Cognitive Systems Research 83 (2024) 101155

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Inductive reasoning in humans and large language models

Simon Jerome Han *, Keith J. Ransom, Andrew Perfors, Charles Kemp

The University of Melbourne, Parkville, Australia

A R T I C L E I N F O

Action editor: Matteo Colombo

Dataset link: https://github.com/S-J-HAN/InductiveReasoningInLargeLanguageModels

Keywords:
Reasoning
Property induction
Category-based induction
Non-monotonicity
Neural networks
GPT-3.5
GPT-4
AI
Large language models
Representation

A B S T R A C T

The impressive recent performance of large language models has led many to wonder to what extent they can serve as models of general intelligence or are similar to human cognition. We address this issue by applying GPT-3.5 and GPT-4 to a classic problem in human inductive reasoning known as property induction. Over two experiments, we elicit human judgments on a range of property induction tasks spanning multiple domains. Although GPT-3.5 struggles to capture many aspects of human behavior, GPT-4 is much more successful: for the most part, its performance qualitatively matches that of humans, and the only notable exception is its failure to capture the phenomenon of premise non-monotonicity. Our work demonstrates that property induction allows for interesting comparisons between human and machine intelligence and provides two large datasets that can serve as benchmarks for future work in this vein.

In late 2022 and early 2023, large language models (LLMs) exploded into the public arena and captured the imagination of academic researchers and the general public alike. Systems such as ChatGPT and GPT-4 are so adept at engaging in natural conversations on a broad range of topics that even sober teams of researchers have concluded that these models show "sparks of artificial general intelligence" (Bubeck et al., 2023). As a result, there is currently intense interest in the scope and limitations of these models and the ways in which they may transform society.

LLMs had been extensively studied even before their recent surge in popularity, and for several years there has been an active research area that aims to carefully evaluate how their abilities compare with those of humans. Many families of tasks are used in this literature (Chang & Bergen, 2023), including some that specifically target linguistic abilities (Hu, Gauthier, Qian, Wilcox, & Levy, 2020) and others that target commonsense knowledge and logical reasoning (Rae et al., 2021). Here we propose that the set of existing tasks can be usefully supplemented by drawing on the extensive psychological literature on inductive reasoning. To support this general claim we explore the extent to which two generations of the GPT model (GPT-3.5 and GPT-4) are able to account for core phenomena in human property induction (also known as category-based or categorical induction).

Inductive reasoning is a fundamental cognitive challenge that requires arriving at plausible conclusions in the face of uncertainty (Holland, Holyoak, Nisbett, & Thagard, 1986; Sloman & Lagnado, 2005). An inference is deductive if the conclusion follows with certainty given the available evidence, but inductive if the conclusion is plausible but not guaranteed. Because most everyday reasoning problems involve sparse, noisy, or uncertain data, most of these problems require inductive rather than deductive reasoning (Chater, Oaksford, Hahn, & Heit, 2011). Within the AI literature, work on inductive reasoning falls under many different headings including non-monotonic reasoning (Brewka et al., 1997), commonsense reasoning (Davis & Marcus, 2015) and natural language inference (Storks, Gao, & Chai, 2019). Psychologists have also studied many varieties of induction, including generalization (Shepard, 1987), categorization (Pothos & Wills, 2011), and analogical reasoning (Vosniadou & Ortony, 1989). Here we focus on property induction because this task is easily formulated using simple linguistic stimuli, and because the literature on this task is relatively rich.

In a property induction task, people are given premises that indicate that a property is shared by one or more categories and must assess whether the property is shared by a different category (Rips, 1975; Sloman & Lagnado, 2005). For example, the top left argument in Fig. 1

* Corresponding author.
E-mail address: simon.jerome.han@gmail.com (S.J. Han).

https://doi.org/10.1016/j.cogsys.2023.101155
Received 24 April 2023; Received in revised form 16 July 2023; Accepted 23 July 2023
Available online 9 August 2023

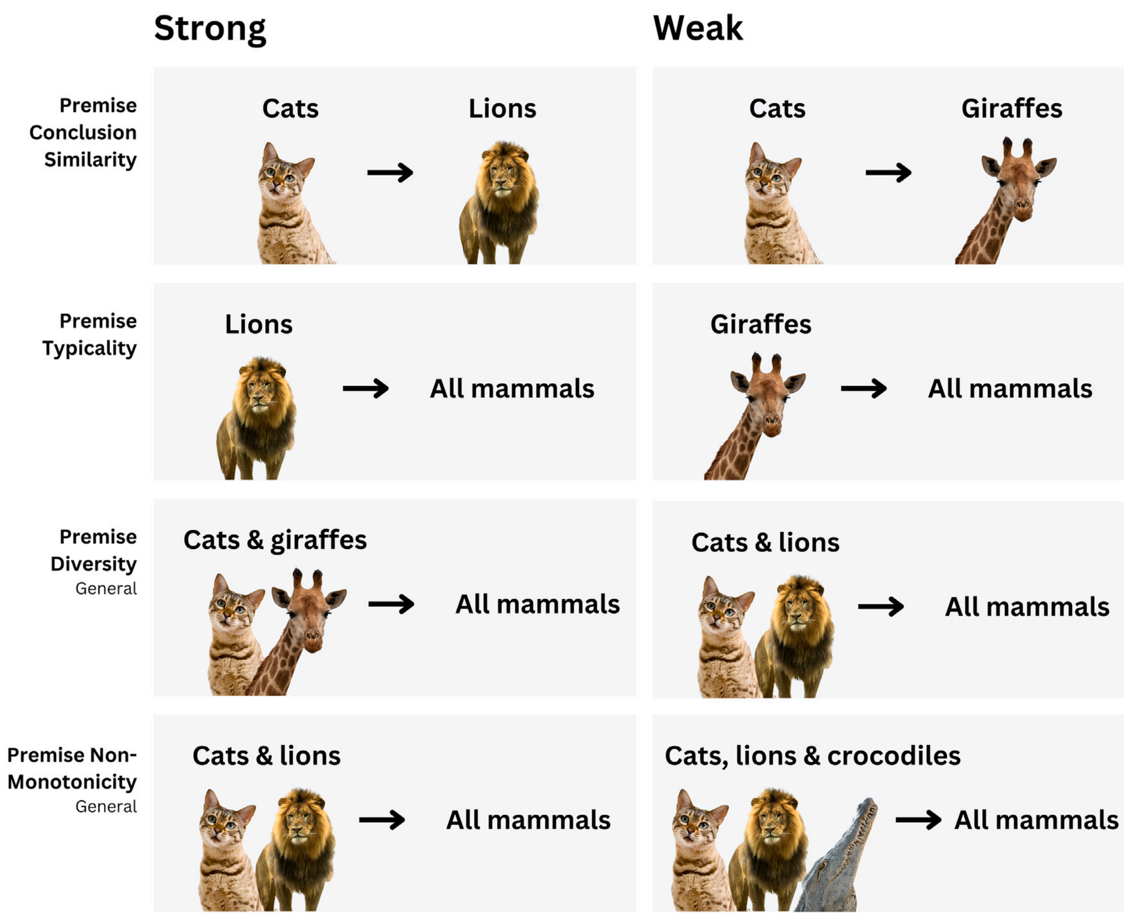

Fig. 1. Schematic illustration of selected property induction phenomena. The column on the left depicts arguments that people perceive as stronger than the corresponding column on the right. For instance, people who are told that Cats have some property are more willing to conclude that similar animals like Lions have the property than they are to conclude that dissimilar ones like Giraffes do. This phenomenon, shown in the top row, is known as Premise-Conclusion Similarity. The figure depicts only four of the eleven phenomena we investigate in this paper.

might be presented by informing participants that cats have sesamoid bones, then asking them whether lions also have sesamoid bones. Fig. 1 compares four pairs of arguments, and the distinctions between strong and weak arguments in the top two rows are relatively obvious. In the first row, cats are more similar to lions than to giraffes, and in the second row, lions are more typical mammals than are giraffes. The final two rows show how property induction tasks can elicit more sophisticated kinds of reasoning. In the third row, the argument on the left is stronger even though giraffes are less typical mammals than lions (see row 2). This difference in typicality between lions and giraffes is dominated by the fact that cats and giraffes form a more diverse set of premises than cats and lions. The final row suggests that adding a premise to an argument can sometimes make an argument weaker. Under many circumstances, observing an additional species with sesamoid bones should provide increased confidence that all mammals have sesamoid bones, so examples of Non-Monotonicity are especially interesting.

Property induction is an appealingly simple task that has been used to study the reasoning of children (Carey, 1985) and adults from a broad range of cultural backgrounds (López, Atran, Coley, Medin, & Smith, 1997). Despite this apparent simplicity, the task yields a rich range of phenomena that draw on many kinds of knowledge (see Hayes & Heit, 2018 for a review). This knowledge includes not just similarity (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990), but also causal relationships (Medin, Coley, Storms, & Hayes, 2003) and assumptions about the process by which the premises were generated (Ransom, Perfors, & Navarro, 2016). This range of inductive phenomena – from simple similarity-based effects to theory-based effects that draw on richer kinds of knowledge – corresponds to a sequence of increasingly difficult challenges for LLMs and other computational models (Kemp & Tenenbaum, 2009; Rogers & McClelland, 2004; Sloman, 1993). As such, property induction tasks could potentially lead to benchmarks that help to drive continued progress in computer science and AI. Some of the benchmarks currently used to evaluate LLMs focus on inductive problems and are directly inspired by psychological research (Jiang et al., 2023; Sap, Rashkin, Chen, LeBras, & Choi, 2019), but to our knowledge none of these benchmarks considers the task of property induction. Here we introduce two property induction data sets that are relatively large by the standards of psychological research, and thus represent an initial step towards a comprehensive property induction benchmark.

For psychologists, property induction is an important tool for assessing LLMs and predecessors such as LSA (Landauer & Dumais, 1997) as computational accounts of the acquisition, use, and representation of semantic knowledge. Recent work has evaluated the extent to which LLMs account for human similarity ratings, typicality ratings, and response times (Bhatia & Richie, 2021; Lake & Murphy, 2021) in semantic verification, and several groups have evaluated LLMs on inductive phenomena inspired by the psychological literature, including analogical reasoning (Webb, Holyoak, & Lu, 2022), pragmatic reasoning (Lipkin, Wong, Grand, & Tenenbaum, 2023), causal reasoning (Kıcıman, Ness, Sharma, & Tan, 2023) and social reasoning (Shapira et al., 2023; Ullman, 2023). Closest to the current paper is a study by Misra, Ettinger, and Taylor Rayz (2021), who focus on typicality and include property induction as one of the tasks that they consider. Typicality is among the phenomena considered here, but we investigate many others as well.

Table 1
Eleven property induction phenomena introduced by Osherson et al. (1990) and investigated in this paper. The second column is based on the levels occupied by premises and conclusion in a category hierarchy. For specific arguments, premises and conclusion lie at the same level, but for general arguments the conclusion lies at a higher level than the premises. Premises are indicated in the brackets and the conclusion is on the right of the arrow.

Phenomenon	Type	Stronger argument	Weaker argument
Similarity	Specific	{ROBIN, BLUEJAY} → SPARROW	{ROBIN, BLUEJAY} → GOOSE
Typicality	General	ROBIN → BIRD	PENGUIN → BIRD
Specificity	General	{BLUEJAY, FALCON} → BIRD	{BLUEJAY, FALCON} → ANIMAL
Monotonicity	General	{SPARROW, EAGLE, HAWK} → BIRD	{SPARROW, EAGLE} → BIRD
Monotonicity	Specific	{PIG, WOLF, FOX} → GORILLA	{PIG, WOLF} → GORILLA
Diversity	General	{HIPPO, HAMSTER} → MAMMAL	{HIPPO, RHINO} → MAMMAL
Diversity	Specific	{LION, GIRAFFE} → RABBIT	{LION, TIGER} → RABBIT
Non-Monotonicity	General	{CROW, PEACOCK} → BIRD	{CROW, PEACOCK, RABBIT} → BIRD
Non-Monotonicity	Specific	FLY → BEE	{FLY, ORANGUTAN} → BEE
Asymmetry	Specific	MOUSE → BAT	BAT → MOUSE
Inclusion Fallacy	Both	ROBIN → BIRD	ROBIN → OSTRICH

The next section introduces the inductive phenomena that we analyze, along with a theoretical account of these phenomena known as the similarity-coverage model (Osherson et al., 1990). We then present two new datasets that we collected to characterize inductive reasoning in humans, and use them to evaluate the performance of both GPT-3.5 and GPT-4. To preview our results, we find that GPT-4 accounts well for all of the phenomena that we investigate except Non-Monotonicity.

Because the literature on LLMs is moving so fast, it seems useful to document the period during which this work was carried out. At the time we developed the study, GPT-3 was the most advanced version available, and we chose to focus on a set of phenomena that included some that were within the reach of GPT-3 but others that seemed more challenging. GPT-4 became available shortly before we submitted this work for publication, and we found that it performed relatively well when we included it in our evaluation. Had GPT-4 been available when we designed the study, we would have aimed to include additional inductive phenomena that seemed clearly out of reach for GPT-3 but that may be feasible for GPT-4. The closing sections of the paper discuss some of these phenomena and suggest additional directions that future work on LLMs and inductive reasoning can pursue.

Inductive phenomena

We follow a long tradition of studies that examine inductive reasoning by focusing on property induction with semantically "blank" or unfamiliar properties. In a typical property induction task, participants are asked to rate the strength of inductive arguments like "ROBINS have property P, therefore BIRDS have property P". We will use the notation ROBIN → BIRD to indicate an argument that involves generalizing a property from a premise (e.g., ROBIN) to a conclusion (e.g., BIRDS). Arguments can also have multiple premises, indicated by putting them in brackets on the left.

Although this task may seem simple, it gives rise to numerous phenomena that are indicative of the complex ways that humans reason inductively. Osherson et al. (1990) presented thirteen such phenomena, eleven of which are shown in Table 1. All eleven involve comparing a stronger argument with a weaker argument; the two phenomena not included in the table or analyzed here are omitted because they are not formulated in terms of a similar comparison.

Some of the phenomena directly capture effects of similarity or typicality. For instance, Premise-Conclusion **Similarity** reflects the finding that people are more likely to generalize a property from one concept to another when the concepts are more similar. Premise **Typicality** is the finding that arguments are stronger if the premises are more typical of the conclusions. A slightly less reliable phenomenon, Premise-Conclusion **Asymmetry**, reflects the fact that an argument that generalizes from a typical category member to a less typical one (e.g. MOUSE → BAT) is often rated as stronger than the reverse argument (e.g. BAT → MOUSE); this is probably because atypical categories are more likely to have atypical properties.

Other phenomena relate to the hierarchical organization of categories. Conclusion **Specificity** reflects the intuition that greater inductive leaps are required to support broader generalizations; arguments are thus stronger if the conclusion category is more specific. The **Inclusion Fallacy** relates to the observation that a general argument that projects from a category to its enclosing class (e.g. ROBIN → BIRD) is often considered stronger than a more specific argument (e.g. ROBIN → OSTRICH) even though the latter is logically entailed by the former. We evaluate the inclusion fallacy for completeness, but because it is normally viewed as a fallacy it may not necessarily be appropriate as a target for AI models like GPT-3.5 and GPT-4.

In addition to these relatively straightforward phenomena, there are also those which appear to reflect more sophisticated or theory-based reasoning about underlying mechanisms. Premise **Diversity** refers to the fact that arguments are often considered stronger if their premises are less similar to one another. This captures the general intuition, based on an understanding of statistical sampling, that diverse evidence is more compelling than narrow evidence. A similar mechanism may underlie systematic violations of Premise **Monotonicity**, in which additional positive premises increase the strength of an argument. Monotonicity often holds if all premises are drawn from the same superordinate category, but adding premises from a different superordinate category can lead to the opposite pattern of reasoning, known as Premise **Non-Monotonicity**. For example, the inclusion of ORANGUTAN in the argument {FLY, ORANGUTAN} → BEE means that the context of the argument (the smallest category which includes the premise and inclusion categories) changes from INSECT to ANIMAL. This suggests that the property in question is not insect-specific, and thus reduces the chance that bees share it. These violations of premise monotonicity have been shown to be influenced by the reasoner's theoretical assumptions about how the premises were generated (Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019; Ransom et al., 2016; Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015).

Similarity-coverage model (SCM)

In addition to characterizing the inductive phenomena just described, (Osherson et al., 1990) presented a theory known as the similarity-coverage model (SCM) that is able to account for all of them. It will be used as part of our evaluation of GPT-3.5 and GPT-4.

The SCM builds on the fact that several inductive phenomena can be derived purely from similarity between categories. For example, ROBIN → SPARROW is stronger than ROBIN → GOOSE because robins are more similar to sparrows than geese. Similarly, ROBIN → BIRD is stronger than PENGUIN → BIRD because robins are more similar to the prototypical bird than penguins are. In both cases, the probability that the premise and conclusion categories share a property increases solely based on the similarity of the two sets of categories.

Although similarity-based accounts of property induction are simple and intuitive, they fail to account for more complex phenomena such as non-monotonicity and diversity. The SCM captures them by incorporating a notion called *coverage*, which denotes the degree to which the premise categories are similar to members of the lowest level category class that encapsulates each of the premise and conclusion categories. Osherson et al. (1990) demonstrate that a weighted combination of coverage and premise-conclusion similarity captures all eleven of the phenomena in Table 1.

Overview

This paper is divided into two experiments, each focused on comparing human inductive judgments with equivalent judgments derived from GPT-3.5 and GPT-4. The design of each experiment was inspired by the experiments of Osherson et al. (1990). In Experiment 1, the task was to pick the stronger argument out of a pair of property induction arguments, and in Experiment 2 it was to rate the strength of an argument in isolation.

Both experiments included stimuli from three category domains: Mammals, Birds, and Vehicles, sourced from the Leuven Natural Concepts Dataset (LNCD) reported in De Deyne et al. (2008). Psychological studies of induction often use arguments about mammals, and examples involving birds are common in the AI literature on defeasible reasoning. Comparing performance across three separate domains ensures that the LLMs are required to reason about domains with different levels of prominence in their training data. We selected 24 categories from each of the three domains by first excluding those that were near-duplicates or unfamiliar, and then randomly removing some until the category sets for each domain were of the same size. The full list of categories in each domain is shown in Table 2.

Table 2
The 24 categories used in each domain. The same categories were used in both experiments except that in Experiment 2, four Mammals (*beaver, giraffe, lion,* and *rabbit*) were replaced with *monkey, hippo, fox* and *wolf* in order to facilitate comparison to Osherson et al. (1990), which used the latter.

Mammals	bat, beaver, camel, cat, cow, deer, dog, donkey, elephant, giraffe, hamster, hedgehog, horse, kangaroo, lion, llama, mouse, pig, rabbit, rhino, sheep, squirrel, tiger, zebra
Birds	blackbird, canary, chicken, crow, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parrot, peacock, penguin, robin, rooster, seagull, sparrow, stork, swan, swallow, turkey, vulture
Vehicles	airplane, bicycle, boat, bus, car, caravan, carriage, cart, helicopter, hovercraft, jeep, moped, motorbike, rocket, skateboard, sled, submarine, taxi, tractor, tram, train, truck, van, zeppelin

1. Experiment 1: Argument pairs

In this experiment, agents were presented with a series of property induction argument pairs and were then asked to pick the stronger argument of each pair.

1.1. Generating argument pairs

With three domains, 24 categories in each domain and the 11 phenomena in Table 1, evaluating all possible combinations of arguments is extremely prohibitive. We therefore generated a subset of argument pairs for each phenomenon and domain, ensuring as far as possible that they were sufficient to capture the range of variation. As a first step, for each phenomenon in each domain we sampled thousands of candidate argument pairs by replacing each slot in the argument template with a randomly sampled category. We then selected the candidate pairs that were most appropriate for each of the phenomena in question. For instance, Similarity, Typicality, Asymmetry, and the Inclusion Fallacy are all stronger if one of the arguments contains a highly typical or similar premise category and the other contains a highly atypical or dissimilar premise category. We thus restricted the candidate pool for those phenomena to the categories (and category pairs) that had typicality (and similarity) ratings of 0.75SD above or below their domain's mean typicality (or similarity) rating as given by the LNCD norms (De Deyne et al., 2008). To control for similarity and typicality effects in Diversity and Monotonicity argument pairs, we ensured that the unique premise category in each pair was no more typical or similar in the stronger argument than it was in the weaker argument. To construct premise categories for Non-Monotonicity, we sampled from three supplementary domains: reptiles for Mammals, insects for Birds and tools for Vehicles.

Once we had candidate pools, our goal was to select the 24 pairs in each phenomenon-domain split that were most likely to capture that phenomenon. We achieved this by using an SCM model based on LNCD similarity ratings to calculate, for each argument pair, a measure of the disparity between the strength of the stronger and weaker arguments in that pair. We then selected the 24 pairs with the highest SCM disparity.[1] This resulted in 792 unique argument pairs in total across the whole experiment (11 phenomena × 3 domains × 24 argument pairs).

1.2. Presenting argument pairs to humans

Participants
We recruited 120 people via Amazon Mechanical Turk who were each paid $1.50USD for the 5-8 min study. All passed a screening for English language competency prior to participation and indicated informed consent via an online consent form. Both experiments were approved by the Human Research Ethics Subcommittee of the University of Adelaide.

Stimuli
Each participant was shown a different set of 66 argument pairs that were randomly sampled from the set of 792 described above. We ensured that everyone saw stimuli from all three domains and all 11 phenomena (presented in a different random order for each person). In addition to the 66 experimental trials, there were four attention check trials, resulting in 70 trials in total. The attention check trials, which occurred after every 16 trials, were the same for all participants. They were designed to look like a standard argument pair but had a relatively unambiguous answer (e.g., Robins → Gorillas vs Robins → Sparrows); we reasoned that responding incorrectly to them would indicate inattentiveness or a failure to understand the task. Participants who did not answer at least three of the four check questions correctly were excluded ($N = 10$), resulting in 110 people in the full dataset. There were a mean number of 9.17 ratings obtained for each of the 33 phenomenon-domain splits (min: 8; max: 10).

Procedure
At the start of every session, participants were shown the following instructions:

> We are interested in how people evaluate arguments. On each trial there will be two arguments labeled 'A' and 'B.' Each will contain one, two, or three statements separated from a claim by a line. Assume that the statements above the line are facts, and choose the argument whose facts provide a better reason for believing the claim. These are subjective judgments; there are no right or wrong answers.

[1] Because the LNCD only includes ratings between pairs of categories within the same domain, we could not obtain SCM scores for Specificity and Non-Monotonicity argument pairs. However, it was straightforward to derive the stronger and weaker arguments for each of our constructed pairs: for Specificity, the stronger argument is always the argument with the more specific conclusion, and for Non-Monotonicity the stronger argument is always the one with a more specific category superset. For these phenomena, we simply took a random sample of 24 pairs in order to generate the final split.

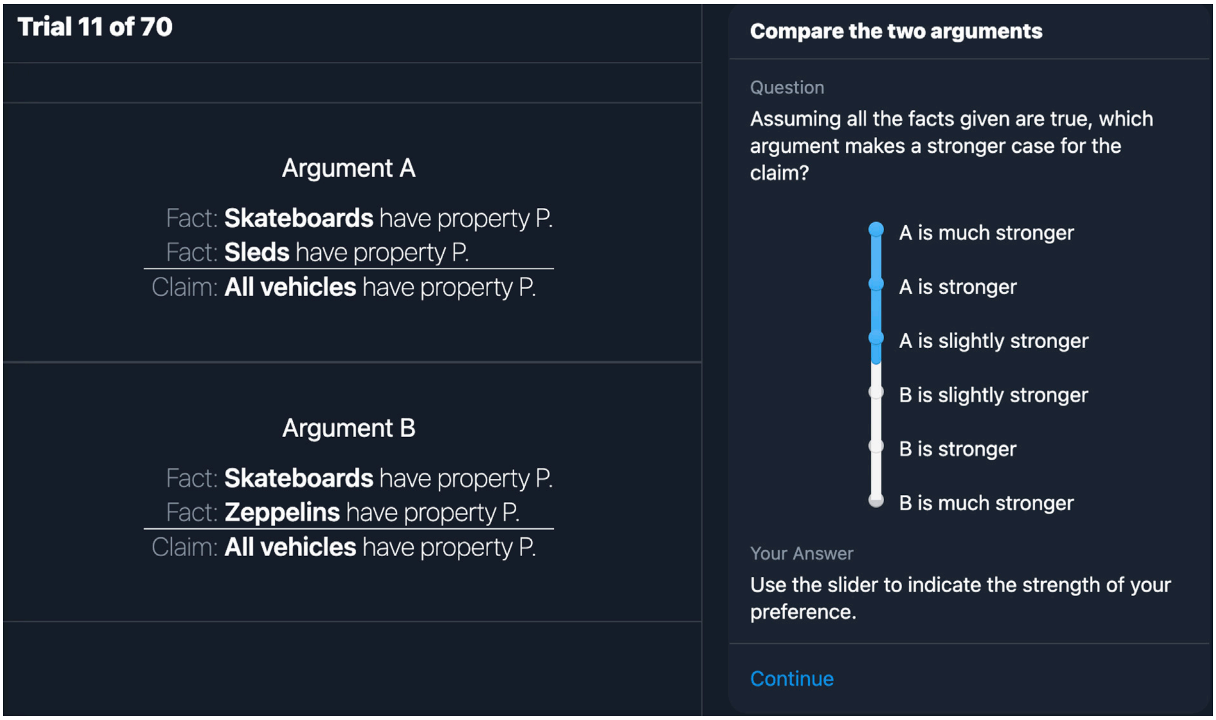

Fig. 2. Experiment 1 screenshot. On each trial, participants were shown two arguments (A and B) on the left, and then asked to rate which was stronger using the 6-point slider on the right (the slider was always initialized to the midpoint but people were required to move it to one of the six options before they could continue to the next trial). All participants saw a different random subset of the 792 possible argument pairs, constrained so that each person saw examples from all three domains and all 11 phenomena. This screenshot contains an argument pair from the domain of Vehicles and the phenomenon of Premise Diversity (which predicts that Argument B is stronger).

Table 3
Best-performing GPT-4 prompt for Experiment 1. It was identified by evaluating multiple prompts varying the components of System, Context, Argument, Question, and Options. The column on the left indicates which variant of that component it corresponds to. For System message S3, X is replaced by `living things` for the Mammals and Birds domains, and `objects` for the Vehicles domain.

S3	`You are an expert on X and the types of real world properties that they have. The questions you'll see don't have right or wrong answers, and you are willing to use your best judgment and commit to a concrete, specific response even in cases where you can't be sure that you are correct.`
C1	`We are interested in how people evaluate arguments. On each trial there will be two arguments labeled 'A' and 'B.' Each will contain one, two, or three statements separated from a claim by a line. Assume that the statements above the line are facts, and choose the argument whose facts provide a better reason for believing the claim. These are subjective judgments; there are no right or wrong answers.`

A1	`Argument A: Fact - Dogs have property P.`
	`Claim - All mammals have property P.`
	`Argument B: Fact - Hedgehogs have property P.`
	`Claim - All mammals have property P.`

Q3	`Question: Assuming all the facts given are true, which argument makes a stronger case for the claim? To get the best answer, first write down your reasoning. Then, based on this,`

O1	`Indicate the strength of your preference by providing one of the following options:`
	`A - Argument A is much stronger`
	`B - Argument A is stronger`
	`C - Argument A is slightly stronger`
	`D - Argument B is slightly stronger`
	`E - Argument B is stronger`
	`F - Argument B is much stronger`

On each of the 70 trials, people were shown two argument pairs labeled `Argument A` and `Argument B`; whether the stronger or weaker argument of each pair was labeled as `Argument A` or `Argument B` was randomly assigned. Each argument pair was presented on its own page, and on the right of the pair people were asked "Assuming all the facts given are true, which argument makes a stronger case for the claim?" They then selected their response using a 6-point Likert scale ranging from 1 (Argument A is much stronger) to 6 (Argument B is much stronger). A screenshot from an example trial is shown in Fig. 2.

1.3. Presenting argument pairs to GPT-4

To obtain analogous judgments from GPT-4, we used the `gpt-4-0314` engine within the Chat API of OpenAI. We set temperature $t = 0$ for all API requests alongside a maximum response length of 400 tokens. While higher temperature values in theory might have allowed us to obtain a response distribution from GPT-4 for every argument pair and prompt design, in practice we found that final response ratings did not vary much even for $t = 1$. We therefore presented the 792 argument pairs to GPT-4 only once for each prompt design.

We designed the prompts for GPT-4 with two aims in mind. First, we wanted them to be as similar as possible to the instructions our human participants saw. Second, given the degree to which the performance of GPT-4 varies based on the prompt, we wanted to make sure that our results could not obviously be improved by better prompts. To accomplish this, we designed a logical space of prompts and selected a subset of the most promising to evaluate. The logic of our design was informed by the observation that a prompt could be constructed by combining task agnostic instructions (the **System** message) with four consecutive components: the explanation of the task (**Context**), the structure of the arguments being rated (**Arguments**), the judgment elicitation (**Question**), and the specification of the answer format (**Options**). There were three variants of the System message (S1–S3), four of Context (C1–C4), three of Arguments (A1–A3), four of Question (Q1–Q4), and two of Options (O1–O2); all are shown in Appendix A.

Because running all 288 possible prompts on all 792 arguments was prohibitively expensive, we identified a subset of prompts that we thought would maximize the ability of GPT-4 to respond well. This included a baseline prompt designed to be as similar as possible to the human experiment (S1-C1-A1-Q1-O1), one that is similar to the baseline except with a more specific system message (S3-C1-A1-Q1-O1), one that also adds a chain-of-thought direction (S3-C1-A1-Q3-O1), and one that elicited domain-specific reasoning in particular (S3-C1-A1-Q4-O1). For each of the 33 phenomena-domain splits, we then evaluated the performance of each prompt using the sign test described in Section 1.5. The best-performing prompt was S3-C1-A1-Q3-O1, shown in Table 3, and all of the results reported in the main text are based on this prompt.

1.4. Presenting argument pairs to GPT-3.5

In addition to GPT-4, we also presented our set of argument pairs to GPT-3.5. This was achieved using the `text-davinci-003` engine. Although the GPT-3.5 series of models includes multiple variants, we chose `text-davinci-003` because it is the best performing model that uses the OpenAI Completions API. Where possible, results for other GPT variants are reported in Appendix D. Like our experiments with GPT-4, we set $t = 0$ for all API requests alongside a maximum response length of 400 tokens. Because we use the Completions API and not the Chat API for GPT-3.5, we adapted our baseline prompt for GPT-4 to GPT-3.5 by including both the `system message` and the `user message` in a single completions prompt.

Because the GPT-3.5 Completions API includes token probabilities in its responses, we assess GPT-3.5 differently from GPT-4. Instead of taking the single most likely completion generated by GPT-3.5, we generate a score for each argument pair using the sum of probabilities for each point on the provided Likert scale, weighted by rank, at the token position at which the provided answer occurs. Thus, our results for GPT-3.5 are continuous rather than discrete, and our ability to access token probabilities for GPT-3.5 affords it a methodological advantage in precision over GPT-4.

1.5. Results

Fig. 3 compares the performance of humans, GPT-3.5, and GPT-4 across all three domains on the 11 inductive reasoning phenomena we investigated. It is immediately apparent that each kind of agent exhibits a qualitatively different pattern of response. Human ratings (dark blue) are almost always unimodal and usually located to one side of the rating scale but closer to the center than either end. Conversely, GPT-3.5 ratings (green) are much more often bimodal and extreme. This reflects the fact that GPT-3.5 preferred to select options A and F rather than any of the other four choices. Direct visual comparisons between GPT-4 (red) and humans (blue) are more difficult because of the discrete nature of the GPT-4 responses, but it is clear that the peaks of the red and blue distributions do not always coincide.

Table 4
Quantitative evaluation of GPT-3.5, GPT-4 and Humans on the 11 phenomena across all three domains. The number represents the p-value on a sign test, with significant p-values indicating a preference for one argument over the other. Most of the time this preference is in the theoretically predicted direction (*), but when the weaker argument is endorsed significantly more (∘) it is in the opposite. Results for GPT-3.5's chat variant are reported in Table D.8.

Phenomenon	Domain	GPT-3.5	GPT-4	Humans
Similarity	Mammals	0.17	<0.001 *	<0.001 *
	Birds	0.69	<0.001 *	<0.001 *
	Vehicles	0.11	<0.001 *	<0.001 *
Typicality	Mammals	0.54	<0.02 *	<0.001 *
	Birds	0.54	<0.001 *	<0.001 *
	Vehicles	1.0	<0.001 *	<0.001 *
Specificity	Mammals	0.84	<0.001 *	<0.001 *
	Birds	0.84	<0.001 *	<0.001 *
	Vehicles	0.54	<0.001 *	<0.001 *
Monotonicity (General)	Mammals	<0.02 *	<0.001 *	<0.001 *
	Birds	0.54	<0.001 *	<0.001 *
	Vehicles	<0.001 *	<0.001 *	<0.001 *
Monotonicity (Specific)	Mammals	0.31	<0.001 *	0.06
	Birds	0.84	<0.001 *	<0.001 *
	Vehicles	0.84	<0.001 *	0.29
Diversity (General)	Mammals	0.84	<0.001 *	0.06
	Birds	0.54	<0.001 *	1.0
	Vehicles	1.0	0.06	< 0.03∘
Diversity (Specific)	Mammals	1.0	0.15	< 0.01∘
	Birds	0.54	0.84	0.68
	Vehicles	0.31	0.15	0.68
Nonmonotonicity (General)	Mammals	< 0.001∘	<0.02 *	<0.001 *
	Birds	0.31	<0.001 *	<0.001 *
	Vehicles	0.84	0.15	<0.01 *
Nonmonotonicity (Specific)	Mammals	1.0	< 0.001∘	0.15
	Birds	0.15	< 0.001∘	<0.001 *
	Vehicles	1.0	< 0.001∘	<0.01 *
Asymmetry	Mammals	0.06	0.68	0.4
	Birds	0.06	1.0	<0.001 *
	Vehicles	0.54	0.82	0.05
Inclusion Fallacy	Mammals	0.84	0.06	0.06
	Birds	0.84	< 0.001∘	1.0
	Vehicles	0.54	0.54	<0.001 *

While Fig. 3 illustrates performance visually, Table 4 characterizes it statistically. Since our theories primarily make predictions about *which* argument (if any) is endorsed more rather than *by how much*, we recode each response as a (-) or a (+), where (+) indicates that the theoretically-predicted stronger argument was actually rated as stronger and (-) indicates the opposite; the few ratings that were exactly even were discarded. For each of the three agents (humans, GPT-3.5, and GPT-4) and each of 33 phenomenon/domain splits, this yields a vector of +/- values; we can then use a one-sample sign test to compare them against the null hypothesis that there will be equal numbers of each. The numbers in each cell of Table 4 correspond to the p-value on the sign test for that agent, domain, and phenomenon; significant p-values indicate a significant preference for one argument or the other. Since the sign test does not capture directionality, some items are significant but in the opposite of the predicted direction, such that the agent favors the theoretically *weaker* argument; these are indicated in Table 4 with the ∘ symbol.

There is a substantial difference across phenomena in the ability of the large language models to capture human behavior. For some phenomena – Specificity, Similarity, Typicality, and Monotonicity – both humans and GPT-4 consistently endorsed the stronger argument more, in accordance with previous findings. For GPT-3.5, performance was poor; it was also usually bimodal, reflecting the fact that it often appeared to react based on superficial features of the prompt like the

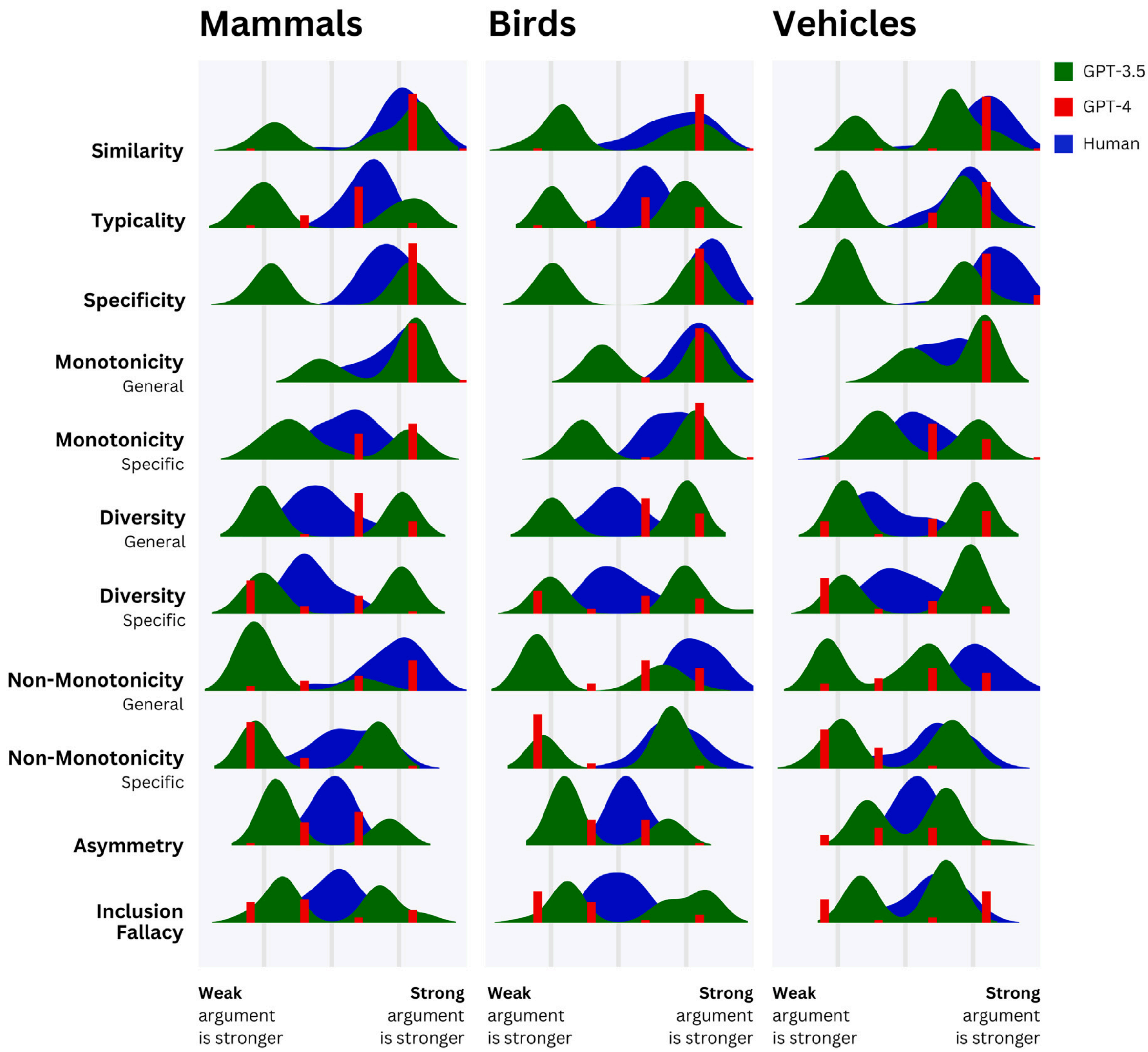

Fig. 3. Inductive reasoning phenomena exhibited by GPT-3.5, GPT-4 and human reasoners. Responses reflect which arguments were rated stronger (by convention this is the argument on the right) for each of the 3 domains (columns) and 11 phenomena (rows). Distributions are continuous for GPT-3.5 and humans and discrete for GPT-4 because of API constraints (multiple bars occur when it gave different answers for specific stimuli within each domain-phenomenon cell). GPT-3.5 performs poorly in general. While GPT-4 captures phenomena involving Similarity, Specificity, Typicality and Monotonicity, it makes the opposite predictions as humans for Non-Monotonicity (Specific).

order the arguments appeared in.[2] Consistent with this, as shown in Appendix C, the reasoning by GPT-4 anecdotally appears more appropriate than that presented by GPT-3.5, referencing relevant concepts (e.g., stating that "robins are more representative of the typical bird" when explaining its answer for Typicality).

There are another set of phenomena – Diversity, Asymmetry, and Inclusion Fallacy – where people in our experiments did not robustly endorse the argument that is theoretically predicted to be stronger. For the most part, GPT-3.5 and GPT-4 did not do so either, with the exception that GPT-4 (unlike our participants) showed the Diversity effect for General arguments. We consider reasons for this puzzling pattern of results in the Discussion.

Perhaps most interestingly, neither GPT-3.5 nor GPT-4 captured human behavior on the Non-Monotonicity phenomena. For General arguments, GPT-3.5 leans towards endorsing the *opposite* argument from the one that is theoretically predicted and endorsed by our participants. GPT-4 performs fairly well for General arguments but shows a strong preference in the opposite direction from people for the Specific arguments. Indeed, GPT-4's stated reasoning often reflects the erroneous idea that additional examples should make a conclusion stronger: it makes an argument *for* Monotonicity and *against* Non-Monotonicity (see Appendix C). We return to Non-Monotonicity and GPT-4's failure to capture this phenomenon in the Discussion.

2. Experiment 2: Individual arguments

Considering inductive phenomena in isolation is a useful starting point, but this approach is limited because multiple phenomena are relevant to some inferences, and these phenomena sometimes conflict. For example, from the perspective of Diversity {FLAMINGO, ALBATROSS} → BIRD is relatively strong because the premise categories are so different from each other. However, it is weak from the perspective of Typicality since the premise categories are atypical of birds.

In Experiment 2, we therefore moved beyond the individual phenomena in Table 1 by assessing the ability of GPT-3.5 and GPT-4 to rate the inductive strength of relatively large sets of arguments. Osherson

[2] In earlier work (Han, Ransom, Perfors, & Kemp, 2022) we reported results suggesting that GPT-3 does capture phenomena including Similarity, Typicality, and Specificity, but in that evaluation we controlled for argument order by including prompts with both possible orders. The difference between our previous and current results exposes how heavily GPT-3.5 is influenced by superficial features of the prompt.

et al. (1990) studied this in humans by asking participants to rank two sets of arguments involving mammals. One set included 36 two-premise Specific arguments such as {COW, CHIMP} → HORSE, where the conclusion in all cases was HORSE. The second included 45 three-premise **General** arguments such as {HORSE, COW, MOUSE} → ALL MAMMALS, where the conclusion category was always ALL MAMMALS. The Osherson et al. (1990) dataset has been used in several subsequent studies (Han et al., 2022; Kemp & Tenenbaum, 2009) but is limited because it includes a relatively small set of arguments from a single domain. As a result, our first step in Experiment 2 was to obtain judgments from humans on a much larger and more varied dataset of arguments. We then compared human judgments with ratings of argument strength elicited from GPT-3.5 and GPT-4 as described below.

Experiment 2 used a design in which agents were presented with a series of individual arguments and asked to rate each argument's strength in isolation. The experiment therefore departs from the approach of Osherson et al. (1990), who asked participants to rank argument sets, but follows the approach of Glick (2011), who also collected strength ratings for individual arguments.

2.1. Generating individual arguments

For each of our three domains, we generated 100 two-premise general arguments (whose conclusion was ALL MAMMALS, ALL BIRDS, or ALL VEHICLES) and 100 two-premise specific arguments (whose conclusion was one of the 24 categories in that domain). We did this by first randomly sampling 10,000 arguments for each of the six domain-conclusion splits (3 domains × 2 conclusions, general or specific). After removing duplicates, we ranked the strength of each argument using the SCM. In order to ensure that our set of arguments contained the full range of strengths, we divided the ranked set of arguments into 25 equal width bins, then sampled four from each bin to construct the full set of 100 arguments. Sampling was random subject to the constraint that no premise set could appear more than three times and no category could appear more than 15 times.

For each set of 100 two-premise arguments we then constructed a corresponding set of single-premise arguments consisting of all of the premise-conclusion mappings from the two-premise set. Thus, for instance, the two-premise argument {CANARY, SEAGULL} → STORK would correspond to the one-premise arguments CANARY → STORK and SEAGULL → STORK.

The precise number of single premise arguments varied depending on the degree of overlap in the categories that were selected, but ranged from 24 for general arguments to 169 for specific mammal and bird arguments. In total, across all six domain-conclusion splits as well as both one-premise and two-premise sets, there were 1168 arguments.

2.2. Presenting individual arguments to humans

Participants

We recruited 610 people via Amazon Mechanical Turk who were each paid $1.00USD for the five minute study. All participants passed a screening for English language competency prior to participation and indicated informed consent via an online consent form.

Stimuli

Each participant was randomly assigned to one of the six domain-conclusion splits and only rated arguments from that split. The arguments were presented in two blocks, one corresponding to one-premise arguments and one corresponding to two-premise arguments. Each participant saw one of 10 versions of each block, each based on a stratified sample of the 100 possible two-premise arguments for that domain and conclusion type.[3] Each participant completed 42 trials consisting of 10 trials for the two-premise block and 32 trials for the one-premise block. The 32 trials of the one-premise block were composed of one-premise arguments that corresponded with the two-premise block's arguments alongside a random sample of additional one-premise arguments to ensure that all participants saw the same number of arguments in the one-premise block. In addition to the 42 trials, participants were also shown two trial arguments at the start of each block as well as 4 attention check arguments every 8 trials. The attention check arguments were handcrafted to have an unambiguous answer (for example: ALL ANIMALS → ALL MAMMALS).

Procedure

At the start of every session, participants were shown the following instructions:

```
We're going to show you a series of claims relating to living things and the properties they share. Rather than mention any specific property (e.g. ''Hyenas have sesamoid bones'') we'll refer to an unspecified property (e.g. ''Hyenas have property P''). Each claim may be true or false, and to help you decide which, we'll provide you with facts about whether or not other living things have the same property (e.g. ''Lions have property P'', and ''Zebras have property P'').
```

People were then shown separate instructions for the one-premise and two-premise blocks. For each block, participants were told how many supporting facts to expect (one or two) and informed that their job was to rate how likely the claim was. They then saw two sample trials using fruit, and upon completing them began the main sequence of trials in that block.

On each trial, people were shown a single argument labeled `Argument A` on the left of the screen. On the right, they were asked "Given the facts presented, how likely is it that the claim is true?" They then selected their response using a slider with a scale ranging from 0 (very unlikely) to 100 (very likely). A screenshot from an example trial is shown in Fig. 4.

Replicating Osherson et al.

To ensure that our study was broadly consistent with the original study by Osherson et al. (1990), we ran a version of our procedure that used the 36 specific arguments used by Osherson et al. rather than our LNCD based arguments. This experiment involved 40 participants who each rated one block of 10 multi-premise and one block of 10 single-premise Osherson arguments. In all other respects, its procedure was identical to what is described above. We found that across the 36 arguments, these rankings had a Spearman correlation of 0.65 ($p < 0.001$) with the original rankings obtained by Osherson et al..

2.3. Presenting individual arguments to GPT-4

Our method for presenting individual arguments to GPT-4 is very similar to that of our first experiment. We again used the `gpt-4-0314` engine within the Chat API of OpenAI with temperature $t = 0$ and a maximum response length of 400 tokens. For each prompt design,

[3] Argument stratification was achieved using human similarity based SCM scores in order to ensure that each block contained arguments with a wide variety of strengths. There were 11 rather than 10 versions of each block in the General Birds split because two of the arguments in that split shared the same SCM score. Consequently some General Bird arguments have 11 rather than 10 participant ratings. All other aspects of this split are identical to the other splits.

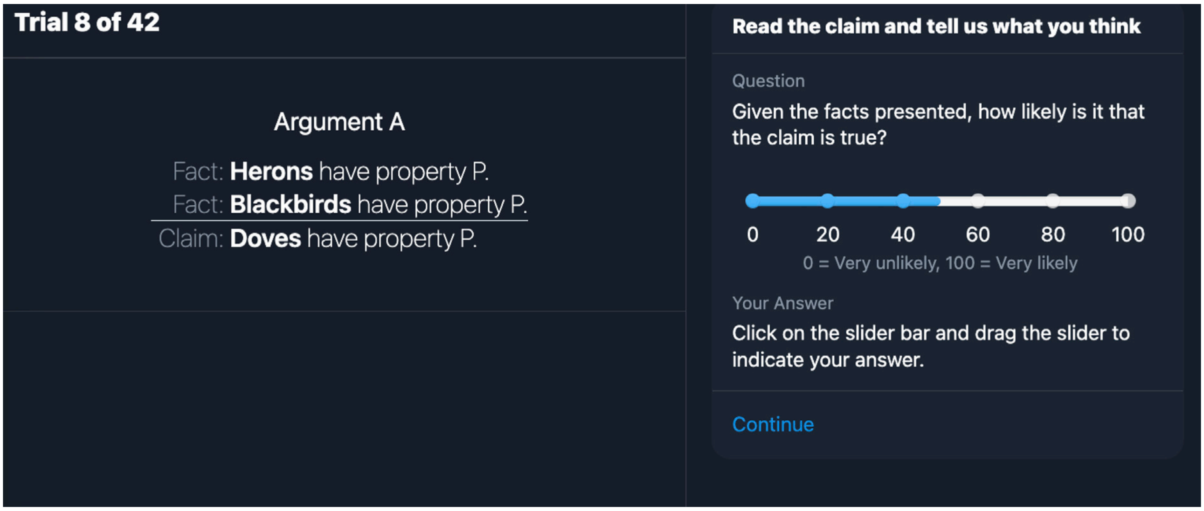

Fig. 4. Experiment 2 screenshot. On each trial, participants were shown one argument on the left, and then asked to rate it using the slider on the right (the slider was always initialized to the midpoint but people were required to move it before they could continue to the next trial). All participants saw 42 arguments from one of the six domain-conclusion splits in two blocks, one for one-premise arguments and one for two-premise arguments. This screenshot contains a two-premise argument from the domain of Birds with the specific conclusion category Dove.

Table 5
Best-performing GPT-4 prompt for Experiment 2. It was identified by evaluating multiple prompts varying the components of System, Context, Trials, Argument, Question, and Options. The column on the left indicates which variant of that component it corresponds to.

S3	`You are an expert on X and the types of real world properties that they have. The questions you'll see don't have right or wrong answers, and you are willing to use your best judgment and commit to a concrete, specific response even in cases where you can't be sure that you are correct.`
C1	`We're going to show you a series of claims relating to living things and the properties they share. Rather than mention any specific property (e.g. "Hyenas have sesamoid bones") we'll refer to an unspecified property (e.g. "Hyenas have property P"). Each claim may be true or false, and to help you decide which, we'll provide you with facts about whether or not other living things have the same property (e.g. "Lions have property P", and "Zebras have property P").`
T1	`This section contains a series of claims that include only one supporting fact. Before we start, we'll give you two examples as practice.` *`[insert two examples following the same format as the main trials]`* `Now that you've practiced you're ready to continue on to the main trials for this section.`
A1	`Argument A: Fact - Dogs have property P.` `Claim - All mammals have property P.`
Q1	`Question: Given the facts presented, how likely is it that the claim is true?`
O1	`Indicate your answer by providing a number between 0 and 100, where 0 means that the claim is very unlikely and 100 means that the claim is very likely.`

we presented the 600 two premise arguments and 568 single premise arguments to GPT-4 only once.

The design of possible prompts was also very similar. The task agnostic instructions (the **System** message) were identical, as was the structure of the arguments being rated (**Arguments**). The explanation of the task (**Context**), judgment elicitation (**Question**), and the specification of the answer format (**Options**) were straightforwardly adapted to correspond to this task. The only major change was the addition of a new component (**Trials**) corresponding to the presence or absence of the two practice trials involving fruit; we added this because the participants in our experiment saw these. All variants are shown in Appendix B using the example of one-premise arguments; prompts for the two-premise arguments were exactly analogous.

As before, for resource reasons we identified a subset of prompts to explore rather than systematically testing hundreds. These corresponded to the four in Experiment 1 plus an additional one (S3-C1-A1-Q1-O1-T1) which included the practice trials, which we sequentially fed to GPT-4 in order to include its own answers to the trials in the final prompt. Performance was evaluated based on the correlation obtained between GPT-4 judgments and human judgments, described more fully in Section 2.7. The best-performing prompt was the one with the practice trials, shown in Table 5.

2.4. Presenting individual arguments to GPT-3.5

Like our first experiment, we used the `text-davinci-003` engine within the OpenAI Completions API to elicit argument ratings from GPT-3.5, setting $t = 0$ alongside a maximum response length of 100 tokens. We again relied on GPT-3.5 token probabilities and presented our top performing prompt for GPT-4 to GPT-3.5, concatenating the system and user messages into a single prompt. To convert GPT-3.5 responses to argument ratings, we took the sum of the top five token completion probabilities, each multiplied by their numeric value. Again, this affords GPT-3.5 a methodological advantage in precision over GPT-4.

2.5. Extracting similarity judgments from GPT-4

Supplementing our analysis of argument ratings, we also extracted similarity judgments from GPT-4 and GPT-3.5. We used the same GPT-4 engine and parameters as our previous experiments, and set the response scale as 0-20 to match the original study by De Deyne et al. (2008). We then presented GPT-4 with every category pair in the LNCD using the `user message` below. In each message, `X` was replaced by the category pair's domain name (e.g., `animals`) while `C1` and `C2` were replaced by the category names (e.g., `rabbits` and `hippos`).

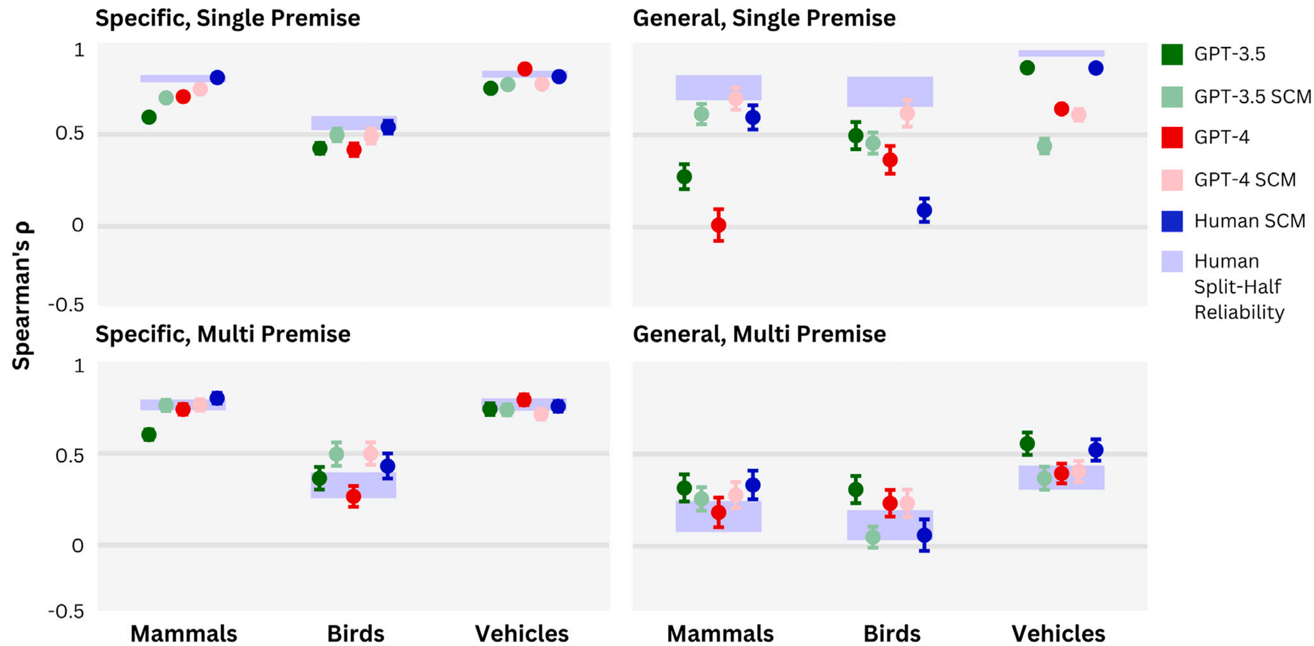

Fig. 5. Correlations (Spearman's ρ) between human argument rankings and rankings of five different models. Human SCM (a special-purpose model that applies the similarity-coverage model to human similarity ratings) performs best overall, but its performance is nearly matched by GPT-4 on Specific arguments and GPT-3.5 on General ones. The remaining two models apply the SCM to similarity ratings derived from GPT-4 and GPT-3.5. GPT-3.5 SCM improves on GPT-3.5 for Specific arguments, and GPT-4 SCM improves on GPT-4 for General arguments. The width of each split-half bar spans one standard error on either side of the mean. Results for GPT-3 and other GPT-3.5 models are shown in Fig. D.7.

```
You are an expert on X and the various properties that they have. With these properties in mind, we will ask you to rate the similarity of two X on a scale of 0 to 20, where 0 means that the X have no similarity and 20 means that they are identical.

Question: With their respective properties in mind, how similar are C1 and C2 on a scale of 0 to 20? Answer with a single number. Answer:
```

2.6. Extracting similarity judgments from GPT-3.5

To extract similarity judgments from GPT-3.5, we followed two approaches. The first was to present GPT-3.5 with the same prompt as GPT-4. As we did when extracting argument ratings, we calculated GPT's similarity rating as the sum of the top five token completion probabilities multiplied by the numeric value of each token; each rating was then converted to a ranking.

In the second approach, we passed each category name to the OpenAI Embeddings API using the `text-embedding-ada-002` endpoint, and then extracted an embeddings-based similarity score by calculating the cosine similarity between each LNCD category pair. Although the `text-embedding-ada-002` model is different from the `text-davinci-003` model that we used elsewhere, we used it because it is the best performing GPT-3 based model for embeddings-based applications (OpenAI, 2022). We thus view its similarity ratings as a noisy indication of what is possible when similarity is measured using internal representations rather than completions.

2.7. Results

Split-half reliabilities across 100 splits for each argument set are shown in Fig. 5. The split-half reliabilities for multi-premise general arguments are relatively low, indicating a high level of variability across participants, and as a result we will not attempt to interpret differences in model performance for these arguments. In contrast, the split-half-reliabilities for single-premise general arguments and for specific arguments seem high enough to use these data for evaluating alternative models.

As a benchmark for comparison, we computed ratings of argument strength for a **Human SCM** model that applies the similarity-coverage model to the human similarity ratings collected by De Deyne et al. (2008). This human SCM model is a special-purpose model from the literature that was specifically designed to capture judgments about the kinds of arguments included in our experiment. A general-purpose system such as GPT-4 has therefore performed relatively well if it accounts for our data as well as the human SCM model.

Fig. 5 shows the correlations between the five different models and our human ratings. We additionally perform quantitative comparisons between model pairs based on 1000 bootstrap samples for each model, and report those in Table 6.

For **Specific** arguments, Fig. 5 suggests that GPT-4 accounts for human ratings nearly as well as Human SCM. Table 6 shows that Human SCM outperforms GPT-4 for all four sets of Specific arguments about Mammals and Birds, but GPT-4 outperforms Human SCM for the two sets about Vehicles. Relative to GPT-4, GPT-3.5 performs worse for Specific arguments. That said, across all six sets, the differences between the models are relatively small.

The results for **General** arguments reveal a more substantial gap between Human SCM and GPT-4. Table 6 shows that Human SCM outperforms GPT-4 on single-premise sets for Mammals and Vehicles, but that GPT-4 is superior for Birds. The difference between the models is especially large for single-premise arguments about Mammals, where GPT-4 is uncorrelated with human ratings. Interestingly, GPT-3.5 performs better than GPT-4 for General arguments and seems roughly comparable to Human SCM for single-premise General arguments.

We did not anticipate the superiority of GPT-3.5 over GPT-4 for General arguments, but in retrospect this finding seems plausible. An important difference between the two models is that GPT-4 alone was trained using reinforcement learning from human feedback (RLHF), which aims to better align the model's output with human expectations. Among other benefits, RLHF makes the model less likely to "hallucinate" responses with no factual basis, but also impairs the model's reasoning ability in some contexts (OpenAI, 2023). Evaluating general arguments may be one of these contexts, because projecting a novel

Table 6
Comparisons between pairs of models for Experiment 2. Each comparison is based on 1000 bootstrap samples per model, and the entries in column "M1 vs M2" indicate the proportion of samples for which M1 correlates better with human ratings than does M2. Proportions greater than 0.95 indicate strong evidence in favor of M1, and smaller than 0.05 indicate strong evidence in favor of M2 (these are bold and colored according to the favored model). Human SCM, GPT-4 and GPT-3.5 are abbreviated as H, 4 and 3.

Conc.	Prems	Domain	H vs 4	H vs 3	4 vs 3	4 vs 4 SCM	3 vs 3 SCM
Specific	Single	Mammals	**1.0**	**1.0**	**1.0**	**0.02**	**0.0**
		Birds	**0.99**	**0.99**	0.38	**0.0**	**0.0**
		Vehicles	**0.0**	**1.0**	**1.0**	**1.0**	0.13
Specific	Multi	Mammals	**0.98**	**1.0**	**1.0**	0.16	**0.0**
		Birds	**0.97**	0.86	0.07	**0.0**	**0.0**
		Vehicles	0.05	0.71	**1.0**	**0.99**	0.57
General	Single	Mammals	**1.0**	**0.99**	**0.0**	**0.0**	**0.0**
		Birds	**0.0**	**0.0**	**0.05**	**0.0**	0.69
		Vehicles	**1.0**	0.46	**0.0**	0.83	**1.0**
General	Multi	Mammals	**0.96**	0.60	**0.02**	0.07	0.84
		Birds	0.05	**0.0**	0.15	0.49	**0.99**
		Vehicles	**0.99**	0.13	**0.0**	0.41	**1.0**

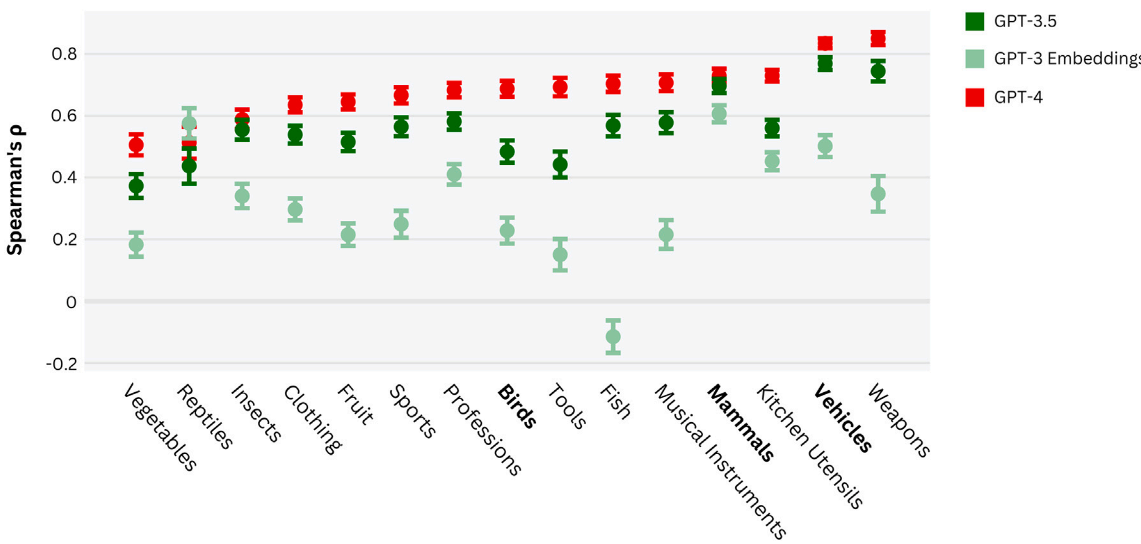

Fig. 6. Correlations (Spearman's ρ) between human similarity ratings and similarity ratings derived from GPT-4, GPT-3.5 and GPT-3 Embeddings. Error bars show standard error of the mean across 1000 bootstrap samples.

property to a superordinate category such as ALL MAMMALS requires an inductive leap that goes well beyond the limited evidence provided by information about two or three individual mammals. Compared to GPT-3.5, GPT-4 may be less likely to take this inductive leap because RLHF encourages it to remain within the bound of established facts. As Fig. D.7 demonstrates, both the GPT-4 and GPT-3.5-chat-turbo models tend to perform worse than non-RLHF GPT variants for general arguments, and for GPT-3.5-chat-turbo in particular we find that almost a quarter of all responses to general arguments explicitly mention neutrality or suggest that the answer is 'difficult to determine' given the provided information.

Although neither GPT-3.5 nor GPT-4 accounts as well as human SCM for both Specific and General arguments, the strong performance of GPT-4 for Specific arguments and GPT-3.5 for General arguments suggests that a single GPT model (perhaps GPT-4 trained with less RLHF) might be able to account for our data as well as Human SCM. Overall, our results suggest that large language models are broadly able to capture inductive inferences about blank properties about as well as the best special-purpose models available in the psychological literature.

2.8. Disentangling representation and reasoning

Evaluating an inductive argument requires some kind of reasoning process that operates over representations of the categories in the argument. For connectionist models such as GPT-4 and GPT-3.5 it may be difficult to establish a sharp division between representation and reasoning (Rogers & McClelland, 2014), but we can still develop analyses that attempt to separate the two.

Representation

To probe the representational abilities of GPT-3.5 and GPT-4 we used the similarity judgments extracted from these models by the procedure described above. Although these similarity judgments are not direct measures of internal representations, we take them as evidence of underlying representations in the same way that human similarity judgments are often taken as evidence of underlying mental representations. Fig. 6 shows correlations between these similarity judgments and the human similarity judgments reported by De Deyne et al. (2008). Most relevant to us are results for the three domains considered in our experiments (Mammals, Birds, and Vehicles), but we include results for all of the domains in De Deyne et al. (2008).

The correlations between GPT-4 and human similarity ratings are relatively high and exceed 0.6 for all three domains in our experiments. The prompt-based correlations achieved by GPT-3.5 are lower across all domains but still exceed 0.4 for all three of the domains in our experiments. In contrast, the embedding-based results are relatively poor, and we therefore use prompt-based GPT-3.5 similarities in our subsequent analyses.

Our results in Fig. 6 parallel previous findings that the internal representations of LLMs can be used to make relatively accurate predictions about human similarity judgments (Bhatia & Richie, 2021).

All of these findings therefore suggest that the category representations of both GPT-4 and GPT-3.5 should be sufficient to allow human-like ratings of inductive strength.

Reasoning

We now ask how well the GPT models exploit their representational abilities when evaluating inductive arguments. To do so we consider **GPT-4 SCM** and **GPT-3.5 SCM**, two hybrid models that apply the SCM (a special-purpose model of property induction) to similarity ratings derived from the GPT models. According to the SCM, the strength of a specific argument with premises P and conclusion c is

$$\mathrm{SCM}_{GPT-N}(P,c) = \alpha \cdot \max_{p\in P} \mathrm{sim}_{GPT-N}(p,c) + \frac{1-\alpha}{|C|}\sum_{c_i\in C}\max_{p\in P}\mathrm{sim}_{GPT-N}(p,c_i), \quad (1)$$

where the premise and conclusion categories are all drawn from a finite set of categories called domain C. In Osherson et al. (1990) $0 < \alpha < 1$ is left as a free parameter, and here we set α to 0.5.

For general arguments with premises P that all belong to a conclusion domain C, the GPT-SCM rates argument strength using:

$$\mathrm{SCM}_{GPT-N}(P,C) = \frac{1}{|C|}\sum_{c_i\in C}\max_{p\in P}\mathrm{sim}_{GPT-N}(p,c_i). \quad (2)$$

In both cases, $\mathrm{sim}_{GPT-N}(c_i, c_j)$ refers to the similarity rating derived from a GPT-N model for categories c_i and c_j.

Comparing against a GPT-SCM model can potentially expose a gap between the representational and reasoning abilities of a GPT model. For example, if GPT-3.5 captures human similarity ratings relatively well but does not use its representations effectively when evaluating inductive arguments, then GPT-3.5 SCM should account for our human data better than GPT-3.5. On the other hand, if GPT-3.5 reasons as well as possible given the representations that contribute to its similarity ratings, then GPT-3.5 SCM should perform no better than GPT-3.5.

Fig. 5 and Table 6 show that GPT-4 SCM does better than GPT-4 on three of the six sets of **Specific** arguments and worse on two: this suggests that there is no major improvement of the hybrid model relative to GPT-4. By contrast, GPT-3.5 SCM performs better than GPT-3.5 across four of the six specific data sets, and is worse on none of them. Overall, these results suggest that combining GPT similarity ratings with the SCM provides more of a boost for GPT-3.5 than GPT-4, which in turn supports the conclusion that GPT-3.5 is more limited in its reasoning capacities than is GPT-4.

For **General** arguments, the pattern of results is again reversed. GPT-4 SCM improves relative to GPT-4 on two of the three single-premise argument sets, but GPT-3.5 SCM is comparable to GPT-3.5 (better on one set and worse on another). These results are consistent with the possibility that GPT-4's ability to handle General arguments is relatively poor and may have been compromised by RLHF. On the other hand, on General arguments GPT-3.5 appears to reason as well as could be expected given its category representations.

Throughout this section we have evaluated the GPT models against human judgments and have implicitly assumed that a model that differs from humans is reasoning relatively poorly. It is possible, however, that the models differ from humans because they are less prone to inferential errors and are therefore reasoning better than our experimental participants. Our primary goal has been to explore the extent to which LLMs make human-like inferences, which means that the normative status of these inferences has not been a major concern. Future work, however, can evaluate both human participants and GPT models against normative accounts of inductive reasoning (Heit, 1998; Kemp & Tenenbaum, 2009; Tenenbaum, Griffiths, & Kemp, 2006) with the aim of identifying ways in which the models may be better and worse than humans.

3. Discussion

We developed two new benchmark datasets that are directly inspired by previous psychological work on property induction but incorporate a much larger set of categories, arguments, and domains. Across two studies, we compared responses from GPT-3.5 and GPT-4 with the human inferences in these datasets. Experiment 1 focused on 11 qualitative phenomena including Similarity, Typicality, Diversity, and Non-Monotonicity; of them, only Non-Monotonicity (Specific) was observed in our participants but not well captured by either GPT model. Experiment 2 asked models and humans to provide quantitative ratings of argument strength, and found that the best results from GPT-3.5 and GPT-4 were comparable with results obtained by the similarity-coverage model, a special-purpose model of inductive reasoning that incorporates human similarity judgments. Taken together, the experiments therefore suggest that GPT-style models broadly perform well at the tasks that we used, and capture everything that we considered except Non-Monotonicity.

We began this project when GPT-3 was the most advanced GPT model available, and expected at the time that it might capture effects related to Similarity and Typicality but would struggle to capture Diversity and Non-Monotonicity. Recall that Non-Monotonicity describes the phenomenon in which adding premises results in *weaker* arguments: the three-premise argument {CROW, PEACOCK, RABBIT} → BIRD is considered less strong than the corresponding two-premise argument {CROW, PEACOCK} → BIRD. One intuitive explanation for why people appear to reason this way is that they implicitly assume that the premises are being generated by a helpful provider of information (Medin et al., 2003; Ransom et al., 2016; Voorspoels et al., 2015). Given a desired conclusion of BIRD, a helpful provider would be much more likely to offer the premises CROW and PEACOCK alone than to also include RABBIT; this is because including RABBIT implicitly changes the context to favor a conclusion like MAMMAL.

The relationship between Non-Monotonic reasoning and assumptions about the data generation process becomes even more obvious when considering other Non-Monotonic argument pairs like TIGER → MAMMAL vs {TIGER, LION} → MAMMAL. If the conclusion truly is ALL MAMMALS it seems a strange coincidence indeed (or evidence of incompetence or unhelpfulness on the part of the provider) for both premises to involve large cats: a helpful provider should offer clues that cover the conclusion category as well as possible. In addition to being intuitively appealing, this chain of thought can be formalized as Bayesian reasoning based on assumptions about how data is sampled (Ransom et al., 2016). GPT-4's failure to capture non-monotonicity (specific) therefore suggests that it may fail to reason about how the premises of an inductive argument were generated.

Diversity is a second phenomenon that has been justified on the basis of sampling assumptions (Hayes et al., 2019), and is partially captured by GPT-4. Our results for Diversity, however, are puzzling in two respects. First, GPT-4 captured Diversity for arguments with general conclusions (e.g., MAMMAL), but did not show Diversity effects when the conclusions were specific (e.g., RABBIT). Second, our human participants did *not* show any Diversity effect at all, and it is important to ask why we did not replicate the work of Osherson et al. (1990) in this respect.

The design for Experiment 1 followed the work of Osherson et al. (1990) relatively closely, and in retrospect we see no obvious problems with the design or the stimuli. That said, one difference between our experiment and that of Osherson et al. (1990) is that our participants not only saw stimuli from more domains (Vehicles and Birds as well as Mammals), but also a wider selection of categories within a single domain than theirs did. It is possible that this experimental context created such a large level of *background* diversity that it dwarfed the much smaller variation within the key Diversity argument pairs we tested. This is consistent with previous work suggesting that Diversity may be more elusive or context-dependent than many of the other phenomena in Experiment 1 (López et al., 1997; López, Gelman, Gutheil, & Smith, 1992; Proffitt, Coley, & Medin, 2000). For example, López et al. (1992) found support for Similarity, Typicality and Specificity in kindergarteners but no evidence for Diversity. If Diversity-based reasoning is relatively fragile, it may have little chance of emerging under circumstances when participants must quickly rate a large number of argument pairs from a wide variety of domains.

4. Strengths and limitations of our data

Compared to existing property induction datasets, our data offer several advantages. First, our data are relatively comprehensive; as far as we know these datasets are the largest set of human property induction judgments ever collected. Second, the arguments in these data sets were generated according to systematic procedures described above, which improves on the common practice of working with a small set of hand-crafted arguments. Finally, our data have been released publicly, which allows future researchers to replicate our results, evaluate future LLMs against our data, and explore aspects of our data that we have not considered ourselves.

Despite these strengths, our data are subject to some important limitations. In addition to our failure to replicate the effect of Diversity, our data are somewhat noisy. For instance, Experiment 2 was inspired by a study of Osherson et al. (1990) that asked participants to rank a set of arguments, but we felt that GPT-3.5 (the best LLM at the time) would be unable to handle this task. We therefore asked participants to rate the strength of individual arguments, which allowed us to give the same task to humans and LLMs. It also made it straightforward to collect data for a relatively large set of arguments per domain. Single argument ratings, however, are intrinsically noisy, which means that our data may not provide enough resolution to identify all of the factors that influence inductive reasoning in humans. Our first version of Experiment 2 followed Osherson et al. (1990) in using general arguments with three premises, but the resulting data were so noisy that we considered them unreliable. We therefore simplified the task and used general arguments with two premises, but as suggested above the split-half reliabilities for multi-premise general arguments are still relatively low.

Because a ranking task allows for direct comparisons between arguments, ranking and rating are fundamentally different tasks that may expose different aspects of inductive reasoning, and future work should consider both tasks. Of the two, rating is the easier target for LLMs because ranking requires information to be integrated across context windows that include at least two arguments, and requires an element of planning: in order to succeed, one must first find the strongest argument, then the next strongest, and so on. Given that GPT-4 appears to provide a good account of human argument ratings, future studies should explore whether it still performs well when evaluated against humans using a ranking task.

A final limitation of our data is that we worked with a convenience sample recruited through Amazon Mechanical Turk. All of our participants previously passed a manually assessed qualification task measuring English proficiency and meticulousness, and are thus consistently more reliable than the general MTurk population. However, this does not mean that this participant pool is completely unproblematic. Property induction phenomena are known to vary across cultures (López et al., 1997), developmental stages (López et al., 1992), and levels of expertise (Proffitt et al., 2000), and it is not entirely clear which group of people provide the most natural comparison to LLMs such as GPT-4. From one perspective, GPT-4 has been trained on a vast body of information about mammals, birds, and vehicles, and may therefore be best compared with human experts. It seems possible, for example, that our participants know relatively little about birds, and that GPT-4's responses would match the responses of domain experts more closely than they matched the responses of our experimental participants. If so, then evaluating GPT-4 against our human data could potentially underestimate its true abilities.

4.1. What are the GPT models doing?

Our analyses characterize the performance of GPT models on property induction tasks in some detail, but we have relatively little insight into why the models behave as they do. At one level we understand these models: we know that they are trained on vast corpora on the task of predicting the next token, and that GPT-4 goes through a second training phase that relies on reinforcement learning from human feedback. But at another level these models remain opaque and mysterious, and it is unclear how this training gives rise to some of the behavior we observe.

At least three approaches can be used to develop a better understanding of large language models. We refer to them here as *LLM psychology*, *LLM neuroscience* and *LLM ecological analysis*. LLM Psychology (Binz & Schulz, 2023; Hagendorff, 2023) studies language models using behavioral experiments, computational analyses and all of the other techniques that psychologists have used to study human cognition and behavior. Our work offers several examples of this. We relied heavily on behavioral experiments to generate comparison data, and used related techniques to explore how GPT models respond to inductive problems as well as how sensitive they are to the way in which they are prompted. Moreover, we also developed computational analyses that build on psychological models such as the similarity-coverage model.

All of the analyses in this paper are thus examples of *LLM cognitive psychology*. However, future work could draw on ideas from other branches of psychology. For example, *LLM developmental psychology* might aim to characterize the order in which inductive phenomena appear over the course of training a language model, and the resulting developmental sequence could be compared with analogous developmental sequences in humans (López et al., 1992).

While *LLM psychology* focuses on data from behavioral studies, *LLM neuroscience* goes beyond these studies by directly probing the mechanisms that give rise to behavior. We took a simple step in this direction by extracting and working with embeddings that are likely to approximate the internal representations of GPT-3. Probing the internal mechanisms that emerge in LLMs is an active area of research, but there are already several approaches that could be used to gain a better understanding of how GPT models evaluate inductive arguments (Li, Nye, & Andreas, 2021; Olsson et al., 2022; Voita, Talbot, Moiseev, Sennrich, & Titov, 2019). For example, since LLMs incorporate a set of attention weights, analyzing these weights may reveal systematic regularities in what parts of the premise(s) are weighted most highly by a LLM while assessing an argument's conclusion. A major challenge in applying LLM neuroscience to GPT-3.5 and GPT-4 is that these models are not open source, but other LLMs such as LLaMA (Touvron et al., 2023) are publicly available and their internal mechanisms can be studied in great detail.

A third approach, *LLM ecological analysis*, focuses less on the computations carried out by LLMs and more on understanding the structure of the data on which they are trained. These data constitute the "environment" of the LLM, and studying a LLM by analyzing its training data is reminiscent of "rational analysis" (Anderson, 1990) and other research programs (Brunswik, 1957; Simon, 1970; Todd & Gigerenzer, 2007) that aim to understand an organism's behavior by characterizing the environment in which the organism is embedded. For studies of property induction, it may be valuable to explore the extent to which premise and conclusion categories co-occur in the training data, and to investigate whether and how LLMs go beyond models of inductive reasoning that rely on co-occurrence alone. A more urgent question is the extent to which the test tasks themselves (or highly similar ones) were available in the training data (Frank, 2023; Magar & Schwartz, 2022; Mitchell, 2023). Most of the arguments in our data sets were generated for this project rather than drawn from the literature, so it seems unlikely that GPT-4 has previously seen these exact arguments. Even so, the training data for GPT-4 presumably contain numerous discussions of property induction, probably including the highly-cited paper (Osherson et al., 1990) that introduced the inductive phenomena we considered in Experiment 1. It seems possible that these components of the training data are at last partially responsible for the high level of performance achieved by GPT-4. Because the training data for GPT-3.5 and GPT-4 are not publicly available, any systematic work on LLM ecological analysis will need to consider models like LLaMA (Touvron et al., 2023) where that data is openly accessible.

4.2. Other inductive reasoning tasks

There is room for multiple future studies to explore how LLMs respond to the items in our two datasets, but future work should also explore how LLMs respond to other inductive tasks. Here we focused on inferences about blank properties, and psychological work in this tradition has identified several phenomena that go beyond the ones we examined (Medin et al., 2003; Sloman, 1993). For example, Medin et al. (2003) show that causal relationships lead to phenomena including causal asymmetry (CARROT → RABBIT is stronger than RABBIT → CARROT), causal violations of similarity (BANANA → MONKEY is stronger than MOUSE → MONKEY) and diversity ({FLEA, BUTTERFLY} → SPARROW is stronger than {FLEA, DOG} → SPARROW). The same authors show that arguments in which all premise categories share a salient property can lead to non-monotonicity (e.g. BROWN BEAR → BUFFALO is stronger than {BROWN BEAR, POLAR BEAR, GRIZZLY BEAR} → BUFFALO). We gave examples including the ones just described to GPT-3.5 and GPT-4, and the results in Appendix E suggest that GPT-4 is sensitive to causal relationships between categories but may often struggle with non-monotonicity.

When evaluating an inductive argument, knowledge about the property in question is usually just as important as knowledge about the premise and conclusion categories, and several studies have documented property effects in inductive reasoning (Gelman & Markman, 1986; Heit & Rubinstein, 1994; Smith, Shafir, & Osherson, 1993). For example, Gelman (2007) shows that even young children reason differently about biological properties (e.g. "has a spleen inside") and non-generalizable properties (e.g. "is dirty"). Heit and Rubinstein (1994) considered both anatomical properties (e.g. "has sesamoid bones") and behavioral properties (e.g. "usually gathers large amounts of food at once"), and found that these properties interacted with anatomical and behavioral similarity between premise and conclusion categories (e.g. GOLDFISH → SHARK is stronger than WOLF → SHARK for anatomical properties but not for behavioral properties). Although there are numerous empirical studies of property effects in inductive reasoning, to our knowledge there are no psychological models that can capture these effects in a general way; perhaps the closest is a formal account of human plausible reasoning developed by Collins and Michalski (1989). LLMs therefore qualify as the first ever models that seem theoretically capable of capturing a broad range of property effects, and future work should compare them with humans using systematic benchmarks that include inductive arguments with a variety of non-blank properties. Appendix E includes example responses for a handful of relevant cases, and suggests that GPT-4 will struggle to capture many of the property effects documented in the literature.

The psychological literature on property induction is somewhat distinct from work on generalization (Shepard, 1987), categorization (Pothos & Wills, 2011), identification (Kemp, Chang, & Lombardi, 2010), and analogical reasoning (Vosniadou & Ortony, 1989), but all of these topics can be viewed as special cases of inductive reasoning (Kemp & Jern, 2014). According to one standard definition, an inference is inductive (or ampliative, or defeasible) if it reaches at a conclusion that does not follow with certainty from the available evidence (Chater et al., 2011; Holland et al., 1986). From this perspective there are many different inductive problems, which opens up the possibility for a line of work that applies LLMs to inductive problems drawn from many parts of the psychological literature. Existing work along these lines includes studies that evaluate LLMs on commonsense reasoning tasks from benchmarks such as BIG-bench (Srivastava et al., 2022), and a study that explores analogical reasoning in LLMs (Webb et al., 2022).

Appendix F includes responses of GPT-3.5 and GPT-4 to a range of other inductive problems, and these examples suggest the possibility of developing multiple new benchmarks that explore different aspects of inductive reasoning. Although the existing literature on LLMs explores a wide range of reasoning problems, our impression is that deductive tasks are currently slightly more prominent than inductive tasks. For example, 59 of the 204 BIG-bench are tagged as "logical reasoning" tasks, and 50 are tagged as "commonsense reasoning" (Srivastava et al., 2022). Chater et al. (2011, p 553), however, suggest that "many, and perhaps even almost all, inferences outside mathematics involves uncertain inductive inference" (p 553). Although inductive and deductive reasoning are both important, future work on LLMs should perhaps prioritize inductive reasoning because the majority of reasoning problems encountered in everyday situations are inductive rather than deductive.

5. Conclusion

We compared inductive inferences in reasoning in humans and large language models, and found that GPT-4 provides a relatively good account of property induction in humans. At the time we began this project GPT-3 was the most advanced model available to us, and we correctly anticipated that this model would struggle to account for many aspects of our data. GPT-4, however, performs substantially better, which motivates future work on property effects and other inductive phenomena that may be more challenging to capture than most of the phenomena considered here.

Our work draws on AI and psychology and holds lessons for both fields. For AI, our work suggests that evaluations of LLMs can draw on psychological work on inductive reasoning. Moreover, our datasets represent a step towards a large scale evaluative benchmark that could be considered alongside other popular benchmarks. For psychology, our work suggests that comprehensive benchmarks similar to those used in evaluating LLMs can also valuable for understanding how humans learn and reason. In comparing LLMs and people we realized that we do not have a convincing theoretical account on *either* side of the comparison: not only do we lack an understanding of how GPT-4 succeeds on these tasks, but we also lack a detailed picture of how humans perform as well. Psychological models such as the similarity-coverage model are useful starting points, but they only account partially for our data, and it is unclear whether they can be extended to handle phenomena such as property effects. We hope that large language models point the way towards psychological models that come closer to capturing the rich intricacy of human inductive reasoning.

CRediT authorship contribution statement

Simon Jerome Han: Wrote code for the project, Wrote the paper, Discussed the models and analyses and commented on the manuscript. **Keith J. Ransom:** Wrote code for the project, Discussed the models and analyses and commented on the manuscript. **Andrew Perfors:** Wrote the paper, Discussed the models and analyses and commented on the manuscript. **Charles Kemp:** Wrote code for the project, Wrote the paper, Discussed the models and analyses and commented on the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and code are available here: https://github.com/S-J-HAN/InductiveReasoningInLargeLanguageModels.

Acknowledgments

This work was supported in part by the Complex Human Data Hub at the University of Melbourne and by Australian Research Council Grant FT190100200.

Appendix A. Prompt variations from Experiment 1

Each of the possible prompts for GPT-4 was constructed by combining a system message with variations of four components. These four components are the **Context**, **Arguments**, **Question**, and **Options**. Each variant for each component is shown below.

System: General, task-agnostic instructions

ID	*Description*	*Format*
S1	Default (blank)	
S2	Be specific	`You are an intelligent language model that responds exactly as people do. The questions you'll see don't have right or wrong answers, and you are willing to use your best judgment and commit to a concrete, specific response even in cases where you can't be sure that you are correct.`
S3	Domain expert	`You are an expert on X and the types of real world properties that they have. The questions you'll see don't have right or wrong answers, and you are willing to use your best judgment and commit to a concrete, specific response even in cases where you can't be sure that you are correct.` (X = `living things` or `objects`, depending on the domain)

Context: Explanation of the task and general task info

ID	*Description*	*Format*
C1	Same as human experiment	`We are interested in how people evaluate arguments. On each trial there will be two arguments labeled 'A' and 'B.' Each will contain one, two, or three statements separated from a claim by a line. Assume that the statements above the line are facts, and choose the argument whose facts provide a better reason for believing the claim. These are subjective judgments; there are no right or wrong answers.`
C2	Same as C1, but used GPT-4 to paraphrase	`We aim to understand how individuals assess arguments. In every trial, you will encounter two arguments marked as 'A' and 'B.' Each argument may have one, two, or three statements, which are followed by a claim and separated by a line. Treat the statements above the line as factual and select the argument with facts that better support the claim. Keep in mind that these evaluations are subjective, and there are no correct or incorrect answers.`
C3	Sparse context	`On each trial there will be two arguments labeled 'A' and 'B.' Each will contain one, two, or three statements separated from a claim by a line. Assume that the statements above the line are facts, and choose the argument whose facts provide a better reason for believing the claim.`
C4	No context	

Arguments: Structure of the arguments being rated

ID	*Description*	*Format*

A1	Same as human experiment	`Argument A: Fact - Dogs have property P.`
		`Claim - All mammals have property P.`
		`Argument B: Fact - Hedgehogs have property P.`
		`Claim - All mammals have property P.`

A2	Similar to A1, but minor rewording	`Argument A: Fact - Dogs possess property P.`
		`Claim - All mammals possess property P.`
		`Argument B: Fact - Hedgehogs possess property P.`
		`Claim - All mammals possess property P.`

A3	More natural	`Argument A: Based on the fact that dogs have property P, we claim that all mammals have property P.` `Argument B: Based on the fact that hedgehogs have property P, we claim that all mammals have property P.`

Question: How we ask GPT-4 to make a judgment

ID	*Description*	*Format*
Q1	Same as human experiment	`Question: Assuming all the facts given are true, which argument makes a stronger case for the claim?`
Q2	Simplest	`Question: Which argument makes a stronger case for the claim?`
Q3	Prompts reasoning	`Question: Assuming all the facts given are true, which argument makes a stronger case for the claim? To get the best answer, first write down your reasoning. Then, based on this,`
Q4	Prompts reasoning for properties	`Question: Assuming all the facts given are true, which argument makes a stronger case for the claim? To get the best answer, first write down your reasoning about which real world properties ‘property P’ might be referring to. Then, based on this,`

Options: How we specify the format of the answer

ID	*Description*	*Format*

O1	Same as human experiment	`Indicate the strength of your preference by providing one of the following options:`
		`A - Argument A is much stronger`
		`B - Argument A is stronger`
		`C - Argument A is slightly stronger`
		`D - Argument B is slightly stronger`
		`E - Argument B is stronger`
		`F - Argument B is much stronger`

O2	Scale from 0 to 100	`Indicate the strength of your preference by providing a number between 0 and 100, where 0 corresponds to argument A being much stronger and 100 corresponds to argument B being much stronger.`

Appendix B. Prompt variations from Experiment 2

As in Experiment 1, each of the possible prompts for GPT-4 was constructed by combining a system message with variations of four components. The system message was the same as before and the variants for each of the other components are shown below.

Context: Explanation of the task and general task info

ID	*Description*	*Format*
C1	Same as human experiment	`We’re going to show you a series of claims relating to living things and the properties they share. Rather than mention any specific property (e.g. ‘‘Hyenas have sesamoid bones’’) we’ll refer to an unspecified property (e.g. ‘‘Hyenas have property P’’). Each claim may be true or false, and to help you decide which, we’ll provide you with facts about whether or not other living things have the same property (e.g. ‘‘Lions have property P’’, and ‘‘Zebras have property P’’).`
C2	Sparse context	`We’re going to show you a series of claims relating to living things and the properties they share. Rather than mention any specific property we’ll refer to an unspecified property. Each claim may be true or false, and to help you decide which, we’ll provide you with facts about whether or not other living things have the same property.`
C3	No context	

Arguments: Structure of the arguments being rated		
ID	*Description*	*Format*
A1	Same as human experiment	`Argument A: Fact - Dogs have property P.` `Claim - All mammals have property P.`
A2	Minor reword of A1	`Argument A: Fact - Dogs possess property P.` `Claim - All mammals possess property P.`
A3	More natural	`Argument A: Based on the fact that dogs have property P, we claim` `that all mammals have property P.`

Question: how we ask GPT-4 to make a judgment		
ID	*Description*	*Format*
Q1	Same as human experiment	`Question: Given the facts presented, how likely is it that the claim is true?`
Q2	Simplest	`Question: How likely is it that the claim is true?`
Q3	Prompts reasoning	`Question: Given the facts presented, how likely is it that the claim is true? To get the best answer, first write down your reasoning. Then, based on this,`
Q4	Prompts reasoning for properties	`Question: Given the facts presented, how likely is it that the claim is true? To get the best answer, first write down your reasoning about which real world properties ‘property P’ might be referring to. Then, based on this,`

Options: How we specify the format of the answer		
ID	*Description*	*Format*
O1	Same as human experiment	`Indicate your answer by providing a number between 0 and 100, where 0 means that the claim is very unlikely and 100 means that the claim is very likely.`

O2	Six choice options	`Indicate your answer by providing one of the following options:`
		`A - The claim is very unlikely`
		`B - The claim is moderately unlikely`
		`C - The claim is slightly unlikely`
		`D - The claim is slightly likely`
		`E - The claim is moderately likely`
		`F - The claim is very likely`

Trials: inserted practice trials after context and before main trials		
ID	*Description*	*Format*

T1	Same as human experiment	`This section contains a series of claims that include only one supporting fact. Before we start, we'll give you two examples as practice.`
		
		`Argument A: Fact - Papayas have property P.`
		`Claim - All fruits have property P.`
		`Question: Given the facts presented, how likely is it that the claim is true? Indicate your answer by providing a number between 0 and 100, where 0 means that the claim is very unlikely and 100 means that the claim is very likely.`
		
		`[GPT-4 completion assistant message]`
		
		`Argument A: Fact - Apples have property P.`
		`Claim - All fruits have property P.`
		
		`Question: Given the facts presented, how likely is it that the claim is true? Indicate your answer by providing a number between 0 and 100, where 0 means that the claim is very unlikely and 100 means that the claim is very likely.`
		
		`[GPT-4 completion assistant message]`

T2	No trials	

Appendix C. Property induction phenomena from Experiment 1

Table C.7 shows responses for all of the argument pairs listed in Table 1. The prompts used the same format shown in Table 3, and Table C.7 shows the first result obtained for each prompt (we did not query GPT-3.5 and GPT-4 multiple times and select among the responses).

Because there is a single prompt for each phenomenon and Table C.7 includes only a single response to each prompt, the results in this table and in the appendices that follow are little more than anecdotal. Even so they suggest that GPT-4 often explains its choices in ways that seem sensible and invoke some of the key notions such as similarity (row 7), typicality (row 2) and diversity (row 6) considered in the psychological literature on property induction. The explanations offered by GPT-4 are also usually more convincing than those provided by GPT-3.5. Because non-monotonicity is the most important phenomenon not consistently captured by GPT-4, the responses for the non-monotonicity pairs are of particular interest. GPT-4's response in row 9 suggests that additional positive examples provide additional support for the conclusion, which is an appropriate explanation of *monotonicity*, but does not capture how people tend to think about *non-monotonicity*. GPT-4's response in row 8 actually shows a non-monotonicity effect, but the explanation provided is unconvincing — according to that explanation both arguments in the pair should be equally strong.

Table C.7
Responses for argument pairs illustrating the eleven inductive phenomena in Experiment 1. In all cases argument A is stronger according to data reported by Osherson et al. (1990).

Phenomenon	GPT-4	GPT-3.5
(1) Similarity *Argument A:* Robins Bluejays → Sparrows *Argument B:* Robins Bluejays → Geese	`Robins, Bluejays, and Sparrows are all small birds that belong to the passerine order. Geese, on the other hand, are larger birds that belong to the Anseriformes order. Since the facts provided in both arguments are the same, we can assume that property P might be more likely to be shared among birds that are more closely related. Based on this reasoning, I would choose option B - Argument A is moderately stronger.`	`Answer: B - Argument A is moderately stronger. Reasoning: Argument A provides two facts that support the claim that sparrows have property P, which is more evidence than Argument B provides for the claim that geese have property P. Therefore, Argument A makes a moderately stronger case for the claim.`

(*continued on next page*)

Table C.7 (*continued*).

Phenomenon	GPT-4	GPT-3.5
(2) Typicality *Argument A:* Robins → All birds *Argument B:* Penguins → All birds	`Robins are more representative of the typical bird, as they can fly and have many features that are commonly associated with birds. Penguins, on the other hand, are flightless and have some unique features that set them apart from other birds. Therefore, if a property is found in robins, it is more likely to be found in other birds as well, compared to if the property is found in penguins. Based on this reasoning, I would choose option B - Argument A is moderately stronger.`	`C - Argument A is slightly stronger.` `The fact that robins have property P provides some evidence that all birds have property P, but the fact that penguins have property P provides stronger evidence that all birds have property P. Therefore, Argument B makes a slightly stronger case for the claim than Argument A.`
(3) Specificity *Argument A:* Bluejays Falcons → All birds *Argument B:* Bluejays Falcons → All animals	`Both arguments are based on the same facts, which are that Bluejays and Falcons have property P. Bluejays and Falcons are both types of birds, so it is more reasonable to generalize the property P to all birds rather than all animals. Generalizing to all animals would require more diverse examples of animals having property P. Based on this reasoning, I would choose option B - Argument A is moderately stronger.`	`Answer: B - Argument A is moderately stronger.` `Reasoning: Argument A makes a stronger case for the claim because it is more specific and limited in scope. It states that all birds have property P, which is supported by the facts given. Argument B states that all animals have property P, which is not supported by the facts given since only birds have been mentioned. Therefore, Argument A is moderately stronger.`
(4) Monotonicity (General) *Argument A:* Sparrows Eagles Hawks → All birds *Argument B:* Sparrows Eagles → All birds	`Argument A provides three examples of birds with property P, while Argument B provides only two examples. Having more examples of birds with property P strengthens the case for the claim that all birds have property P. Based on this reasoning, I would choose option B - Argument A is moderately stronger.`	`Answer: B - Argument A is moderately stronger.` `Reasoning: Argument A provides more evidence for the claim that all birds have property P, as it includes three facts that support the claim. Argument B only includes two facts, which is not as strong of a case for the claim. Therefore, Argument A is moderately stronger than Argument B.`
(5) Monotonicity (Specific) *Argument A:* Pigs Wolves Foxes → Gorillas *Argument B:* Pigs Wolves → Gorillas	`Argument A provides more evidence for the claim by including an additional animal (foxes) that has property P. This additional evidence strengthens the argument by showing that property P is more widespread among animals, making it more likely that gorillas might also have property P. Based on this reasoning, I would choose option B - Argument A is moderately stronger.`	`Answer: B - Argument A is moderately stronger.` `Reasoning: Argument A provides more evidence for the claim that gorillas have property P, as it includes three facts that all support the claim. Argument B only includes two facts, so it provides less evidence for the claim. Therefore, Argument A is moderately stronger than Argument B.`
(6) Diversity (General) *Argument A:* Hippos Hamsters → All mammals *Argument B:* Hippos Rhinos → All mammals	`Argument A provides evidence that both a large aquatic mammal (hippo) and a small terrestrial mammal (hamster) have property P. Argument B provides evidence that two large, semi-aquatic mammals (hippo and rhino) have property P. Since Argument A includes a wider range of mammals in terms of size and habitat, it provides a slightly better case for the claim that all mammals have property P. Answer: C - Argument A is slightly stronger`	`Answer: B - Argument A is moderately stronger.` `Reasoning: Argument A provides two examples of mammals that have property P, while Argument B provides only one. This suggests that Argument A provides a better case for the claim that all mammals have property P.`

(*continued on next page*)

Table C.7 (*continued*).

Phenomenon	GPT-4	GPT-3.5
(7) Diversity (Specific) *Argument A:* Lions Giraffes → Rabbits *Argument B:* Lions Tigers → Rabbits	In Argument A, we have two animals with property P - lions and giraffes. Lions are carnivorous mammals, while giraffes are herbivorous mammals. In Argument B, we have lions and tigers with property P, both of which are carnivorous mammals. Since rabbits are herbivorous mammals, they share a more similar trait with giraffes than with lions or tigers. Therefore, the presence of an herbivorous mammal (giraffe) in Argument A makes it slightly more likely that rabbits might also have property P. Answer: C - Argument A is slightly stronger	Answer: B - Argument A is moderately stronger. Reasoning: Argument A provides two facts that are both relevant to the claim, which makes it a moderately stronger argument than Argument B, which only provides one relevant fact.
(8) Non-monotonicity (General) *Argument A:* Crows Peacocks → All birds *Argument B:* Crows Peacocks Rabbits → All birds	Both arguments A and B have the same claim that all birds have property P. Argument A provides two examples of birds (crows and peacocks) having property P, while Argument B provides the same two examples of birds plus an additional example of a non-bird (rabbits) having property P. The inclusion of rabbits in Argument B does not strengthen the claim about birds, as rabbits are not birds. Therefore, the additional fact in Argument B does not contribute to the claim about birds. Based on this reasoning, I would choose option C - Argument A is slightly stronger. This is because both arguments have the same bird examples, but Argument B includes an irrelevant fact about rabbits that does not contribute to the claim about birds.	Answer: E - Argument B is moderately stronger. Reasoning: Argument B provides more evidence for the claim that all birds have property P by providing an additional fact that rabbits have property P. This additional fact strengthens the argument that all birds have property P, making Argument B moderately stronger than Argument A
(9) Non-monotonicity (Specific) *Argument A:* Flies → Bees *Argument B:* Flies Orangutans → Bees	Argument A only provides one example of an animal (flies) having property P. Argument B, on the other hand, provides two examples of animals (flies and orangutans) having property P. Although neither flies nor orangutans are closely related to bees, the fact that two different animals have property P makes it more likely that bees might also have property P. Based on this reasoning, my preference is: E - Argument B is moderately stronger	Answer: B - Argument A is moderately stronger. Reasoning: Argument A provides one fact that supports the claim that bees have property P, while Argument B provides two facts that support the claim. Therefore, Argument B makes a stronger case for the claim that bees have property P.
(10) Asymmetry *Argument A:* Mice → Bats *Argument B:* Bats → Mice	Mice and bats are both mammals, but they belong to different orders (Rodentia and Chiroptera, respectively). While they share some common features, they also have many differences. In this case, it is not clear what property P is, so it is difficult to determine how likely it is that one species having property P would imply the other species also has it. However, bats are more closely related to other mammals like primates and carnivores than they are to rodents like mice. This means that if a property is found in bats, it might be more likely to be found in other mammals, including mice. On the other hand, if a property is found in mice, it might be less likely to be found in more distantly related mammals like bats. Based on this reasoning, I would choose option E - Argument B is moderately stronger.	Answer: F - Argument B is much stronger. The fact given in Argument B directly supports the claim, while the fact given in Argument A does not. Therefore, Argument B makes a much stronger case for the claim than Argument A.

(*continued on next page*)

Table C.7 (*continued*).

Phenomenon	GPT-4	GPT-3.5
(11) Inclusion Fallacy *Argument A:* Robins → All birds *Argument B:* Robins → Ostriches	`Argument A makes a general claim about all birds, while Argument B makes a specific claim about ostriches. Since robins and ostriches are both birds, knowing that robins have property P provides some evidence that ostriches might also have property P. However, it does not provide strong evidence that all birds have property P, as there are many different types of birds with varying properties. Based on this reasoning, I would choose option D - Argument B is slightly stronger.`	`Answer: B - Argument A is moderately stronger. Reasoning: Argument A makes a stronger case for the claim because it states that all birds have property P, which is supported by the fact that robins have property P. Argument B states that ostriches have property P, but does not provide any evidence to support this claim. Therefore, Argument A is moderately stronger.`

Appendix D. Results for other OpenAI models

We began this project when GPT-3 (`davinci`) was the best performing OpenAI model available, and updated our work as subsequent versions of GPT-3, GPT-3.5 and GPT-4 were released. Our primary analysis focuses on GPT-3.5 (`text-davinci-003`) and GPT-4 because they are the best performing completion and chat based OpenAI models respectively, but here we provide results for all models that we analyzed.

For Experiment 1, Table D.8 provides a quantitative evaluation of GPT-3.5's chat variant (`gpt-3.5-turbo-0613`) alongside the other models that we report in Table 4 for the argument pair rating task. Although we ran this analysis on other GPT-3 and GPT-3.5 variants including `davinci`, `text-davinci-001` and `text-davinci-002`, we found that these models largely failed to respond appropriately to our prompt, which may not be surprising because this prompt was optimized for GPT-4.

Like GPT-4, GPT-3.5's chat variant captures the phenomena of specificity and monotonicity, and it also leans towards predicting the opposite argument from humans for non-monotonicity. GPT-3.5's chat variant, however, fails to capture similarity and typicality as robustly as does GPT-4, and also displays more sensitivity towards the inclusion fallacy.

For Experiment 2, Fig. D.7 displays model correlations with humans for our rating task across five models: `text-davinci-001`, `text-davinci-002`, `text-davinci-003`, `gpt-3.5-turbo` and GPT-4. Like our analysis for Experiment 1, we also attempted to gather results for `davinci`, but found that our GPT-4 optimized prompt did not work on models that lacked instruction tuning.

In general, we found that the differences between the GPT models are relatively small. The most striking difference is the degradation that can be observed across the models for single premise general arguments. We believe that this finding may reflect recent alignment efforts to prevent models from making broad generalizations ('Claim - *All* X have property P') based on user prompts that contain limited information.

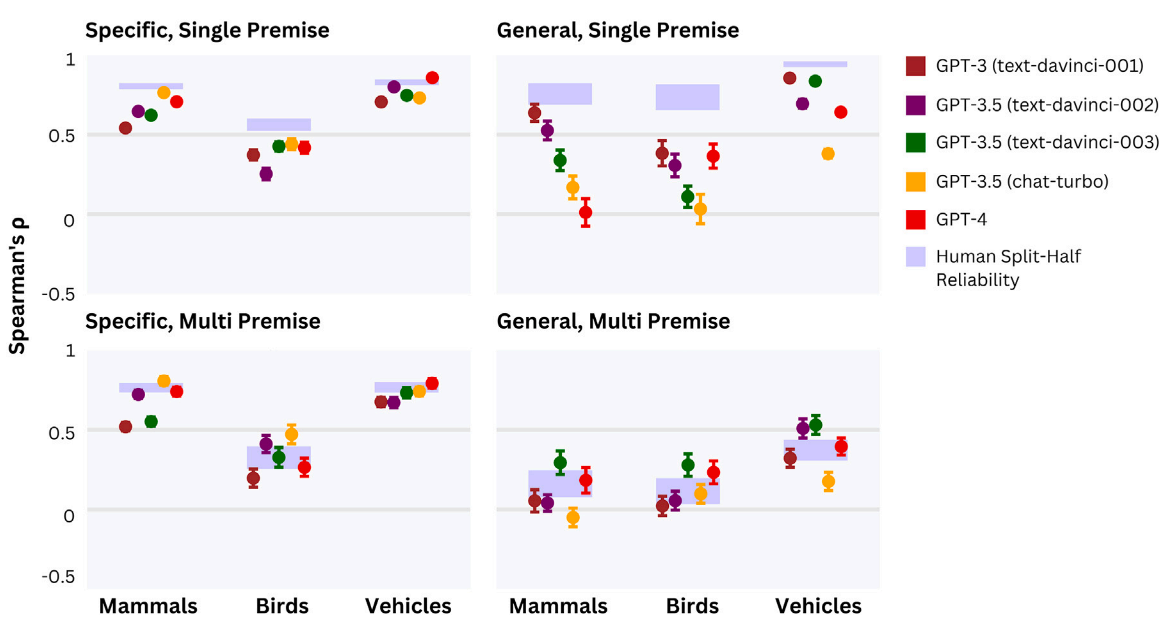

Fig. D.7. Correlations (Spearman's ρ) between human argument rankings and rankings of GPT-3 (text-davinci-001), GPT-3.5 (text-davinci-002, text-davinci-003 and chat-turbo), and GPT-4.

Table D.8
Quantitative evaluation of GPT-3.5 (text-davinci-003), GPT-3.5 (chat-turbo), GPT-4 and Humans on the 11 phenomena across all three domains.

Phenomenon	Domain	GPT-3.5 (text-davinci-003)	GPT-3.5 (chat-turbo)	GPT-4	Humans
Similarity	Mammals	0.17	0.17	<0.001 *	<0.001 *
	Birds	0.69	0.4	<0.001 *	<0.001 *
	Vehicles	0.11	<0.001 *	<0.001 *	<0.001 *
Typicality	Mammals	0.54	0.29	<0.02 *	<0.001 *
	Birds	0.54	0.15	<0.001 *	<0.001 *
	Vehicles	1.0	<0.02 *	<0.001 *	<0.001 *
Specificity	Mammals	0.84	<0.001 *	<0.001 *	<0.001 *
	Birds	0.84	<0.001 *	<0.001 *	<0.001 *
	Vehicles	0.54	<0.001 *	<0.001 *	<0.001 *
Monotonicity (General)	Mammals	<0.02 *	<0.001 *	<0.001 *	<0.001 *
	Birds	0.54	<0.001 *	<0.001 *	<0.001 *
	Vehicles	<0.001 *	<0.001 *	<0.001 *	<0.001 *
Monotonicity (Specific)	Mammals	0.31	<0.001 *	<0.001 *	0.06
	Birds	0.84	<0.001 *	<0.001 *	<0.001 *
	Vehicles	0.84	<0.001 *	<0.001 *	0.29
Diversity (General)	Mammals	0.84	0.31	<0.001 *	0.06
	Birds	0.54	0.15	<0.001 *	1.0
	Vehicles	1.0	0.84	0.06	< 0.03∘
Diversity (Specific)	Mammals	1.0	1.0	0.15	< 0.01∘
	Birds	0.54	0.68	0.84	0.68
	Vehicles	0.31	0.4	0.15	0.68
Nonmonotonicity (General)	Mammals	< 0.001∘	1.0	<0.02 *	<0.001 *
	Birds	0.31	0.06	<0.001 *	<0.001 *
	Vehicles	0.84	< 0.02∘	0.15	<0.01 *
Nonmonotonicity (Specific)	Mammals	1.0	< 0.001∘	< 0.001∘	0.15
	Birds	0.15	< 0.001∘	< 0.001∘	<0.001 *
	Vehicles	1.0	< 0.001∘	< 0.001∘	<0.01 *
Asymmetry	Mammals	0.06	0.82	0.68	0.4
	Birds	0.06	0.84	1.0	<0.001 *
	Vehicles	0.54	0.09	0.82	0.05
Inclusion Fallacy	Mammals	0.84	<0.001 *	0.06	0.06
	Birds	0.84	<0.02 *	< 0.001∘	1.0
	Vehicles	0.54	0.06	0.54	<0.001 *

Appendix E. Other property induction phenomena

Table E.9 illustrates a range of property induction phenomena from the literature that were not explored in our experiments but can be systematically explored in future work. As in Appendix C, the prompts matched the format shown in Table 3, and Table E.9 shows the first model response we obtained for each argument pair.

The first four rows of the table illustrate phenomena described by Medin et al. (2003) that involve inferences about blank properties. GPT-4 provides relatively compelling responses to the first two argument-pairs, but the explanation given for the third pair is unconvincing (birds and mammals are both vertebrates so it is not clear that birds are more closely related to mammals than to insects). The fourth pair illustrates that non-monotonicity can arise for reasons that differ from those captured by our non-monotonicity stimuli in Experiment 1. In that experiment non-monotonicity was achieved by adding a premise drawn from a different superordinate category, but the fourth row shows that non-monotonicity can also arise when all premises share a very salient feature (e.g. the feature of being bears) which makes it less likely that “Property P” extends outside the set of items with that feature. GPT-4’s response to the fourth pair suggests that it may struggle with non-monotonicity effects in general, not just with the variety of non-monotonicity explored in Experiment 1.

The final six rows in Table E.9 all involve arguments with non-blank predicates. They include four cases (rows 5,6,8 and 10) in which the specific predicate used overturns inferences that would be expected on the basis of overall similarity between the premise and conclusion categories. For example, german shepherds are more similar to dobermanns than to poodles, but learning that poodles “bite through wire” provides more evidence that german shepherds have the same property than learning that dobermanns bite through wire. Similarly, mosquitoes are more similar to ladybugs than to vampire bats, but learning that vampire bats “travel at speeds of twice their body length per second” after eating provides more evidence that mosquitoes have the same property than learning that ladybugs have the property. Across all four cases, GPT-4 gives responses that are consistent with overall similarity between premise and conclusion categories rather than the specific properties used in the arguments, suggesting that the model provides a relatively poor account of inferences about non-blank properties. In contrast, people appear to rely on similarity only when reasoning about properties (e.g. rows 7 and 9) that align well with overall similarity relationships.

Table E.9
Phenomena from the property induction literature that go beyond those tested in our analyses. All argument pairs are either identical to or modeled on pairs for which average human ratings indicate that argument A is stronger than argument B.

Phenomenon	GPT-4	GPT-3.5
(1) Causal asymmetry (Medin et al., 2003) Argument A: F – Carrots have property P. C – Rabbits have property P. Argument B: F – Rabbits have property P. C – Carrots have property P.	Carrots are a type of vegetable, while rabbits are animals. It is more likely that a property found in a vegetable could be present in an animal that consumes it, rather than a property found in an animal being present in the vegetable it consumes. This is because animals can obtain certain properties from their diet, while vegetables do not obtain properties from the animals that consume them. Based on this reasoning, I would choose option B – Argument A is moderately stronger	Answer: E – Argument B is moderately stronger. Reasoning: Argument B provides direct evidence that rabbits have property P, while Argument A does not provide any direct evidence that rabbits have property P. Therefore, Argument B makes a stronger case for the claim that carrots have property P
(2) Causal violations of similarity (Medin et al., 2003) Argument A: F – Bananas have property P. C – Monkeys have property P. Argument B: F – Mice have property P. C – Monkeys have property P.	Monkeys are more closely related to bananas in terms of their diet and habitat, as they are known to consume bananas and live in regions where bananas grow. Mice, on the other hand, have a more distant relationship with monkeys in terms of their diet and habitat. Therefore, it is more likely that a property found in bananas could also be found in monkeys, as opposed to a property found in mice. Based on this reasoning, I would choose option B – Argument A is moderately stronger.	Answer: B – Argument A is moderately stronger. Reasoning: Argument A provides a direct connection between the fact and the claim, as bananas are a type of fruit that monkeys eat. Argument B provides an indirect connection, as mice are not typically associated with monkeys. Therefore, Argument A provides a moderately stronger case for the claim that monkeys have property P.
(3) Causal nondiversity (Medin et al., 2003) Argument A: F – Fleas have property P. F – Butterflies have property P. C – Sparrows have property P. Argument B: F – Fleas have property P. F – Dogs have property P. C – Sparrows have property P.	Fleas and butterflies are both insects, while dogs are mammals. Sparrows are birds, which are more closely related to insects than mammals in terms of taxonomy. Therefore, the fact that both fleas and butterflies have property P might be more relevant to the claim that sparrows have property P than the fact that fleas and dogs have property P. Based on this reasoning, I would choose option B – Argument A is moderately stronger.	Answer: B – Argument A is moderately stronger. Reasoning: Argument A provides two facts that are both relevant to the claim, which makes it a moderately stronger argument than Argument B, which only provides one relevant fact.
(4) Non-monotonicity via property reinforcement (Medin et al., 2003) Argument A: F – Brown bears have property P. C – Buffalos have property P. Argument B: F – Brown bears have property P. F – Polar bears have property P. F – Grizzly bears have property P. C – Buffalos have property P.	Argument A provides only one fact, that brown bears have property P. Argument B provides three facts, that brown bears, polar bears, and grizzly bears all have property P. Although none of these animals are closely related to buffalos, Argument B provides more evidence of property P being present in a wider range of animals, which could potentially increase the likelihood that buffalos also have property P. Based on this reasoning, I would choose option D – Argument B is slightly stronger	Answer: E – Argument B is moderately stronger. Reasoning: Argument B provides more evidence for the claim that buffalos have property P by providing three facts about other animals that have property P. This makes the argument stronger than Argument A, which only provides one fact.
(5) Property effects (Smith et al., 1993) Argument A: F – Poodles can bite through wire. C – German shepherds can bite through wire. Argument B: F – Dobermanns can bite through wire. C – German shepherds can bite through wire.	Both Poodles and Dobermanns are different breeds of dogs, just like German Shepherds. While Poodles are known for their intelligence and agility, Dobermanns are known for their strength and guarding abilities. German Shepherds are also known for their strength, intelligence, and guarding abilities. Since Dobermanns and German Shepherds share more similar characteristics in terms of strength and function, it is more likely that if Dobermanns can bite through wire, German Shepherds can as well. Preference: E – Argument B is moderately stronger	Answer: E – Argument B is moderately stronger. Reasoning: Argument B provides a more direct connection between the fact and the claim, as it states that a Dobermann, which is a type of German shepherd, can bite through wire. This is a more specific example than the fact provided in Argument A, which only states that poodles can bite through wire. Therefore, Argument B makes a moderately stronger case for the claim that German shepherds can bite through wire.

(continued on next page)

Table E.9 (*continued*).

Phenomenon	GPT-4	GPT-3.5
(6) Property effects (Smith et al., 1993) Argument A: F - House cats have skins that are more resistant to penetration than most synthetic fibers. C - Hippos have skins that are more resistant to penetration than most synthetic fibers. Argument B: F - Elephants have skins that are more resistant to penetration than most synthetic fibers. C - Hippos have skins that are more resistant to penetration than most synthetic fibers.	Both house cats and elephants have skin that is more resistant to penetration than most synthetic fibers. However, elephants are more similar to hippos in terms of size, habitat, and overall biology than house cats are. Therefore, it is more likely that the skin properties of elephants would be a better indicator of the skin properties of hippos. Based on this reasoning, I would choose option F - Argument B is much stronger.	Answer: E - Argument B is moderately stronger. Reasoning: Argument B provides a more direct comparison to the claim, as it states that elephants have skin that is more resistant to penetration than most synthetic fibers. This is a more direct comparison to the claim that hippos have skin that is more resistant to penetration than most synthetic fibers. Argument A only states that house cats have skin that is more resistant to penetration than most synthetic fibers, which is not as direct of a comparison. Therefore, Argument B makes a moderately stronger case for the claim.
(7) Property effects (Heit & Rubinstein, 1994) Argument A: F - Chickens have livers with two chambers that act as one. C - Hawks have livers with two chambers that act as one. Argument B: F - Tigers have livers with two chambers that act as one. C - Hawks have livers with two chambers that act as one.	Chickens and hawks are both birds, while tigers are mammals. Since the claim is about hawks, it is more likely that the liver structure of another bird (chicken) would be similar to that of a hawk than the liver structure of a mammal (tiger). Based on this reasoning, I would choose option B - Argument A is moderately stronger.	Answer: B - Argument A is moderately stronger. Reasoning: Argument A provides a direct comparison between chickens and hawks, which are both birds, and suggests that hawks likely have livers with two chambers that act as one. Argument B provides a comparison between tigers and hawks, which are not closely related, and does not provide as strong of evidence that hawks have livers with two chambers that act as one.
(8) Property effects (Heit & Rubinstein, 1994) Argument A: F - Tigers usually gather large amounts of food at once. C - Hawks usually gather large amounts of food at once. Argument B: F - Chickens usually gather large amounts of food at once. C - Hawks usually gather large amounts of food at once.	Both tigers and chickens are not closely related to hawks, as tigers are mammals and chickens are birds but not birds of prey. However, chickens might be slightly more related to hawks than tigers, as they are both birds. Therefore, the fact that chickens gather large amounts of food at once might be a slightly better indicator of hawks' behavior than the fact that tigers do the same. Based on this reasoning, my preference is: D - Argument B is slightly stronger	Answer: B - Argument A is moderately stronger. Reasoning: Argument A provides a more relevant fact than Argument B, as tigers and hawks are both predatory animals that hunt for food, while chickens are not. Therefore, Argument A provides a better reason for believing the claim that hawks usually gather large amounts of food at once.
(9) Property effects (Heit & Rubinstein, 1994) Argument A: F - Ladybugs have some cells in their respiratory systems that require carbon dioxide to function. C - Mosquitoes have some cells in their respiratory systems that require carbon dioxide to function. Argument B: F - Vampire bats have some cells in their respiratory systems that require carbon dioxide to function. C - Mosquitoes have some cells in their respiratory systems that require carbon dioxide to function.	Ladybugs and mosquitoes are both insects, while vampire bats are mammals. Since insects are more closely related to each other than they are to mammals, it is more likely that they share similar respiratory system properties. Based on this reasoning, I would choose option B - Argument A is moderately stronger.	Answer: E - Argument B is moderately stronger. Reasoning: Argument B provides a more specific example of an animal with cells in its respiratory system that require carbon dioxide to function, which makes it a stronger argument for the claim that mosquitoes have some cells in their respiratory systems that require carbon dioxide to function.

(*continued on next page*)

Table E.9 (*continued*).

Phenomenon	GPT-4	GPT-3.5
(10) Property effects (Heit & Rubinstein, 1994) Argument A: F – After eating, vampire bats travel at speeds of twice their body length per second. C – After eating, mosquitoes travel at speeds of twice their body length per second. Argument B: F – After eating, ladybugs travel at speeds of twice their body length per second. C – After eating, mosquitoes travels at speeds of twice their body length per second.	Vampire bats, ladybugs, and mosquitoes are all different types of animals with distinct characteristics. Vampire bats are mammals, ladybugs are insects, and mosquitoes are also insects. Since ladybugs and mosquitoes are both insects, they share more similarities in terms of their body structure and movement patterns compared to vampire bats. Therefore, the fact about ladybugs might provide a better basis for making a claim about mosquitoes. Based on this reasoning, I would choose option E – Argument B is moderately stronger.	Answer: B – Argument A is moderately stronger. Reasoning: Argument A provides a fact about vampire bats, which are closely related to mosquitoes, and thus provides some evidence that the claim about mosquitoes may be true. Argument B provides a fact about ladybugs, which are not closely related to mosquitoes, and thus provides less evidence that the claim about mosquitoes may be true. Therefore, Argument A provides a moderately stronger case for the claim.

Appendix F. Other inductive phenomena

Property induction is sometimes considered separately from other inductive problems such as categorization and generalization, but all of these problems are so closely related that it makes little sense to separate them when developing theories of induction (Kemp & Jern, 2014). Table F.10 includes model responses to an illustrative set of problems from the broader literature on inductive reasoning.

Table F.10
Phenomena from the broader literature on inductive reasoning.

Phenomenon	GPT-4	GPT-3.5
(1) Feature inference (Kemp, 2011) Animal A has scales and can fly. How likely is it that Animal A has wings?	It is likely that Animal A has wings, as the ability to fly is typically associated with the presence of wings. However, the presence of scales is unusual for a flying animal, as most animals with scales (such as reptiles) do not have wings. This suggests that Animal A might be an atypical or fictional creature. Nonetheless, given the information provided, I would estimate that there is a 70% chance that Animal A has wings, considering that wings are a common adaptation for flight.	It is very likely that Animal A has wings, as most animals with scales and the ability to fly have wings. However, there are some exceptions, such as some species of fish that can fly, so it is not certain that Animal A has wings.
(2) Feature inference (Kemp, 2011) Animal A has scales and two legs. How likely is it that Animal A has wings?	It is moderately likely that Animal A has wings. The presence of scales suggests that Animal A could be a reptile, and the two legs could indicate that it is a bipedal creature. Some reptiles, like certain species of lizards, have wing-like structures that allow them to glide or fly short distances. However, it is not guaranteed that Animal A has wings, as there are also many reptiles with scales and two legs that do not have wings. Without more information, it is difficult to determine the exact likelihood, but based on the given characteristics, I would estimate that there is a moderate chance that Animal A has wings.	It is difficult to say how likely it is that Animal A has wings without knowing more information about the animal. If Animal A is a reptile, then it is unlikely that it has wings, as most reptiles do not have wings. However, if Animal A is a bird, then it is more likely that it has wings, as most birds have wings. Ultimately, without more information, it is impossible to say how likely it is that Animal A has wings.
(3) Plausible reasoning (Collins & Michalski, 1989) Do you think they might grow rice in Florida?	Yes, I think they might grow rice in Florida. Florida has a warm, subtropical climate with plenty of rainfall, which is suitable for growing rice. Additionally, the state has large areas of flat land and abundant water resources, which are essential for rice cultivation. However, rice production in Florida might not be as extensive as in other major rice-growing states like California, Arkansas, or Texas. The primary crops in Florida are citrus fruits, sugarcane, and vegetables, but it is still possible that rice is grown on a smaller scale or for local consumption.	It is possible to grow rice in Florida, but it is not common. Rice requires a lot of water and a long growing season, both of which are not typically found in Florida. Additionally, the soil in Florida is not ideal for growing rice. While it is possible to grow rice in Florida, it is not likely to be successful.

(*continued on next page*)

Table F.10 (*continued*).

Phenomenon	GPT-4	GPT-3.5
Plausible reasoning (4) (Collins & Michalski, 1989) Can a goose quack?	A goose does not quack like a duck. Geese make a variety of vocalizations, including honking, hissing, and cackling. The primary sound a goose makes is a honk, which is different from a duck's quack. The honk is a louder and more assertive sound, often used by geese to communicate with each other or to warn off potential threats. While both ducks and geese are waterfowl and share some similarities, their vocalizations are distinct from one another.	No, a goose cannot quack. Geese make a honking sound, while ducks make a quacking sound.
(5) Theory-based categorization (Rips, 1989) I have a circle-shaped object that is 3 inches in diameter. Do you think it's more likely to be a pizza or a quarter?	I think it's more likely to be a quarter. A 3-inch diameter is quite small for a pizza, as most personal-sized pizzas are at least 6 inches in diameter. On the other hand, a quarter is approximately 0.955 inches in diameter, which is closer to 3 inches than a typical pizza size. While a 3-inch diameter is still larger than a standard quarter, it's more plausible that it could be an oversized novelty quarter or a similarly sized coin, rather than an unusually small pizza.	Based on the size of the object, it is more likely to be a quarter than a pizza. A quarter is approximately 1.2 inches in diameter, so the object is more than twice the size of a quarter. A pizza, on the other hand, is typically much larger than 3 inches in diameter.
(6) Theory-based categorization (Murphy & Medin, 1985) You are at a party and you learn that one of the guests fell into the pool. Why do you think this happened?	I think the guest fell into the pool due to a combination of factors, such as being distracted, possibly consuming alcohol, and being too close to the pool's edge. It's common for people to be engaged in conversations or other activities at a party, which can lead to a lack of awareness of their surroundings. Additionally, if the guest had been drinking, their coordination and balance might have been impaired, making it more likely for them to accidentally fall into the pool. Finally, standing or walking too close to the pool's edge increases the risk of slipping or losing balance and falling in.	It is difficult to say why the guest fell into the pool without knowing more information. It is possible that the guest was not paying attention and accidentally stepped into the pool, or that they were pushed or slipped. It is also possible that the guest was intoxicated and lost their balance.
(7) Theory-based categorization (Keil, 1989) The doctors took a raccoon and shaved away some of its fur. They dyed what was left all black. Then they bleached a single stripe all white down the center of its back. Then, with surgery, they put in its body a sac of super smelly odor, just like a skunk has. When they were done, the animal looked just like a skunk. After the operation was this a skunk or a raccoon?	After the operation, the animal is still a raccoon. Despite the changes made to its appearance and the addition of a smelly odor sac, its genetic makeup and species remain that of a raccoon. The modifications made by the doctors only altered its external appearance and gave it a skunk-like feature, but these changes do not transform the raccoon into a different species. The animal's identity is determined by its genetic makeup and biological characteristics, not just its appearance.	This is a difficult question to answer definitively because it depends on how you define a skunk and a raccoon. From a physical standpoint, the animal looks like a skunk because it has the same physical features. However, genetically, it is still a raccoon because it has not been altered in any way to change its species. Therefore, it is both a skunk and a raccoon.
(8) Theory-based categorization (Keil, 1989) The doctors took a coffeepot. They sawed off the handle, sealed the top, took off the top knob, sealed closed the spout, and sawed it off. They also sawed off the base and attached a flat piece of metal. They attached a little stick, cut a window in it, and filled the metal container with bird food. After the operation was this a coffeepot or a bird feeder?	After the operation, this object can be considered a bird feeder. The reason for this classification is that the original function and features of the coffeepot have been significantly altered and repurposed to serve as a container for bird food. The handle, top knob, and spout have been removed and sealed, and a window has been cut to allow birds access to the food. The addition of a flat piece of metal and a stick further supports its new function as a bird feeder. While it may have originated as a coffeepot, its current form and purpose align more closely with that of a bird feeder.	No, this is not a coffeepot anymore. It has been modified so much that it is now a bird feeder. The handle has been sawed off, the top knob has been removed, the spout has been sealed closed, the base has been sawed off, and a flat piece of metal has been attached. A stick has been attached and a window has been cut in it, and it has been filled with bird food. All of these modifications have changed the object from a coffeepot to a bird feeder.

In all cases, the system message supplied to GPT-4 was

`You will be asked a series of questions that don't have right or wrong answers. You are willing to use your best judgment and commit to a concrete, specific response even in cases where you can't be sure that you are correct.`

The user message included one of the exact questions shown in Table F.10 followed by `Please explain your answer carefully.` As for all other experiments, GPT-3.5 was given a single prompt that included both the system and user messages just described.

The first two rows in the table consider arguments which involve a single category (Animal A) and multiple properties (e.g. "has scales", "can fly" and "has wings" in row 1). They therefore differ from the arguments used in our experiments, all of which involved a single property and multiple categories. In the most general case an argument may involve multiple categories and properties, and these arguments have also been studied in the literature (Kemp, Shafto, & Tenenbaum, 2012). GPT-4 gives a fairly good response to the first question, but in row 2 it makes the questionable claim that there are "many reptiles with scales and two legs that do not have wings".

Rows 3 and 4 consider two examples from the work of Collins and Michalski (1989) on plausible reasoning. Most people have never encountered these questions previously, but can nevertheless construct a chain of reasoning that arrives at a plausible answer. In contrast, these questions are presumably discussed explicitly in the corpora used when training GPT-3.5 and GPT-4, which makes them less than ideal as a test of reasoning in LLMs.

Rows 5 through 8 consider examples from the literature on categorization. In some settings it may be useful to distinguish category labels (e.g. "is a wug") from properties (e.g. "is wuggish") but the problem of projecting a property from several items to a novel item seems deeply related to the problem of projecting a category label. Ultimately, then, it seems best to consider these problems together.

Row 5 is a case that requires reasoners to go beyond similarity (the 3 inch object is more similar to a pizza than a quarter) to infer that the object is likely to be a pizza. GPT-3.5 fails the test, but GPT-4 suggests that it may be an "oversized novelty quarter" and therefore identifies a plausible way in which it could actually be considered a quarter.

Row 6 is inspired by an example in which a person who falls into a pool is classified as "intoxicated" even though falling into a pool is not typically associated with the concept of intoxication. Both GPT-4 and GPT-3.5 give good responses by pointing out that alcohol may have been involved.

Rows 7 and 8 consider questions from a classic line of work that explores how inferences about category membership can go beyond surface appearances. In Row 7, GPT-4 gives an excellent response and argues that the creature is still a raccoon even though it looks identical to a skunk. In Row 8, however, GPT-4 suggests that an analogous transformation applied to a coffeepot can indeed transform this object into a bird feeder. A more extensive analysis of examples like these is provided by Zhang, She, Gerstenberg, and Rose (2023).

Although GPT-4 provides good responses to most of the questions in Table F.10, these questions are even more likely than those in Tables C.7 and Tables E.9 to have appeared in its training data in some form. Finding appropriate ways to test the underlying phenomena is therefore a major challenge for future work on inductive reasoning in large language models.

References

Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bhatia, S., & Richie, R. (2021). Transformer networks of human conceptual knowledge. *Psychological Review*, http://dx.doi.org/10.1037/rev0000319.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), Article e2218523120.

Brewka, G., Dix, J., Konolige, K., et al. (1997). *Nonmonotonic reasoning: an overview, vol. 73*. CSLI publications Stanford.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In H. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary approaches to cognition* (pp. 5–31). Harvard University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. arXiv preprint arXiv:2303.11504.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2011). Inductive logic and empirical psychology. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic. volume 10: inductive logic* (pp. 553–624). Amsterdam: North Holland.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science*, *13*(1), 1–49.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 1–2.

Gelman, S. (2007). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65–95.

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

Glick, J. J. P. (2011). *Uncovering the organization of semantic structure with similarity and inductions* (Ph.D. thesis), Stanford University.

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988.

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2022). Human-like property induction is a challenge for large language models. In *Proceedings of the annual meeting of the cognitive science society*.

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

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, *26*, 1043–1050.

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: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(2), 411–422.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1725–1744).

Jiang, G., Xu, M., Xin, S., Liang, W., Peng, Y., Zhang, C., et al. (2023). MEWL: Few-shot multimodal word learning with referential uncertainty. arXiv preprint arXiv:2306.00503.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kemp, C. (2011). Inductive reasoning about chimeric creatures. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. C. N. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, vol. 24* (pp. 316–324).

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*, 216–233.

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

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *1–2*(64), 35–73.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*, http://dx.doi.org/10.1037/rev0000297.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

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

Lipkin, B., Wong, L., Grand, G., & Tenenbaum, J. B. (2023). Evaluating statistical language models as pragmatic reasoners. arXiv preprint arXiv:2305.01020.

López, A., Atran, S., Coley, J. D., Medin, D., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, *32*(3), 251–295.

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*(5), 1070–1090.

Magar, I., & Schwartz, R. (2022). Data contamination: From memorization to exploitation. arXiv preprint arXiv:2203.08242.

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

Misra, K., Ettinger, A., & Taylor Rayz, J. (2021). Do language models learn typicality judgments from text? In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science society* (pp. 216–222).

Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, *381*(6654).

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context learning and induction heads. arXiv preprint arXiv:2209.11895.

OpenAI (2022). New and improved embedding model. URL https://openai.com/blog/new-and-improved-embedding-model, (Accessed 23 April 2023).

OpenAI (2023). *GPT-4 Technical Report*.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200.

Pothos, E. M., & Wills, A. J. (2011). *Formal Approaches in Categorization*. Cambridge University Press.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 811–828.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: Methods, analysis & insights from training Gopher. arXiv:2112.11446.

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

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

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 21–59). Cambridge: Cambridge University Press.

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

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science*, *38*(6), 1024–1077.

Sap, M., Rashkin, H., Chen, D., LeBras, R., & Choi, Y. (2019). SocialIQA: Commonsense reasoning about social interactions. In *Proc. of the 2019 EMNLP-IJCNLP*.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., et al. (2023). Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. arXiv preprint arXiv:2305.14763.

Shepard, R. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Simon, H. A. (1970). *The sciences of the artificial*. MIT Press.

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

Sloman, S. A., & Lagnado, D. A. (2005). The problem of induction. In R. Morrison, & K. Holyoak (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 95–116). New York: Cambridge University Press.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, *49*, 67–96.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Storks, S., Gao, Q., & Chai, J. Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*(7), 309–318.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current directions in psychological science*, *16*(3), 167–171.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. arXiv preprint arXiv:2302.08399.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual meeting of the association for computational linguistics* (pp. 5797–5808). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1580, URL https://aclanthology.org/P19-1580.

Voorspoels, W., Navarro, D., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25.

Vosniadou, S., & Ortony, A. (Eds.), (1989). *Similarity and analogical reasoning*. Cambridge University Press.

Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent analogical reasoning in large language models. arXiv preprint arXiv:2212.09196.

Zhang, S., She, S., Gerstenberg, T., & Rose, D. (2023). You are what you're for: Essentialist categorization in large language models.