DeepMind

Language models show human-like content effects on reasoning

Ishita Dasgupta[*,1], **Andrew K. Lampinen**[*,1], **Stephanie C. Y. Chan**[1], **Antonia Creswell**[1], **Dharshan Kumaran**[1], **James L. McClelland**[1,2] **and Felix Hill**[1]
[*]Equal contributions, listed alphabetically, [1]DeepMind, [2]Stanford University

Abstract reasoning is a key ability for an intelligent system. Large language models achieve above-chance performance on abstract reasoning tasks, but exhibit many imperfections. However, human abstract reasoning is also imperfect, and depends on our knowledge and beliefs about the content of the reasoning problem. For example, humans reason much more reliably about logical rules that are grounded in everyday situations than arbitrary rules about abstract attributes. The training experiences of language models similarly endow them with prior expectations that reflect human knowledge and beliefs. We therefore hypothesized that language models would show human-like content effects on abstract reasoning problems. We explored this hypothesis across three logical reasoning tasks: natural language inference, judging the logical validity of syllogisms, and the Wason selection task (Wason, 1968). We find that state of the art large language models (with 7 or 70 billion parameters; Hoffmann et al., 2022) reflect many of the same patterns observed in humans across these tasks — like humans, models reason more effectively about believable situations than unrealistic or abstract ones. Our findings have implications for understanding both these cognitive effects, and the factors that contribute to language model performance.

1. Introduction

A hallmark of abstract reasoning is the ability to systematically perform algebraic operations over variables that can be bound to any entity (Newell, 1980; Fodor and Pylyshyn, 1988): the statement: 'X is bigger than Y' logically implies that 'Y is smaller than X', no matter the values of X and Y. That is, abstract reasoning is ideally content-independent (Newell, 1980). The capacity for reliable and consistent abstract reasoning is frequently highlighted as a crucial missing component of current AI (Marcus, 2020; Mitchell, 2021; Russin et al., 2020). For example, while large language models exhibit *emergent* behaviors, including some abstract reasoning performance (Brown et al., 2020; Ganguli et al., 2022; Nye et al., 2021a; Kojima et al., 2022; Wei et al., 2022a), they have been criticized for inconsistencies in their abstract reasoning (e.g. Rae et al., 2021; Razeghi et al., 2022; Patel et al., 2021; Valmeekam et al., 2022).

However, these commentaries often overlook the fact that humans — our standard for intelligent behavior — are far from perfectly rational abstract reasoners (Gigerenzer and Gaissmaier, 2011; Kahneman et al., 1982; Marcus, 2009). The human ability to perform abstract reasoning is heavily influenced by our knowledge and beliefs about the content over which we are reasoning (Johnson-Laird et al., 1972; Wason, 1968; Wason and Johnson-Laird, 1972; Evans, 1989; Evans et al., 1983; Cohen et al., 2017). Humans reason more readily and more accurately about familiar, believable situations, compared to unfamiliar ones where their prior beliefs no longer apply, or situations that contradict their beliefs.

For example, when presented with a syllogism like the following:

> All students read.
> Some people who read also write essays.
> Therefore some students write essays.

humans will often classify it as a valid argument. However, when presented with:

> All students read.
> Some people who read are professors.
> Therefore some students are professors.

If seas are bigger than puddles, then puddles are **smaller than seas**

(a) Natural language inference.

Argument:
All diamonds are gems.
Some gems are transparent things.
Conclusion: Some diamonds are transparent things.
Answer: **This argument is invalid.**

(b) Syllogism validity.

A parent needs to check that their children are following the rules. The following cards represent the children; they each have a current activity on one side and a homework completion on their other side. The rule is that if the children are playing games then they must have finished their homework. Which of the cards would you need to flip over to help them to ensure that the rule is being followed?
card: playing video games
card: eating dinner
card: homework is not done
card: homework is done
Answer: **You need to flip over the ones showing "playing video games" and "homework is not done"**

(c) Wason selection task.

Figure 1 | Examples of the three logical reasoning tasks we evaluate: (a) simple single-step natural language inferences, (b) assessing the validity of logical syllogisms, and (c) the Wason selection task. In each case, the model must choose the answer (blue and bold) from a set of possible answer choices.

humans are much less likely to say it is valid (Evans et al., 1983; Evans and Perry, 1995; Klauer et al., 2000) — despite the fact that the arguments above are logically equivalent (both invalid). In short, human judgements of logical validity are biased by the believability of the conclusion.

Similarly, humans struggle to reason about rules involving abstract attributes (Wason, 1968; Johnson-Laird, 1999), but reason readily about logically-equivalent rules grounded in realistic situations (Cheng and Holyoak, 1985; Cosmides, 1989; Cosmides and Tooby, 1992). This human tendency also extends to other forms of reasoning e.g. probabilistic reasoning where humans are notably worse when problems do not reflect intuitive expectations (Cohen et al., 2017). Human reasoning is therefore not a content-independent, algebraic computation; instead, our reasoning is fundamentally entangled with our preexisting knowledge and beliefs.

Language models also have prior knowledge — expectations over the likelihood of particular sequences of tokens — that are shaped by their training. Indeed, the goal of the "pre-train and adapt" or the "foundation models" (Bommasani et al., 2021) paradigm is to endow a model with broadly accurate prior knowledge that enables learning a new task rapidly. Thus, language model predictions often *reflect* human knowledge and beliefs about the world (Trinh and Le, 2018; Petroni et al., 2019; Liu et al., 2021; Jiang et al., 2021). These predictions can percolate into the way language models answer a reasoning problem. We therefore hypothesized that language models might reflect human content effects on reasoning. That is, while past work has shown that large language models exhibit biases and imperfections on abstract reasoning tasks (e.g. Rae et al., 2021; Razeghi et al., 2022; Patel et al., 2021; Valmeekam et al., 2022), we ask a more specific question: are these biases similar to those observed in humans?

To investigate this possibility, we explored how the content of logical reasoning problems affects the performance of large language models (with 7 or 70 billion parameters, Hoffmann et al., 2022). We find that large models reproduce a variety of human content effects from the cognitive literature, across three different logical reasoning tasks. We first explore a simple Natural Language Inference (NLI) task, and show that model answers to questions are influenced by both prior knowledge as well as abstract reasoning. We then examine the more challenging task of judging whether a syllogism is a valid argument, and show that models are biased by the believability of the conclusion. We finally evaluate models on realistic and abstract/arbitrary versions of the Wason selection task (Wason, 1968), and show that models perform better with a realistic framing. These results reflect the findings in the cognitive literature.

These results have implications for both cognitive science and machine learning. From the cognitive perspective, investigating the reasoning biases of language models can offer baseline hypotheses for human reasoning biases. Our results illustrate that a unitary model can exhibit

many of the varied, context-sensitive patterns of human reasoning behavior. Relatedly, an influential line of work argues that 'illogical' human behavior is actually normative given bounded human resources (Lieder and Griffiths, 2020; Simon, 1990); but computing the utility to implement this trade-off can be impractically expensive (Horvitz et al., 2013). Our findings complement other recent works that use neural networks to implement this context-sensitive trade-off (Binz et al., 2022; Dasgupta et al., 2018). Finally, our behavioral results complement other research investigating similarities between transformer-based language models and humans at the neural processing level (Schrimpf et al., 2021; Goldstein et al., 2022; Kumar et al., 2022), and could generate exciting hypotheses for further research on the computational basis of human reasoning.

From the machine-learning perspective, building reliable AI systems requires understanding the factors that affect their behaviour. Our findings highlight the role that content plays in the reasoning processes of language models, and help to determine the situations in which LMs are likely to reason incorrectly — for example, by reverting to human-like content biases rather than following the rules of abstract logical reasoning. These reasoning errors may be especially difficult for humans to detect, because they reflect our own biases. More broadly, we demonstrate the value of using established theories and experimental paradigms from cognitive science to probe the increasingly complex behaviors of machine learning models (cf. Binz and Schulz, 2022; Ritter et al., 2017). Our work is therefore of both practical and theoretical significance to research in artificial intelligence.

2. Content effects on logical reasoning

2.1. Natural Language Inference

The first task we consider has been studied extensively in the natural language processing literature (MacCartney and Manning, 2007). In the classic natural language inference problem, a model receives two sentences, a 'premise' and a 'hypothesis', and has to classify them based on whether the hypothesis 'entails', 'contradicts', or 'is neutral to' the premise. Traditional datasets for this task were crowdsourced (Bowman et al., 2015) leading to sentence pairs that don't strictly follow logical definitions of entailment and contradiction. To make this a more strictly logical task, we follow Dasgupta et al. (2018) to generate comparisons (e.g. `Premise: X is smaller than Y, Hypothesis: Y is bigger than X` is an entailment).

Content effects are generally more pronounced in difficult tasks that require extensive logical reasoning, and are stronger in children or adults under cognitive load (Evans, 1989; Evans and Perry, 1995). The relatively simple logical reasoning involved in this task means that adult humans would likely exhibit high performance. Perhaps for this reason, content effects on human reasoning have not been examined on this task. We therefore consider two more challenging tasks where human content effects have directly been studied.

2.2. Syllogisms

Syllogisms (Smith, 2020) are multi-step reasoning problems with a simple argument form in which two true statements necessarily imply a third. For example, the statements "All humans are mortal" and "Socrates is a human" together imply that "Socrates is mortal". But human syllogistic reasoning is not purely abstract and logical; instead it is affected by our prior beliefs about the contents of the argument (Evans et al., 1983; Klauer et al., 2000; Tessler et al., 2022).

For example, Evans et al. (1983) showed that if participants were asked to judge whether a syllogism was logically valid or invalid, they were biased by whether the conclusion was consistent with their beliefs. Participants were much more likely (90% of the time) to mistakenly say an invalid syllogism was valid if the conclusion was believable, and thus mostly relied on belief rather than abstract reasoning. Participants would also sometimes say that a valid syllogism was invalid if the conclusion was not believable, but this effect was somewhat weaker (but cf. Dube et al., 2010). These "belief-bias" effects have been repli-

cated and extended in various subsequent studies (Klauer et al., 2000; Dube et al., 2010; Trippas et al., 2014; Tessler, 2015).

2.3. The Wason Selection Task

The Wason Selection Task (Wason, 1968) is a logic problem that can be challenging even for subjects with substantial education in mathematics or philosophy. Participants are shown four cards, and told a rule such as: "if a card has a 'D' on one side, then it has a '3' on the other side." The four cards respectively show 'D', 'F', '3', and '7'. The participants are then asked which cards they need to flip over to check if the rule is true or false.

The correct answer is to flip over the cards showing 'D' and '7'. However, Wason (1968) showed that while most participants correctly chose 'D', they were much more likely to choose '3' than '7'. That is, the participants should check the *contrapositive* of the rule ("not 3 implies not D", which is logically implied), but instead they confuse it with the *converse* ("3 implies D", which is not logically implied). This is a classic task in which reasoning according to the rules of formal logic does not come naturally for humans, and thus there is potential for prior beliefs and knowledge to affect reasoning.

Indeed, the difficulty of the Wason task depends substantially on the content of the problem. If an identical logical structure is instantiated in a common situation, such as a social rule, participants are much more accurate (Wason and Shapiro, 1971; Cheng and Holyoak, 1985; Cosmides, 1989; Cosmides and Tooby, 1992). For example, if the cards show ages and beverages, and the rule is "if they are drinking alcohol, then they must be 21 or older" and shown cards with 'beer', 'soda', '25', '16', the vast majority of participants correctly choose to check the cards showing 'beer' and '16'.

These biases are also affected by background knowledge; education in mathematics appears to be associated with improved reasoning in abstract Wason tasks (Inglis and Simpson, 2004; Cresswell and Speelman, 2020). However, even those experienced participants were far from perfect — undergraduate mathematics majors and academic mathematicians achieved less than 50% accuracy at the Wason task (ibid). This illustrates the challenge of abstract logical reasoning, even for humans who are clearly capable of executing equivalent reasoning it less abstract contexts.

3. Methods

Creating datasets While many of these tasks have been extensively studied in cognitive science, the stimuli used in cognitive experiments are often online, and thus may be present in the training data of large language models, which could compromise results (e.g. Emami et al., 2020; Dodge et al., 2021). To reduce these concerns, we generate new datasets, by following the approaches used in prior work. We briefly outline this process here; see Appx. A.1 for full details.

For each of the three tasks above, we generate multiple versions of the task stimuli. Throughout, the logical structure of the stimuli remains fixed, we simply manipulate the entities over which this logic operates (Fig. 2). We generate propositions that are:
Consistent with human beliefs and knowledge (e.g. ants are smaller than whales).
Violate beliefs by inverting the consistent statements (e.g. whales are smaller than ants).
Nonsense tasks about which the model should not have strong beliefs, by swapping the entities out for nonsense words (e.g. kleegs are smaller than feps).

For the Wason tasks, we slightly alter our approach to fit the different character of the tasks. We generate questions with:
Realistic rules involving plausible relationships (e.g. "if the passengers are traveling outside the US, then they must have shown a passport").
Arbitrary rules (e.g. "if the cards have a plural word, then they have a positive emotion").
Nonsense rules relating nonsense words ("if the cards have more bem, then they have less stope").
However, there are other ways that a rule can be unrealistic. The component propositions can be realistic even if the relationship between them is not. We therefore generate two control variations on realistic rules:
Shuffled realistic rules, which combine realistic

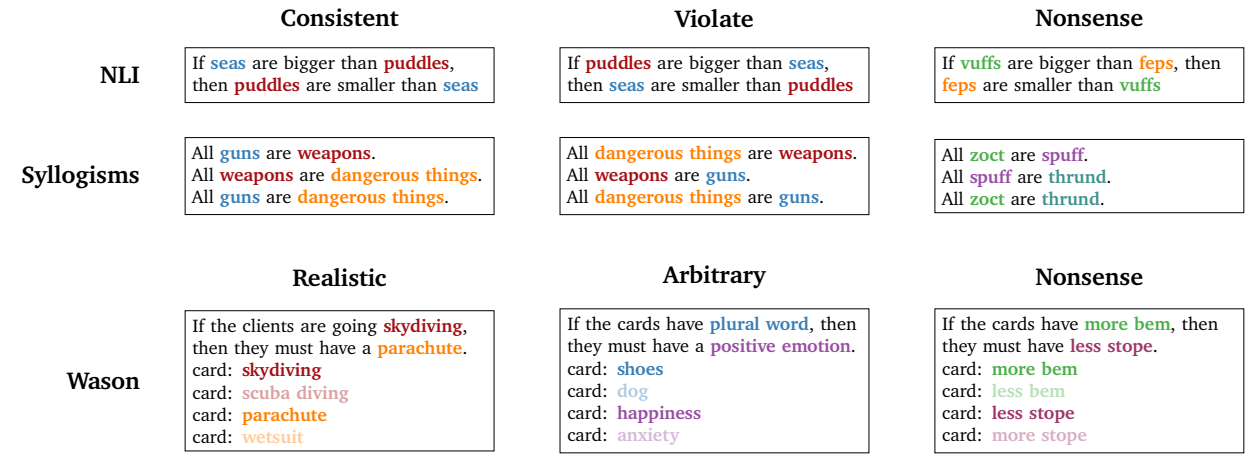


Figure 2 | Manipulating content within fixed logical structures. In each of our three datasets (rows), we instantiate different versions of the logical problems (columns). Different versions of a problem have the same logical structure, but instantiated with different entities or relationships between those entities.

components in nonsensical ways (e.g. "if the passengers are traveling outside the US, then they must have received an MD").
Violate realistic rules, which directly violate the expected relationship (e.g. "if the passengers are flying outside the US, then they must have shown a drivers license [not a passport]").

Models & evaluation We evaluate one language model — Chinchilla (Hoffmann et al., 2022) — across all datasets and conditions. This large model (with 70 billion parameters) has been previously demonstrated to achieve some non-trivial performance across various logical reasoning tasks, and is therefore a natural candidate for investigating how reasoning is affected by beliefs in a large model; by contrast, smaller models would likely perform poorly across all conditions. However, for the NLI task, which is substantially easier, we also consider the performance of a model an order of magnitude smaller — the 7 billion parameter version of the same model. This comparison allows us to evaluate how scale interacts with content to affect performance.

For each task, we assess the model by evaluating the likelihood of completing the question with each of a set of possible answers. These answers are evaluated independently; the model does not see all answers at once when choosing. We apply the DC-PMI correction proposed by Holtzman et al. (2021) — i.e., we measure the change in likelihood of each answer in the context of the question relative to a baseline context, and choosing the answer that has the largest increase in likelihood in context. This scoring approach is intended to reduce the possibility that the model would simply phrase the answer differently than the available choices; for example, answering "this is not a valid argument" rather than "this argument is invalid". For the NLI task, however, the direct answer format means that the DC-PMI correction would control for the very bias we are trying to measure. We therefore choose the answer that receives the maximum likelihood among the set of possible answers, but report the DC-PMI results in Appx. B.2. We also report syllogism and Wason results with maximum likelihood scoring in Appx. B.7; the model reproduces the qualitative effects with either scoring method.

When we present a few-shot prompt of examples of the task to the model, the examples are presented with correct answers, and each example (as well as the final probe) is separated from the previous example by a single blank line.

4. Results

In presenting the results, we analyze performance using hierarchical logistic regressions that account for the dependency structure of the data; see Appx. C for full regression specifications and results.

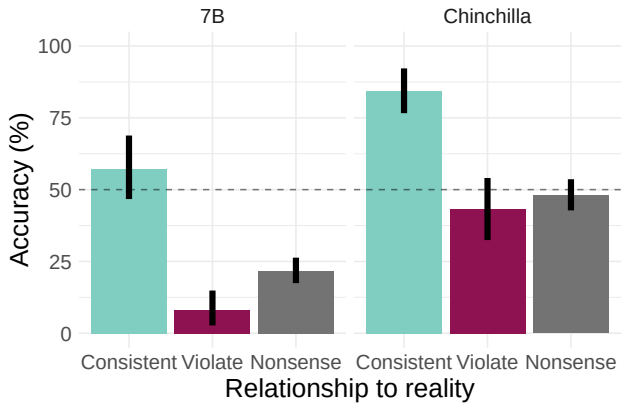


Figure 3 | NLI results zero-shot, for a smaller (7B) and larger (Chinchilla) language model. Both models exhibits a substantial belief bias — their accuracy is significantly higher for conclusions that are consistent with their expectations. (Throughout, errorbars are bootstrap 95%-CIs, and dashed lines are chance performance.)

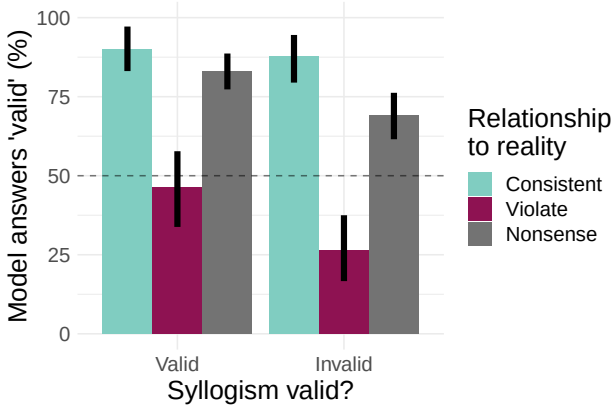


Figure 4 | Syllogism validity judgements zero-shot (for Chinchilla). The model exhibits substantial belief bias — it is strongly biased towards saying an argument is valid if the conclusion is consistent with expectations or the argument contains nonsense words, but strongly biased towards saying the argument is invalid if the conclusion violates expectations. (Note that this figure plots the proportion of the time the model answers 'valid' rather than raw accuracy, to more clearly illustrate the bias. To see accuracy, simply reverse the vertical axis for the invalid arguments.)

4.1. Zero shot performance

We first examine the language models' ability to solve logical reasoning tasks zero-shot — that is, without any solved examples in the prompt.

Natural Language Inference We evaluate two different language models — Chinchilla and 7B (Fig. 3). We find that performance is at or below chance[1] for the 'violate' or 'nonsense' types, but is significantly higher for 'consistent' types than the other conditions ($z = 5.868$, $p = 4.4 \cdot 10^{-9}$). Only the largest model exhibits above chance performance in any condition, and only on the 'consistent' condition. This indicates a strong content bias: the models prefer to complete the sentence in a way consistent with prior expectations rather than in a way consistent with the rules of logic. We find similar results from a continuous analysis, where we combined the consistent and violate categories, and instead estimated "consistency" directly from the model's unconditional likelihood of the hypothesis (App B.1). Following findings from past literature (Kojima et al., 2022, e.g.), we also examine the effect of different prompt type that might elicit more logical reasoning. We find that some prompts can in fact significantly boost performance in the 'violate' condition (see Appx. B.3), although performance on the 'consistent' condition remains better. Further, the language model always perform poorly on the nonsense versions of the task, indicating clear content effects.

As noted above, in this task prior expectations are embedded in the actual answers. Are LMs 'reasoning' *per se*, or are these effects driven solely by the probability of the hypotheses? We run two controls to disentangle these possibilities. First, we apply the DC-PMI correction with respect to a mismatched control (details in Appendix B.2). We find significantly greater than chance performance across conditions even with the correction, indicating that the model is sensitive to logical validity. Second, we evaluate a more standard NLI format, where both the premise and hypothesis are provided and the model simply has to say whether the inference is valid or invalid. In this case, the model should not have a strong prior over the responses (and we also apply a DC-PMI correction). We find the same pattern of effects in this setting, where performance in the consistent

[1] LMs naive preference towards repeating words in the prompt may contribute to this effect.

condition is higher than in the nonsense and violate conditions (Appx. B.2). Thus, language models can be sensitive to the logical structure of natural language inferences, as well as the content.

While human belief bias has not been investigated in this particular task, this pattern of behavior is qualitatively similar to humans' belief bias effects observed in other tasks, such as the syllogisms task which we investigate next.

Syllogisms For syllogisms, belief bias drives almost all zero-shot decisions (Fig. 4). In particular, the model responds that an argument is valid in most cases where the conclusion is consistent with reality, or if the argument uses novel (nonsense) words. However, if the conclusion is inconsistent with reality, the model is strongly biased toward saying the argument is invalid, regardless of its actual logical validity. The model does consider both consistency and validity to some extent — there is a significant effect of both belief-consistency ($z = -8.6$, $p < 2 \cdot 10^{-16}$) and logical validity ($z = -3.3$, $p = 8.5 \cdot 10^{-4}$). (These effects are not substantially affected by different task instructions, see Appx. B.3.)

These results closely reflect the pattern of human results observed by Evans et al. (1983). Human subjects endorsed arguments with a believable conclusion around 90% of the time, regardless of their actual validity, as does the model. Humans generally rejected arguments with an unbelievable conclusion, but were more sensitive to logical validity in this case; the model shows a qualitatively similar pattern (although the interaction does not rise to statistical significance). Like humans, the model's responses seem to be determined primarily by believability, with secondary effects of logical validity.

The above results should *not* be interpreted to mean that the model is unable to perform syllogistic reasoning zero-shot. Above, we focus on the paradigm that has generally been used in cognitive work on content effects — presenting a complete syllogism and asking whether that syllogism is valid. However, if we instead ask the model to *complete* a syllogism by choosing a valid conclusion to an argument, from among all possible conclusions (like the NLI tasks above), the model

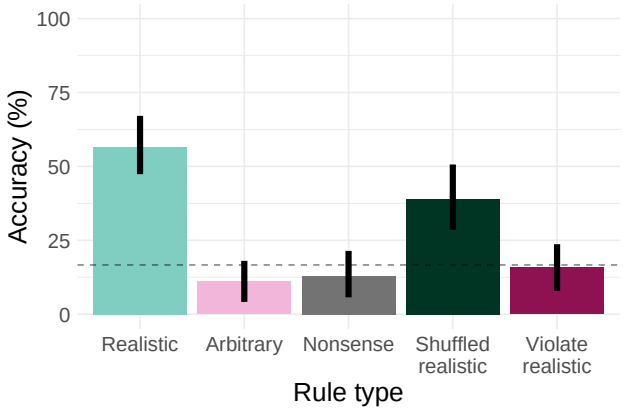


Figure 5 | Wason selection task results zero-shot (for Chinchilla). While the model performs poorly on arbitrary or nonsense rules, it performs substantially above chance for realistic rules. For shuffled realistic rules, it exhibits performance above chance, but lower than that for realistic rules. (Chance is a random choice of two cards among the four shown.)

achieves 70-80% accuracy (chance is 8%) across conditions, with only minimal differences in accuracy depending on whether the conclusions are consistent with reality (Appx. B.6). Intriguingly, humans are also much less biased when comparing arguments and deciding which is valid than when making forced-choice assessments of whether a single syllogism is valid (Trippas et al., 2014; but cf. Johnson-Laird, 1999). Thus, the particular instantiation of the problem can substantially alter the accuracy and biases of models, just as it can for humans.

Wason For Wason tasks (Fig. 5), the model exhibits chance-level accuracy on arbitrary and nonsense rules zero-shot, but fairly high accuracy on realistic ones ($z = 3.3$, $p = 1.0 \cdot 10^{-3}$). This reflects findings in the human literature: humans are much more accurate at answering the Wason task when it is framed in terms of realistic situations than arbitrary rules about abstract attributes. To verify that these effects are due to familiarity with the rule, and not due to the difficulty of evaluating their component propositions, we compute forced choice evaluation of the base propositions (Appx. B.4) and found no significant differences.

As noted above, the Wason task rules can be

realistic or unrealistic in multiple ways. We therefore also compared shuffled realistic rules and violate realistic rules. For shuffled rules, results are well above chance, but lower than realistic rules. For violate rules, by contrast, performance is at chance. It appears that the model reasons more accurately about rules formed from realistic propositions, particularly if the relationships between propositions in the rule are also realistic, but even to some degree if they are shuffled in nonsensical ways that do not directly violate expectations. However, if the rules strongly violate beliefs, performance is low.

4.2. Few-shot training on the logic task

Logical reasoning is non-trivial and previous work has shown significant improvement from few-shot prompting (Brown et al., 2020), to help the model understand the logic task at hand. In this section, we first examine the effect of communicating the logic task explicitly using nonsense examples, over which the model should not have prior beliefs. We evaluate on probe task performance with objects over which the model does have prior beliefs. We then explore whether models learn more from different types of examples in the prompt.

Natural language inference Few-shot training on the purely logical task with nonsense entities improves performance in all cases, even on tasks with entities that are consistent with or violate expectations. All conditions reach comparable performance with five shots. Directly communicating the logical task with nonsense examples can wash out the content effect on the probe task (at least for the larger model), by encouraging the LM to focus on the logical structure rather than prior expectations. We did not find substantial differences if we used other entity types in the few-shot prompts (Appx. B.8).

Syllogisms Few-shot examples somewhat improve the performance of the model on the syllogisms task (Fig. 7).[2] With examples, the models are somewhat better calibrated than zero-shot (number of shots by validity interaction $z = -2.9$, $p = 3.5 \cdot 10^{-3}$) — responding "valid" more frequently to valid than invalid syllogisms in every condition — but still exhibit notable belief bias effects. Again, we observed similar benefits with other types of entities in the prompt (Appx. B.8).

[2]Note that we use 2-shot evaluation for these tasks rather than 1-shot, to avoid biasing the models' answers strongly towards repeating the same answer to the next question.

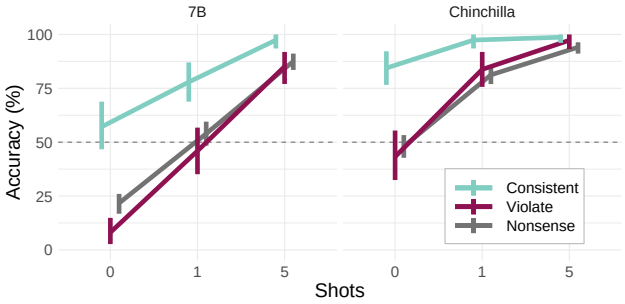


Figure 6 | NLI with 0, 1, or 5 prompt examples using nonsense entities. With one nonsense example in the prompt, performance for all entity types go up significantly for the larger model, and belief bias effects reduce for both models. With five examples, only the smaller model shows a significant content effect, and performance for the larger model is near ceiling.

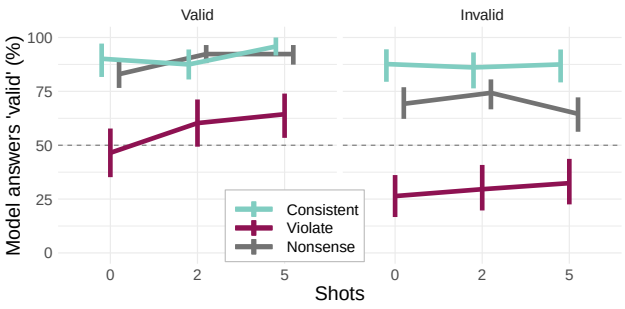


Figure 7 | Syllogism results few-shot, with nonsense prompt examples. With examples, responses are slightly better calibrated, and exhibit reduced — but still substantial — belief bias effects.

Wason While we used nonsense examples for the previous two tasks, we find that few-shot prompting with nonsense examples results in only weak improvements on the Wason tasks (Fig. 8). Instead, we find that the realistic examples were the most effective at conveying the task across conditions (all $z \in [-4.1, -2.2]$, $p \in [4.8 \cdot 10^{-5}, 0.026]$); qualitatively, while other types of prompt examples result in comparable performance with certain types of probes, each

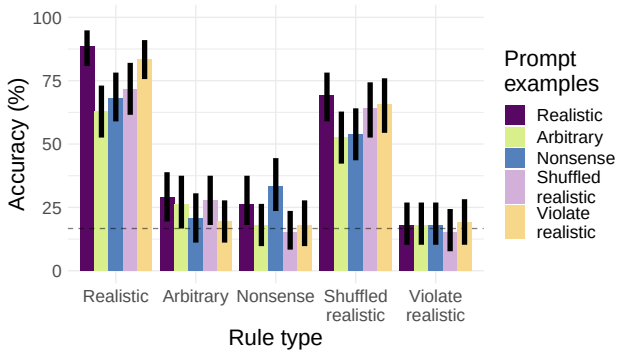


Figure 8 | Wason results 5-shot, with different types of prompt examples. Realistic prompt examples appear to be most beneficial overall — especially for realistic and shuffled realistic probes. Other types of prompts are generally helpful in a more limited set of conditions; there may be an overall benefit to prompts matching probes.

other condition is worse than realistic examples in some probe conditions. (Fig. 8). Examples that are realistic might be easier to understand, and thereby make the problem more accessible; similar arguments have been used in favor of beginning with concrete examples when teaching abstract mathematical concepts to humans (Fyfe et al., 2014). Investigating similar strategies with language models could be an exciting future direction.

However, regardless of example type, performance on the arbitrary and nonsense probes remains low. Similarly, Wason (1968) found that it was hard to improve human response patterns on the abstract task, even after explicitly showing them which examples can falsify a rule. Likewise, as noted above, even mathematics undergraduates can struggle with the abstract task (Inglis and Simpson, 2004; Cresswell and Speelman, 2020). Language models reflect human failures on the abstract selection task.

4.3. Details of Wason response patterns

Because each answer to the Wason problems selects a pair of cards, we further analyzed the individual cards chosen. The card choices in each problem are designed so that two cards respectively match and violate the antecedent, and similarly for the consequent. In Fig. 9 we show the proportion of choices the model is making of each card type. The correct answer is to choose the card that matches the antecedent, and the card that violates the consequent.

While the model produces substantially fewer errors 5-shot — especially in the realistic condition — the patterns of errors are similar 0-shot and 5-shot. The model matches most humans Wason (1968) in consistently choosing one antecedent and one consequent card most of the time, across conditions (Appx. B.5). However, the humans and the model both make errors in their choices within each category. For humans, most errors correspond to choosing the card which matches the consequent. However, the model's errors are relatively more evenly distributed between the antecedent and consequent in most conditions. Thus, the model appears *less likely* than humans to prefer superficial matching to the rule (although human error patterns depend on education; Inglis and Simpson, 2004; Cresswell and Speelman, 2020).

The violate realistic condition results in a unique pattern of errors: few errors in the antecedent, but frequent errors in the consequent. This pattern corresponds to the correct answers for the more-believable, realistic rule — the violate rules reverse which consequent answer is correct, but the model appears to be giving similar consequent answers to realistic or violate conditions. These results suggest that the model is ignoring the subtle change to the rule. It would be interesting in future work to evaluate whether humans exhibit similar answers to violate rules — e.g., because they read quickly and assume the rule will fit prior expectations, or because they assume the change is a typo or mistake.

5. Discussion

Humans are imperfect reasoners. We reason most effectively about entities and situations that are consistent with our understanding of the world. Our experiments show that language models mirror these patterns of behavior. Language models perform imperfectly on logical reasoning tasks, but this performance depends on content and context. Most notably, such models often fail in situa-

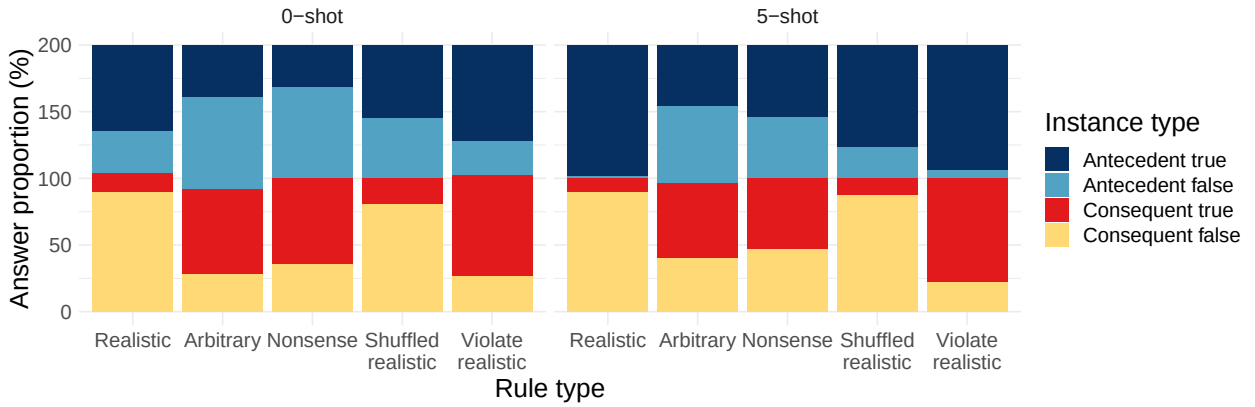


Figure 9 | Answer patterns for the Wason tasks, broken down by answer component. Correct choices are antecedent true and consequent false. Model errors are distributed relatively equally between choosing the antecedent-false instance and choosing the consequent-true instance in most cases. However, the violate realistic rules have a high error rate for the consequent — i.e., the model is effectively ignoring the belief violations and answering as though the more realistic form of the rule held. The error pattern is similar 0-shot and 5-shot, although 5-shot has substantially fewer errors, especially for realistic rules. (Answer percentages sum to 200% because all answers include two instances; 5-shot results use examples of the same rule type as the probe.)

tions where humans fail — when stimuli become too abstract or conflict with prior understanding of the world.

Beyond these parallels, we also observed reasoning effects in language models that to our knowledge have not been previously investigated in the human literature. For example, the patterns of errors on the 'violate realistic' rules, or the relative ease of 'shuffled realistic' rules in the Wason tasks. Likewise, language model performance on the Wason tasks increases most when they are demonstrated with realistic examples; benefits of concrete examples have been found in cognitive and educational contexts (Sweller et al., 1998; Fyfe et al., 2014), but remain to be explored in the Wason problems. Investigating whether humans show similar effects is a promising direction for future research.

Prior research on language model reasoning. Since Brown et al. (2020) showed that large language models could perform moderately well on some reasoning tasks, there has been a growing interest in language model reasoning (Binz and Schulz, 2022). Typical methods focus on prompting for sequential reasoning (Nye et al., 2021a; Wei et al., 2022b; Kojima et al., 2022), altering task framing (Khashabi et al., 2022; Lampinen et al., 2022) or iteratively sampling answers (Wang et al., 2022).

In response, some researchers have questioned whether these language model abilities qualify as "reasoning". The fact that language models sometimes rely on "simple heuristics" (Patel et al., 2021), or reason more accurately about frequently-occurring numbers (Razeghi et al., 2022), have been cited to "rais[e] questions on the extent to which these models are *actually reasoning*" (ibid, emphasis ours). The implicit assumption in these critiques is that reasoning should be a purely algebraic, syntactic computations over symbols from which "all meaning had been purged" (Newell, 1980; cf. Marcus, 2003). In this work, we emphasize how *both* humans and language models rely on content when reasoning — using simple heuristics in some contexts, and reasoning more accurately about frequently-occurring situations (Mercier and Sperber, 2017; Dasgupta et al., 2020). Thus, abstract reasoning may be a graded, content-sensitive capacity in both humans and models.

Dual systems? The idea that humans possess dual reasoning systems — an implicit, intuitive

system "system 1", and an explicit reasoning "system 2' — was motivated in large part by belief bias and Wason task effects (Evans, 1984, 2003; Oaksford and Chater, 2003). The dual system idea has more recently become popular (Kahneman, 2011; Evans and Over, 2013), including in machine learning (e.g. Bengio, 2017). It is often claimed that current ML (including large language models) behave like system 1, and that we need to augment this with a classically-symbolic process to get system 2 behaviour (e.g. Nye et al., 2021b).

Our results show that a unitary system — a large transformer language model — can mirror this dual behavior in humans, demonstrating both biased and consistent reasoning depending on the context. In the NLI tasks, a few examples takes Chinchilla from highly content-biased performance to near ceiling performance, and even a simple instructional prompt can substantially reduce bias. These findings integrate with prior works showing that language models can be prompted to exhibit sequential reasoning, and thereby improve their performance in domains like mathematics (Nye et al., 2021a; Wei et al., 2022b; Kojima et al., 2022).

These observations suggest a unitary model of reasoning in language models, where a single system deploys a context-dependent attention mechanism to arbitrate between conflicting responses. Such models have been developed in the literature on human cognitive control (Cohen et al., 1990; Botvinick and Cohen, 2014), and have been suggested as central processes underlying the variations in human reasoning (Duncan et al., 2020; Li and McClelland, 2022). An exciting future direction is examining whether the internal computations of LMs on these tasks can be captured by such models. Further, from the unitary perspective, augmenting language models with a second system might not be necessary. Instead, like humans, these models could be further developed towards abstract reasoning by altering context and training, as we discuss below.

Neural mechanisms of human reasoning. Deep learning models are increasingly used as models of neural processing in biological systems (e.g. Yamins et al., 2014; Yamins and DiCarlo, 2016), as they often develop similar patterns of representation. These findings have led to proposals that deep learning models capture *mechanistic* details of neural processing at an appropriate level of description (Cao and Yamins, 2021a,b), despite the fact that aspects of their information processing clearly differ from biological systems. More recently, large language models have been similarly shown to accurately predict neural representations in the human language system — large language models "predict nearly 100% of the explainable variance in neural responses to sentences" (Schrimpf et al., 2021; see also Kumar et al., 2022; Goldstein et al., 2022). Language models also predict low-level behavioral phenomena; e.g. surprisal predicts reading time (Wilcox et al., 2020). In the context of these works, our observation of behavioral similarities in reasoning patterns between humans and language models raise important questions about possible similarities of the underlying reasoning processes between humans and language models, and the extent of overlap between neural mechanisms for language and reasoning in humans.

Towards a normative account of content effects? Various accounts of human cognitive biases frame them as 'normative' according to some objective. Some explain biases as the application of processes — such as information gathering or pragmatics — that are broadly rational under a different model of the world (e.g. Oaksford and Chater, 2003; Tessler et al., 2022). Others interpret them as a rational adaptation to reasoning under constraints such as limited memory or time (e.g. Lieder and Griffiths, 2020; Gershman et al., 2015; Simon, 1990) — where content effects actually support fast and effective reasoning in commonly encountered tasks (Mercier and Sperber, 2017; Dasgupta et al., 2020). Our results show that content effects can emerge from simply training a large transformer to imitate language produced by human culture, without incorporating these human-specific internal mechanisms. In other words, language models and humans both arrive at these content biases — but from seemingly very different architectures, experiences, and training objectives. A promising direction for future enquiry would be to causally manipulate features of the training objective and experience, to explore which features contribute

to the emergence of content biases in language models, and investigate whether these features could offer insight into the origins of human patterns of reasoning.

Why might model reasoning patterns differ from human ones? The language model reasoning patterns do not perfectly match all aspects of the human data. For example, the error patterns on the Wason tasks are somewhat different than those observed in humans (although human errors depend on education; Inglis and Simpson, 2004; Cresswell and Speelman, 2020). Similarly, the model does not show the significant interaction between believability and validity on the syllogism tasks that humans do (Evans et al., 1983), although the pattern is qualitatively in the same direction (and this interaction may be an artifact; Dube et al., 2010). Various factors could contribute to differences between model and human behaviors.

First, it is difficult to know how to prompt a language model in order to evaluate a particular task. Language model training blends many tasks into a homogeneous soup, which makes controlling the model difficult. For example, presenting task instructions might not actually lead to better performance (cf. Webson and Pavlick, 2021); indeed, instructions did not substantially affect performance on the harder syllogisms task (Appx. B.3). Similarly, presenting negative examples can help humans learn, but is generally detrimental to model performance (e.g. Mishra et al., 2021) — presumably because the model infers that the task is to sometimes output wrong answers. It is possible that idiosyncratic details of our task framing may have caused the model to infer the task incorrectly. However, where we varied these details we generally did not observe substantial differences.

More fundamentally, language models do not directly experience the situations to which language refers (McClelland et al., 2020); grounded experience (for instance the capacity to simulate the physical turning of cards on a table) presumably underpins some human beliefs and reasoning. Furthermore, humans sometimes use physical or motor processes such as gesture to support logical reasoning (Alibali et al., 2014; Nathan et al., 2020). Finally, language models experience language passively, while humans experience language as an active, conventional system for social communication (e.g. Clark, 1996); active participation may be key to understanding meaning as humans do (Santoro et al., 2021; Schlangen, 2022). Some differences between language models and humans may therefore stem from differences between the rich, grounded, interactive experience of humans and the impoverished experience of the models.

How can we achieve more abstract, context-independent reasoning? If language models exhibit some of the same reasoning biases as humansm could some of the factors that reduce content dependency in human reasoning be applied to make these models less content-dependent? In humans, formal education is associated with an improved ability to reason logically and consistently (Luria, 1971; Lehman and Nisbett, 1990; Attridge et al., 2016; Inglis and Simpson, 2004; Cresswell and Speelman, 2020; Nam and McClelland, 2021).[3] Could language models learn to reason more reliably with targeted formal education?

Several recent results indicate that this may not be as far-fetched as it sounds. Pretraining on synthetic logical reasoning tasks can improve model performance on reasoning and mathematics (Clark et al., 2020; Wu et al., 2021). More broadly, language models can be trained or tuned to better follow instructions (Wei et al., 2021; Ouyang et al., 2022; Gupta et al., 2022). In some cases language models can either be prompted or can learn to verify, correct, or debias their own outputs (Schick et al., 2021; Cobbe et al., 2021; Saunders et al., 2022; Kadavath et al., 2022). Finally, language model reasoning can be bootstrapped through iterated fine-tuning on successful instances (Zelikman et al., 2022). These results suggest the possibility that a model trained with instructions to perform logical reasoning, and to check and correct the results of its work, might move closer to the logical reasoning capabilities of formally-educated humans. Perhaps logical reasoning is a graded competency that is supported by a range of different environmen-

[3] Causal evidence is scarce, because years of education are difficult to experimentally manipulate.

tal and educational factors (Santoro et al., 2021; Wang, 2021), rather than a core ability that must be built in to an intelligent system.

Acknowledgements

We thank Michiel Bakker, Michael Henry Tessler, and Adam Santoro for helpful comments and suggestions.

Author contributions

ID initiated the investigation into interactions between language model knowledge and reasoning, formulated and lead the project, performed the experiments for the NLI tasks, and wrote the paper.
AKL lead the project, developed the syllogisms and Wason datasets, performed the experiments for the syllogisms and Wason tasks, performed the main analyses and created the figures, and wrote the paper.
SCYC created the NLI dataset, contributed ideas, contributed to experiments, and contributed to writing the paper.
TC contributed ideas, created technical infrastructure, and contributed to writing the paper.
DK, JLM and **FH** advised throughout and contributed to writing the paper.

References

M. W. Alibali, R. Boncoddo, and A. B. Hostetter. Gesture in reasoning: An embodied perspective. *The Routledge handbook of embodied cognition*, page 150, 2014. 5

N. Attridge, A. Aberdein, and M. Inglis. Does studying logic improve logical reasoning? 2016. 5

Y. Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017. 5

M. Binz and E. Schulz. Using cognitive psychology to understand gpt-3, Jun 2022. URL psyarxiv.com/6dfgk. 1, 5

M. Binz, S. J. Gershman, E. Schulz, and D. Endres. Heuristics from bounded meta-learned inference. *Psychological review*, 2022. 1

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

M. M. Botvinick and J. D. Cohen. The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive science*, 38(6):1249–1285, 2014. 5

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 2.1

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 4.2, 5

R. Cao and D. Yamins. Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*, 2021a. 5

R. Cao and D. Yamins. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021b. 5

P. W. Cheng and K. J. Holyoak. Pragmatic reasoning schemas. *Cognitive psychology*, 17(4): 391–416, 1985. 1, 2.3

H. H. Clark. *Using language*. Cambridge university press, 1996. 5

P. Clark, O. Tafjord, and K. Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020. 5

K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 5

A. L. Cohen, S. Sidlowski, and A. Staub. Beliefs and bayesian reasoning. *Psychonomic Bulletin & Review*, 24(3):972–978, 2017. 1

J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, 97(3):332, 1990. 5

L. Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276, 1989. 1, 2.3

L. Cosmides and J. Tooby. Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, 163:163–228, 1992. 1, 2.3

C. Cresswell and C. P. Speelman. Does mathematics training lead to better logical thinking and reasoning? a cross-sectional assessment from students to professors. *PLOS ONE*, 15(7):1–21, 07 2020. doi: 10.1371/journal.pone.0236153. URL https://doi.org/10.1371/journal.pone.0236153. 2.3, 4.2, 4.3, 5

I. Dasgupta, D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*, 2018. 1, 2.1

I. Dasgupta, E. Schulz, J. B. Tenenbaum, and S. J. Gershman. A theory of learning to infer. *Psychological review*, 127(3):412, 2020. 5

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98. 3

C. Dube, C. M. Rotello, and E. Heit. Assessing the belief bias effect with rocs: it's a response bias effect. *Psychological review*, 117(3):831, 2010. 2.2, 5

J. Duncan, M. Assem, and S. Shashidhara. Integrated intelligence from distributed brain activity. *Trends in Cognitive Sciences*, 24(10):838–852, 2020. 5

A. Emami, A. Trischler, K. Suleman, and J. C. K. Cheung. An analysis of dataset overlap on winograd-style tasks. *arXiv preprint arXiv:2011.04767*, 2020. 3

J. Evans, J. L. Barston, and P. Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306, 1983. 1, 2.2, 4.1, 5, A.1.2

J. S. B. Evans. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468, 1984. 5

J. S. B. Evans. *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc, 1989. 1, 2.1

J. S. B. Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459, 2003. 5

J. S. B. Evans and D. E. Over. *Rationality and reasoning*. Psychology Press, 2013. 5

J. S. B. Evans and T. S. Perry. Belief bias in children's reasoning. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1995. 1, 2.1

J. S. B. Evans, J. Clibbens, and B. Rood. The role of implicit and explicit negation in conditional reasoning bias. *Journal of Memory and Language*, 35(3):392–409, 1996. A.1.2

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

E. R. Fyfe, N. M. McNeil, J. Y. Son, and R. L. Goldstone. Concreteness fading in mathematics and science instruction: A systematic review. *Educational psychology review*, 26(1):9–25, 2014. 4.2, 5

D. Ganguli, D. Hernandez, L. Lovitt, N. DasSarma, T. Henighan, A. Jones, N. Joseph, J. Kernion, B. Mann, A. Askell, et al. Predictability and surprise in large generative models. *arXiv preprint arXiv:2202.07785*, 2022. 1

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006. C

S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015. 5

G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62 (1):451–482, 2011. 1

A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022. 1, 5

P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. P. Bigham. Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*, 2022. 5

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. (document), 1, 3

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. A.1.1

A. Holtzman, P. West, V. Shwartz, Y. Choi, and L. Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*, 2021. 3, A.2, B.6, B.7

E. J. Horvitz, J. Suermondt, and G. F. Cooper. Bounded conditioning: Flexible inference for decisions under scarce resources. *arXiv preprint arXiv:1304.1512*, 2013. 1

A. Hosseini, S. Reddy, D. Bahdanau, R. D. Hjelm, A. Sordoni, and A. Courville. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.102. URL https://aclanthology.org/2021.naacl-main.102. A.1.2, A.1.3

M. Inglis and A. Simpson. Mathematicians and the selection task. *International Group for the Psychology of Mathematics Education*, 2004. 2.3, 4.2, 4.3, 5

Z. Jiang, J. Araki, H. Ding, and G. Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. 1

P. N. Johnson-Laird. Deductive reasoning. *Annual review of psychology*, 50(1):109–135, 1999. 1, 4.1

P. N. Johnson-Laird, P. Legrenzi, and M. S. Legrenzi. Reasoning and a sense of reality. *British journal of Psychology*, 63(3):395–400, 1972. 1

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. H. Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221. 5

D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011. 5

D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982. 1

D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi. Reframing instructional prompts to gptk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, 2022. 5

K. C. Klauer, J. Musch, and B. Naumer. On belief bias in syllogistic reasoning. *Psychological review*, 107(4):852, 2000. 1, 2.2

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. 1, 4.1, 5, B.3.1

S. Kumar, T. R. Sumers, T. Yamakoshi, A. Goldstein, U. Hasson, K. A. Norman, T. L. Griffiths, R. D. Hawkins, and S. A. Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*, 2022. 1, 5

A. K. Lampinen, I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, and F. Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022. 5, C

D. R. Lehman and R. E. Nisbett. A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, 26(6): 952, 1990. 5

Y. Li and J. L. McClelland. A weighted constraint satisfaction approach to human goal-directed decision making. *PLOS Computational Biology*, 18(6):e1009553, 2022. 5

F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43, 2020. 1, 5

L. Z. Liu, Y. Wang, J. Kasai, H. Hajishirzi, and N. A. Smith. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*, 2021. 1

A. K. Luria. Towards the problem of the historical nature of psychological processes. *International Journal of Psychology*, 6(4):259–272, 1971. 5

B. MacCartney and C. D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, 2007. 2.1

G. Marcus. *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt, 2009. 1

G. Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020. 1

G. F. Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2003. 5

J. L. McClelland, F. Hill, M. Rudolph, J. Baldridge, and H. Schütze. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974, 2020. 5

H. Mercier and D. Sperber. *The Enigma of Reason*. Harvard University Press, 2017. 5

S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 5

M. Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021. 1

A. J. Nam and J. L. McClelland. What underlies rapid learning and systematic generalization in humans. *arXiv preprint arXiv:2107.06994*, 2021. 5

M. J. Nathan, K. E. Schenck, R. Vinsonhaler, J. E. Michaelis, M. I. Swart, and C. Walkington. Embodied geometric reasoning: Dynamic gestures during intuition, insight, and proof. *Journal of Educational Psychology*, 2020. 5

A. Newell. Physical symbol systems. *Cognitive science*, 4(2):135–183, 1980. 1, 5

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021a. 1, 5

M. Nye, M. Tessler, J. Tenenbaum, and B. M. Lake. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204, 2021b. 5

M. Oaksford and N. Chater. Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, 10(2):289–318, 2003. 5

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 5

A. Patel, S. Bhattamishra, and N. Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021. 1, 5

F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019. 1

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 1, A.1.1

Y. Razeghi, R. L. Logan IV, M. Gardner, and S. Singh. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*, 2022. 1, 5

S. Ritter, D. G. Barrett, A. Santoro, and M. M. Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR, 2017. 1

J. Russin, R. C. O'Reilly, and Y. Bengio. Deep learning needs a prefrontal cortex. *Work Bridging AI Cogn Sci*, 107:603–616, 2020. 1

A. Santoro, A. Lampinen, K. Mathewson, T. Lillicrap, and D. Raposo. Symbolic behaviour in artificial intelligence. *arXiv preprint arXiv:2102.03406*, 2021. 5

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022. 5

T. Schick, S. Udupa, and H. Schütze. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00434. URL https://doi.org/10.1162/tacl_a_00434. 5

D. Schlangen. Norm participation grounds language. *arXiv preprint arXiv:2206.02885*, 2022. 5

M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021. 1, 5

H. A. Simon. Bounded rationality. In *Utility and probability*, pages 15–18. Springer, 1990. 1, 5

R. Smith. Aristotle's Logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020. 2.2

J. Sweller, J. J. Van Merrienboer, and F. G. Paas. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296, 1998. 5

M. H. Tessler. Understanding belief bias by measuring prior beliefs for a bayesian model of syllogistic reasoning. In *Proceedings of ESSLLI*, pages 225–237, 2015. 2.2

M. H. Tessler, J. B. Tenenbaum, and N. D. Goodman. Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, 2022. 2.2, 5

T. H. Trinh and Q. V. Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018. 1

D. Trippas, M. F. Verde, and S. J. Handley. Using forced choice to test belief bias in syllogistic reasoning. *Cognition*, 133(3):586–600, 2014. 2.2, 4.1, B.6

K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change), 2022. URL https://arxiv.org/abs/2206.10498. 1

J. X. Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95, 2021. 5

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 5

P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20:273 – 281, 1968. (document), 1, 2.3, 4.2, 4.3, B.5

P. C. Wason and P. N. Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press, 1972. 1

P. C. Wason and D. Shapiro. Natural and contrived experience in a reasoning problem. *Quarterly journal of experimental psychology*, 23(1):63–71, 1971. 2.3

A. Webson and E. Pavlick. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*, 2021. 5, A.1.2

J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 5

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a. 1

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b. 5

E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. Levy. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*, 2020. 5

Y. Wu, M. N. Rabe, W. Li, J. Ba, R. B. Grosse, and C. Szegedy. LIME: Learning inductive bias for primitives of mathematical reasoning. In *International Conference on Machine Learning*, pages 11251–11262. PMLR, 2021. 5

D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. 5

D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. 5

T. Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45, 2022. C

E. Zelikman, Y. Wu, and N. D. Goodman. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. 5

In Appendix A we provide more details of the methods and datasets, in Appendix B we provide supplemental analyses, and in Appendix C we provide full results of statistical models for the main results.

A. Supplemental methods

A.1. Datasets

As noted in the main text, we generated new datasets for each task to avoid problems with training data contamination. In this section we present further details of dataset generation.

A.1.1. NLI task generation

In the absence of existing cognitive literature on generating belief-aligned stimuli for this task, we used a larger language model (Gopher, 280B parameters, from Rae et al., 2021) to generate 100 comparison statements automatically, by prompting it with 6 comparisons that are true in the real world. The exact prompt used was:

```
The following are 100 examples of comparisons:

1. mountains are bigger than hills

2. adults are bigger than children

3. grandparents are older than babies

4. volcanoes are more dangerous than cities

5. cats are softer than lizards
```

We prompted the LLM multiple times, until we had generated 100 comparisons that fulfilled the desired criteria. The prompt completions were generated using nucleus sampling (Holtzman et al., 2019) with a probability mass of 0.8 and a temperature of 1. We filtered out comparisons that were not of the form "[entity] is/are [comparison] than [other entity]". We then filtered these comparisons manually to remove false and subjective ones, so the comparisons all respect real-world facts. An example of the generated comparisons includes "puddles are smaller than seas".

We generated a natural inference task derived from these comparison sentences as follows. We began with the *consistent* version, by taking the the raw output from the LM, "puddles are smaller than seas" as the hypothesis and formulating a premise "seas are bigger than puddles" such that the generated hypothesis is logically valid. We then combine the premise and hypothesis into a prompt and continuations. For example:

```
If seas are bigger than puddles, then puddles are
A. smaller than seas
B. bigger than seas
```

where the logically correct (A) response matches real-world beliefs (that 'puddles are smaller than seas'). Similarly, we can also generate a *violate* version of the task where the logical response violates these beliefs. For example,

```
If seas are smaller than puddles, then puddles are
A. smaller than seas
B. bigger than seas
```

here the correct answer, (B), violates the LM's prior beliefs. Finally, to generate a *nonsense* version of

No flowers are animals.
All reptiles are animals.
Conclusion: No flowers are reptiles.

(a) Valid, consistent

No flowers are animals.
All reptiles are flowers.
Conclusion: No reptiles are animals.

(b) Valid, violate

No flowers are reptiles.
All reptiles are animals.
Conclusion: No flowers are animals.

(c) Invalid, consistent

No flowers are animals.
All flowers are reptiles.
Conclusion: No reptiles are animals.

(d) Invalid, violate

Figure 10 | Example syllogism cluster, showing 2×2 design of valid (top row), invalid (bottom row), and consistent (left column) and violate (right column) arguments.

the task, we simply replace the nouns ('seas' and 'puddles') with nonsense words. For example:

```
If vuffs are smaller than feps, then feps are
A. smaller than vuffs
B. bigger than vuffs
```

Here the logical conclusion is B. For each of these task variations, we evaluate the log probability the language model places on the two options and choose higher likelihood one as its prediction.

As detailed in B.2, we also generated a version of the task where the model did not have to actually produce a sentence that is either consistent or inconsistent with real-world beliefs, for example:

```
Premise: seas are bigger than puddles.
Hypothesis: puddles are smaller than seas.
This hypothesis is:
A. valid
B. invalid
```

To generate logically invalid queries, we keep the hypothesis the same, and alter the premise as follows:

```
Premise: seas are smaller than puddles.
Hypothesis: puddles are smaller than seas.
This hypothesis is:
A. valid
B. invalid
```

A.1.2. Syllogisms data generation

We generated a new set of problems for syllogistic reasoning. Following the approach of Evans et al. (1983), in which the syllogisms were written based on the researchers intuitions of believability, we hand-authored these problems based on beliefs that seemed plausible to the authors. See Fig. 1b for an example problem. We built the dataset from clusters of 4 arguments that use the same three entities, in a 2×2 combination of valid/invalid, and belief-consistent/violate. For example, in Fig. 10 we present a full cluster of arguments about reptiles, animals, and flowers.

By creating the arguments in this way, we ensure that the low-level properties (such as the particular entities referred to in an argument) are approximately balanced across the relevant conditions. In total there are twelve clusters. We avoided using the particular negative form ("some X are not Y") to avoid substantial negation, which complicates behavior both for language models and humans (cf. Hosseini et al., 2021; Evans et al., 1996). We then sampled an identical set of nonsense arguments by simply replacing the entities in realistic arguments with nonsense words.

```
Some librarians are happy people
All happy people are healthy people
Conclusion: Some librarians are healthy people

All guns are weapons
All weapons are dangerous things
Conclusion: All guns are dangerous things

Some electronics are computers
All computers are expensive things
Conclusion: Some electronics are expensive things

All trees are plants
Some trees are tall things
Conclusion: Some plants are tall things

No flowers are animals
All reptiles are animals
Conclusion: No flowers are reptiles

All diamonds are gems
Some diamonds are transparent things
Conclusion: Some gems are transparent things
```

```
All dragons are mythical creatures
No mythical creatures are things that exist
Conclusion: No dragons are things that exist

Some politicians are dishonest people
All dishonest people are people who lie
Conclusion: Some politicians are people who lie

All whales are mammals
Some whales are big things
Conclusion: Some mammals are big things

All vegetables are foods
Some vegetables are healthy things
Conclusion: Some foods are healthy things

All famous actors are wealthy people
Some famous actors are old people
Conclusion: Some old people are wealthy people

All vehicles are things that move
No buildings are things that move
Conclusion: No buildings are vehicles
```

Figure 11 | One argument (valid, consistent) from each of the 12 argument clusters we used for the syllogisms tasks, showing the entities and argument forms covered.

We present the arguments to the model, and give a forced choice between "The argument is valid." or "The argument is invalid." Where example shots are used, they are sampled from distinct clusters, and are separated by a blank line. We prompt the model before any examples with either instructions about evaluating logical arguments, answering logic problems, or with no prompt, but we aggregate across these in the main analyses as effects are similar (Appx. B.3; cf. Webson and Pavlick 2021). We also tried some minor variations as follow-up experiments (such as prefixing the conclusion with "Therefore:" or omitting the prefix before the conclusion), but observed qualitatively similar results so we omit them.

A.1.3. Wason data generation

As above, we generated a new dataset of Wason problems to avoid potential for dataset contamination (see Fig. 1c for an example). The final response in a Wason task does not involve a declarative statement (unlike completing a comparison as in NLI), so answers do not directly 'violate' beliefs. Rather, in the cognitive science literature, the key factor affecting human performance is whether the entities are 'realistic' and follow 'realistic' rules (such as people following social norms) or consist of arbitrary relationships between abstract entities such as letters and numbers. We therefore study the effect of realistic and arbitrary scenarios in the language models.

We created 12 realistic rules and 12 arbitrary rules. Each rule appears with four instances, respectively matching and violating the antecedent and consequent. Each realistic rule is augmented with one sentence of context for the rule, and the cards are explained to represent the entities in the context. The model is presented with the context, the rule, and is asked which of the following instances it needs to flip over, then the instances. The model is then given a forced choice between sentences of the form "You need to flip over the ones showing "X" and "Y"." for all subsets of two items from the instances. There are two choices offered for each pair, in both of the possible orders, to eliminate possible biases if the model prefers one ordering or another. (Recall that the model scores each answer choice independently; it does not see all answers at once.)

See Figs. 12 and 13 for the realistic and arbitrary rules and instances used — but note that problems were presented to the model with more context and structure, see Fig. 1c for an example. We demonstrate in Appx. B.4 that the difficulty of basic inferences about the propositions involved in each rule type is similar across conditions.

We also created 12 rules using nonsense words. Incorporating nonsense words is less straightforward in the Wason case than in the other tasks, as the model needs to be able to reason about whether

```
An airline worker in Chicago needs to check passenger documents. The rule is that if the passengers are traveling outside the US then
↪  they must have showed a passport.
Buenos Aires / San Francisco / passport / drivers license

A chef needs to check the ingredients for dinner. The rule is that if the ingredients are meat then they must not be expired.
beef / flour / expires tomorrow / expired yesterday

A lawyer for the Innocence Project needs to examine convictions. The rule is that if the people are in prison then they must be
↪  guilty.
imprisoned / free / committed murder / did not commit a crime

A medical inspector needs to check hospital worker qualifications. The rule is that if the workers work as a doctor then they must
↪  have received an MD.
surgeon / janitor / received an MD / received a GED

A museum curator is examining the collection. The rule is that if the artworks are in the museum then they must be genuine.
displayed in the museum / not in the museum / genuine / forgery

An adventure trip organizer needs to ensure their clients have the appropriate gear. The rule is that if the clients are going
↪  skydiving then they must have a parachute.
skydiving / mountain biking / parachute / wetsuit

A parent needs to check that their children are following the rules. The rule is that if the children are playing games then they must
↪  have finished their homework.
playing video games / eating dinner / homework is done / homework is not done

A priest needs to check if people are ready for marriage. The rule is that if the people are engaged then they must be adults.
engaged / single / 25 years old / 7 years old

A traffic enforcement officer needs to check that people are following the law. The rule is that if the people in vehicles are driving
↪  then they must have a driver license.
driver / passenger / has a license / does not have a license

A gardener needs to take care of their plants. The rule is that if the plants are flowers then they must be fertilized.
rose / oak / fertilized / not fertilized

A farmer is getting equipment ready for the day. The rule is that if the pieces of equipment have an engine then they must have fuel.
tractor / shovel / has gasoline / does not have gasoline

A person is cleaning out and organizing his closet. The rule is that if the clothes are going to the thrift store then they must be
↪  old.
thrift store / keep / worn out / brand new

An employer needs to check that their business is following health regulations. The rule is that if the employees are working then
↪  they must not be sick.
working / on vacation / healthy / has a cold
```

Figure 12 | Realistic Wason rules and instances used.

instances match the antecedent and consequent of the rule. We therefore use nonsense rules of the form "If the cards have less gluff, then they have more caft" with instances being more/less gluff/caft. The more/less framing makes the instances roughly the same length regardless of rule type, and avoids using negation which might confound results (Hosseini et al., 2021).

Finally, we created two types of control rules based on the realistic rules. First, we created shuffled realistic rules by combining the antecedents and consequents of different realistic rules, while ensuring that there is no obvious rationale for the rule. We then created violate-realistic rules by taking each realistic rule and reversing its consequent. For example, the realistic rule "If the clients are skydiving, then they must have a parachute" is transformed to "If the clients are skydiving, then they must have a wetsuit", but "parachute" is still included among the cards. The violate condition is designed to make the rule especially implausible in context of the examples, while the rule in the shuffled condition is somewhat more arbitrary/belief neutral.

To rule out a possible specific effect of cards (which were used in the original tasks) we also sampled versions of each problem with sheets of paper or coins, but results are similar so we collapse across these conditions in the main analyses.

A.2. Evaluation

DC-PMI correction: We use the DC-PMI correction (Holtzman et al., 2021) for the syllogisms and Wason tasks; i.e., we choose an answer from the set of possible answers ($\mathcal{A}$) as follows:

$$\text{argmax}_{a \in \mathcal{A}} \; p(a \mid \text{question}) - p(a \mid \text{baseline prompt})$$

```
The rule is that if the cards have a plural word then they must have a positive emotion.
crises / dog / happiness / anxiety

The rule is that if the cards have a soft texture then they must have a polygon.
soft / rough / hexagon / circle

The rule is that if the cards have a French word then they must have a positive number.
chapeau / sombrero / 4 / -1

The rule is that if the cards have a prime number then they must have a secondary color.
11 / 12 / purple / red

The rule is that if the cards have a European country then they must have something hot.
Germany / Brazil / furnace / ice cube

The rule is that if the cards have the name of a famous book then they must have the name of an elementary particle.
Moby Dick / Citizen Kane / neutrino / atom

The rule is that if the cards have a type of plant then they must have the name of a philosopher.
cactus / horse / Socrates / Napoleon

The rule is that if the cards have the name of a web browser then they must have a type of pants.
Internet Explorer / Microsoft Word / jeans / sweatshirt

The rule is that if the cards have a beverage containing caffeine then they must have a material that conducts electricity.
coffee / orange juice / copper / wood

The rule is that if the cards have something electronic then they must have a hairy animal.
flashlight / crescent wrench / bear / swan

The rule is that if the cards have a verb then they must have a Fibonacci number.
walking / slowly / 13 / 4

The rule is that if the cards have a text file extension then they must have a time in the morning.
.txt / .exe / 11:00 AM / 8:00 PM
```

Figure 13 | Arbitrary Wason rules and instances used.

Where the baseline prompt is “Answer:” and $p(x \mid y)$ denotes the model’s evaluated likelihood of continuation x after prompt y.

B. Supplemental analyses

B.1. Directly estimating model beliefs via log likelihood, in the NLI task

In the natural language inference (NLI) analyses of 3, we found the models had strong content bias towards NLI completions that are consistent with the models’ prior expectations. We inferred which completions were consistent with the model’s beliefs by using comparisons generated by the model. In these analyses, we use a more graded measure of the models’ beliefs.

Here, instead of using the consistent/violate categories we assumed, we pool together the comparison statements that were generated by the model (taken to be "belief consistent" in other analyses; e.g. “ants are smaller than whales”) and the inverted versions (taken to be “belief violating”; e.g. “whales are smaller than ants”). We then directly estimate the model’s beliefs about the plausibility of a comparison statement by computing the model’s average per-token log likelihood on the statements standing alone (not conditioned on any prior input).

For each NLI problem, we took the difference between [the unconditioned model log likelihood for the correct conclusion] and [the unconditioned model log likelihood for the incorrect conclusion]. This difference, along with a bias term, was used as the input to a logistic regression against whether the model answered the NLI problem correctly. We see in Fig 14 that model accuracy on zero-shot NLI problems is highly positively related with its pre-existing, unconditioned log likelihood on the conclusion. This is consistent with our previous NLI results (3). The positive relationship is maintained but diminishes for 1-shot and 5-shot problems, echoing our other results showing that few-shot examples mitigate belief bias in NLI (4.2).

These analyses were performed using the larger model (Chinchilla).

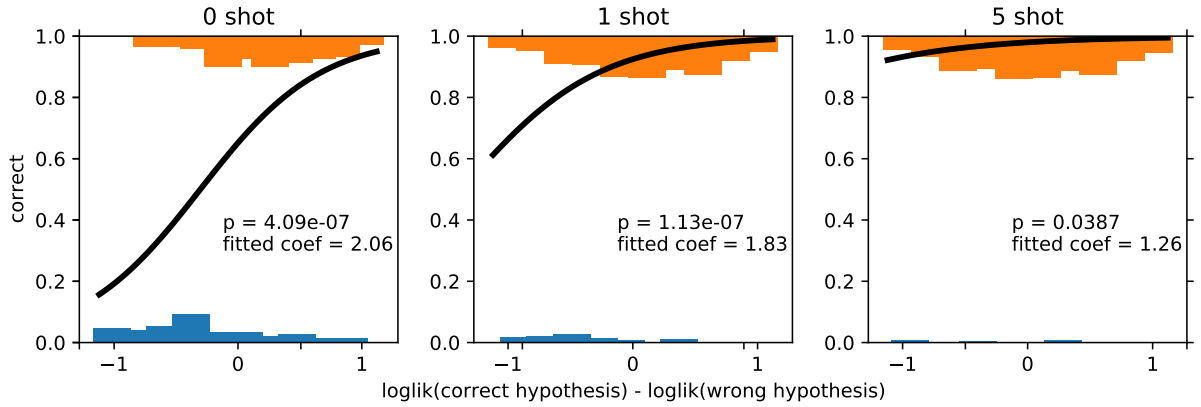
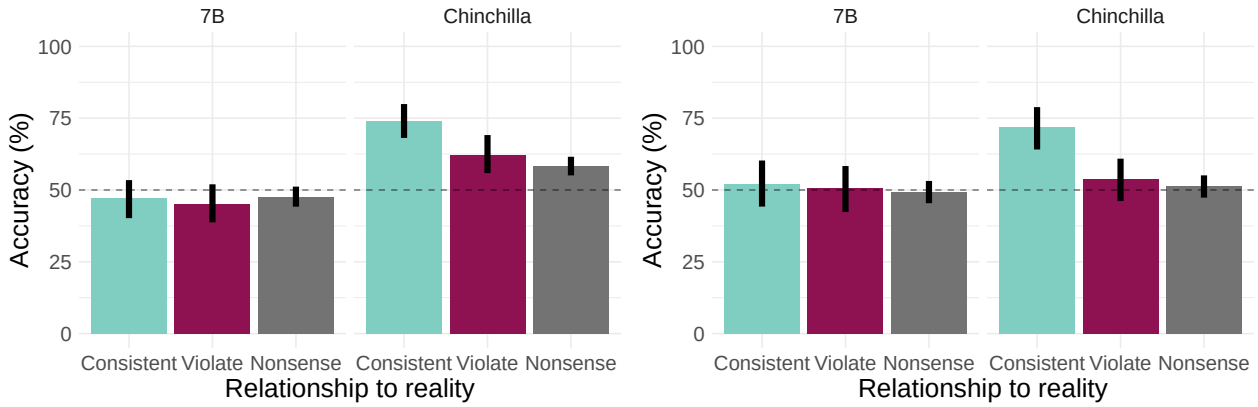


Figure 14 | Logistic regression of Chinchilla accuracy on NLI vs. likelihoods. Histograms at top and bottom respectively show the distributions of correct and incorrect answers; curves are logistic regression fits.



(a) No correction

(b) With correction

Figure 15 | NLI results zero-shot, for a smaller (7B)and larger (Chinchilla) language model, where the models have to classify validity of natural language inferences rather than completing the valid hypothesis. The smaller model is unable to do this task at all and shows no belief bias. The larger model does show a belief bias, details in main text.

B.2. Classifying natural language inferences as valid or invalid produces a belief-bias effect.

We evaluate whether the language model can classify provided premise and hypothesis pairs as a valid or invalid inference (Figure 15). We find that the smaller model is unable to perform the task above chance accuracy, and shows no effects of belief bias. The larger model does show a belief bias effect: accuracy on this task is highest when the hypothesis is consistent with beliefs. We find lower performance in the violate condition where the hypothesis is inconsistent with beliefs but the entities are still realistic. We find the lowest performance on the nonsense entities, over which the model has no prior beliefs.

Note that the 'consistent' and 'violate' condition in this setting are slightly confounded by the requirement to include 'invalid' logical inference. For an invalid logical inference, if we want the hypothesis to violate real world beliefs, the premise must be consistent with real world belief, and vice versa. This might explain the attenuated difference between violate and consistent in this condition.

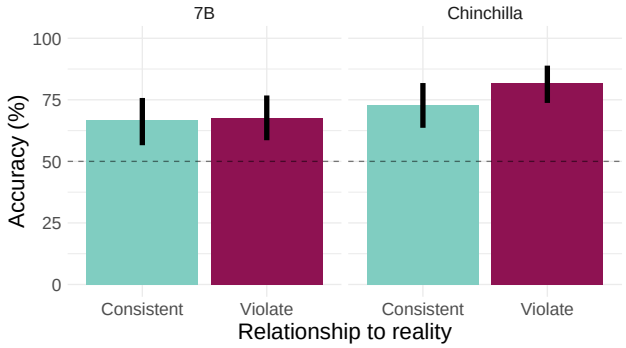


Figure 16 | Above chance performance on NLI with DC-PMI correction implies that the model is accounting for the premise in its responses.

As discussed in the main text, we also applied the DC-PMI correction to the original framing of the NLI queries to verify that the the language models are in fact accounting for the premise in their responses. We do this with a control condition where the hypotheses-premise pairs are mismatched such that the hypothesis is no longer logically implied by the premise – but the hypotheses themselves are still consistent with or violate prior expectations. Normalizing for performance in this condition checks that the model is and not just classifying on the basis of strong beliefs about the hypothesis irrespective of the premise. We found significantly above chance performance after DC-PMI with a mismatched premise, indicating that the premise is playing a role in performance in both consistent and violate conditions (Figure 16). We didn't run this control for nonsense examples because the model doesn't have strong a priori beliefs about these hypotheses.

B.3. The effects of different initial instruction prompts

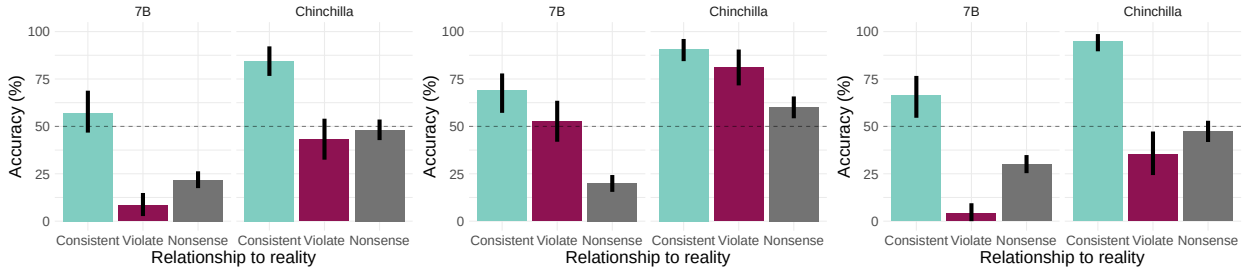
One common strategy for improving language model zero-shot performance is to prompt the model with a task instruction. To investigate whether instructions would affect results, we evaluated three different task instructions on the NLI and syllogism tasks.

B.3.1. NLI

For the NLI task, we tried the following prompts:

```
INITIAL_PROMPTS = {
"none": "If {seas are smaller than puddles}, then {puddles are}",
"evaluate_arguments": "Carefully evaluate these logical arguments. If {seas are smaller
    than puddles}, then {puddles are}",
"pretend": "Let's pretend that {seas are smaller than puddles}, then {puddles are}",
}
```

The explicit instruction to evaluate arguments strongly improves performance on the violate condition. This mirrors previous findings (Kojima et al., 2022) that explicit instructions can make language models more logical. Both models however still struggle with the abstract 'nonsense' version of the task, and performance is still best on the 'consistent' condition. The 'pretend' prompt gives very similar findings. These results illustrate that the belief bias effects we find are not overfit to the specific prompt we use; these effects show up across other prompts as well.



(a) None (b) Evaluate Arguments (c) Pretend

Figure 17 | Performance on NLI with different prompt types.

B.3.2. Syllogism

For the Syllogism task, we tried the following prompts:

```
INITIAL_PROMPTS = {
"none": "",
"evaluate_arguments": "Carefully evaluate these logical arguments, and determine
    whether each is valid or invalid.\n\n",
"logic_problems": "Answer these logic problems carefully, by determining whether each
    argument is valid or invalid.\n\n"
}
```

In Fig. 18 we show the results. While the instructions do slightly change model behavior — in particular, shifting the overall response tendency across conditions — the patterns of bias remain similar; we therefore collapse across the prompts in other figures (and include them as a random effect in regressions).

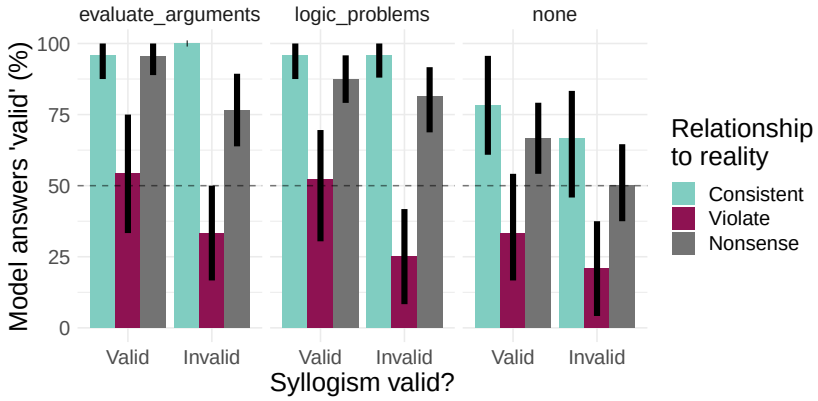


Figure 18 | Syllogism 0-shot performance with different initial instruction prompts; the belief-bias effects are similar across conditions.

B.4. The Wason rule propositions have similar difficulty across conditions

One possible confounding explanation for our Wason results would be that the base propositions that form the antecedents and consequents of the rules have different difficulty across conditions—this could potentially explain why the realistic rules and shuffled realistic rules are both easier than abstract or nonsesnse ones. To investigate this possibility, we tested the difficulty of identifying which of the options on the cards matched the corresponding proposition. Specifically, for the antecedent

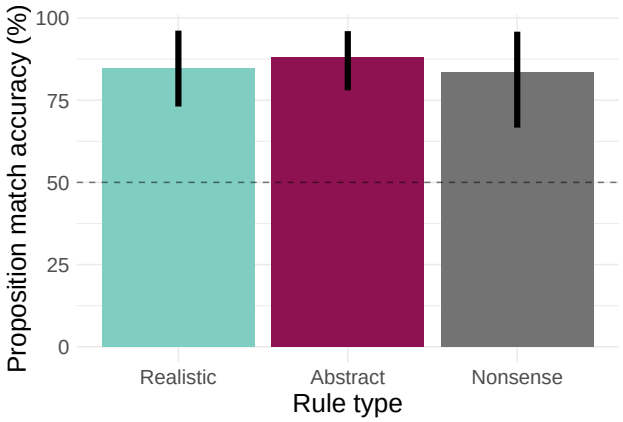


Figure 19 | The component propositions (antecedents and consequents) of the Wason rules have similar difficulty across conditions. This plot shows model accuracy on forced choices of which instance matches a proposition, across conditions. (Note that the shuffled realistic rules use the same component propositions as the realistic rules.)

of the rule "if the workers work as a doctor then they must have received an MD" we prompted the language model with a question like:

```
Which choice better matches "work as a doctor"?
choice: surgeon
choice: janitor
Answer:
```

And then gave a two-alternative forced choice between 'surgeon' and 'janitor'. We repeated this process for both possible answer choice orderings in the prompt, and then aggregated likelihoods across these and chose the highest-likelihood answer.

By this metric, we find that there are no substantial differences in difficulty across the rule types (Fig. 19)—in fact, arbitrary rule premises are numerically slightly easier, though the differences are not significant. Thus, the effects we observed are not likely to be explained by the base difficulty of verifying the component propositions.

B.5. The model consistently chooses one antecedent card and one consequent card on the Wason task

In Fig. 20 we show that the language model is consistently choosing one antecedent and one consequent card. Thus, the model is matching the majority of experimental participants in Wason (1968) in testing both parts of the rule. However, like the humans, it is not always making the correct choice within each category, as the main analyses show.

B.6. Models can identify the valid conclusion of a syllogism from among all possible conclusions with high accuracy

In Fig. 21 we show the accuracy of the model when choosing from among all possible predicates containing one of the quantifiers used and two of the entities appearing in the premises of the syllogism. The model exhibits high accuracy across conditions, and relatively little bias (though bias increases few shot). This observation is reminiscent of the finding of Trippas et al. (2014) that humans exhibit less bias when making a forced choice among two possible arguments (one valid and

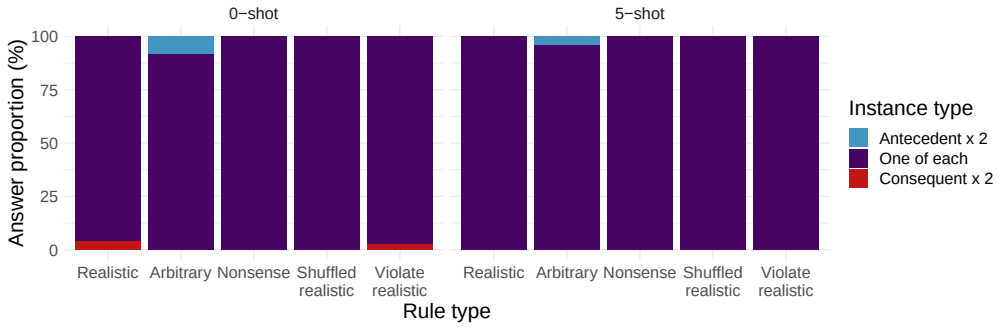
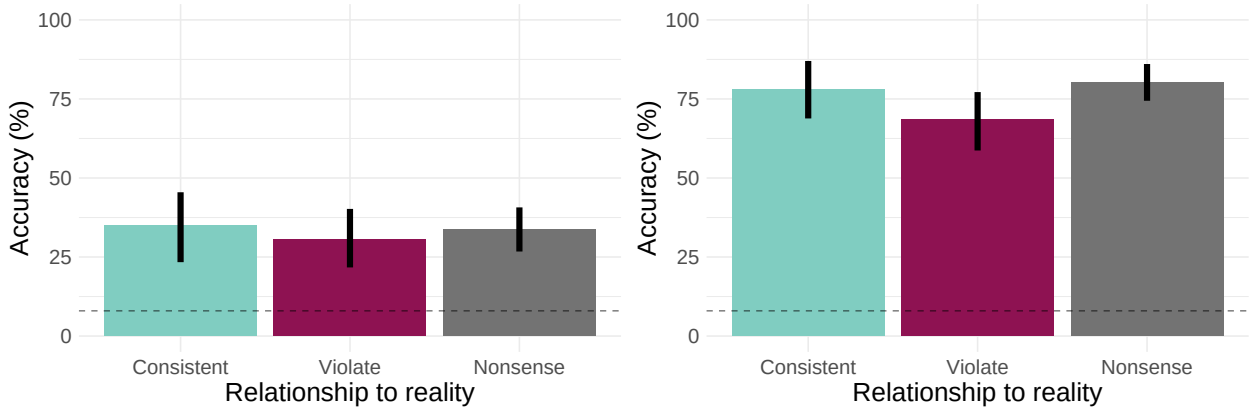


Figure 20 | On the Wason tasks, the model correctly chooses one antecedent card and one consequent card the vast majority of the time, across experimental conditions—by chance, it would only do so 50% of the time.

one invalid) rather than deciding if a single syllogism is valid or invalid.

Note that in this case scoring with the Domain-Conditional PMI (Holtzman et al., 2021)—which we used for the main Syllogisms and Wason results—produces much *lower* accuracy than the raw likelihoods, and minor differences in bias. The patterns are qualitatively similar with or without the correction, but accuracy is lower without (around 35-40%) regardless of belief consistency.



(a) Domain-conditional PMI scoring

(b) Raw log-likelihood scoring

Figure 21 | Zero-shot accuracy at identifying the correct conclusion to a syllogism among all possible conclusions. The model exhibits far above chance performance (especially when scoring with raw log-likelihoods), and relatively weaker bias with this task design.

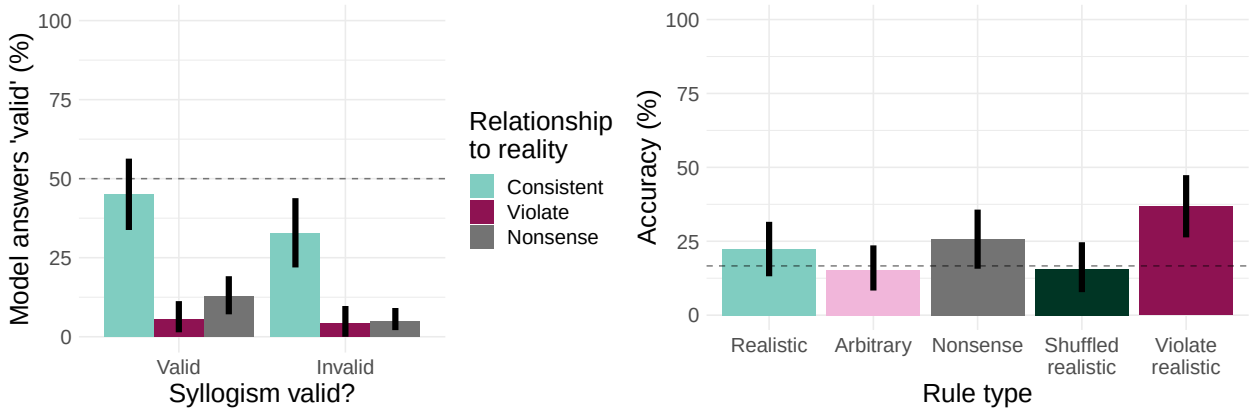
B.7. Using raw likelihoods rather than Domain-Conditional PMI on the Syllogisms and Wason tasks

In the main results for the Syllogisms and Wason tasks, we scored the model using the Domain-Conditional PMI (Holtzman et al., 2021). However, the above syllogisms conclusion results raise the possibility that raw likelihoods would result in better performance. In Fig. 22 we show that comparing the raw likelihoods of the answers instead of DC-PMI results in much worse zero-shot performance on both tasks, and greater bias toward answering 'invalid' on the Syllogisms tasks. Thus, the DC-PMI correction is consistently beneficial for both of the more complex reasoning tasks zero-shot.

For Syllogisms, we find similar results few-shot (Fig. 23a)—raw-likelihood scoring results in more

bias towards answering 'invalid', though this bias decreases somewhat with 5 example shots. However, the belief bias towards responding 'valid' more for consistent statements than inconsistent ones is present in all conditions, and becomes larger as the 'invalid' bias reduces with more examples.

For the Wason tasks (Fig. 23b), the effects are somewhat more complicated. For the realistic rules, using the raw-likelihood scoring results in worse performance than DC-PMI (compare to Fig. **??**). For the arbitrary and nonsense rules, using the raw-likelihood scoring actually improves performance relative to DC-PMI—however, this performance is still substantially worse than realistic rule performance by either metric. For the shuffled realistic rules, raw-likelihood scoring slightly increases performance (at least 5-shot), which, together with the decrease in realistic performance using this metric, makes the realistic and shuffled realistic rules competitive. However, scoring each rule type by the most favorable metric for that type would still result in the realistic rules performing substantially better than the shuffled realistic rules. In summary, while the numerical Wason results are somewhat altered by the raw-likelihood metrics, the pattern of advantage for realistic rules over arbitrary or nonsense ones is preserved.



(a) Syllogisms

(b) Wason

Figure 22 | Scoring using the raw answer likelihoods—rather than the Domain-Conditional PMI—results in poor zero-shot performance on the Syllogisms and Wason tasks, with low accuracy and high bias. (Compare to Figs. 4 and 5, respectively, which use DC-PMI scoring.))

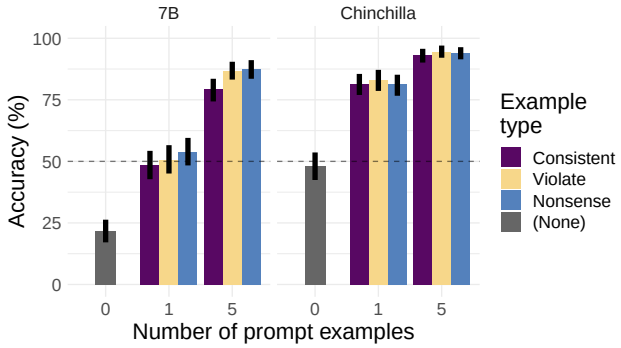
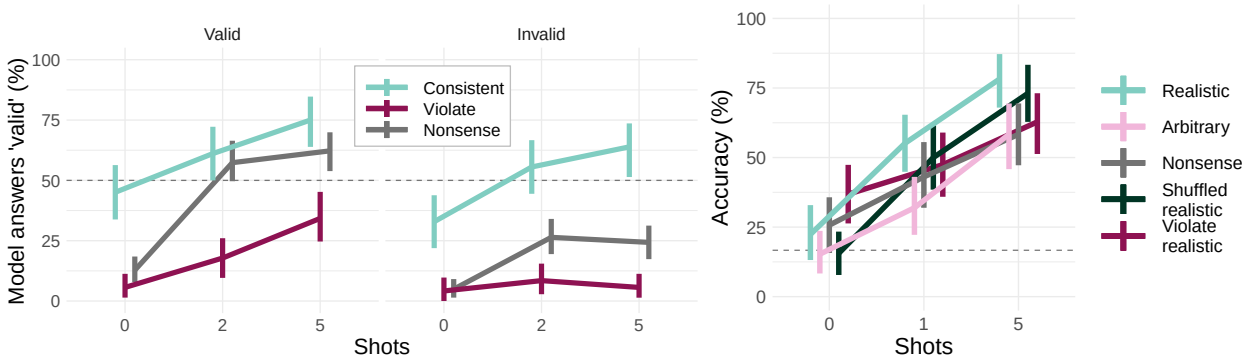


Figure 24 | NLI evaluated on nonsense probes, with 0, 1, or 5 example shots of different types.



(a) Syllogisms (b) Wason

Figure 23 | Scoring using the raw answer likelihoods—rather than the Domain-Conditional PMI—for the Syllogisms and Wason tasks few-shot. (a) On the syllogisms tasks, using raw likelihoods results in poorer performance, with high bias towards answering 'invalid'. However, belief bias is still present, and performance does not improve as much with few-shot examples (in fact, it deteriorates on invalid arguments). (b) On the Wason tasks (using matched examples), the realistic rules score worse with raw likelihood scoring, but intriguingly the other rule types score better, especially 5-shot. However, the advantage for realistic over arbitrary or nonsense rules remains. Compare to Figs. 7 and 26b, respectively, which use DC-PMI scoring.))

B.8. Varying few-shot examples

For the NLI tasks, there are not substantial effects of the prompt example type on probe performance (Fig. 24), all kinds of prompts give close to ceiling performance with five shots. Similarly, for the syllogism tasks, there are not substantial effects of the prompt example type on probe performance (Fig. 25)—overall, different types of prompts offer similar probe performance in most cases.

For the Wason tasks, there is limited benefit from nonsense examples (Fig. 26a). There appears to be a benefit to having realistic examples in the prompt for either realistic or shuffled-realistic probes (Fig. 26b). However, for arbitrary and nonsense rules, realistic examples do not appear substantially beneficial. Instead, arbitrary and nonsense rules generally benefit more from matching prompt examples. We therefore use matching examples in other follow-up experiments on the Wason tasks, such as the error analysis.

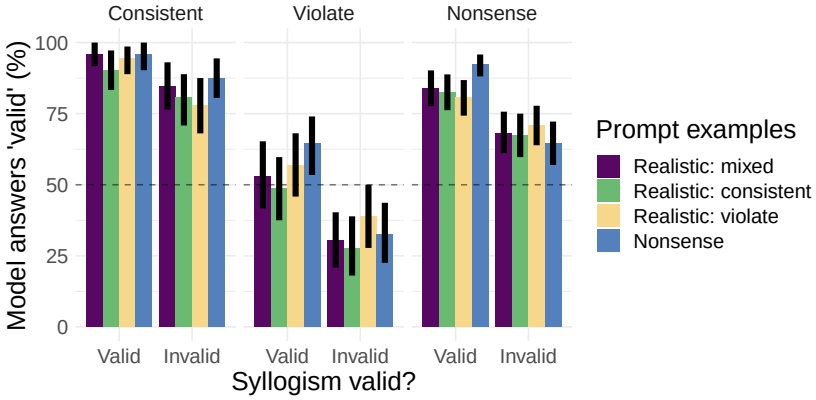
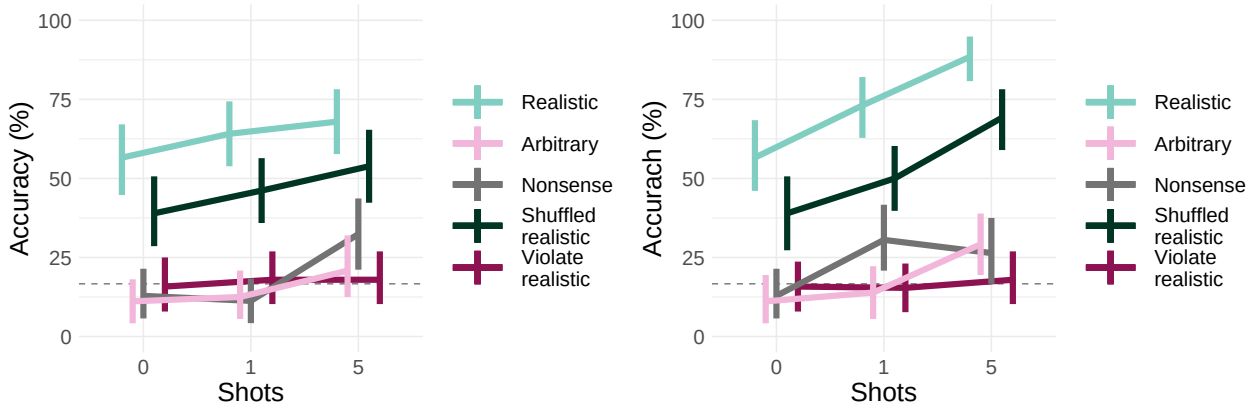


Figure 25 | Syllogism results few-shot, with different types of prompt examples. The “Realistic: mixed” condition includes realistic examples from both the consistent and violate subsets.



(a) Nonsense examples

(b) Realistic examples

Figure 26 | Wason selection task results in different probe conditions, after a few-shot prompt with (a) nonsense or (b) realistic example shots. Realistic examples result in much more benefit than nonsense examples, at least for realistic or shuffled realistic tasks. In all conditions the overall human pattern of an advantage of realistic probes remains.

C. Statistical analyses

In this section, we present the full output of statistical models for the main results we report. We analyze all results using hierarchical (multilevel) logistic regressions that account for the statistical dependency structure of the data (e.g. Gelman and Hill, 2006); such analytic methods are broadly used in the behavioral sciences because they allow drawing generalizable inferences from data (Yarkoni, 2022), and they are similarly useful tools for analyzing the behaviors of large language models (Lampinen et al., 2022).

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: multiple_choice_grade ~ noun_type_factor * model_name + (1 |
    hypothesis)
   Data: nli_df %>% filter(n_prepended_shots %in% c(0))

     AIC      BIC   logLik deviance df.resid
  1030.4   1064.1   -508.2   1016.4      903

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3340 -0.5622 -0.2448  0.5579  3.5608

Random effects:
 Groups     Name        Variance Std.Dev.
 hypothesis (Intercept) 2.715    1.648
Number of obs: 910, groups:  hypothesis, 443

Fixed effects:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -3.4172     0.5823  -5.868 4.40e-09 ***
noun_type_factorconsistent                         3.8310     0.6940   5.520 3.39e-08 ***
noun_type_factornonsense                           1.4804     0.5666   2.613  0.00898 **
model_nameprod_chinchilla                          2.9612     0.6012   4.926 8.40e-07 ***
noun_type_factorconsistent:model_nameprod_chinchilla -0.9536     0.7347  -1.298  0.19427
noun_type_factornonsense:model_nameprod_chinchilla   -1.1564     0.6107  -1.894  0.05829 .

(Intercept)
noun_type_factorconsistent
noun_type_factornonsense
model_nameprod_chinchilla
noun_type_factorconsistent:model_nameprod_chinchilla
noun_type_factornonsense:model_nameprod_chinchilla
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 1 | NLI zero-shot regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
answer_with_valid_prior_corrected ~ consistent_plottable + valid_factor +
    (1 | initial_prompt) + (1 | syllogism_name)
   Data: syl_df %>% filter(num_shots %in% c(0))

     AIC      BIC   logLik deviance df.resid
   512.4    538.5   -250.2    500.4      565

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6454 -0.3729  0.2145  0.3970  3.2771

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 2.1165   1.4548
 initial_prompt (Intercept) 0.7366   0.8582
Number of obs: 571, groups:  syllogism_name, 60; initial_prompt, 3

Fixed effects:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                             3.5530     0.7755   4.581 4.62e-06 ***
consistent_plottableviolate_vs_consistent  -3.8417     0.4458  -8.617  < 2e-16 ***
consistent_plottablenonsense_vs_consistent -1.2874     0.6070  -2.121  0.03395 *
valid_invalid                           -1.0184     0.3053  -3.336  0.00085 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2 | Syllogisms zero-shot regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
answer_with_valid_prior_corrected ~ consistent_plottable * valid_factor +
    (1 | initial_prompt) + (1 | syllogism_name)
   Data: syl_df %>% filter(num_shots %in% c(0))

     AIC      BIC   logLik deviance df.resid
   514.5    549.3   -249.3    498.5      563

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.1789 -0.3592  0.2127  0.4282  3.0836

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 2.1504   1.4664
 initial_prompt (Intercept) 0.7455   0.8634
Number of obs: 571, groups:  syllogism_name, 60; initial_prompt, 3

Fixed effects:
                                                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)                                             3.1487     0.8206    3.837 0.000125 ***
consistent_plottableviolate_vs_consistent              -3.2822     0.5767   -5.691 1.26e-08 ***
consistent_plottablenonsense_vs_consistent             -0.8723     0.7586   -1.150 0.250167
valid_invalid                                          -0.3202     0.6097   -0.525 0.599517
consistent_plottableviolate_vs_consistent:valid_invalid  -1.0401     0.7538   -1.380 0.167624
consistent_plottablenonsense_vs_consistent:valid_invalid -0.7076     0.8304   -0.852 0.394154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3 | Syllogisms zero-shot regression results with an interaction term—note that this model is reported only for comparison, since the information criteria (AIC and BIC) suggest that the simpler model without an interaction is preferable.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
prior_corrected_multiple_choice_grade ~ probe_type_factor + card_type_factor +
    (1 | wason_name)
   Data: wason_df %>% filter(num_shots %in% c(0))

     AIC      BIC   logLik deviance df.resid
   330.0    361.4   -157.0    314.0      363

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5011 -0.3332 -0.1604  0.3255  3.1830

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 4.715    2.171
Number of obs: 371, groups:  wason_name, 50

Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.5013     0.9416  -3.719  0.00020 ***
probe_type_factorrealistic     3.7535     1.1465   3.274  0.00106 **
probe_type_factornonsense      0.3959     1.1846   0.334  0.73826
probe_type_factorshuffledreal  2.5229     1.1465   2.201  0.02777 *
probe_type_factorviolatereal   0.8755     1.1499   0.761  0.44642
card_type_factorcoins         -0.2652     0.2318  -1.144  0.25251
card_type_factorsheets_of_paper  0.1657     0.2248   0.737  0.46111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4 | Wason zero-shot regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
multiple_choice_grade ~ (noun_type_factor + model_name) * num_shots_regressable +
    (1 | hypothesis)
   Data: nli_df %>% filter(n_prepended_shots %in% c(0, 1, 5), shots_type %in%
    c("nonsense", "None"))
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))

     AIC      BIC   logLik deviance df.resid
  2264.9   2341.8  -1119.4   2238.9     2717

Scaled residuals:
     Min       1Q   Median       3Q      Max
-14.3350  -0.2842   0.1174   0.2986   4.6978

Random effects:
 Groups     Name        Variance Std.Dev.
 hypothesis (Intercept) 3.576    1.891
Number of obs: 2730, groups:  hypothesis, 443

Fixed effects:
                                                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)                                               -0.02624    0.28042   -0.094   0.9255
noun_type_factorconsistent                                 2.56592    0.45903    5.590 2.27e-08 ***
noun_type_factornonsense                                   0.31384    0.30673    1.023   0.3062
model_nameprod_chinchilla                                  1.88943    0.15361   12.300  < 2e-16 ***
num_shots_regressableany_shots                             2.84223    0.26137   10.874  < 2e-16 ***
num_shots_regressableone_or_five                           1.50063    0.22817    6.577 4.80e-11 ***
noun_type_factorconsistent:num_shots_regressableany_shots -0.77785    0.38116   -2.041   0.0413 *
noun_type_factornonsense:num_shots_regressableany_shots   -0.39870    0.25541   -1.561   0.1185
noun_type_factorconsistent:num_shots_regressableone_or_five -0.06510  0.43184   -0.151   0.8802
noun_type_factornonsense:num_shots_regressableone_or_five -0.11444    0.23900   -0.479   0.6321
model_nameprod_chinchilla:num_shots_regressableany_shots   -0.17412    0.17924   -0.971   0.3313
model_nameprod_chinchilla:num_shots_regressableone_or_five -0.45226    0.18825   -2.402   0.0163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 5 | NLI 0, 1, or 5 nonsense shot regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
answer_with_valid_prior_corrected ~ (consistent_plottable + valid_factor) *
    scale(num_shots, scale = F) + (1 | initial_prompt) + (1 |
    syllogism_name)
   Data: syl_df %>% filter(num_shots %in% c(0, 2, 5), prompt_condition %in%
    c("nonsense", "zero_shot"))
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))

     AIC      BIC   logLik deviance df.resid
  1536.8   1591.3   -758.4   1516.8     1711

Scaled residuals:
    Min      1Q  Median      3Q     Max
-9.6556 -0.4020  0.2882  0.4668  2.8407

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0.8831   0.9397
 initial_prompt (Intercept) 0.1970   0.4439
Number of obs: 1721, groups:  syllogism_name, 60; initial_prompt, 3

Fixed effects:
                                                                  Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                                                        3.221998   0.431995    7.458  8.76e-14 ***
consistent_plottableviolate_vs_consistent                         -2.960377   0.218646  -13.540   < 2e-16 ***
consistent_plottablenonsense_vs_consistent                        -0.896209   0.366319   -2.447    0.0144 *
valid_invalid                                                     -1.211116   0.166914   -7.256  3.99e-13 ***
scale(num_shots, scale = F)                                        0.214230   0.095510    2.243    0.0249 *
consistent_plottableviolate_vs_consistent:scale(num_shots, scale = F)  -0.001357   0.099954   -0.014    0.9892
consistent_plottablenonsense_vs_consistent:scale(num_shots, scale = F) -0.065002   0.094666   -0.687    0.4923
valid_invalid:scale(num_shots, scale = F)                         -0.196119   0.067163   -2.920    0.0035 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6 | Syllogisms few-shot regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
multiple_choice_grade ~ model_name * shots_type_factor * num_shots_nonsense +
    (1 | hypothesis)
   Data: nli_df %>% filter(n_prepended_shots %in% c(1, 5), noun_type ==
    "nonsense")
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))

     AIC      BIC   logLik deviance df.resid
  3901.9   3986.3  -1938.0   3875.9     4871

Scaled residuals:
     Min       1Q   Median       3Q      Max
-12.0392   0.0401   0.1859   0.3609   7.4393

Random effects:
 Groups     Name        Variance Std.Dev.
 hypothesis (Intercept) 3.248    1.802
Number of obs: 4884, groups:  hypothesis, 393

Fixed effects:
                                                                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                                                        2.7285    0.1716    15.899   < 2e-16 ***
model_name7b                                                      -2.4164    0.1689   -14.305   < 2e-16 ***
shots_type_factorViolate                                           0.6090    0.2034     2.994   0.00276 **
shots_type_factorNonsense                                          0.4678    0.2001     2.338   0.01937 *
num_shots_nonsenseone_or_five                                      0.9525    0.1359     7.010  2.38e-12 ***
model_name7b:shots_type_factorViolate                              0.2242    0.2431     0.922   0.35642
model_name7b:shots_type_factorNonsense                             0.2958    0.2417     1.224   0.22088
model_name7b:num_shots_nonsenseone_or_five                        -0.1420    0.1624    -0.875   0.38174
shots_type_factorViolate:num_shots_nonsenseone_or_five            -0.2093    0.2030    -1.031   0.30244
shots_type_factorNonsense:num_shots_nonsenseone_or_five           -0.1313    0.1998    -0.657   0.51117
model_name7b:shots_type_factorViolate:num_shots_nonsenseone_or_five  0.4048    0.2431     1.665   0.09594 .
model_name7b:shots_type_factorNonsense:num_shots_nonsenseone_or_five 0.6201    0.2419     2.563   0.01037 *

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 7 | NLI with different type of example shots, when evaluating on nonsense examples, regression results.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula:
prior_corrected_multiple_choice_grade ~ probe_type_factor * prompt_condition_plottable +
    card_type_factor + (1 | wason_name)
   Data: wason_five_shot_df
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))

     AIC      BIC   logLik deviance df.resid
  1555.5   1710.7   -749.7   1499.5     1863

Scaled residuals:
     Min       1Q   Median       3Q      Max
-13.2799  -0.4181  -0.0978   0.4297   5.9983

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 6.917    2.63
Number of obs: 1891, groups:  wason_name, 50

Fixed effects:
                                                                      Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                                                          -0.44266     0.43153   -1.026   0.30499
probe_type_factorArbitrary                                           -1.77578     0.81098   -2.190   0.02855 *
probe_type_factorNonsense                                            -1.71260     0.78895   -2.171   0.02995 *
probe_type_factorShuffled_real                                        1.99525     0.79148    2.521   0.01170 *
probe_type_factorViolate_real                                        -1.40470     0.60368   -2.327   0.01997 *
prompt_condition_plottableArbitrary                                  -0.89641     0.22056   -4.064  4.82e-05 ***
prompt_condition_plottableNonsense                                   -0.70453     0.22064   -3.193   0.00141 **
prompt_condition_plottableShuffled_real                              -0.64101     0.22475   -2.852   0.00434 **
prompt_condition_plottableViolate_real                               -0.50862     0.22701   -2.241   0.02506 *
card_type_factorcoins                                                 0.11061     0.09481    1.167   0.24333
card_type_factorsheets_of_paper                                      -0.04185     0.09490   -0.441   0.65923
probe_type_factorArbitrary:prompt_condition_plottableArbitrary        0.61708     0.45483    1.357   0.17487
probe_type_factorNonsense:prompt_condition_plottableArbitrary         0.23707     0.42270    0.561   0.57490
probe_type_factorShuffled_real:prompt_condition_plottableArbitrary   -0.90202     0.47447   -1.901   0.05729 .
probe_type_factorViolate_real:prompt_condition_plottableArbitrary     0.89641     0.43303    2.070   0.03844 *
probe_type_factorArbitrary:prompt_condition_plottableNonsense        -0.16637     0.46736   -0.356   0.72186
probe_type_factorNonsense:prompt_condition_plottableNonsense          1.21030     0.40776    2.968   0.00300 **
probe_type_factorShuffled_real:prompt_condition_plottableNonsense    -0.97851     0.47448   -2.062   0.03918 *
probe_type_factorViolate_real:prompt_condition_plottableNonsense      0.70452     0.43307    1.627   0.10378
probe_type_factorArbitrary:prompt_condition_plottableShuffled_real    0.50209     0.45528    1.103   0.27011
probe_type_factorNonsense:prompt_condition_plottableShuffled_real    -0.26972     0.43369   -0.622   0.53398
probe_type_factorShuffled_real:prompt_condition_plottableShuffled_real -0.00703   0.48460   -0.015   0.98843
probe_type_factorViolate_real:prompt_condition_plottableShuffled_real  0.39202     0.44688    0.877   0.38036
probe_type_factorArbitrary:prompt_condition_plottableViolate_real    -0.52253     0.47560   -1.099   0.27191
probe_type_factorNonsense:prompt_condition_plottableViolate_real     -0.15072     0.42614   -0.354   0.72357
probe_type_factorShuffled_real:prompt_condition_plottableViolate_real  0.02127   0.48752    0.044   0.96520
probe_type_factorViolate_real:prompt_condition_plottableViolate_real   0.62147     0.43172    1.440   0.15000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 8 | Wason 5-shot regression results; how performance on probe (target) questions of different rule types is affected by the type of examples in the prompt. To evaluate the overall effects of prompt conditions, together with individual interactions, we dummy code prompt condition with Realistic prompt examples as the reference, while effect code probe type (so that probe type effects are overall differences from the grand mean)—in this way, the base prompt condition effects represent the “main” effect of switching from Realistic examples to that other prompt condition across all probe types, while the interaction terms represent how the effect of a given prompt condition differs for a specific probe type from the overall effect of that prompt condition. The significant negative effects of each prompt condition show that overall, using non-realistic examples makes performance worse; the strong interaction effects for Nonsense examples show that in that specific instances prompt examples of the same type as the probe may be beneficial, or at least less detrimental; etc.