P15-2119

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

Safavi, T., & Koutra, D. (2021). Relational World Knowledge Representation in Contextual Language Models: A Review. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1053–1067. https://doi.org/10.18653/v1/2021.emnlp-main.81

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2699–2712. https://doi.org/10.18653/v1/2020.acl-main.240

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Schuster, S., & Linzen, T. (2022). When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 969–982. https://doi.org/10.18653/v1/2022.naacl-main.71

Schwanenflugel, P. J., & Rey, M. (1986). The relationship between category typicality and concept familiarity: Evidence from spanish-and english-speaking monolinguals. *Memory & Cognition*, *14*, 150–163.

Semrush. (2023). Reddit.com website traffic, ranking, analytics [may 2023]. https://www.semrush.com/website/reddit.com/overview/

Shafto, P., Kemp, C., Baraff, E., Coley, J., & Tenenbaum, J. B. (2005). Context-sensitive induction. *Proceedings of the 27th annual conference of the cognitive science society*, 2003–2008.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., & Tenenbaum, J. B. (2005). Context-sensitive induction. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. *International Conference on Machine Learning*, 31210–31227.

Shwartz, V., & Choi, Y. (2020). Do neural language models overcome reporting bias? *Proceedings of the 28th International Conference on Computational Linguistics*, 6863–6870. https://doi.org/10.18653/v1/2020.coling-main.605

Sinha, K., Gauthier, J., Mueller, A., Misra, K., Fuentes, K., Levy, R., & Williams, A. (2023). Language model acceptability judgements are not always robust to context. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6043–6063. https://doi.org/10.18653/v1/2023.acl-long.333

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

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Smith, E. E., & Estes, W. K. (1978). Theories of semantic memory. *Handbook of learning and cognitive processes*, *6*, 1–56.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts* (Vol. 9). Harvard University Press Cambridge, MA.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(1), 1–23.

Smolensky, P. (1995). Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. *Connectionism: Debates on Psychological Explanation*, *2*, 221–290.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Sommerauer, P. (2022). Diagnosing semantic properties in distributional representations of word meaning.

Sommerauer, P., & Fokkens, A. (2018). Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. https://doi.org/10.18653/v1/W18-5430

Spärck-Jones, K. (1964). *Synonymy and semantic classification*. Edinburgh University Press.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

Taylor, W. L. (1953). cloze procedure: A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 167–204). Cambridge University Press.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *7th International Conference on Learning Representations, ICLR 2019*.

Thrun, S., & Pratt, L. (2012). *Learning to learn*. Springer Science & Business Media.

Toneva, M. (2021). *Bridging language in machines with language in the brain* (Doctoral dissertation). Carnegie Mellon University.

Trinh, T. H., & Le, Q. V. (2018). A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4704–4710.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008. https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, A., & Cho, K. (2019). BERT has a mouth, and it must speak: BERT as a Markov random field language model. *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 30–36. https://doi.org/10.18653/v1/W19-2304

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 3266–3280.

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. *International Conference on Learning Representations*.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://doi.org/10.1162/tacl_a_00321

Webson, A., Loo, A. M., Yu, Q., & Pavlick, E. (2023). Are language models worse than humans at following prompts? it's complicated. *arXiv preprint arXiv:2301.07085*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2020). Neural text generation with unlikelihood training. *International Conference on Learning Representations*. https://openreview.net/forum?id=SJeYe0NtvH

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Wittgenstein, L. (1953). *Philosophical investigations*. John Wiley & Sons.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP 2020: Demos*, 38–45.

Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. *32nd Annual Meeting of the Association for Computational Linguistics*, 133–138. https://doi.org/10.3115/981732.981751

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yu, L., & Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4896–4907. https://doi.org/10.18653/v1/2020.emnlp-main.397

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bousquet, O., Le, Q., & Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, 19–27.

# A. SUPPLEMENTARY MATERIAL FOR Misra et al. (2021)

**Table A.1.** Superordinate and their 5 most and least typical subordinate concepts, as rated by native English speakers, extracted from Rosch (1975).

| **Superordinate** | **Subordinate** | |
|---|---|---|
| | **Most Typical** | **Least Typical** |
| TOY | DOLL, TOP, JACK-IN-THE-BOX, TOY SOLDIER, YO-YO | HORSE, GUN, ANIMALS, TENNIS RACKET, BOOK |
| BIRD | ROBIN, SPARROW, BLUE-JAY, BLUEBIRD, CANARY | OSTRICH, TITMOUSE, EMU, PENGUIN, BAT |
| SPORT | FOOTBALL, BASEBALL, BASKETBALL, TENNIS, SOFTBALL | CHESS, DANCING, CHECKERS, CARDS, SUNBATHING |
| VEGETABLE | PEA, CARROT, GREEN BEAN, STRING BEAN, SPINACH | SEAWEED, GARLIC, DANDELION, PEANUT, RICE |
| TOOL | SAW, HAMMER, RULER, SCREWDRIVER, DRILL | ANVIL, HATCHET, RAG, SCISSOR, CRANE |
| FRUIT | ORANGE, APPLE, BANANA, PEACH, PEAR | NUT, GOURD, OLIVE, PICKLE, SQUASH |
| CLOTHING | PANT, SHIRT, DRESS, SKIRT, BLOUSE | WATCH, CUFF LINK, NECKLACE, BRACELET, CANE |
| VEHICLE | AUTOMOBILE, STATION WAGON, TRUCK, CAR, BUS | SKI, SKATEBOARD, WHEELBARROW, SURFBOARD, ELEVATOR |
| FURNITURE | CHAIR, SOFA, COUCH, TABLE, EASY CHAIR | CLOSET, VASE, ASHTRAY, FAN, TELEPHONE |
| WEAPON | GUN, PISTOL, REVOLVER, MACHINE GUN, RIFLE | FOOT, CAR, SCREWDRIVER, GLASS, SHOE |

# B. SUPPLEMENTARY MATERIAL FOR Ch. 4 AND Misra et al. (2023)

**Table B.1.** Summary of the 22 models that we evaluate for chapter.
**Legend for Corpora:** Wiki: Wikipedia; bc: BookCorpus (Zhu et al., 2015); cw: ClueWeb (Callan et al., 2009); cc: CommonCrawl Giga: Gigaword (Graff et al., 2003); owtc: OpenWebTextCorpus (Gokaslan & Cohen, 2019); cc-news: CommonCrawl News (Nagel, 2016); Stories: Stories corpus (Trinh & Le, 2018); WebText: WebText corpus (Radford et al., 2019); Pile: The Pile (Gao et al., 2020).

| Family | Model (Abbrev.) | Parameters | Vocab Size | Tokenization | Corpora | Tokens |
|---|---|---|---|---|---|---|
| ALBERT | `albert-base-v2` (A-b) | 11M | 30,000 | SentencePiece | Wiki and bc | 3.3B |
| | `albert-large-v2` (A-l) | 17M | | | | |
| | `albert-xlarge-v2` (A-xl) | 59M | | | | |
| | `albert-xxlarge-v2` (A-xxl) | 206M | | | | |
| BERT | `distilbertbase-uncased` (dB-b) | 67M | 30,522 | WordPiece | Wiki and bc | 3.3B |
| | `bert-base-uncased` (B-b) | 110M | | | | |
| | `bert-large-uncased` (B-l) | 345M | | | | |
| ELECTRA | `electra-small` (E-s) | 13M | 30,522 | WordPiece | Wiki and bc | 3.3B |
| | `electra-base` (E-b) | 34M | | | | |
| | `electra-large` (E-l) | 51M | | | Wiki, bc, cw, cc, and Giga | 33B |
| RoBERTa | `distilroberta-base` (dR-b) | 82M | 50,265 | Byte-pair encoding | owtc | 2B |
| | `roberta-base` (R-b) | 124M | | | bc, cc-news, owtc, and Stories | 30B |
| | `roberta-large` (R-l) | 355M | | | | |
| GPT2 | `distilgpt2` (dGPT2) | 82M | 50,257 | Byte-pair encoding | owtc | 2B |
| | `gpt2` (GPT2) | 124M | 50,257 | | WebText | 8B* |
| | `gpt2-medium` (GPT2-m) | 355M | | | | |
| | `gpt2-large` (GPT2-l) | 774M | | | | |
| | `gpt2-xl` (GPT2-xl) | 1.5B | | | | |
| EleutherAI | `gpt-neo-125M` (Neo-125M) | 125M | 50,257 | Byte-pair encoding | pile | 300B |
| | `gpt-neo-1.3B` (Neo-1.3B) | 1.3B | | | | 380B |
| | `gpt-neo-2.7B` (Neo-2.7B) | 2.7B | | | | 420B |
| | `gpt-j-6B` (GPT-J) | 6B | | | | 402B |

*As estimated by Warstadt et al. (2020).

# C. SUPPLEMENTARY MATERIAL FOR Ch. 5 AND Misra et al. (2022)

## C.1 Property Knowledge Re-annotation

### C.1.1 Premise

Datasets such as the CSLB (Devereux et al., 2014) naturally lend themselves to investigations that probe the conceptual knowledge of computational models and their representations. The CSLB dataset was collected by tasking 123 human participants to generate properties of a total of 638 concepts. For each property the authors then calculated its production frequency for all concepts for which it was generated, i.e., if the property *can fly* was generated for the concept ROBIN by 20 out of the 30 participants who were shown the concept, then its production frequency is 20. Note that the CSLB data set contains only positive property-concept associations. To construct negative samples, prior works that use CSLB as ground-truth to probe word representations typically use the set of concepts for which a given property was not generated, as negative (e.g. Lucy & Gauthier, 2017; Forbes et al., 2019; Da & Kasai, 2019; Bhatia & Richie, 2021). That is, negative samples are usually generated using concepts that have a production frequency of 0 for each property. Once a sufficient number of negative samples have been generated, the authors then train a probing classifier for every property, which predicts 1 if the production frequency of the property for that concept is nonzero, and 0 otherwise.

### C.1.2 Limitation

Since the task that was employed to construct the CSLB dataset was that of generation as opposed to validation, it is possible—and perhaps likely—that it resulted in inconsistent annotations, where some humans might have forgotten to generate *obvious* properties for certain concepts, or simply ignored them. For instance, the property *can breathe*, which is obviously applicable for all animals, was missing in 146 animal concepts within the dataset. This means that if one were to follow the standard negative-sampling method described earlier, they would consider all 146 of these animals as concepts for which the property *can* *breathe* does not hold true, which is incorrect. We conjecture that humans fail to generate features that are *obviously valid* for certain concepts (e.g., *can breathe, can grow, is a living thing* for animals) because they may be operating under Grice's maxim of quantity (Grice, 1989), by only eliciting non-trivial or *truly* informative properties for concepts in order to avoid redundancy. While we leave the testing of the hypotheses within this conjecture for future work, this limitation of incomplete data raises questions about the extent to which we should trust the results and conclusions of prior work which are crucially affected by this problem, which we summarize using the aphorism: *absence of evidence is not evidence of absence.*

### C.1.3 Manual re-annotation of missing property-concept pairs

To mitigate the limitation discussed above, we first selected the categories (hand-annotated by Devereux et al., 2014; e.g., BIRD, VEHICLE, TREE, etc.) that had at least 9 concepts in the dataset and were not labeled as "miscellaneous," resulting in 23 different categories with a total of 529 unique noun concepts, and 4,970 unique properties. Next, we manually removed concepts and properties that contained proper nouns (e.g., ROLLS-ROYCE, *is in Harry Potter*), stereotypical or subjective data (e.g., *is meant for girls*, *is ugly*), and explicit mentions of similarity or relatedness (e.g., *is similar to horse*). We further normalized properties that were paraphrases of each other (e.g., *is used to flavor, does flavor* → *is used to flavor*). This resulted in 521 concepts and 3,643 properties. Again through manual search, we further identified a total of 726 properties that were incompletely annotated (i.e., those that were associated with certain concepts but were omitted for many relevant concepts during data collection—e.g., the property *can grow* was missing for all invertebrates, despite being associated with all of them). We manually extended the coverage for these properties by adding in entries for concepts for which they had not been elicited. For instance, for the property *can breathe*, which was generated for 6 out of 152 animals in the original dataset, we further add the remaining 146 concepts as additional positively associated concepts, increasing its coverage from 6 to 152. While the total number of incompletely annotated properties is small (10% of the valid properties), our re-annotation process greatly increases the total number of concept-property pairs (from 13,355 pairs in the original, unmodified dataset, to 30,076: an increase of 125%) since many of the incompletely labeled properties were applicable across several categories (e.g., *has a mouth, can grow,* etc). After applying this process to the CSLB dataset, we are left with 30,076 property-concept pairs, which we use in subsequent experiments. The re-annotated data can be found in the file `post_annotation_all.csv`[1] in the github repository.

### C.1.4 Final thoughts

The re-annotation process described above was performed manually due to resource, time, and financial constraints. However, we recommend running a large-scale empirical validation studies for datasets such as CSLB and McRae, before using them for probing experiments. While this is non-ideal in terms of resource use, it is necessary in order to draw faithful and appropriate conclusions about the correspondence between conceptual knowledge in humans and machines.

## C.2 Negative Sample generation using Taxonomies

Here we describe our algorithm to generate negative samples for our first experiment in the paper—the property judgment task, where LMs are fine-tuned to classify as True or False sentences that attribute properties to concepts. For instance, the sentence *a cat can fly* is labeled as False as CAT is a negative sample for the property *can fly*, whereas, *a robin can fly* is labeled as True. Briefly, for the set of positive samples for a given property, we sample an equal-sized set of negative samples that are maximally similar to the positive samples. We use a taxonomic similarity (described below) as our similarity measure as it is model-free. Below we describe useful notation involved in the process, and then describe the full algorithm.

[1] ↑https://github.com/kanishkamisra/lm-induction/data/post_annotation_all.csv

### C.2.1 Notation and Preliminaries

Table C.1 describes the notation we follow to construct our property judgment dataset. Our goal here is to generate 30,076 negative samples and then take the entire set of 60,152 concept-property pairs and their labels to carry out the property-judgment experiment.

**Table C.1.** Notation for various artifacts involved in the paper.

| Notation | Meaning | Remarks |
|---|---|---|
| $\mathcal{C}$ | The set of all concepts in our experiments. These are also at the lowest level of the taxonomy—i.e., its leaf nodes. | $\lvert\mathcal{C}\rvert = 521$ |
| $\mathcal{P}$ | The set of all unique properties used in our experiments. | $\lvert\mathcal{P}\rvert = 3735$ |
| $\mathcal{Q}_{P_i}$ | The set of concepts that possess the property $P_i$. | $\mathcal{Q}_{P_i} \subset \mathcal{C}, \lvert\mathcal{Q}_{P_i}\rvert = k$ |
| $\neg\mathcal{Q}_{P_i}$ | The set of concepts that do not possess the property $P_i$, i.e., $\neg\mathcal{Q}_{P_i} = \mathcal{C} - \mathcal{Q}_{P_i}$ | $\lvert\neg\mathcal{Q}_{P_i}\rvert = 521 - k$ |
| $\delta(\neg\mathcal{Q}_{P_i}, k)$ | A function that extracts $k$ negative samples from $\neg\mathcal{Q}_{P_i}$ using the method described in Algorithm 1 (lines 6–9). | $\lvert\delta(\neg\mathcal{Q}_{P_i}, k)\rvert = k,$ |

In order to generate negative samples, we first tag the senses of all our 521 concepts using the WordNet (Miller, 1995) taxonomy, and also retrieve the sub-tree from WordNet that perfectly contains our concepts and use this as our ground-truth taxonomy on the basis of which we carry out subsequent experiments. We generate our negative samples by choosing a measure derived primarily from the Wu-Palmer similarity (Wu & Palmer, 1994). This similarity can be computed over any taxonomy using the following operations:

$$sim_{\mathsf{wup}}(c_i, c_j) = \frac{2 \times \mathtt{depth}(\mathtt{lcs}(c_i, c_j))}{\mathtt{depth}(c_i) + \mathtt{depth}(c_j)}, \tag{C.1}$$

where $\mathtt{lcs}(x_1, x_2)$ is a function that computes the least-common subsumer[2] of the two[3] concepts, and $\mathtt{depth}(x)$ computes the length of the path between the input concept and

[2] ↑a node in the hierarchy that is a hypernym/parent of the input concepts with minimum depth. For instance, $\mathtt{lcs}(\text{ROBIN}, \text{BAT}) = \text{VERTEBRATE}$.

[3] ↑although in practice it can be applied for multiple concepts.

the root node of the hierarchy. We consider a generalized form of this measure (denoted as $sim_{\mathsf{gwup}}$), to compute the similarity of a single concept to a set of concepts:

$$sim_{\mathsf{gwup}}(c_1, \dots, c_n) = \frac{n \times \mathsf{depth}(\mathtt{lcs}(c_1, \dots, c_n))}{\mathsf{depth}(c_1) + \cdots + \mathsf{depth}(c_n)} \tag{C.2}$$

For every property $P_i$, we use this measure in algorithm 1 to sample $k$ concepts from $\neg\mathcal{Q}_{P_i}$, based on their $sim_{gwup}$ with $\mathcal{Q}_{P_i} = \{c_1, \dots, c_k\}$. For example, consider the property *has striped patterns on its body*, the corresponding artifacts would be:

$$\begin{aligned} \mathcal{Q} &= \{\textsc{zebra}, \textsc{tiger}, \textsc{bee}, \textsc{wasp}\} \\ \neg\mathcal{Q} &= \mathcal{C} - \mathcal{Q} \\ &= \{\textsc{accordion}, \dots, \textsc{yo-yo}\} \\ \text{NS} = \delta(\neg\mathcal{Q}, 4) &= \{\textsc{horse}, \textsc{lion}, \textsc{ant}, \textsc{beetle}\} \end{aligned}$$

$$\begin{aligned} \mathcal{D} = \{&[\textit{a zebra has striped patterns on its body}, \mathsf{True}], \\ &\dots, \\ &[\textit{a beetle has striped patterns on its body}, \mathsf{False}]\} \end{aligned}$$

Note that we follow the method outlined by Bhatia and Richie (2021) to convert concept-property pairs into sentences, which we denote as *sentencizer*() in Algorithm 1.

**Algorithm 1** Algorithm to generate the dataset, $\mathcal{D}$, for the property judgment task

```
Input: $\mathcal{C} = \{c_1, \dots, c_n\}$: Set of all concepts, $n = 521$.
       $\mathcal{P} = \{P_1, \dots, P_m\}$: Set of all properties, $m = 3643$.
1.  $\mathcal{D} \leftarrow [\,]$ $\triangleright$ the final set of stimuli for the property judgment task.
2.  for $i = 1, \dots, m$ :
3.      $\mathcal{Q}_{P_i} \leftarrow [c_1, \dots, c_k]$ $\triangleright$ set of $k$ concepts that possess the property $P_i$
4.      $\neg\mathcal{Q}_{P_i} \leftarrow \mathcal{C} - \mathcal{Q}_{P_i}$
5.      $\triangleright$ Lines 6–9 compute $\delta(\neg\mathcal{Q}_{P_i}, k)$
6.      $\text{NS}_{P_i} \leftarrow [\,]$ $\triangleright$ set of negative samples for the property $P_i$
7.      $\neg\tilde{\mathcal{Q}}_{P_i} \leftarrow \text{argsort}(\neg\mathcal{Q}_{P_i}, sim_{\mathsf{gwup}})$ $\triangleright$ sort $\neg\mathcal{Q}_{P_i}$ based on $sim_{\mathsf{gwup}}(c_1, \dots, c_k, x_j)\, \forall x_j \in \neg\mathcal{Q}_{P_i}$
8.      for $j = 1, \dots, k$ :
9.          $\text{NS}_{P_i}.\text{append}(\neg\tilde{\mathcal{Q}}_{P_i}[j])$ $\triangleright$ take the top $k$ concepts from $\neg\mathcal{Q}_{P_i}$ as negative samples
10.     $\triangleright$ the following pairs the positive and negative samples with their labels, and appends them to $\mathcal{D}$
11.     for $j = 1, \dots, k$ :
12.         $\triangleright$ sentencize() constructs a sentence using a concept and a property-phrase (see Bhatia & Richie, 2021).
13.         $\mathcal{D}.\text{append}([sentencize(\mathcal{Q}_{P_i}[j], P_i), \mathsf{True}])$
14.         $\mathcal{D}.\text{append}([sentencize(\text{NS}_{P_i}[j], P_i), \mathsf{False}])$
15. return $\mathcal{D}$
```

### C.3 Computing typicality from feature representations

Inspired by Saxe et al. (2019), we compute typicality judgments for category members by considering a low dimensional subspace of the concept-property matrix ($\mathcal{M}$), defined as part of our world $\mathcal{W}$, computed using Singular-value decomposition. That is, we first decompose $\mathcal{M} \in \mathbb{R}^{152 \times 726}$ into three matrices, each of which has its own semantic interpretation:

$$\mathcal{M} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}, \tag{C.3}$$

where $\mathbf{U}$ is a $152 \times k$ matrix, $\Sigma$ is a $k \times k$ diagonal matrix, and $\mathbf{V}$ is a $726 \times k$ matrix. The rows of $\mathbf{U}$ can be thought to represent *concept-analyser vectors*, where the $i^{th}$ row represents the $i^{th}$ concept's position in each of the $k$ different semantic components. Each component here could denote an important semantic distinction. For instance, there could exist a component $a \leq k$ that effectively differentiates mammals from non-mammals, where all mammals may have positive value, and all non-mammals could have a negative value. Similarly, the columns of $\mathbf{V}^{\mathsf{T}}$ denote *property-synthesizer vectors*,[4] where each value represents the extent to which that particular property (column) is present in the various semantic components. For instance the third component could represent the differences between bird and non-bird, so it is likely that values in this row corresponding to bird-like properties (*lays eggs, can fly, has wings*) hold similar values, while values corresponding to non-bird properties might also hold similar values, and at the same time substantially different from bird-properties. Finally, $\Sigma$ is a diagonal matrix whose diagonal elements represent the strength of each component, and are ordered descendingly according to their strength.

To extract typicality measures, we find the component in $\mathbf{U}$ that maximally separates category members from non-category members. For instance, if the category of interest is FISH ($N = 14$), we take each column of $\mathbf{U}$, representing a different component. We then compute the difference between the values that are mapped to fishes and those that are mapped to non-fishes, and select the component where this difference is maximal. We then take the values of fish-concepts within this column vector as estimates for their typicality. table C.2 shows the FISH members contained in $\mathcal{W}$, organized by their typicality estimates.

[4] ↑both *concept-analyzer vectors* and *property-synthetizer vectors* are names borrowed from Saxe et al. (2019).

**Table C.2.** Typicality measures for FISH concepts extracted from the component of **U** with greatest difference in value for fish vs. non fish concepts (here, component = 8). Typicality values are unmodified post-extraction, except for being rounded up to three decimal places.

| **Concept** | **Typicality** | **Concept** | **Typicality** |
|---|---|---|---|
| TUNA | 0.162 | TROUT | 0.079 |
| MACKEREL | 0.153 | EEL | 0.077 |
| HERRING | 0.152 | FLOUNDER | 0.068 |
| SARDINE | 0.137 | MINNOW | 0.067 |
| SALMON | 0.129 | CARP | 0.054 |
| COD | 0.119 | SHARK | 0.043 |
| GOLDFISH | 0.106 | SEAHORSE | 0.029 |