Psychological Review

Inductive Reasoning in Minds and Machines

Sudeep Bhatia

Online First Publication, September 21, 2023. https://dx.doi.org/10.1037/rev0000446

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000446

Inductive Reasoning in Minds and Machines

Sudeep Bhatia
Department of Psychology, University of Pennsylvania

Induction—the ability to generalize from existing knowledge—is the cornerstone of intelligence. Cognitive models of human induction are largely limited to toy problems and cannot make quantitative predictions for the thousands of different induction arguments that have been studied by researchers, or to the countless induction arguments that could be encountered in everyday life. Leading large language models (LLMs) go beyond toy problems but fail to mimic observed patterns of human induction. In this article, we combine rich knowledge representations obtained from LLMs with theories of human inductive reasoning developed by cognitive psychologists. We show that this integrative approach can capture several benchmark empirical findings on human induction and generate human-like responses to natural language arguments with thousands of common categories and properties. These findings shed light on the cognitive mechanisms at play in human induction and show how existing theories in psychology and cognitive science can be integrated with new methods in artificial intelligence, to successfully model high-level human cognition.

Keywords: induction, reasoning, computational modeling, artificial intelligence, large language models

Supplemental materials: https://doi.org/10.1037/rev0000446.supp

Our ability to learn and reason about the world relies on successful induction: We often have to generalize from we know, in order to form beliefs and make predictions about new observations. Thus, unsurprisingly, induction has been the focus of considerable scholarly enquiry in cognitive science and psychology (Heit, 2000; Osherson et al., 1990; Sloman, 1993; see Hayes & Heit, 2018, for a review). Over the past 3 decades, this work has uncovered a large set of systematic regularities in how people evaluate the strength of induction arguments, particularly those in which the properties of some concepts and categories are induced from others. Here, researchers have found that people more easily generalize the properties of an item to its superordinate category if it is highly *typical* of the superordinate category. Thus, for example, the argument robins have a higher potassium concentration in their blood than humans, therefore birds have a higher potassium concentration in their blood than humans is judged to be stronger than the argument penguins have a higher potassium concentration in their blood than humans therefore birds have a higher potassium concentration in their blood than humans (Osherson et al., 1990). Another finding involves the *diversity* of the items in the premise: People find it easier to generalize from premises that are dissimilar to each other than from premises that are similar to each other. For example, the argument that lions and giraffes use norepinephrine as a neurotransmitter therefore rabbits use norepinephrine as a neurotransmitter is judged to be stronger than the argument lions and tigers use norepinephrine as a neurotransmitter therefore rabbits use norepinephrine as a neurotransmitter (Osherson et al., 1990).

The premise typicality and diversity effects, along with several related effects, have been used to motivate cognitive theories of human induction (Heit, 2000; Kemp & Tenenbaum, 2009; Medin et al., 2003; Osherson et al., 1990; Sloman, 1993). These theories attempt to describe the reasoning processes that people use when generalizing across items. For example, one prominent theory, the *feature-based model* (Sloman, 1993), proposes that people judge the strength of an induction argument by measuring the extent to which the known properties of the premise item are shared with the known properties of the conclusion item. The feature-based model can explain the premise typicality effect as highly typical items (robins) have more shared features with other members of their superordinate categories (birds) than do atypical items (penguins). Likewise, it can explain the premise diversity effect as the common features of dissimilar premises (lions and giraffes) are more likely to be shared with a conclusion item (rabbits) relative to the common features of similar premises (lions and tigers).

Theories of inductive reasoning, like the feature-based model, are some of the most prominent and influential accounts of high-level cognition. Yet currently, these theories are typically applied to toy problems involving a small number of concepts and categories and are unable to predict human responses for the large and diverse stimuli sets used in empirical research. Of course, these theories are also unable to make predictions for the types of induction problems that children and adults frequently encounter in everyday life. This is because theories of inductive reasoning in cognitive science lack the *knowledge representations* necessary to judge arguments involving natural concepts and categories. For example, the feature-based model specifies reasoning algorithms for judging induction

Sudeep Bhatia https://orcid.org/0000-0001-6068-684X

Sudeep Bhatia received funding from Grant 1847794 from the Directorate for Social, Behavioral and Economic Sciences. Code and data are available at https://osf.io/gebqv/.

Correspondence concerning this article should be addressed to Sudeep Bhatia, Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, United States. Email: bhatiasu@sas.upenn.edu

arguments but does not specify the underlying features of items, and thus cannot assess the actual extent of feature overlap in arguments.

This limitation has important consequences for theory development. It is, for example, unclear whether existing theories can be used to model the diverse types of data that have been collected over decades of empirical work. Likewise, the precise set of assumptions necessary to accurately predict human responses and capture empirical regularities is obscured by the idiosyncratic ontologies that researchers use to illustrate the properties of their models. Of course, the development of quantitative models capable of a priori prediction is itself one of the fundamental goals of cognitive science (see Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010, for discussions). In addition to their theoretical value, such models are also necessary for translating research into real-world applications capable of improving people's lives. In the case of induction, these applications involve topics in cognitive development, social and political reasoning, decision making, as well as the study of cognitive impairments and disfunctions.

Recently, a new class of models have been developed in statistical natural language processing. These models encode representations for words and sentences in the layers of deep neural networks, which are trained on language statistics obtained from vast amounts of text data (Brown et al., 2020; Devlin et al., 2018; He et al., 2021). Unlike models developed in cognitive science, these large language models (LLMs) have rich knowledge representations that can be used to solve several many types of natural language processing tasks. These include reasoning tasks like natural language inference (NLI), in which LLMs attempt to predict the extent to which a premise sentence entails or contradicts a conclusion sentence (Bowman et al., 2015; Wang et al., 2019; Williams et al., 2018). The inductive reasoning tasks examined in this article are a special type of NLI, suggesting that these tasks may be within the descriptive scope of leading LLMs. However, Han et al. (2022) have tested this and have found that LLMs do quite poorly. For example, these models fail to generate premise diversity effects with the stimuli used in prior psychology experiments. They also fail at replicating many of the other effects documented in the psychology and cognitive science literatures. This indicates that even though LLMs may possess the type of knowledge necessary for inductive reasoning, for example, knowledge of category membership relations (Misra et al., 2022), they do not possess the reasoning algorithms necessary to generate human-like behavior.

The goal of this article is to develop computational models capable of human-like inductive reasoning by combining the knowledge representations of LLMs with realistic reasoning algorithms previously proposed in psychology and cognitive science. Specifically, we fine-tune a LLM (Devlin et al., 2018) on a large data set of participant-generated category and feature norms (Devereux et al., 2014; McRae et al., 2005; Van Overschelde et al., 2004). In our prior work, we have shown that this model predicts human judgments about the features of concepts with high accuracy (Bhatia & Richie, 2022). We extend this model to inductive reasoning tasks by passing its knowledge base through a second model which calculates the degree to which the features of the premise items overlap with those of the conclusion item in an induction argument (building on the propositions of Sloman, 1993). We compare our feature overlap model's assessments of argument strength with the judgments of human participants, and additionally evaluate whether it is able to replicate different empirical regularities, such as the premise typicality and the premise diversity effects (Hampton & Cannon, 2004; Heit & Rubinstein, 1994; Medin et al., 2003; Osherson et al., 1990; Rips, 1975; Sloman, 1993; Sloman, 1998). We also test our model against several state-of-the-art LLMs for NLI, which judge entailment relations between premises and conclusions without explicitly calculating feature overlap (Brown et al., 2020; He et al., 2021; Lewis et al., 2020; Liu et al., 2020). Our analyses use over 16,000 reasoning problems taken from prior psychology studies, as well as new studies. These problems involve hundreds of concepts and categories from several common domains, including animals, fruits and vegetables, clothing items, furniture, and vehicles. In this way, they test both the reasoning capabilities of our models, as well as the ability of these models to apply their reasoning algorithms to rich real-world knowledge structures.

Models

Feature Overlap Model

Overview

Our feature overlap model takes, as inputs, arguments that generalize a property from one or more premise items to a conclusion item. It generates, as outputs, a continuous assessment of argument strength in range [0,1]. In order to generate these assessments, the feature overlap model relies on a specialized Bidirectional Encoder Representations from Transformers (BERT) network (Devlin et al., 2018) that we refer to as Feature-BERT (Bhatia & Richie, 2022). To query Feature-BERT, we concatenate a concept word (e.g., `cats`) and a feature or property phrase (e.g., `have fur`) into a natural language sentence (e.g., `cats have fur`). Feature-BERT generates a probability assessment of the input sentence being true or false; that is the probability that the feature in the sentence applies to the concept in the sentence. Here, we use Feature-BERT to obtain, for an item, a high-dimensional feature vector that specifies the probabilities of thousands of different features applying to the item.

Building on the work of Sloman (1993), our feature overlap model judges the strength of an argument by calculating the overlap of the features of its premise and conclusion items (e.g., the overlap of Feature-BERT's feature vectors for `robins` and `birds` for the premise typicality argument discussed at the start of the article). It quantifies overlap using the *cosine similarity* of the feature vectors of the items. In the case of arguments with multiple premise items (e.g., `lions` and `giraffes`), our model simply sums the feature vector of the premise items to get a single feature vector for the premise, which is compared with the feature vector of the conclusion (e.g., `rabbits`) using cosine similarity. This summation operation, combined with the normalization inherent in cosine similarity, implies that features that are shared by the premise items receive a higher weight when assessing feature overlap. When the argument involves a "nonblank" property with semantic content (e.g., `have a higher potassium concentration in their blood`), the model uses the simple semantic similarity between the argument property and each of the features that make up the feature vector, in order to modify the weight on the vector dimensions for the calculation of cosine similarity. In this way, features that have similar semantics to the argument property are more salient. This influences the dimensions along which feature overlap is assessed, biasing the assessment of argument strength in favor of premise and conclusion

items that share those features. Figure 1A provides an overview of the feature overlap model.

It is worth noting that our implementation of the feature overlap model using the cosine similarity of feature vectors is different to that proposed in Sloman (1993), who emphasized *feature coverage* (i.e., the extent to which the features in the conclusion category are also in the premise categories) in addition to feature overlap. Coverage, crucially, explains observed asymmetries in induction, which cannot be captured by the cosine similarity implementation introduced above. Formally, a coverage model would take the form of a *projection* of the premise item's feature vector onto the conclusion item's feature vector. We have implemented the projection variant of our model but have found that it does poorly when the premise contains multiple items. The projection model is also unable to capture nonmonotonicity effects (Medin et al., 2003; Osherson et al., 1990), which we discuss in detail below. For this reason, we have retained cosine similarity as the main metric for comparing the features of the premise and conclusion items (this is why we call our model *Feature Overlap*), but present the results of its projection variant in the Supplemental Materials and in the Asymmetry and Projection section below. Here, we also show that a hybrid model that strikes a balance between projection and cosine similarity successfully captures all effects and resolves many of the issues of the pure projection approach. Since this hybrid model was tested (and calibrated) post hoc, we retain the focus of the main text on the cosine similarity implementation.

Obtaining Item Features

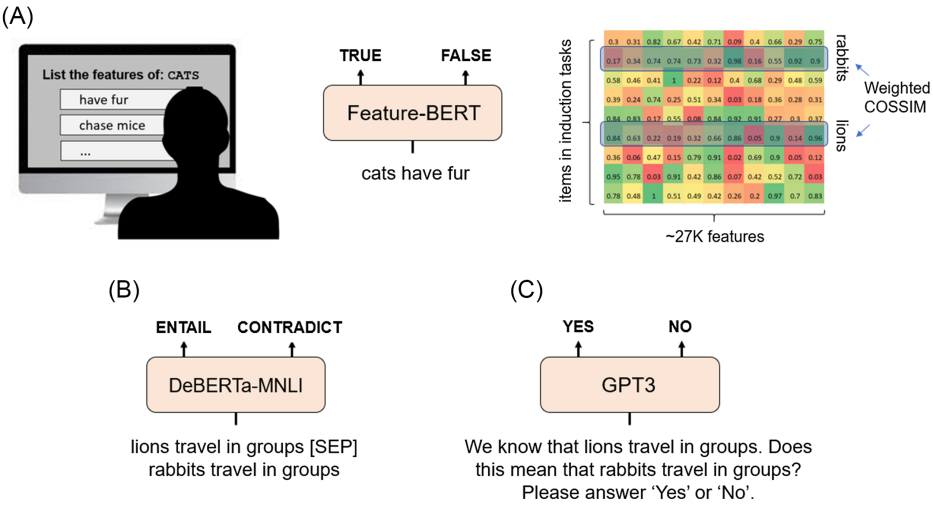
Feature-BERT provides our feature overlap model with the knowledge representations necessary to solve induction problems with natural concepts and features. This model was developed using participant-generated category norms collected by Van Overschelde et al. (2004) and feature norms collected by McRae et al. (2005) and Devereux et al. (2014). In Bhatia and Richie (2022), we compiled these norms into a training data set of 245,642 "true" sentences (sentences that combine concepts with features that were actually generated by participants for those concepts) and 245,642 "false" sentences (sentences that combine concepts with features that were not generated by participants for those concepts), with a total of 2,066 unique concepts and 29,048 unique participant-generated features. We then trained a BERT model to classify each sentence as true or false.

Once trained, this model can take in any sentence (composed of a natural concept and feature) as an input, and output the probability of that sentence being true or false, which is the probability that the feature in the sentence applies to the concept in the sentence. In Bhatia and Richie (2022), we have shown that these predictions are generally accurate. For example, Feature-BERT achieves an accuracy rate of 92% on novel sentences. Simpler models that use only the GloVe similarities of features and concepts or do not fully fine-tune the BERT model do much worse.

We have also shown that Feature-BERT replicates many observed patterns in human semantic verification. For example, this model is

Figure 1

Overview of Models Shown in Main Text



Note. In this example, all models are asked to evaluate the argument `lions travel in groups therefore rabbits travel in groups`. The feature overlap model (A) is based on a data set of human-generated features for common concepts (top left panel). Feature-BERT is fine-tuned on this data set (top middle panel), and then used to extract a high-dimensional feature vector for the items in the judged argument (e.g., `lions` and `rabbits`). The dimensions of these vectors are weighted based on the similarity of the features to the target property (`travel in groups`). Finally, cosine similarity on the resulting weighted feature vectors is used to generate an assessment of argument strength (top right panel). The remaining two models, DeBERTa-MNLI (B) and GPT3 (C), evaluate the argument without the explicit calculation of features or feature overlap. These models output "entailment" and "contradiction" judgments or "yes" and "no" judgments, respectively. BERT = Bidirectional Encoder Representations from Transformers; GPT = Generative Pretrained Transformer. See the online article for the color version of this figure.

able to capture the classic level-of-hierarchy effect (Collins & Quillian, 1969) by assigning higher probabilities to sentences composed of Level 0 features (e.g., `canaries are yellow`) than Level 1 or Level 2 features (e.g., `canaries have skin`) as well as reversals of these effects due to semantic relatedness (Rips et al., 1973; Smith et al., 1974). Feature-BERT also predicts observed response time differences as a function of sentence truth, sentence relatedness, item category membership, feature correlation, and feature distinctiveness (Anderson & Reder, 1974; Cree et al., 2006; Glass et al., 1974; Hampton, 1984; McRae et al., 1997). Additionally, its predictions for category membership judgments are proportional to the typicality of concepts in superordinate categories (Rosch, 1975), allowing it to capture observed inconsistencies across participants, transitivity violations, and violations of set membership relations in semantic judgment, previously attributed to typicality (Hampton, 1982; McCloskey & Glucksberg, 1978; Roth & Mervis, 1983). Feature-BERT can also predict concept typicality (Rosch & Mervis, 1975), and feature correlations uncovered by BERT are similar to those obtained in previous experimental data (Malt & Smith, 1984). Finally, Feature-BERT is able to predict patterns in similarity judgment that are problematic for existing distributional semantics models, such as asymmetry, the distinction between association and similarity, and the measurement of similarity within (rather than across) categories (Hill et al., 2015; Richie & Bhatia, 2021; Whitten et al., 1979). The training data set and algorithm are certainly not developmentally or psychologically realistic, since children learning spoken language hear far less than billions of words of language and are not exposed to hundreds of thousands of statements with true-false labels. Nonetheless, these results show that Feature-BERT proxies people's knowledge for simple concept-feature pairings thus is a useful tool for implementing and testing cognitive theories like the feature-based model in general populations (see Bhatia & Richie, 2022, for a detailed discussion of this point). In the Supplemental Materials, we describe the neural network architecture, training data set, and implementational assumptions of Feature-BERT in greater detail.

Calculating Feature Overlap

The feature overlap model in this article compares items on the 25,797 unique features in Devereux et al. (2014). Specifically, for a target item, we pass sentences composed of that item and each of the 25,797 features through the Feature-BERT model, to obtain a 25,797-dimensional vector of probabilities of features applying to that item. For item i, we write this feature vector as $\boldsymbol{f_i}$. The jth element of $\boldsymbol{f_i}$, $f_{i,j}$, is the probability that Feature-BERT attaches to the jth Devereux et al. (2014) feature being true for item i. Since they are probabilities, the elements of $\boldsymbol{f_i}$ are in range [0,1].

We also use a GloVe (Pennington et al., 2014) bag-of-word model to calculate the simple semantic similarity between a target property and each of the 25,797 features. For this, we first average the 300-dimensional GloVe vector representations for each of the words in the target property to get a 300-dimensional GloVe vector for the target property, which we write as $\boldsymbol{g_p}$. We do the same for each of the 25,797 Devereux et al. (2014) features. For feature j, we write this vector as $\boldsymbol{g_j}$. We then calculate the cosine similarity of $\boldsymbol{g_p}$ and each $\boldsymbol{g_j}$, with $\text{COSSIM}(\boldsymbol{g_p}, \boldsymbol{g_j}) = \boldsymbol{g_p} \cdot \boldsymbol{g_j}/(\|\boldsymbol{g_p}\| \cdot \|\boldsymbol{g_j}\|)$, resulting in a 25,797-dimensional similarity vector $\boldsymbol{s_p} = [\text{COSSIM}(\boldsymbol{g_p}, \boldsymbol{g}_1), \text{COSSIM}(\boldsymbol{g_p}, \boldsymbol{g}_2), \ldots, \text{COSSIM}(\boldsymbol{g_p}, \boldsymbol{g}_{25797})]$. The elements of $\boldsymbol{s_p}$ are in range $[-1,1]$. We transform $\boldsymbol{s_p}$ to obtain feature similarity weights $\boldsymbol{w_p} = (1 + \boldsymbol{s_p})/2$, in range [0,1].

To calculate the feature overlap between a premise item i and a conclusion item k, for a property p, in a single-premise argument, we first perform an element wise multiplication operation between $\boldsymbol{w_p}$ and $\boldsymbol{f_i}$ to get a 25,797-dimensional weighted feature vector for the premise $\boldsymbol{w_p} \odot \boldsymbol{f_i} = [w_{p,1} \cdot f_{i,1}, w_{p,2} \cdot f_{i,2}, \ldots, w_{p,25797} \cdot f_{i,25797}]$. Then, using cosine similarity, we calculate the feature overlap between the premise and conclusion as $\text{COSSIM}(\boldsymbol{w_p} \odot \boldsymbol{f_i}, \boldsymbol{f_k})$. When the premise has multiple items, we simply sum the feature vectors for those items to get a single feature vector for the premise. For example, in an argument with two-premise items, i and i' (as well as a conclusion item k and a property p) the model's judged argument strength is $\text{COSSIM}(\boldsymbol{w_p} \odot (\boldsymbol{f_i} + \boldsymbol{f_{i'}}), \boldsymbol{f_k})$. In the general case, with a set of premise items P, the model's judged argument strength is $\text{COSSIM}(\boldsymbol{w_p} \odot \Sigma_{i \in P} \boldsymbol{f_i}, \boldsymbol{f_k})$. Finally, when the argument involves a blank property without semantic content, we simply set $\boldsymbol{w_p}$ to a vector of ones. Note that all vectors used in the cosine similarity calculation are in range [0,1], which is why the cosine similarity output is also in the range [0,1].

The above sequence of operations implicitly overweigh features that are shared by the premise items, causing conclusion items that share those features to generate higher judgments of argument strength. For example, consider a simplified two-premise three-feature setting, with premise feature vectors $\boldsymbol{f_i} = [1,1,0]$ and $\boldsymbol{f_{i'}} = [1,0,1]$, and a blank property resulting in $\boldsymbol{w_p} = [1,1,1]$. Our weighted premise vector would thus be $\boldsymbol{w_p} \odot (\boldsymbol{f_i} + \boldsymbol{f_{i'}}) = \boldsymbol{f_i} + \boldsymbol{f_{i'}} = [2,1,1]$. This premise structure would lead to the strongest judgment of argument strength if the conclusion feature vector loads primarily onto the first feature, which is the feature shared by the premises. Thus for example, a conclusion item with $\boldsymbol{f_k} = [1,0,0]$ would be given a high judgment of argument strength of $\text{COSSIM}(\boldsymbol{w_p} \odot (\boldsymbol{f_i} + \boldsymbol{f_{i'}}), \boldsymbol{f_k}) = 0.81$. By contrast, a conclusion item with $\boldsymbol{f_k} = [0,1,0]$ or $\boldsymbol{f_k} = [0,0,1]$ would have a lower judgment of argument strength of $\text{COSSIM}(\boldsymbol{w_p} \odot (\boldsymbol{f_i} + \boldsymbol{f_{i'}}), \boldsymbol{f_k}) = 0.41$.

The feature overlap model also overweighs features that are have a high GloVe bag-of-words similarity to the target property. This causes premise and conclusion items that overlap on those features to be given higher judgments of argument strength. For example, imagine a simplified single-premise and three-feature setting with a nonblank property that results in $\boldsymbol{w_p} = [1,0.5,0.5]$. This setting involves a target property that is semantically similar to the first of the three features but not the second or third. In this case, premise and conclusion items that overlap on the first feature, like $\boldsymbol{f_i} = [1,1,0]$ and $\boldsymbol{f_k} = [1,0,1]$, would be given a high judgment of argument strength of $\text{COSSIM}(\boldsymbol{w_p} \odot \boldsymbol{f_i}, \boldsymbol{f_k}) = 0.63$. By contrast, premise and conclusion items that overlap on the second feature, like $\boldsymbol{f_i} = [1,1,0]$ and $\boldsymbol{f_k} = [0,1,1]$, would be given a low judgment of argument strength of $\text{COSSIM}(\boldsymbol{w_p} \odot \boldsymbol{f_i}, \boldsymbol{f_k}) = 0.31$.

In the Supplemental Materials, we also provide results for a variant of the feature overlap model which replaces Feature-BERT's probabilities with their associated logits, according to the formula $\text{PROBABILITY} = 1/(1 + e^{-\text{LOGIT}})$. In this way, it computes feature overlap (cosine similarity) on 25,797-dimensional vectors of logits. Logits provide a more continuous assessment of the truth or falsehood of sentences (probabilities, by contrast are often very close to 0 or 1) and have been shown in our past work (Bhatia & Richie, 2022) to do slightly better at predicting human semantic verification than the associated probabilities. Code for applying the feature overlap model is available at

https://osf.io/gebqv/. Here, we also provide the feature overlap models' predictions for all induction problems tested in this article.

Natural Language Inference Models

We compare our two feature overlap models to five leading LLMs for NLI. The first three of these are BART-MNLI (Lewis et al., 2020), RoBERTa-MNLI (Liu et al., 2020), and DeBERTa-MNLI (He et al., 2021). These are LLMs that were fine-tuned on a large multigenre natural language inference (MNLI) corpus (Williams et al., 2018). We query these models using the HuggingFace API (model names are facebook/bart-large-mnli, roberta-large-mnli, and microsoft/deberta-large-mnli, respectively). We provide BART-MNLI with the premise sentence and the conclusion sentence as inputs and give it "entailment" and "contradiction" labels with which to classify the sentence. RoBERTa-MNLI model and DeBERTa-MNLI models are also given the premise and conclusion sentences as inputs, but since these models were explicitly built to classify the inputs into entailment and contradiction (as well as neutral) classes, no further classification labels are necessary. All three models output scores for entailment and contradiction, and we use the entailment score, a number in range [0,1], to predict argument strength.

We also use the GPT3-DaVinci-002 and GPT3-Babbage-001 models (Generative Pretrained Transformer; Brown et al., 2020). We query these models on the OpenAI API with the prompt: *We know that* [PREMISE SENTENCE]. *Does this mean that* [CONCLUSION SENTENCE]*? Please answer* "Yes" *or* "No." This prompt was shown to generate the most human-like performance out of all the prompts in Han et al. (2022). We test whether the model's first five output tokens include "Yes," "yes," "YES," "No," "no," and "NO" and subsequently calculate the probability attached to "Yes," "yes" and "YES" versus "No," "no" and "NO." This is the relative probability attached to GPT's yes versus no outputs, ignoring the case in which GPT chooses to give that output. In other words, it considers an output of "Yes," "yes," or "YES" to be the same (i.e., a yes response), and the output of "No," "no," or "NO" to be the same (i.e., a no response), and simply calculates the relative probabilities of the yes versus the no response. This probability, which is in range [0,1], is used to predict argument strength.

Intuitively, these NLI models take the premise and conclusion sentences of an argument as inputs and, as outputs, provide judgments of argument strength. Crucially, they do not explicitly assess the features (or the extent of feature overlap) for the premise and conclusion items. Instead, their reasoning processes are inbuilt into the layers of the network architecture and have been developed through training on hundreds of thousands of NLI problems (as well other linguistic reasoning problems). In the main text, we will present the results of only the DeBERTa-MNLI and GPT3-DaVinci models. Results of the remaining NLI models are in the Supplemental Materials. An illustration of the two models examined in the main text is provided in Figure 1B and 1C. Code for querying the LLMs is available at https://osf.io/gebqv/. Here, we also provide the LLMs' predictions for all induction problems tested in this article.

Predictive Accuracy

Existing Data Sets

We began by evaluating the accuracy of the seven models introduced above in predicting human assessments of argument strength collected in the prior work. Our first data set was obtained from Rips (1975) and involves 60 pairs of animal species. Participants in the Rips data set were told that one animal species has a given disease and were asked to estimate the proportion of instances in the second species that also have the disease. Our second and third data sets were obtained from Experiments 2 and 4 of Osherson et al. (1990), respectively. In Experiment 2 of their article, participants were asked to rank 45 arguments that generalized a property from three mammal species to all mammals. In Experiment 4, participants were asked to rank 36 arguments that generalized a property from two mammal species to horses. We used our feature overlap and NLI models to assess argument strength for the items in these three data sets and evaluated them based on their correlations with participant responses in the data sets.

These correlations are shown in Figure 2A and Supplemental Figure S1A. Here, we can see that our two feature overlap models achieved high correlations and consistently outperformed the NLI models. The feature overlap models also performed similarly to each other indicating that using logits instead of probabilities does not alter performance. Overall, the average correlation, across data sets, of the main feature overlap model was 0.63, whereas the best NLI model, DeBERTA-MNLI, achieved an average correlation of only 0.32.

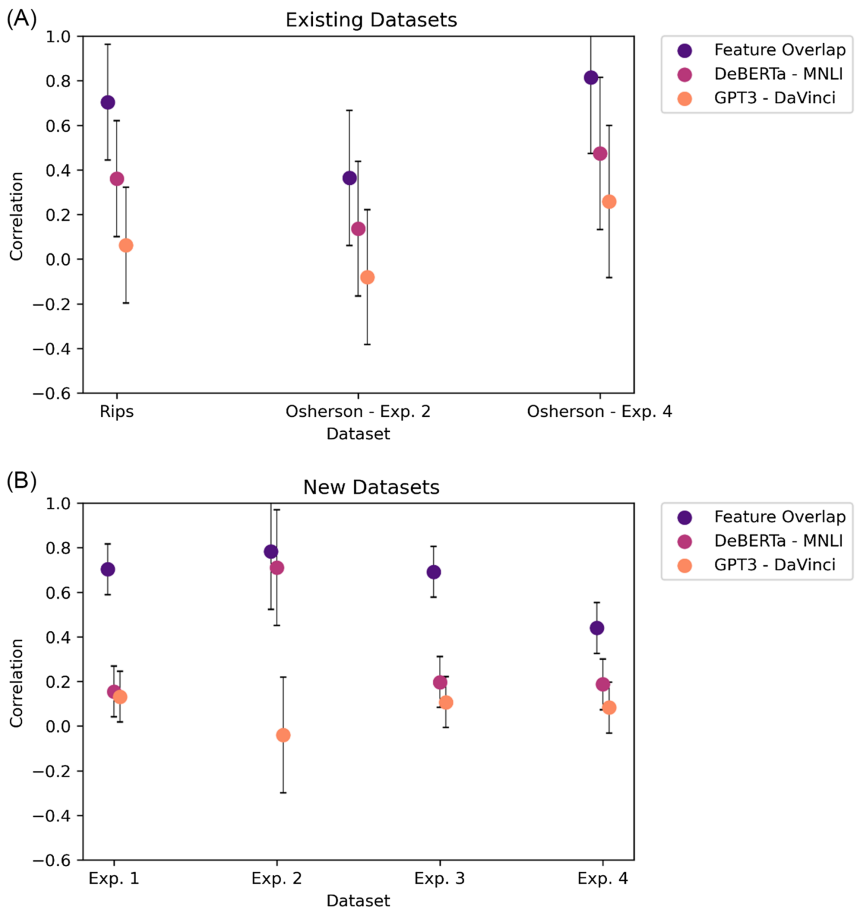
New Experiments

The above data sets use a small number of induction problems involving only animal species. For a more rigorous test of our approach, we conducted four new experiments with 960 distinct arguments taken from six superordinate categories (birds, fruits, vegetables, clothing, furniture, and vehicles). Each experiment offered participants a set of arguments and asked them to provide a continuous rating of argument strength. In Experiment 1, there were 300 arguments that generalized a blank property from one item (e.g., `tables`) to another member of its superordinate category (e.g., `chairs`); in Experiment 2, there were 60 arguments that generalized a blank property from one item (e.g., `tables`) to all members of its superordinate category (e.g., `furniture`); in Experiment 3, 300 arguments generalized a blank property from two items (e.g., `tables` and `bookshelves`) to a third item in their superordinate category (e.g., `chairs`); and in Experiment 4, 300 arguments generalized a blank property from two items (e.g., `tables` and `bookshelves`) to all members of their superordinate category (e.g., `furniture`). The experiments received approval from the University of Pennsylvania Institutional Review Board (Title: "Everyday judgments and decisions"; Institutional Review Board No.: 823184). Additional details about our experiments are presented in the Supplemental Material and the stimuli used and average ratings are stored in the Open Science Framework repository (https://osf.io/gebqv/) for the project.

We correlated our models' predictions with average participant ratings for each of the arguments in the four experiments. These correlations are shown in Figure 2B and Supplemental Figure S1B. Again, our feature overlap model achieved high correlations and outperformed the NLI models. Additionally, the feature overlap models performed similarly to each other showing that using logits instead of probabilities does not alter performance. Overall, the average correlation, across data sets, of the main feature overlap model was 0.65, whereas the best NLI model, DeBERTA-MNLI, achieved an average correlation of only 0.31.

Figure 2

Accuracy Rates Obtained by the Feature Overlap Model, and Two Competing NLI Models on Existing (A) and New (B) Data Sets



Note. In each data set, the models attempt to predict argument strength. These predictions are correlated with participant responses. Error bars are 95% confidence interval of correlations. These bars are larger for existing data sets due to their small sample sizes. NLI = natural language inference; Exp. = experiment; GPT = Generative Pretrained Transformer. See the online article for the color version of this figure.

Empirical Regularities

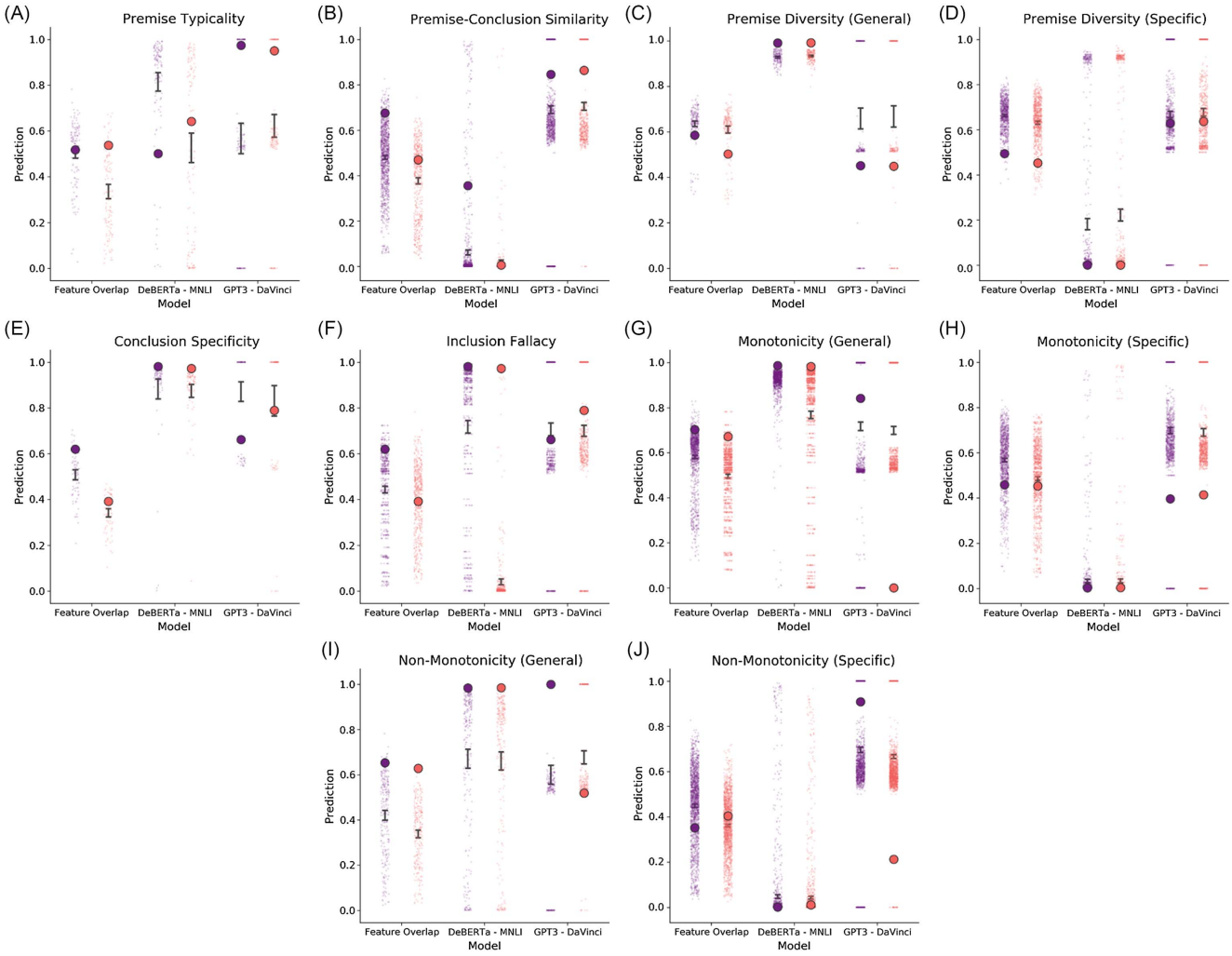
Premise Typicality

Next, we tested whether our models generated empirical patterns documented in the psychology and cognitive science literatures. We began with the premise typicality effect, which was introduced in the first paragraph of this article. We tested this effect using typicality ratings for items in eight different superordinate categories (birds, clothing, fruits, furniture, toys, vegetables, vehicles, and weapons), collected by Rosch (1975). We used this data to generate 254 arguments in which a property was generalized from a premise item (e.g., `sparrows`) to all members of its superordinate category (e.g., `birds`). Using a median split, we divided these arguments into two groups: high premise typicality and low premise typicality and offered the arguments in each group to our models. The predictions of these models are shown in Figure 3A and Supplemental Figure S2A (the large markers are the models' predictions for the example arguments used in Osherson et al., 1990, and presented in the introduction). Here, we can see that the feature overlap model generated higher argument strength judgments for high typicality arguments (purple points) relative to low-typicality arguments (orange points). A separate analysis correlating model predictions with continuous typicality ratings also found this positive relationship (see Table 1 and Supplemental Table S1, for all statistical tests). The feature overlap model was able to generate this effect because the feature vector for the superordinate category is more similar to that of highly typical subordinate items than to that of atypical subordinate items (Bhatia & Richie, 2022).

Most of the NLI models also captured this effect, likely because they are able to encode typicality relations and use these relations for generalization (Han et al., 2022; Misra et al., 2022). However, GPT-DaVinci and BART-MNLI failed to do so. The former model typically generated extreme predictions, that is, it generally outputted either a yes response or a no response, and seldom outputted both. This meant that the probability assignment for yes versus no was 1 or 0 in many cases (see e.g., Figure 3A). This

Figure 3

Models' Predictions for Empirical Regularities Documented by Osherson et al. (1990)



Note. A: premise typicality, B: premise-conclusion similarity, C: premise diversity (general), D: premise diversity (specific), E: conclusion specificity, F: inclusion fallacy, G: monotonicity (general), H: monotonicity (specific), I: nonmonotonicity (general), J: nonmonotonicity (specific). The small semiopaque points in this panel correspond to individual arguments, and the error bars display ±1 *SE* of the mean of the models' predictions for these arguments. Each empirical regularity has two groups of arguments, shown in purple and orange, and discussed in detail in the text. For example, in Panel A, purple arguments have high typicality premises, whereas orange arguments have low-typicality premises. All empirical regularities involve higher human assessments for purple arguments relative to orange arguments. Finally, the large circles indicate model predictions for the specific arguments used to illustrate the empirical regularities in the main text. The arguments are taken from Osherson et al. (1990). For example, the large purple points in Panel A are the models' predictions for `robins have a higher potassium concentration in their blood than humans, therefore birds have a higher potassium concentration in their blood than humans`. *SE* = standard error; GPT = Generative Pretrained Transformer. See the online article for the color version of this figure.

could have obscured the typicality effect, which involves a more subtle shift in judgment as typicality is varied. Interestingly, GPT-Babbage did not have this issue as it often gave a combination of yes and no responses, and thus had judgments that were not at the extreme ends of the [0,1] probability interval (Supplemental Figure S2A). For this reason, GPT-Babbage was able to capture the typicality effect.

It is not completely clear why BART-MNLI failed. This could be because, unlike RoBERTa-MNLI and DeBERTa-MNLI, it was not specifically trained for entailment/contradiction judgment. Instead, it takes in flexible label tokens and performs classification based on the token semantics (in our case, the label tokens offered were "entailment" and "contradiction"), which may not be the best way to model induction.

Premise-Conclusion Similarity

A variant of the premise typicality effect involves premise-conclusion similarity. This effect describes people's tendency to generalize to conclusion items that are highly similar to the premise items. Thus, for example, the argument `robins` and `blue jays`

Table 1
Statistical Tests of the Feature Overlap Model's Argument Strength Predictions for Empirical Regularities

	Regression			Correlation		
Effect name	Coef.	*t*	95% CI	*r*	95% CI	*N*
Premise typicality	0.17	8.64	[0.13, 0.21]	0.53	[0.44, 0.61]	254
Premise-conclusion similarity	0.10	12.97	[0.09, 0.12]	0.45	[0.41, 0.49]	1,662
Premise diversity (general)	0.03	2.61	[0.01, 0.04]	0.17	[0.07, 0.27]	356
Premise diversity (specific)	0.03	7.71	[0.02, 0.04]	0.18	[0.14, 0.23]	1,592
Conclusion specificity	0.16	11.39	[0.13, 0.19]	NA	NA	154
Inclusion fallacy	0.05	9.96	[0.03, 0.07]	NA	NA	1,070
Inclusion fallacy (Sloman)	0.14	4.04	[0.07, 0.21]	NA	NA	72
Monotonicity (general)	0.08	14.94	[0.07, 0.09]	NA	NA	2,356
Monotonicity (specific)	0.08	7.86	[0.07, 0.09]	NA	NA	2,166
Nonmonotonicity (general)	0.08	5.66	[0.05, 0.11]	NA	NA	508
Nonmonotonicity (specific)	0.09	17.64	[0.08, 0.10]	NA	NA	3,324
Conclusion typicality	0.09	10.64	[0.07, 0.10]	NA	NA	1,392
Property type	0.07	10.06	[0.05, 0.08]	0.37	[0.31, 0.43]	784
Property relevance	0.06	1.35	[−0.03, 0.15]	0.29	[−0.04, 0.58]	34

Note. The first three columns describe the output of a single regression of the model's predictions on a binary variable (corresponding to the purple vs. orange bars in Figures 2 and 3). A positive effect here shows that the model generates higher responses for purple versus orange arguments. The next two columns show the Pearson correlation (and associated CIs) between the model's prediction and a continuous variable corresponding to the effect in question. Not all effects have continuous variables. *N* refers to the total number of observations (arguments) in the analysis. Coef. = coefficient; CI = confidence interval; NA = not applicable.

use serotonin as a neurotransmitter therefore sparrows use serotonin as a neurotransmitter is judged to be stronger than the argument robins and blue jays use serotonin as a neurotransmitter, therefore geese use serotonin as a neurotransmitter (Osherson et al., 1990). We tested this effect using similarity ratings collected in Richie and Bhatia (2021). These ratings measure the similarity of pairs of items, with each pair taken from one of six superordinate categories (birds, clothing, fruits, furniture, vegetables, and vehicles). We used this data to generate 1,662 arguments in which a property was generalized from a premise item (e.g., sparrows) to a conclusion item that was a member of the premise's superordinate category (e.g., robins). Using a median split, we divided these arguments into two groups: high and low similarity and offered the arguments in each group to our models. The predictions of these models are shown in Figure 3B and Supplemental Figure S2B. Here, we can see that the feature overlap model generated higher argument strength predictions for high similarity arguments (purple points) relative to low similarity arguments (orange points). A separate analysis correlating model predictions with continuous similarity ratings further illustrates this positive relationship (Table 1 and Supplemental Table S1 present all statistical tests). The feature overlap model generated this effect because similar items also overlap on their features.

By contrast, most of the NLI models failed to generate this effect. Although most of these models are able to generalize from typical items to their superordinate categories, as shown in the previous section, they seem to not be sensitive to the similarity of items within a given superordinate category. We speculate that this could be because these models are trained largely on deduction tasks. In such tasks, a fact being true for an item also means that it is true for the superordinate category but does not imply that it is true for other similar items in the superordinate category. Thus, for example, the premise sentence a blue jay is in the forest implies the conclusion a bird is in the forest but does not necessarily imply the conclusion a robin is in the forest, despite blue jay and robin being highly similar.

Premise Diversity (General and Specific)

The next two effects involve the role of premise diversity: People find it easier to generalize from premise items that are dissimilar to each other than from premise items that are similar to each other. This effect has been tested both with inductions to a general superordinate category as well as inductions to a specific member of the superordinate category. An example of the former (general effect) is the finding that the argument hippos and hamsters have a higher sodium concentration in their blood than humans, therefore mammals have a higher sodium concentration in their blood than humans is judged to be stronger than the argument hippos and rhinos have a higher sodium concentration in their blood than humans, therefore mammals have a higher sodium concentration in their blood than humans (Osherson et al., 1990). An example of the latter (specific effect) is in the first paragraph of the introduction of this article.

We tested the general premise diversity effect by pooling the data sets in Rosch (1975) and Richie and Bhatia (2021). From the latter, we obtained similarity ratings of pairs of items, whereas from the former, we obtained typicality ratings of these items for their superordinate category. We excluded premise items pairs that were highly atypical (e.g., penguins and ostriches) as those pairs were almost always dissimilar to each other, generating a multicollinearity problem. This resulted in a total of 356 arguments with pairs of (largely typical) items in the premise, that vary in terms of similarity, as well as a superordinate category in the conclusion. We also used this approach to test the specific premise diversity effect, except that we replaced the superordinate category in the conclusion

(e.g., birds) with a highly typical member of that category (e.g., sparrows). This generated a total of 1,592 arguments with pairs of items in the premise and an item (that is typical of the premise items' superordinate category) in the conclusion. Using a median split, we divided the above arguments into two groups: high and low premise diversity (inverse of similarity) and offered the arguments in each group to our models. There were a total of six superordinate categories in this analysis.

The predictions of these models are shown in Figure 3C and 3D in the main text and Supplemental Figure S2C and S2D. Here, we can see that the feature overlap model generated higher argument strength predictions for high diversity premises (purple points) relative to low diversity premises (orange points). A separate analysis correlating model predictions with the (continuous) inverse similarity rating further demonstrated a positive effect of diversity (Table 1 and Supplemental Table S1 present statistical tests). The feature overlap model generated this effect because it gives a higher weight to the overlapping features of premise items when assessing the cosine similarity of the premise with the conclusion. These overlapping features are more likely to be shared with the superordinate category (in the general case) and with members of the superordinate category (in the specific case), when the premise items are diverse, generating higher argument strength predictions. By contrast, nondiverse premises overlap on many idiosyncratic features (e.g., hippos and rhinos are both quite large and both are found mainly in Africa), that may not be shared with other members of their superordinate category.

It is interesting to note that the NLI models failed to capture this effect. Even though these models are able to encode typicality and category membership relations, they do not use this information in a manner that considers the differing types of information provided by diverse versus nondiverse premises. In this way, their reasoning processes are limited to simplistic assessments of typicality and are unable to mimic the richness of human induction.

Conclusion Specificity

The more specific the conclusion, the more likely people are to generalize a premise to the conclusion. Thus, for example, people judge the argument blue jays and falcons require vitamin k for the liver to function therefore birds require vitamin K for the liver to function to be stronger than the argument blue jays and falcons require vitamin K for the liver to function, therefore animals require vitamin K for the liver to function (Osherson et al., 1990). We tested this effect by extracting animal species and their immediate superordinate categories (birds, fishes, or invertebrates) from Devereux et al. (2014). We then constructed 154 argument pairs in which a property was generalized from the animal species to either the immediate superordinate category (generating a specific conclusion) or the distal superordinate category of animals (generating a nonspecific or general conclusion).

The predictions of these models on these argument pairs are shown in Figure 3E in the main text and Supplemental Figure S2E in the Supplemental Materials. Here, we see that the feature overlap model generated higher argument strength predictions for arguments with specific conclusions (purple points) relative to arguments with distal conclusions (orange points; Table 1 and Supplemental Table S1 present statistical tests). This is because items have more features in common with their proximate superordinate categories than with distal superordinate categories (e.g., blue jays share many of the features of birds but not many of the features of animals, as judged by Feature-BERT).

All of the NLI models failed to capture this effect. Some even generated statistically significant predictions in the opposite direction. It is not clear why this is the case. It could, for example, reflect the tendency of these models to favor nonspecific conclusions (composed of more general categories) in NLI tasks. Such a heuristic could lead to good performance in deduction, for example, a conclusion sentence an animal is in the forest is more likely to be true than a bird is in the forest. However, this heuristic is not appropriate for the induction of properties across concepts.

Inclusion Fallacy

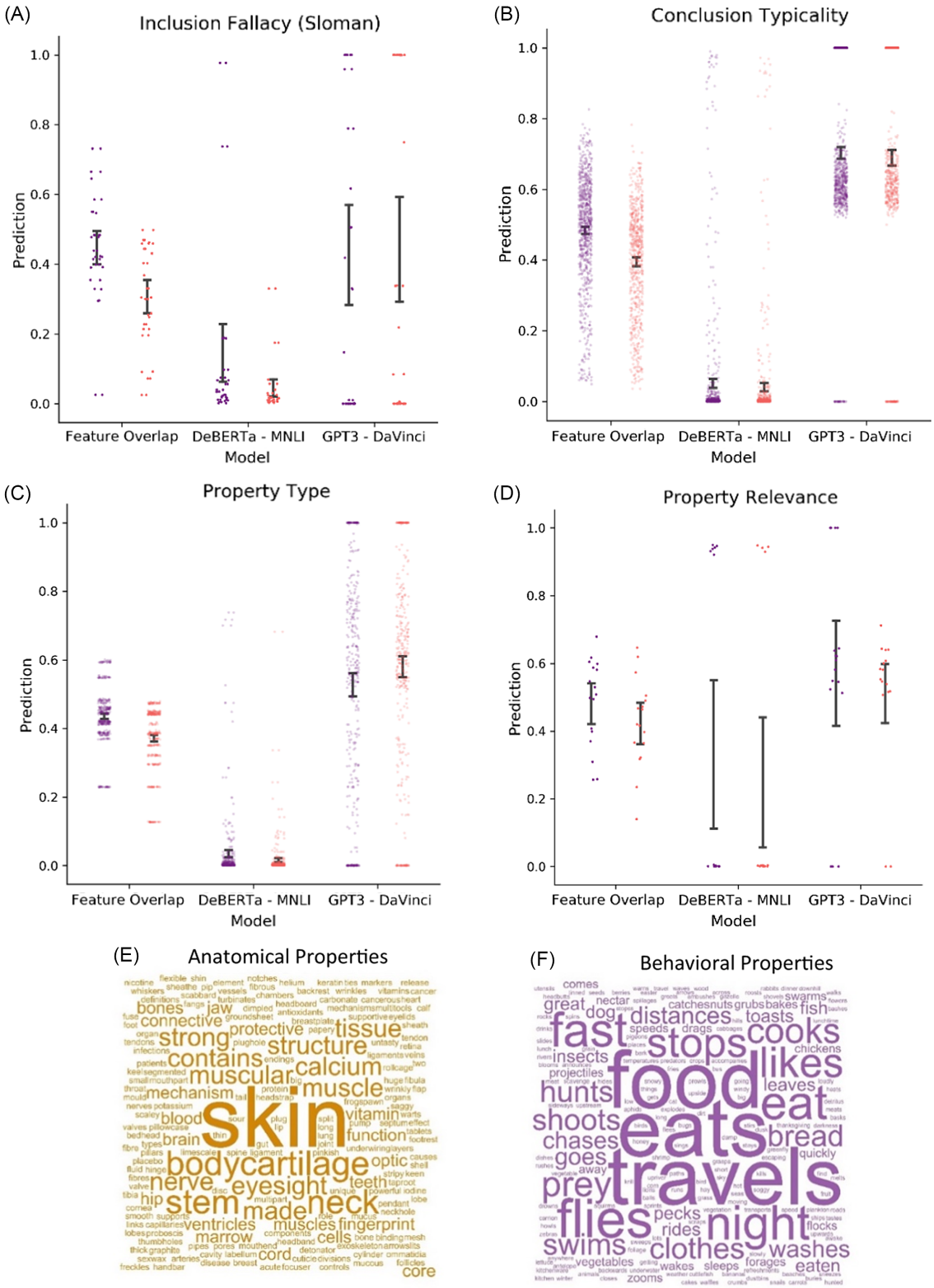
The conclusion specificity effect shows that human induction is sensitive to category hierarchy. However, people do not always generalize information from a superordinate category to all of its members. This is the case with the inclusion fallacy effect, according to which atypical conclusion items can lead to the apparent neglect of category hierarchies. For example, the argument robins have an ulnar artery therefore birds have an ulnar artery is judged to be stronger than robins have an ulnar artery therefore ostriches have an ulnar artery (Osherson et al., 1990), despite the fact that ostriches are types of birds, and therefore the conclusion of the first argument should (through deduction) imply the conclusion of the second. To test if the feature overlap model explained this effect, we used the data sets collected by Rosch (1975) and Richie and Bhatia (2021) to algorithmically generate 1,070 arguments which generalized a property from a premise items to either the superordinate category or to an atypical member of the superordinate category. There were a total of six superordinate categories in this analysis.

We also applied our models to the data sets collected by Sloman (1998), which involve several variants of this effect, including variants that involve manipulating the category membership relations of the premise items instead of the conclusion items. The problems used in this work contain pairs of arguments in which one argument generates higher assessments of argument strength, despite being logically entailed by the other. There are a total of 72 different arguments, spanning several distinct superordinate categories (including plants, tools, musical instruments, and occupations).

The predictions of our models on the algorithmically generated argument pairs are shown in Figure 3F and Supplemental Figure S2F, and predictions on the Sloman (1998) argument pairs are shown in Figure 4A and Supplemental Figure S3A. Each point in these figures corresponds to a single argument. Purple arguments are logically entailed by orange arguments but are nonetheless given higher ratings than the orange arguments by participants. We can see the feature overlap model generated the inclusion fallacy on all data sets (Table 1 and Supplemental Table S1 present statistical tests). This is because atypical conclusion categories (like ostriches) often share few features with the premise categories (as assessed by Feature-BERT), leading to weaker argument strength judgments. The competing NLI models mostly replicated the effect for the algorithmically generated data sets, but failed to do so for the Sloman data set. We suspect that their success for algorithmically

Figure 4

Illustration of Additional Empirical Regularities



Note. Panels A, B, C, and D show model predictions for the inclusion fallacy arguments in Sloman (1998), conclusion typicality arguments in Hampton and Cannon (2004), property type augments in Heit and Rubinstein (1994), and property relevance arguments in Medin et al. (2003), respectively. These use the same format as the plots in Figure 3. Panels E and F show word clouds of features that are most associated with the anatomical properties and behavioral properties in Heit and Rubinstein (1994). GPT = Generative Pretrained Transformer. See the online article for the color version of this figure.

generated data sets stems from the fact that these models favor nonspecific conclusions (e.g., favoring conclusions with birds instead of ostriches) for reasons discussed in the previous section. They failed in the Sloman data set, as that data set also involves argument pairs in which the conclusion is held constant (rather, the inclusion fallacy is generated by manipulating the category membership relations of the premise items).

Monotonicity (General and Specific)

In most settings, knowing that a property is shared by many items makes it easier to generalize that property to new items. This effect is known as monotonicity and persists with induction to the superordinate category shared by the premise items (the general case) and with induction to a specific item that is in the same superordinate category as the premise items (the specific case). An example of the former is the finding that the argument hawks, sparrows, and eagles have sesamoid bones therefore birds have sesamoid bones is judged to be stronger than the argument that sparrows and eagles have sesamoid bones therefore birds have sesamoid bones (Osherson et al., 1990). An example of the latter is the finding that foxes, pigs, and wolves use vitamin K to produce clotting in their blood, therefore gorillas use vitamin K to produce clotting agents in their blood is judged to be stronger than the argument that pigs and wolves use vitamin K to produce clotting in their blood, therefore gorillas use vitamin K to produce clotting agents in their blood (Osherson et al., 1990).

We tested this effect using the stimuli from Richie and Bhatia (2021). Here, we generated argument pairs consisting of a one-item premise and a two-item premise. The conclusions of the arguments involved a superordinate category in the general case, and a randomly chosen item from the premise superordinate category in the specific case. There were a total of 2,356 arguments for the general monotonicity effect and 2,166 arguments for the specific monotonicity effect. These spanned six superordinate categories. We offered these arguments to our models, whose predictions are shown in Figure 3G and 3H and Supplemental Figure S2G and S2H. Here, we can see that the feature overlap model generated higher argument strength predictions for two-premise arguments (purple points) relative to one-premise arguments (orange points) in both the general and specific cases (Table 1 and Supplemental Table S1 present statistical tests). The feature overlap model captured this effect because the overlapping features of multiple items are more likely to be shared with other members of the superordinate category. In other words, adding additional items to the premise leads to a premise feature vector that is, on average, closer to that of other members of the superordinate category. This mechanism is the same mechanism responsible for the premise diversity effect described above. Thus unsurprisingly, as with the premise diversity effect, the NLI models were unable to robustly generate the monotonicity effect. Again, this reflects the fact that these models do not explicitly take into account feature overlap relationships between premise and conclusion items.

Nonmonotonicity (General and Specific)

Although providing additional premise items increases argument strength when premise categories share superordinate categories, this is not necessarily the case when additional premise items are taken from different superordinate categories. This effect is known as nonmonotonicity and persists with induction to the superordinate category shared by the premise items (the general case) and with induction to a specific item that is in the same superordinate category as the premises (the specific case). An example of the former is the finding that the argument crows and peacocks secrete uric acid crystals therefore birds secrete uric acid crystals is judged to be stronger than the argument that crows, peacocks, and rabbits secrete uric acid crystals therefore birds secrete uric acid crystals (Osherson et al., 1990). An example of the latter is the finding that flies require trace amounts of magnesium for reproduction therefore bees require trace amounts of magnesium for reproduction is judged to be stronger than the argument that flies and orangutans require trace amounts of magnesium for reproduction therefore bees require trace amounts of magnesium for reproduction (Osherson et al., 1990).

We tested this effect using the stimuli from Richie and Bhatia (2021), which spanned six superordinate categories. Here, we generated argument pairs consisting of a one-item premise and a two-item premise. Unlike the monotonicity arguments in the prior section, the two-item premises in this analysis involved distinct superordinate categories, for example, sparrows (from category birds) and rabbits (from category mammals). The conclusions of the arguments involved the superordinate category of the first premise (e.g., birds) in the general case, and a randomly chosen item from the superordinate category of the first premise (e.g., ducks) in the specific case. There were a total of 508 arguments for the general nonmonotonicity effect and 3,324 arguments for the specific monotonicity effect. We offered these to our models, whose predictions are shown in Figure 3I and 3J and Supplemental Figure S2I and S2J. Here, we can see that the feature overlap model generated higher argument strength predictions for one-premise arguments (purple points) relative to two-premise arguments (orange points) in both the general and specific cases (Table 1 and Supplemental Table S1 present statistical tests). This effect emerges because the addition of a new premise item that is highly dissimilar to the conclusion item leads to a premise feature vector with much less overlap with the conclusion feature vector. Again, the NLI models were unable to robustly generate nonmonotonicity, as they do not reason over the properties of categories when judging argument strength.

Conclusion Typicality

The effects discussed thus far were initially documented by Osherson et al. (1990). However, since their seminal article, researchers have found several other empirical regularities in human induction. One of these pertains to the typicality of the conclusion item: The more typical the conclusion is of its superordinate category, the more likely people are to generalize a premise to the conclusion. Thus, for example, people judge the argument koalas require vitamin K for the liver to function therefore tigers require vitamin K for the liver to function to be stronger than the argument koalas require vitamin K for the liver to function therefore guinea pigs require vitamin K for the

liver to function (Hampton & Cannon, 2004). We tested this effect by generating pairs of arguments using stimuli from Rosch (1975) and Richie and Bhatia (2021), which involved six superordinate categories. All arguments used a single item premise as well as either a high- or low-typicality conclusion item taken from the same superordinate category as the premise. This led to 1,392 arguments, which we offered to our models.

The predictions of these models on these argument pairs are shown in Figure 4B and Supplemental Figure S3B. Here, we see that the feature overlap model generated higher argument strength predictions for arguments with highly typical conclusions (purple points) relative to arguments with atypical conclusions (orange points). A separate analysis correlating model predictions with the conclusion's (continuous) typicality ratings also found a strong positive relationship (Table 1 and Supplemental Table S1 present statistical tests). This is because premise items have more features in common with typical conclusions than with atypical conclusions (as assessed by FeatureBERT). The competing NLI models did not all significantly generate this result, though all of their predictions were in the direction of human participants (indicating that they may be able to capture the effect with statistical significance with additional data).

Property Type

Another effect captured by our model is the effect of property type on induction. When generalizing properties from one item to another, people are sensitive to the semantic content of the property itself, and find it easier to generalize when that property is similar to the other properties shared by the premise and conclusion items. For example, people judge the argument bears have a liver with two chambers that act as one therefore whales have a liver with two chambers that act as one to be stronger than the argument tuna have a liver with two chambers that act as one therefore whales have a liver with two chambers that act as one. By contrast, people judge the argument tuna usually travel in a zig-zag trajectory therefore whales travel in a zig-zag trajectory to be stronger than the argument bears usually travel in a zig-zag trajectory therefore whales usually travel in a zig-zag trajectory (Heit & Rubinstein, 1994). Overall, generalization from bears to whales is easier than the generalization from tuna to whales when the property is *anatomical*, but harder when the property is *behavioral*.

Unlike our prior effects, which can be tested by algorithmically generating a large set of arguments spanning categories with varying levels of typicality and similarity, the property type effect requires a hand-curated data set with premise and conclusion categories that vary systematically in terms of their anatomical versus behavioral similarity (e.g., bear/whale vs. tuna/whale, mouse/bat vs. sparrow/bat, lizard/snake vs. worm/snake, etc.). Fortunately, Heit and Rubinstein (1994) have collected such a data set. This has 784 arguments with 28 different properties (14 anatomical and 14 behavioral), and 28 pairs of items (e.g., bear/whale) involving assessments on each of these arguments. Their data set also has average participant ratings of the likelihood of the conclusion given the premise for each item pair on each property type. We performed a median split on these ratings to generate arguments with high versus low participant ratings and offered these arguments to our models.

The predictions of these models on these arguments are shown in Figure 4C and Supplemental Figure S3C. Here, we see that the feature overlap model generated higher argument strength predictions for arguments given high (purple points) versus low (orange points) ratings by Heit and Rubinstein's (1994) participants. A separate analysis correlating model predictions with average participant ratings further demonstrated this positive relationship (Table 1 and Supplemental Table S1 present statistical tests). Our models were able to capture this effect as they place higher weights on the dimensions of the feature vector that have similar words to the argument property. Thus, arguments with premise and conclusion items that overlap on features that are similar to the argument property tend to get higher assessments. Most of the competing NLI models did not generate this result, as they do not explicitly use feature overlap to assess argument strength.

It is worth noting that the above tests evaluated our model's predictions on the responses of Heit and Rubinstein's (1994) participants and not on the match between the properties and the item pairs for which Heit and Rubinstein predicted an anatomic versus behavioral relationship. We did this because we wanted to compare our model's predictions to observed data and not to Heit and Rubinstein's predictions for this data. Indeed, participant responses to five of the 28 item pairs used in the experiment were in the opposite direction to that predicted by Heit and Rubinstein. After excluding these five item pairs, we tested whether our model captured the predicted "manipulation effect" in the data. In particular, we calculated, for each of the remaining 23 item pairs, the difference between our model's average predictions for anatomical versus behavioral properties in the case of a predicted anatomical relationship, or the difference between our model's average predictions for behavioral versus anatomical properties in the case of a predicted behavioral relationship. This gave us 23 predictions (one for each item pair), which should be positive if our model captures Heit and Rubinstein's manipulation effect. Indeed we did find that these predictions were on average positive, though not in a statistically significant manner, likely because of the small sample size in this analysis, average = 0.0002, $t(22) = 0.825$, $p = .42$. Although these findings are promising, they should be interpreted with caution, and further work is necessary to conclusively establish whether our approach can predict the property type effect.

To better understand how our model captures the property type effect, we calculated the GloVe bag-of-words similarity between Heit and Rubinstein's (1994) anatomical and behavioral features and the 25,797 unique participant-generated features (from Devereux et al., 2014) that are the basis of our model. We then extracted the 200 participant-generated features that were most similar to Heit and Rubinstein's 14 anatomical features and the 200 participant-generated features that were most similar to Heit and Rubinstein's 14 behavioral features. Word clouds showing the most frequent words in these two sets of features are shown in Figure 4E and 4F (with word size corresponding to word frequency). As can be seen here, participant-generated features that are most similar to Heit and Rubinstein's anatomical properties involve parts of the body, as well as biologically related words like "mechanism," "vitamin," and "function." These are the dimensions of the feature vector that are prioritized in induction with anatomical properties, making the model more likely to induce these properties to conclusion items that are anatomically similar to the premise items (e.g., bears and whales). By contrast, participant-generated features that are most

similar to Heit and Rubinstein's behavioral properties typically involve verbs, as well as behaviorally related words like "food" and "fast." These are the dimensions of the feature vector that are prioritized in induction with behavioral properties, making the model more likely to induce these properties to conclusion items that are behaviorally similar to the premise items (e.g., tuna and whales).

Property Relevance

The property type effect is one instantiation of a general tendency to use background knowledge, rather than simple assessments of item similarity or feature overlap, in induction. This tendency can take on many forms and can lead to violations of the premise diversity and monotonicity effects when the overlapping features of the premises (which are the relevant features for induction) do not apply to the conclusion. For example, people judge the argument skunks and deer have a given property therefore animals have that property to be stronger than the argument skunks and stink bugs have a given property therefore animals have that property (Medin et al., 2003). This violates the premise diversity effect as skunks and stink bugs are judged to be less similar than skunks and deer. The reason why we observe this violation is because skunks and stink bugs have a salient overlapping property (create a foul odor) that is relevant to the induction problem but is not shared with other animals, making it harder to generalize when they are the premise items.

The diversity and monotonicity violations caused by the property relevance effect require carefully curated stimuli which cannot be algorithmically generated as with prior findings. Medin et al. (2003) have collected one such data set with 34 arguments. Their data set also has average participant ratings of the strength of each argument. We performed a median split on these ratings to generate arguments with high versus low participant ratings and offered these arguments to our models. The predictions of these models on these arguments are shown in Figure 4D and Supplemental Figure S3C. Here, we see that the feature overlap model generated higher argument strength predictions for arguments given high (purple points) versus low (orange points) ratings by Medin et al.'s participants, though these differences do not cross the threshold for significance, likely due to small sample of arguments used in this exercise. A separate analysis correlating model predictions with continuous participant ratings further demonstrated this positive relationship (Table 1 and Supplemental Table S1 present statistical tests). The reason that the Feature Overlap model generated correct directional predictions for these effects is because overlapping features of the premise categories play a larger role in the feature overlap assessment. Thus, premises with overlapping features not shared with the conclusion item are given lower assessments by our model. The competing NLI models do not all generate this result, and the ones that do typically have much smaller *t*-values.

The above tests compared our model's predictions to observed data and not to Medin et al.'s (2003) predictions for this data. Indeed, participant responses to four of the 17 argument pairs used in the experiment were in the opposite direction to that predicted by Medin et al. After excluding these four argument pairs, we tested whether our model replicated the predicted "manipulation effect" in the data. In particular, we calculated, for each of the remaining 13 argument pairs, the difference between our model's average predictions for arguments with relevant versus irrelevant properties. This gave us 13 predictions (one for each argument pair), which should be positive if our model captures Medin et al.'s manipulation effect. Indeed we did find that these predictions were on average positive, though not in a statistically significant manner, likely because of the small sample size, average = 0.006, $t(12) = 0.300$, $p = .77$. Although these findings are promising, they should be interpreted with caution, and further work is necessary to conclusively establish whether our approach can predict the property relevance effect.

Limitations and Extensions

Asymmetry and Projection

This is one important effect that lies outside of the descriptive scope of the feature overlap model, as formalized using the cosine similarity operator. This has to do asymmetries in generalizing from a premise item to a conclusion item. For example, people judge the argument mice have a lower body temperature at infancy than at maturity therefore bats have a lower body temperature at infancy than at maturity to be stronger than the argument bats have a lower body temperature at infancy than at maturity therefore mice have a lower body temperature at infancy than at maturity (Osherson et al., 1990). More generally, people find it easier to generalize from a common item to an uncommon item than vice versa. The feature overlap model introduced above does not generate asymmetries since the cosine similarity metric is symmetric. This issue can easily be remedied by replacing cosine similarity with an asymmetric metric, for example, one in which the premise item's feature vector is projected onto the conclusion item's feature vector (as initially suggested by Sloman, 1993). As common items typically have richer feature representations, the feature vector projection from common to uncommon item will generate a higher overlap measurement than vice versa. In Supplemental Figure S4A, we show our feature overlap model's predictions for the asymmetry effects documented in Sloman (1993) when such a projection mechanism is used (see Supplemental Materials, for additional technical details). This figure shows that the feature projection model successfully captured all observed asymmetries in Sloman (1993).

We also tested the feature projection model on the Rips and Osherson data sets, our new experimental data sets, and the problems used in the Empirical Regularities section above. The results of this are shown in Supplemental Table S3. Here, we can see that the projection model is unable to capture nonmonotonicity effects (and associated property relevance effects): More premise items always lead to larger premise vectors which causes higher projections onto the conclusion vector, and thus higher assessments of argument strength. Additionally, the normalization inherent in cosine similarity helps regulate the effect of the premise feature vector on model predictions; without normalization the magnitude of this vector can greatly distort vector projection. This is why the projection model performed poorly on the predictive accuracy tests for Experiment 2 of Osherson et al. (1990), which had arguments with multiple premise items (and thus very large premise feature vectors).

The failure of the projection metric is the main reason why we chose to focus on the cosine similarity implementation in the main text. However, thanks to a suggestion of a reviewer, we also tried out

a hybrid model that combines the assumptions of the cosine similarity and projection metrics. Intuitively this model uses a flexible weight to regulate the effect of the magnitude of the premise item's feature vector on the resulting judgment. With a correctly calibrated weight, we find that it is possible to avoid the problems of cosine similarity (which, by normalizing the premise item's feature vector, completely ignores its magnitude, and thus does not generate asymmetry effects) as well as the problems of projection (which, by not normalizing the premise item's feature vector at all, leads to an oversensitivity to this vector's magnitude, creating issues with multiple premises and with nonmonotonicity effects). The Supplemental Materials provide technical details of this model, Supplemental Figure S4B shows its predictions for the asymmetry effects in Sloman (1993), and Supplemental Table S3 shows its predictions for the Rips and Osherson data sets, our new experimental data sets, and the problems used in the empirical regularities section above. Here, we can see that the hybrid model captures all effects (though its asymmetry predictions are a weaker than those of the projection model). Since we calibrated the weights of the hybrid model post hoc, there is an additional level of flexibility in this model that makes comparisons to cosine similarity and projection difficult to interpret (the cosine similarity and projection metrics do not have any flexible parameters and are not "fit" to the empirical data in any way). Nonetheless, these results show that better models are possible and that our modeling framework could be improved with additional assumptions. We provide a detailed discussion of this in the next section.

Causal and Ecological Knowledge

Asymmetry is not the only effect that lies outside the scope of the model put forth in this article. In the past 2 decades, much of the focus of inductive reasoning research has shifted to the study of relational and ecological factors at play in induction. For example, Medin et al. (2003) have shown that people use causal relationships between premise and conclusion items to generalize properties. This can lead to asymmetries in induction, so that people are more likely to generalize properties from prey to predators than vice versa. For example, gazelles have property X12 therefore lions have property X12 is judged to be stronger than lions have property X12 therefore gazelles have property X12 (Medin et al., 2003). Other work has found that participant beliefs about how the items in the premise were sampled by the experimenter influence their endorsement of the conclusion. When people believe that premises have been sampled randomly, effects like premise diversity tend to be diminished (Ransom et al., 2016).

The knowledge base used in this article involves single place predicates. This knowledge base is passed through a fairly simple algorithm that calculates the extent of featural similarity. Thus, we are not able to explain the effects of Medin et al. (2003), Ransom et al. (2016), or others (e.g., Bright & Feeney, 2014; Hayes et al., 2019; Rehder, 2006; see Hayes & Heit, 2018, for a summary). Indeed, this may also be why we are unable to make strong predictions for property relevance effects. Although our tests are underpowered (there are only 34 arguments used to test for property relevance), it is likely that human responses depend not only on biased assessments of feature overlap (as put forth by our model) but also on more structured computations involving complex relationships and participant beliefs.

That said, this does not imply that our modeling framework is fundamentally incompatible with structured theories of inductive reasoning. For example, our model currently gives a higher weight to anatomical features if the nonblank property is anatomical (vs. behavioral), explaining Heit and Rubinstein's (1994) property type effects. It could be possible that a similar property similarity bias could be implemented if the model detects a prey–predator relationship between the premise and conclusion items, as with the gazelle/lion example given above. Such a model would retain the core assumptions of Sloman's model (i.e., that judgments of argument strength depend on a comparison of the premise and conclusion's features) while also implementing the feature relevance insights of Medin et al. (2003; i.e., that the features that are used in induction depend on more complex relations between the premise and conclusion items). It would also make new predictions, for example, that the predator/prey asymmetry would emerge for biological and chemical properties (e.g., has a higher potassium concentration in their blood than humans) but not behavioral properties (e.g., travels in groups).

Of course, in order to do this, the current framework would need to be supplemented with knowledge of the relations between items. Fortunately, there have been recent advances that solve this problem using a combination of LLM representations and psychologically plausible reasoning rules (Lu et al., 2019; Snefjella et al., 2022). Combining these advances with the framework advanced in the present article is an exciting direction for future work.

The Feature-BERT knowledge base could also be integrated into Bayesian updating rules, such as those put forth by Ransom et al. (2016), in order to explain premise sampling effects. Here, Feature-BERT's outputs would specify the probabilities of hypotheses at play in the argument. These probabilities would be integrated with the participant's beliefs about how the stimuli were selected using a Bayesian reasoning module. Such a module may also be uniquely suited to extracting latent structures (including casual structures) that guide and constrain induction (Kemp & Tenenbaum, 2009). It could also be used to explain individual differences in reasoning such as the effect of expertise (Medin et al., 2003; also see Hayes & Heit, 2018, for a discussion): Domain experts generalize based on property inheritance relations and causal relations that are often different to the more superficial assessments of feature overlap used by nonexperts. We encourage researchers to explore the applicability of such *neurosymbolic* models of naturalistic cognition (Mao et al., 2019; Marcus, 2020; Nye et al., 2020), and, by doing so, extend our approach to more complex reasoning problems.

Theory and Prediction

The past section has implicitly assumed that there is some theoretical value to the development of quantitative models capable of predicting naturalistic high-level cognition. But why might this be the case? Why should psychologists care that we can predict human responses to thousands of diverse induction problems and formally replicate several observed empirical regularities? After all, the kind of modeling pipeline used in this article does not involve the discovery of radically new theories of cognitive processes: The core reasoning algorithms at play in our model—algorithms that compare the features of premise and conclusion items to assess the strength of

an induction argument—are largely the same as in Sloman (1993). Additionally, the knowledge representations that make up our model are certainly not obtained from realistic learning processes: Feature-BERT is not (and does not attempt to be) a cognitive theory of how item-feature knowledge is acquired. It is more like an automated coder, that cheaply and judges the truth values of millions of simple sentences in a human-like manner. So, in this sense, the core cognitive processes in our model are largely the same as those in previous toy models and verbal theories.

Although these are legitimate points, we believe that psychological theorizing involves more than just the discovery of new cognitive processes. It is just as important to use existing theories to predict behavior, as it is to formulate these theories in the first place. After all, without prediction, we cannot rigorously assess the descriptive scope of the theory, that is, the amount of the variation in the data that the theory can explain. Prediction can also help researchers determine the set of assumptions that are necessary to best describe data, and by doing so, can lead to the refinement and improvement of theories (see Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010, for discussions). Current models of induction are unable to make a priori quantitative predictions for the thousands of arguments that have been used in empirical induction research. In this way, the real explanatory scope of research on human induction remains unknown. This is a fundamentally theoretical problem.

The present article solves this problem, and by doing so shows that Sloman's (1993) feature-based model provides a good account of human data, though it needs to be altered slightly to do so. For example, the projection metric initially proposed by Sloman performs poorly in our quantitative tests on the Rips (1975) and Osherson et al. (1990) data sets. This projection metric is also unable to generate the nonmonotonicities documented by Osherson et al. (1990) and Medin et al. (2003). By contrast, a cosine similarity variant of this model performs much better, at the expense of explaining asymmetry in human induction (Sloman, 1993). Our preliminary tests also show that a hybrid between the cosine similarity and projection metrics can explain all effects simultaneously. Finally, we have found that activating features based on their similarity to the central property in the induction argument allows us to explain property type effects (Heit & Rubinstein, 1994). In this way, we have shown how the core insights of other leading models can be implemented in the feature-based framework, synthesizing multiple theoretical perspectives in induction research.

Several researchers have already highlighted the importance of prediction for psychological research (Hofman et al., 2021; Yarkoni & Westfall, 2017). Indeed many other areas of cognitive science, including perception, categorization, decision making, semantic cognition, and memory research have moved from verbal theories and toy problems, to quantitative cognitive models (Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010), to powerful computational models capable making quantitative predictions over large naturalistic stimuli sets (Battleday et al., 2021; Bhatia, 2019; Bhatia & Stewart, 2018; Gandhi et al., 2022; Hebart et al., 2020; Hills et al., 2012; Richie et al., 2022; Sanders & Nosofsky, 2020; Trueblood et al., 2021; Zou & Bhatia, 2021; see Bhatia & Aka, 2022, for a review and discussion). Typically, these new models apply a similar pipeline to ours: Artificial intelligence tools like deep neural networks to extract representations from language or image data, combined with psychologically plausible theories for manipulating and processing this information. We are also seeing similar developments in the study of other reasoning tasks, such as analogy (Ichien et al., 2022; Lu et al., 2019, 2022). Some of this work has been published in this very journal, which has also published articles on purely statistical problems in model fitting and the evaluation of model predictions. Thus, there is no doubt that prediction is (rightfully) a central focus of much of contemporary theoretical psychology and cognitive science.

Why is it that so much psychological theorizing has focused on the modeling of human information processing mechanisms rather than the knowledge representations to which these mechanisms are applied? We believe that this mindset stems from the computer revolution in the 1950s and 1960s, which kickstarted the study of human cognition. As Simon (1979) describes it in the first chapter of Models of Thought:

> The information processing revolution that has occurred during these years has completely changed the face of cognitive psychology. It has introduced computer programming languages as formal ("mathematical") languages for expressing theories of human mental processes; and has introduced the computers themselves to simulate these processes and thereby make behavioral predictions for testing the theories. (p. 9)

It turns out that we are currently in the middle of a second computer revolution, one that is enabled by the rapid growth of digital data sets and the development of new technologies for extracting information from these data sets. In this new era, the theorist's tool kit has expanded, and we can use computers not only to specify information processing algorithms but also aspects of the world knowledge on which these algorithms operate. Although it is currently unclear whether new artificial intelligence (AI) models encode structured representations (multiplace predicates organized into frames and scripts), they can accurately mimic knowledge of simple concept-feature pairings (one-place predicates involving basic concepts and categories; Bhatia & Richie, 2022), which can be used to model aspects of high-level cognition, as shown in this article.

Discussion

The study of inductive reasoning has been one of the most active areas of research in cognitive science and psychology. Researchers have documented several empirical regularities in human induction of properties across concepts and categories and have developed formal theories to account for these regularities (Hayes & Heit, 2018; Heit, 1998, 2000; Kemp & Tenenbaum, 2009; Medin et al., 2003; Osherson et al., 1990; Sloman, 1993). Here, we show how one theory, the feature-based model (Sloman, 1993), can be combined with leading LLMs (Brown et al., 2020; Devlin et al., 2018; He et al., 2021) to successfully model human induction. We have demonstrated the power of our approach in two ways. First, we have correlated our feature overlap model's predictions with human assessments of argument strength obtained in prior work (Osherson et al., 1990; Rips, 1975) as well as in four new experiments. Here, we have found that the feature overlap model achieves consistently high correlations with human responses, and greatly outperforms LLMs that do not use explicit inductive reasoning algorithms. Secondly, we have tested whether the feature overlap model replicates observed empirical regularities using both original experimental stimuli, as well as large sets of new algorithmically generated inductive reasoning arguments. Here, we have found that feature overlap model behaves in a human-

like manner; that is, it is sensitive to the typicality, similarity, specificity, and category membership relationship of items, the number of premises, and the semantic content of the argument properties, in the same way that human participants are (Hampton & Cannon, 2004; Heit & Rubinstein, 1994; Medin et al., 2003; Osherson et al., 1990; Rips, 1975; Sloman, 1993, 1998). Again, leading LLMs fail to mimic these behavioral patterns.

Our approach is the first computational model capable of making accurate quantitative predictions for arbitrary natural language induction arguments. It is successful because it combines the relative strengths of two influential research programs. Psychological theories describe intelligent human-like reasoning processes, whereas LLMs possess the knowledge representations necessary to use these reasoning processes in everyday induction. For this reason, our integrative approach—which feeds knowledge representations from LLMs into human-like inductive reasoning processes—is able to generate sophisticated and realistic responses to arguments involving arbitrary concepts and properties. In fact, the tests in this article involve over 16,000 existing and new induction problems, spanning several distinct domains, greatly exceeding the size and diversity of data sets used in prior psychological research. Psychological theories, by themselves, are not be able to make predictions for these problems as they have been developed on hand-coded ontologies with only a small set of concepts and properties. Additionally, even though LLMs for NLI are able to process and respond to induction problems, we find that they do not do so in a human-like manner. Cognitively plausible reasoning algorithms, like the feature overlap model, are necessary to manipulate and transform LLM representations, in order to mimic human behavior.

We are not the first to highlight the value of combining existing cognitive models with newer AI systems trained on large-scale data. Previously, researchers have shown that integrative approaches, like ours, are useful for modeling human perception, categorization, semantic cognition, memory, decision making, and analogical reasoning for natural concepts and categories (Battleday et al., 2021; Bhatia, 2019; Bhatia & Stewart, 2018; Gandhi et al., 2022; Hebart et al., 2020; Hills et al., 2012; Lu et al., 2019, 2022; Richie et al., 2022; Sanders & Nosofsky, 2020; Trueblood et al., 2021; Zou & Bhatia, 2021; see Bhatia & Aka, 2022, for a review and discussion). It is clear that if psychologists wish to model naturalistic cognition and behavior, they need to equip their theories with rich world knowledge. The present article (along with the work summarized in this paragraph) shows how the knowledge representations of LLMs can be used to solve this important research problem.

LLMs trace their intellectual lineage to older models of human linguistic and semantic cognition (Hinton, 1986; Rogers & McClelland, 2004). Thus, unsurprisingly, researchers have shown that LLMs are able to capture aspects of human linguistic and semantic processing (Goldstein et al., 2022; Linzen & Baroni, 2021; Manning et al., 2020; McClelland et al., 2020) and even mimic some types of reasoning (Bhatia, 2017; Dasgupta et al., 2022). The Feature-BERT model is one example of this (Bhatia & Richie, 2022). This model does not only predict the features that people associate with different concepts; it also captures several core patterns of human semantic verification, and by doing so shows how these patterns are the natural byproducts of semantic processing in deep neural networks. By using Feature-BERT in the present article, we are illustrating one way in which cognitively plausible high-level reasoning algorithms can interface with realistic semantic representations obtained from deep neural networks. We speculate that people may also be engaging in a similar set of operations. In other words, people may use statistical patterns in language data (as well as perhaps perceptual data) to approximate the distribution of features across concepts. This distribution may then be fed into a second, higher level set of reasoning processes, for induction with new features and concepts. We have not made any concrete claims about how these reasoning processes are implemented in the mind. However, it is worth noting that the feature-based model was initially proposed as a neural network (Sloman, 1993), and the types of vector multiplication operations at play in the present article are best interpreted as interactions between interconnected nodes in a large parallel distributed processing system.

It may also be possible to use closely related neural network architectures for semantic cognition, such as the model put forth by Rogers and McClelland (2004). This model takes items as inputs and generates, as outputs, the features that it believes the items to possess. At its core, this model generalizes features from one item to another based on the structure of covariance of features across items in its training data. Although, to our knowledge, this model has not been designed for complex premises consisting of multiple items (it is primarily a model of realistic feature learning rather than inductive reasoning), it is likely that some variant could be used to account for the set of empirical regularities discussed in this article. Of course, such a model would need to possess featural representations for thousands of common concepts and categories in order for it to make quantitative predictions. One way to do this is to train it on Feature-BERT's underlying knowledge base. If successful, the resulting model would be able to provide a single comprehensive account of both feature learning and inductive reasoning with learnt features.

This article has also offered us the opportunity to reflect on the theoretical value of our modeling framework. Unlike many theory articles in psychology, we are not proposing new cognitive processes for solving inductive reasoning tasks. Rather our goal is to show how existing models of reasoning can be extended to make good predictions over large and unconstrained stimuli sets. This is a central goal of psychological theorizing, one that we hope will play a larger role in our field as it responds to the challenges posed by (and opportunities generated by) increasingly powerful AI models of high-level cognition (see Hofman et al., 2021; Yarkoni & Westfall, 2017 for a discussion). Of course, our modeling framework also opens up several new practical applications that would not be possible with verbal theories or toy models of inductive reasoning. For example, with our framework it may be possible to formally model inductive reasoning in developmental, social, and clinical contexts, to influence and improve human cognition in important real-world domains.

We would like to conclude by highlighting one empirical regularity that is outside the scope of the feature overlap model, as specified in the main text of this article. This involves asymmetries in generalizing from a premise item to a conclusion item. For example, people are more likely to generalize a property from `mice` to `bats` than vice versa (Osherson et al., 1990; Sloman, 1993). The feature overlap model used in this article does not generate asymmetries since the cosine similarity metric is symmetric. However, we have shown how an alternate metric that is based on vector projection instead of cosine similarity (and is closer in spirit to Sloman's, 1993, original proposal) can provide a good account of asymmetry effects. We have also shown that a hybrid model that

combines the properties of the cosine similarity and projection approaches can capture all empirical regularities simultaneously. The success of this exercise illustrates that our general modeling pipeline can be used to develop and test new cognitive process theories of induction. We anticipate that future work will implement more complex feature overlap operations, as well as other mechanisms that reflect the use of causal or ecological beliefs, thereby improving upon the performance in the present article.

Conclusion

How do people generalize from what they know to make predictions in new settings, and how can we build models that perform this type of generalization in a human-like manner? We address these questions by integrating psychological theories of human induction (which specify intelligent, cognitively plausible, reasoning algorithms) with leading models from AI research (which possess the world knowledge necessary for everyday reasoning). We find that by combining these two approaches, we are able to generate better predictions than by using each approach by itself. In doing so, we show how existing cognitive theories can be combined with knowledge representations derived from LLMs, to better understand and predict high-level human cognition.

References

Anderson, J. R., & Reder, L. M. (1974). Negative judgments in and about semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(6), 664–681. https://doi.org/10.1016/S0022-5371(74)80054-X

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, *1505*(1), 55–78. https://doi.org/10.1111/nyas.14593

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823. https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Bhatia, S., & Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000319

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88. https://doi.org/10.1016/j.cognition.2018.05.025

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Bright, A. K., & Feeney, A. (2014). The engine of thought is a hybrid: Roles of associative and structured knowledge in reasoning. *Journal of Experimental Psychology: General*, *143*(6), 2082–2102. https://doi.org/10.1037/a0037653

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). *Language models are few-shot learners*. Neural Information Processing Systems.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. SAGE Publications.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 643–658. https://doi.org/10.1037/0278-7393.32.4.643

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. https://doi.org/10.48550/arXiv.2207.07051

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. https://doi.org/10.3758/s13428-013-0420-4

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pretraining of deep bidirectional transformers for language understanding. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science*, *33*(4), 579–594. https://doi.org/10.1177/0956797621104342 6

Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974). Production frequency and the verification of quantified statements. *Journal of Verbal Learning and Verbal Behavior*, *13*(3), 237–254. https://doi.org/10.1016/S0022-5371(74)80061-7

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, *12*(2), 151–164. https://doi.org/10.1016/0010-0277(82)90010-5

Hampton, J. A. (1984). The verification of category and property statements. *Memory & Cognition*, *12*(4), 345–354. https://doi.org/10.3758/BF03198294

Hampton, J. A., & Cannon, I. (2004). Category-based induction: An effect of conclusion typicality. *Memory & Cognition*, *32*(2), 235–243. https://doi.org/10.3758/BF03196855

Han, S. J., Ransom, K., Perfors, A., & Kemp, C. (2022). *Human-like property induction is a challenge for large language models* [Conference session]. Proceedings of the 44th Annual Conference of the Cognitive Science Society, Toronto, Canada.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3), Article e1459. https://doi.org/10.1002/wcs.1459

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, *26*(3), 1043–1050. https://doi.org/10.3758/s13423-018-1562-2

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention* [Conference session]. International Conference on Learning Representations. https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*(4), 569–592. https://doi.org/10.3758/BF03212996

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 411–422. https://doi.org/10.1037/0278-7393.20.2.411

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://aclanthology.org/J15-4004/

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Hinton, G. E. (1986). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). Clarendon Press.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Ichien, N., Lu, H., & Holyoak, K. J. (2022). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 108–121. https://doi.org/10.1037/xlm0001010

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. https://doi.org/10.1037/a0014282

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE Publications.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2020). *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. Association for Computational Linguistics.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2020). *Roberta: A robustly optimized bert pretraining approach* [Conference session]. International Conference on Learning Representations.

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*, *129*(5), 1078–1103. https://doi.org/10.1037/rev0000358

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269. https://doi.org/10.1016/S0022-5371(84)90170-1

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). *The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision* [Conference session]. International Conference on Learning Representations.

Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2002.06177

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517–532. https://doi.org/10.3758/BF03196515

Misra, K., Rayz, J. T., & Ettinger, A. (2022). *A property induction framework for neural language models* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society.

Nye, M., Solar-Lezama, A., Tenenbaum, J., & Lake, B. M. (2020). Learning compositional rules via neural program synthesis. *Advances in Neural Information Processing Systems*, *33*, 10832–10842.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200. https://doi.org/10.1037/0033-295X.97.2.185

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796. https://doi.org/10.1111/cogs.12308

Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory & Cognition*, *34*(1), 3–16. https://doi.org/10.3758/BF03193382

Richie, R., Aka, A., & Bhatia, S. (2022). Free association in a neural network. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000396

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), Article e13030. https://doi.org/10.1111/cogs.13030

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 665–681. https://doi.org/10.1016/S0022-5371(75)80055-7

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. https://doi.org/10.1016/S0022-5371(73)80056-8

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press. https://doi.org/10.7551/mitpress/6161.001.0001

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Roth, E. M., & Mervis, C. B. (1983). Fuzzy set theory and class inclusion relations in semantic categories. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 509–525. https://doi.org/10.1016/S0022-5371(83)90310-9

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category

domain. *Computational Brain & Behavior*, *3*(3), 229–251. https://doi.org/10.1007/s42113-020-00073-z

Simon, H. A. (1979). *Models of thought*. Yale University Press.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*(2), 231–280. https://doi.org/10.1006/cogp.1993.1006

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33. https://doi.org/10.1006/cogp.1997.0672

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. https://doi.org/10.1037/h0036351

Snefjella, B., Ichien, N., Holyoak, K., & Lu, H. (2022). Predicting human judgments of relational similarity: A comparison of computational models based on vector representations of meaning. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the Cognitive Science Society*. Cognitive Science Society.

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., & Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition*, *212*, Article 104713. https://doi.org/10.1016/j.cognition.2021.104713

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *GLUE: A multi-task benchmark and analysis platform for natural language understanding* [Conference session]. International Conference on Learning Representations.

Whitten, W. B., II, Suter, W. N., & Frank, M. L. (1979). Bidirectional synonym ratings of 464 noun pairs. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 109–127. https://doi.org/10.1016/S0022-5371(79)90604-2

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In M. A. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (long papers)* (Vol. 41, pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, W., & Bhatia, S. (2021). *Learning new categories for natural objects* [Conference session]. Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.

Received January 19, 2023
Revision received June 28, 2023
Accepted July 14, 2023 ■