Language models show human-like content effects on reasoning tasks

Andrew K. Lampinen,[1*] Ishita Dasgupta,[2*]
Stephanie C. Y. Chan,[2] Hannah R. Sheahan,[1] Antonia Creswell,[3]
Dharshan Kumaran,[1] James L. McClelland,[2,4] Felix Hill,[1]
[1]Google DeepMind, London, UK
[2]Google DeepMind, Mountain View, CA, USA
[3]Work performed at Google DeepMind, London, UK
[4]Department of Pscyhology, Stanford University, Stanford, CA, USA
[*]Equal contribution; to whom correspondence should be addressed: {lampinen,idg}@google.com

Abstract reasoning is a key ability for an intelligent system. Large language models (LMs) achieve above-chance performance on abstract reasoning tasks, but exhibit many imperfections. However, human abstract reasoning is also imperfect. For example, human reasoning is affected by our real-world knowledge and beliefs, and shows notable "content effects"; humans reason more reliably when the semantic content of a problem supports the correct logical inferences. These content-entangled reasoning patterns play a central role in debates about the fundamental nature of human intelligence. Here, we investigate whether language models — whose prior expectations capture some aspects of human knowledge — similarly mix content into their answers to logical problems. We explored this question across three logical reasoning tasks: natural language inference, judging the logical validity of syllogisms, and the Wason selection task (1). We evaluate state of the art large language models,

as well as humans, and find that the language models reflect many of the same patterns observed in humans across these tasks — like humans, models answer more accurately when the semantic content of a task supports the logical inferences. These parallels are reflected both in answer patterns, and in lower-level features like the relationship between model answer distributions and human response times. Our findings have implications for understanding both these cognitive effects in humans, and the factors that contribute to language model performance.

Introduction

A hallmark of abstract reasoning is the ability to systematically perform algebraic operations over variables that can be bound to any entity (2, 3): 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 (2). The capacity for reliable and consistent abstract reasoning is frequently highlighted as a crucial missing component of current AI (4, 5, 6). For example, while large Language Models (LMs) exhibit some impressive *emergent* behaviors, including some performance on abstract reasoning tasks (7, 8, 9, 10, 11; though cf. 12), they have been criticized for failing to achieve systematic consistency in their abstract reasoning (e.g. 13, 14, 15, 16).

However, humans — arguably the best known instances of general intelligence — are far from perfectly rational abstract reasoners (17, 18, 19). Patterns of biases in human reasoning have been studied across a wide range of tasks and domains (18). Here, we focus in particular on 'content effects' — the finding that humans are affected by the semantic content of a logical reasoning problem. In particular, humans reason more readily and more accurately about familiar, believable, or grounded situations, compared to unfamiliar, unbelievable, or abstract ones.

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.
```

humans are much less likely to say it is valid (20, 21, 22) — despite the fact that the arguments above are logically equivalent (both are invalid). Similarly, humans struggle to reason about how to falsify conditional rules involving abstract attributes (1, 23), but reason more readily about logically-equivalent rules grounded in realistic situations (24, 25, 26). 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 (27).

The literature on human cognitive biases is extensive, but many of these biases can be idiosyncratic and context-dependent. For example, even some of the seminal findings in the influential work of Kahneman et al. (18), like 'base rate neglect', are sensitive to context and experimental design (28, 29), with several studies demonstrating exactly the opposite effect in a different context (30). However, the content effects on which we focus have been a notably consistent finding and have been documented in humans across different reasoning tasks and domains: deductive and inductive, or logical and probabilistic (31, 1, 32, 33, 20, 27). This ubiquity is notable and makes these effects harder to explain as idiosyncracies. This ubiquitous sensitivity to content is in direct contradiction with the definition of abstract reasoning: that it is independent of content, and speaks directly to longstanding debates over the fundamental nature

of human intelligence: are we best described as algebraic symbol-processing systems (2, 34), or emergent connectionist ones (35, 36) whose inferences are grounded in learned semantics? Yet explanations or models of content effects in the psychological sciences often focus on a single (task and content-specific) phenomenon and invoke bespoke mechanisms that only apply to these specific settings (e.g. 25). Could content effects be explained more generally? Could they emerge from simple learning processes over naturalistic data?

In this work, we address these questions, by examining whether language models show this human-like blending of logic with semantic content effects. Language models possess 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" (37) paradigm is to endow a model with broadly accurate prior knowledge that enables learning a new task rapidly. Thus, language model representations often *reflect* human semantic cognition; e.g., language models reproduce patterns like association and typicality effects (38, 39), and language model predictions can reproduce human knowledge and beliefs (40, 41, 42, 43). In this work, we explore whether this prior knowledge impacts a language model's performance in logical reasoning tasks. While a variety of recent works have explored biases and imperfections in language models' performance (e.g. 13, 14, 15, 44, 16), we focus on the specific question of whether content interacts with logic in these systems as it does in humans. This question has significant implications not only for characterizing LMs, but potentially also for understanding human cognition, by contributing new ways of understanding the balance, interactions, and trade-offs between the abstract and grounded capabilities of a system.

We explore how the content of logical reasoning problems affects the performance of a range of large language models (45, 46, 47). To avoid potential dataset contamination, we create entirely new datasets using designs analogous to those used in prior cognitive work, and we also collect directly-comparable human data with our new stimuli. We find that language models

reproduce human content effects across three different logical reasoning tasks (Fig. 1). We first explore a simple Natural Language Inference (NLI) task, and show that models and humans answer fairly reliably, with relatively modest influences of content. We then examine the more challenging task of judging whether a syllogism is a valid argument, and show that models and humans are biased by the believability of the conclusion. We finally consider realistic and abstract/arbitrary versions of the Wason selection task (1) — a task introduced over 50 years ago that demonstrates a failure of systematic human reasoning — and show that models and humans perform better with a realistic framing. Our findings with human participants replicate and extend existing findings in the cognitive literature. We also report novel analyses of item-level effects, and the effect of content and items on continuous measures of model and human responses. We close with a discussion of the implications of these findings for understanding human cognition as well as language models.

Evaluating content effects on logical tasks

In this work, we evaluate content effects on three logical reasoning tasks, which are depicted in Fig. 1. These three tasks involve different types of logical inferences, and different kinds of semantic content. However, these distinct tasks admit a consistent definition of content effects: the extent to which reason is facilitated in situations in which the semantic content supports the correct logical inference, and correspondingly the extent to which reasoning is harmed when semantic content conflicts with the correct logical inference (or, in the Wason tasks, when the content is simply arbitrary). We also evaluate versions of each task where the semantic content is replaced with nonsense non-words, which lack semantic content and thus should neither support nor conflict with reasoning performance. (However, note that in some cases, particularly the Wason tasks, changing to nonsense content requires more substantially altering the kinds of inferences required in the task; see Methods.)

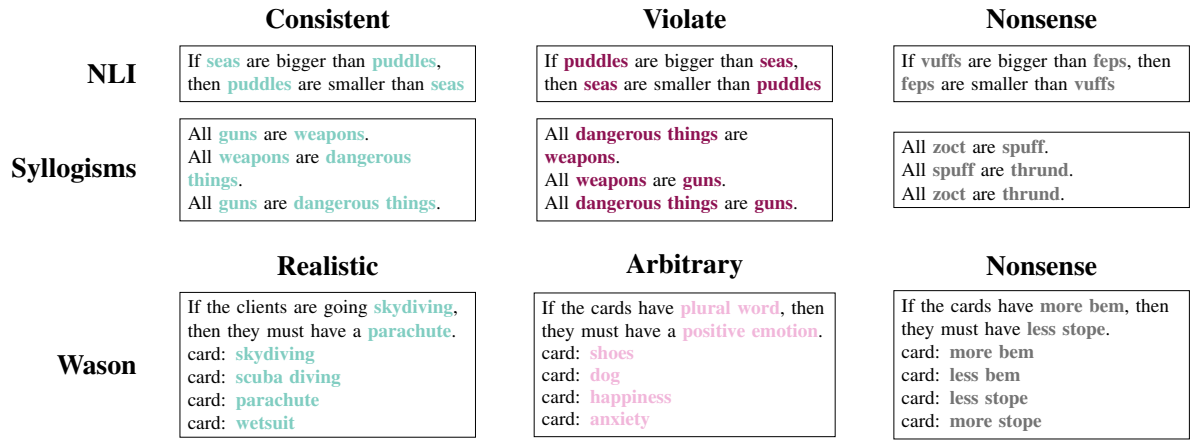


Figure 1: Manipulating content within fixed logical structures. In each of our three datasets (rows), we instantiate different versions of the logical problems (columns). Different versions of a problem offer the same logical structures and tasks, but instantiated with different entities or relationships between those entities. The relationships in a task may either be consistent with, or violate real-world semantic relationships, or may be nonsense, without semantic content. In general, humans and models reason more accurately about belief-consistent or realistic situations or rules than belief-violating or arbitrary ones.

Natural Language Inference The first task we consider has been studied extensively in the natural language processing literature (48). In the classic Natural Language Inference (NLI) 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 crowd-sourced (49) 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. (50) to generate comparisons (e.g. `X is smaller than Y`). We then give participants an incomplete inference such as "If puddles are bigger than seas, then..." and give them a forced choice between two possible hypotheses to complete it: "seas are bigger than puddles" and "seas are smaller than puddles." Note that one of these completions is consistent with real-world semantic beliefs i.e. 'believable' while the other is logically consistent with the premise but contradicts real world beliefs. We can then evaluate whether models and humans answer more accurately when the logically correct hypothesis is

believable; that is, whether the content affects their logical reasoning.

However, content effects are generally more pronounced in difficult tasks that require extensive logical reasoning (33, 21). We therefore consider two more challenging tasks where human content effects have been observed in prior work.

Syllogisms Syllogisms (51) are 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 (20, 22, 52). For example, Evans et al. (20) 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 very 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. 53). These "belief-bias" effects have been replicated and extended in various subsequent studies (22, 53, 54, 55). We similarly evaluate whether models and humans are more likely to endorse an argument as valid if its conclusion is believable, or to dismiss it as invalid if its conclusion is unbelievable.

The Wason Selection Task The Wason Selection Task (1) 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 (1) 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 upon the content of the problem. Past work has found that if an identical logical structure is instantiated in a common situation, particularly a social rule, participants are much more accurate (56, 24, 25, 26). For example, if participants are told the cards represent people, and the rule is “if they are drinking alcohol, then they must be 21 or older” and the cards show ‘beer’, ‘soda’, ‘25’, ‘16’, then many more participants correctly choose to check the cards showing ‘beer’ and ‘16’. We therefore similarly evaluate whether language models and humans are facilitated in reasoning about realistic rules, compared to more-abstract arbitrary ones. (Note that in our implementations of the Wason task, we forced participants and language models to choose exactly two cards, in order to most closely match answer formats between the humans and language models.)

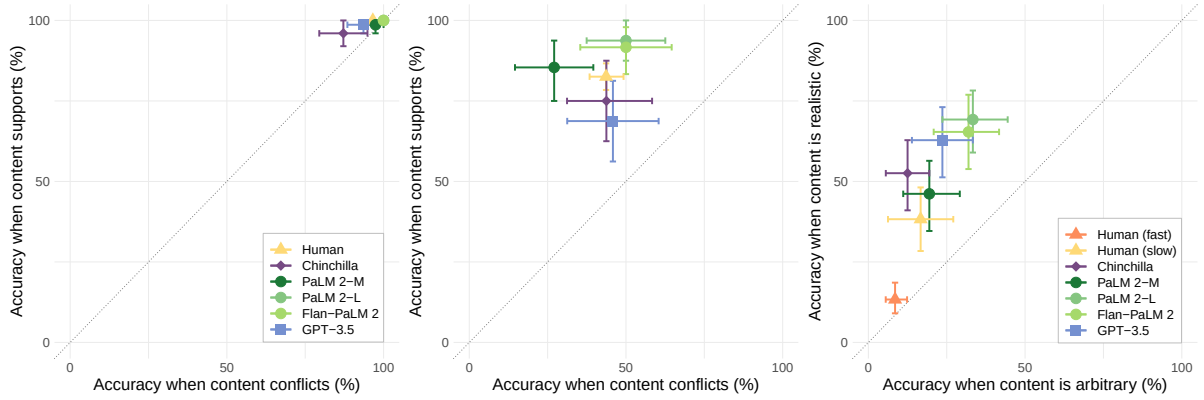
The extent of content effects on the Wason task are also affected by background knowledge; education in mathematics appears to be associated with improved reasoning in abstract Wason tasks (57, 58). However, even those experienced participants were far from perfect — undergraduate mathematics majors and academic mathematicians achieved less than 50% accuracy at the arbitrary Wason task (57). This illustrates the challenge of abstract logical reasoning, even for experienced humans. As we will see in the next section, many human participants did struggle with the abstract versions of our tasks.

Results

Content effects on accuracy

We summarize our primary results in Fig. 2. In each of our three tasks, humans and models show similar levels of accuracy across conditions. Furthermore, humans and models show similar content effects on each task, which we measure as the degree of advantage when reasoning about logical situations that are consistent with real-world relationships or rules. In the simplest Natural Language Inference task, humans and all models show high accuracy and relatively minor effects of content. When judging the validity of syllogisms, both humans and models show more moderate accuracy, and significant advantages when content supports the logical inference. Finally, on the Wason selection task, humans and models show even lower accuracy, and again substantial content effects. We describe each task, and the corresponding results and analyses, in more detail below.

Natural Language Inference The relatively simple logical reasoning involved in this task means that both humans and models exhibit high performance, and correspondingly show relatively little effect of content on their reasoning (Fig. 3). Specifically, we do not detect a statistically-significant effect of content on accuracy in humans or any of the language models in mixed-effects logistic regressions controlling for the random effect of items (or χ^2 tests where regressions did not converge due to ceiling effects; all $z < 1.21$ or $\chi^2 < 0.1$, all $p > 0.2$; see Appx. C.1 for full results). However, we do find a statistically significant relationship between human and model accuracy at the item level ($t(832) = 3.49$, $p < 0.001$; Appx. 27) — even when controlling for condition. Furthermore, as we discuss below, further investigation into the model confidence shows evidence of content effects on this task as well.



(a) Natural language inference. (b) Syllogisms. (c) Wason selection task.

Figure 2: Across the three tasks we consider, various language models and humans show similar patterns of overall accuracy and content effects on reasoning. The vertical axis shows accuracy when the content of the problems supports the logical inference. The horizontal axis shows accuracy when the content conflicts (or, in the Wason task, when it is arbitrary). Thus, points above the diagonal indicate an advantage when the content supports the logical inference. (a) On basic natural language inferences, both humans and LMs demonstrate fairly high accuracy across all conditions, and thus relatively little effect of content. (a) When identifying whether syllogisms are logically valid or invalid, both humans and LMs exhibit moderate accuracy, and substantial content effects. (c) On the Wason selection task, the majority of humans show fairly poor performance overall. However, the subset of subjects who are take the longest to answer show somewhat higher accuracy, but primarily on the realistic tasks — i.e. substantial content effects. On this difficult task, language models generally exceed humans in both accuracy and magnitude of content effects. (Throughout, errorbars are bootstrap 95%-CIs, and dashed lines are chance performance.)

Syllogisms Syllogism validity judgements are significantly more challenging than the NLI task above; correspondingly, we find lower accuracy in both humans and language models. Nevertheless, humans and most language models are sensitive to the logical structure of the task. However, we find that both humans and language models are strongly affected by the content of the syllogisms (Fig. 4), as in the past literature on syllogistic belief bias in humans (21). Specifically, if the semantic content supports the logical inference — that is, if the conclusion is believable and the argument is valid, or if the conclusion is unbelievable and the argument is invalid — both humans and all language models tend to answer more accurately (all $z \geq 2.25$

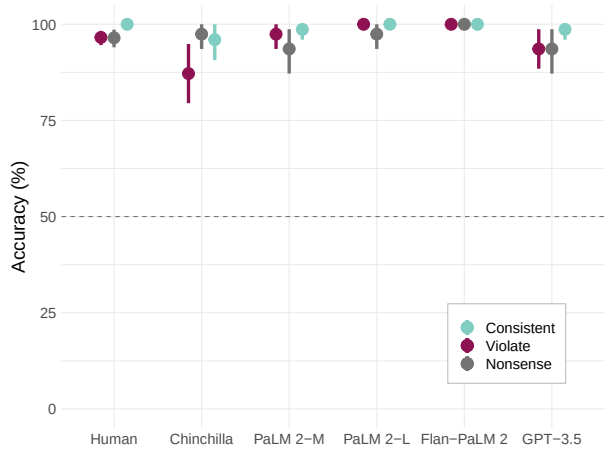


Figure 3: Detailed results on the Natural Language Inference tasks. Both humans (left) and all models show relatively high performance, and relatively little difference in accuracy between belief-consistent and belief-violating inferences, or even nonsense ones.

or $\chi^2 > 6.39$, all $p \leq 0.01$; see Appx. C.2 for full results).

Two simple effects contribute to this overall content effect — that belief-consistent conclusions are judged as logically valid and that belief-inconsistent conclusions are judged as logically invalid. As in the past literature, we find that the dominant effect is that humans and models tend to say an argument is valid if the conclusion is belief-consistent, regardless of the actual logical validity. If the conclusion is belief-violating, humans and models do tend to say it is invalid more frequently, but most humans and models are more sensitive to actual logical validity in this case. Specifically, we observe a significant interaction between the content effect and believability in humans, PaLM 2-L, Flan-PaLM 2, and GPT-3.5 (all $z > 5.9$ or $\chi^2 > 14.3$, all $p < 0.001$); but do not observe a significant interaction in Chinchilla or PaLM 2-M (both $\chi^2 < 0.001$, $p > 0.99$). Both humans and models appear to show a slight bias towards saying syllogisms with nonsense words are valid, but again with some sensitivity to the actual logical structure.

Furthermore, even when controlling for condition, we observe a significant correlation between item-level accuracy in humans and language models ($t(345) = 4.98$, $p < 0.001$), sug-

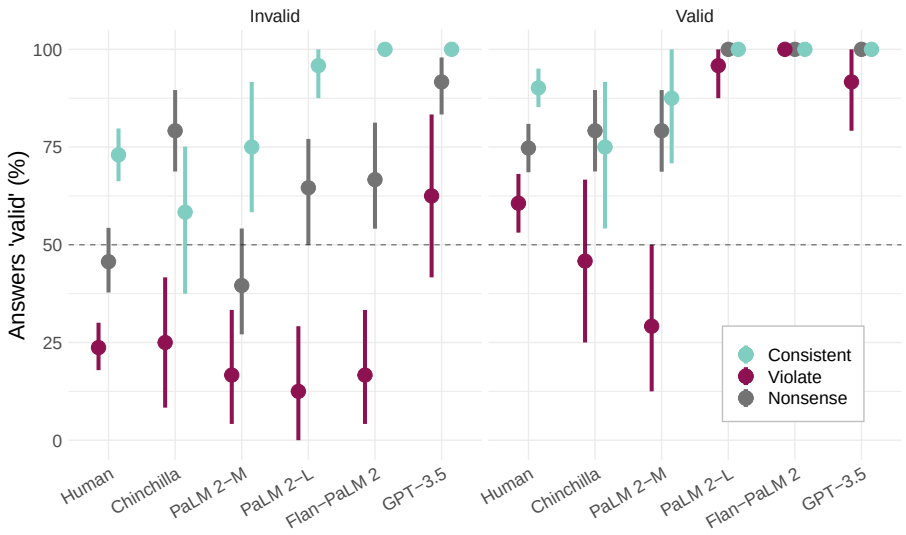


Figure 4: Detailed results on syllogism validity judgements. The vertical axis shows the proportion of the time that each system answers that an argument is valid. Both humans and models exhibit substantial content effects — they are strongly biased towards saying an argument is valid if the conclusion is consistent with expectations (cyan), and somewhat biased towards saying the argument is invalid if the conclusion violates expectations (maroon). If the argument contains nonsense words (grey), both humans and models show a slight bias towards saying "valid." (Note that this figure plots the proportion of the time the humans or models answer 'valid' rather than raw accuracy, to more clearly illustrate the bias. To see accuracy, simply reverse the vertical axis for the invalid arguments.)

gesting shared patterns in the use of lower-level details of the logic or content.

The Wason Selection Task As in the prior human literature, we found that the Wason task was relatively challenging for humans, as well as for language models (Fig. 5). Nevertheless, we observed significant content advantages for the Realistic tasks in humans, and in Chinchilla, PaLM 2-L, and GPT-3.5 (all $z > 2.2$, all $p < 0.03$; Appx. C.3). We only observed marginally significant advantages of realistic rules in PaLM 2-M and Flan-PaLM 2 (both $z \geq 1.78$, both $p \leq 0.08$), due to stronger item-level effects in these models (though the item-level variance does not seem particularly unusual; see Appx. B.7.3 for further analysis). Intriguingly, some language models also show better performance at the versions of the tasks with Nonsense nouns

compared to the Arbitrary ones, though generally Realistic rules are still easier. We also consider several variations on these rules in Appx. 22.

Our human participants struggled with this task, as in prior research, and did not achieve significantly higher-than-chance performance overall — although their behavior is not random, as we discuss below, where we analyze answer choices in more detail. However, spending longer on logical tasks can improve performance (59, 60), and thus many studies split analyses by response time to isolate participants who spend longer, and therefore show better performance (61, 62). Indeed, we found that human accuracy was significantly associated with response time ($z = 4.44$, $p < 0.001$; Appx. C.3.1). We depict this relationship in Fig. 6. To visualize the the performance of discrete subjects in our Figures 2c and 5, we split subjects into 'slow' and 'fast' groups. The distribution of times taken by subjects is quite skewed, with a long tail. We separate out the top 15% of subjects that take the longest, who spent more than 80 seconds on the problem, as the slow group. These subjects showed above chance performance in the Realistic condition, but still performed near chance in the other conditions. We also dig further into the predictive power of human response times in the other tasks in the following sections (and Appx. B.6.1).

We collected the data for the Wason task in two different experiments; after observing the lower performance in the first sample, we collected a second sample where we offered a performance bonus for this task. We did not observe significant differences in overall performance or content effects between these subsets, so we collapse across them in the main analyses; however, we present results for each experiment and some additional analyses in Appx. B.5.

Robustness of results to factors like removing instructions, few-shot prompting and scoring methods Language model behavior is frequently sensitive to details of the evaluation. Thus, we performed several experiments to confirm that our results were robust to details of the

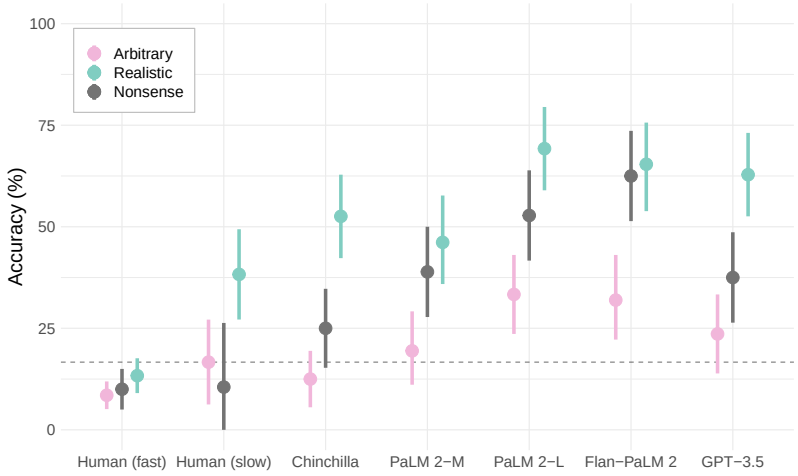


Figure 5: Detailed results on the Wason selection task. Human performance is low, even on the Realistic rules. In particular, the majority of the subjects show at-or-below-chance accuracy in all conditions (though this behavior is not random; see below). However, the subset of subjects who answer more slowly show above chance accuracy for the realistic rules (cyan), but not for the arbitrary ones (pink). This pattern matches the prior results in the cognitive literature. Furthermore, each of the language models reproduces this pattern of advantage for the realistic rules. In addition, two of the larger models perform above chance at the arbitrary rules. (The dashed line corresponds to chance — a random choice of two cards among the four shown. Both models and humans were forced to choose exactly two cards.)

methods used. We present these results in full in Appx. B.2, but we outline the key experiments here. First, we show that removing the pre-question instructions does not substantially alter the overall results (Appx.B.2.1). Next, we show that our use of the DC-PMI correction for scoring is not the primary driver of content effects (Appx. B.2.2). On the syllogisms tasks, raw likelihood scoring with the instruction prompt yields strong answer biases — several models say every argument is valid irrespective of actual logical validity or content. However, the models that don't uniformly say valid show content effects as expected. Furthermore, if the instructions before the question are removed, raw likelihood scoring results in less validity bias, and again strong content effects. For the Wason task, raw likelihood scoring actually improves the accuracy of some models; however, again the content effects are as found with the DC-PMI scoring. Thus, although overall model accuracy and response biases change with uncorrected

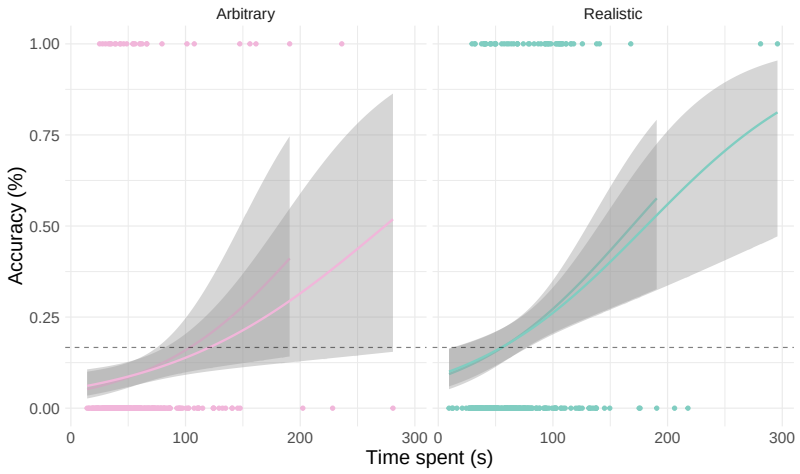


Figure 6: There is a strong relationship between response time and answer accuracy in the Wason tasks; subjects who take longer to answer are more accurate on average. Participants who take sufficiently long to answer perform above chance in the Realistic tasks. There are hints of a similar effect in the Arbitrary condition, but we do not have the power to detect it. (Curves are logistic regression fits, with 95% CIs. We also plot regressions dropping outliers with time greater than 180 seconds, to show that the effect is not driven solely by outliers.)

likelihood scoring, the content effects are similar. Finally, we consider few-shot evaluation, and show that giving few-shot examples yields some mild improvements in accuracy (with greater improvement in the simpler tasks), but does not eliminate the content effects (Appx. B.2.4). Together, these results suggest that our findings are not strongly driven by idiosyncratic details of our evaluation, and thus support the robustness of our findings.

Variability across different language models While we generally find similar content effects across the various models we evaluate, there are some notable differences among them. First, across tasks the larger models tend to be more accurate overall (e.g., comparing the large vs. the medium variants of PaLM 2); however, this does not necessarily mean they show weaker content effects. While it might be expected that instruction-tuning would affect performance, the instruction-tuned models (Flan-PaLM 2 and GPT-3.5-turbo-instruct) do not show consistent differences in overall accuracy or content effects across tasks compared to the base language

models—in particular, Flan-PaLM 2 performs quite similarly to PaLM 2-L overall. (However, there are some more notable differences in the distributions of log-probabilities the instruction-tuned models produce; Appx. B.8.)

On the syllogisms task in particular, there are some noticeable difference among the models. GPT-3.5, and the larger PaLM 2 models, have quite high sensitivity for identifying valid arguments (they generally correctly identify valid arguments) but relatively less specificity (they also consider several invalid conclusions valid). By contrast, PaLM 2-M and Chinchilla models answer more based on content rather than logical validity i.e. regularly judging consistent conclusions as more valid than violating ones, irrespective of their logical validity. The sensitivity to logical structure in the nonsense condition also varies across models – the PaLM models are fairly sensitive, while GPT 3.5 and Chinchilla both having a strong bias toward answering valid to all nonsense propositions irrespective of actual logical validity.

On the Wason task, the main difference of interest is that the PaLM 2 family of models show generally greater accuracy on the Nonsense problems than the other models do, comparable to their performance on the Realistic condition in some cases.

Model confidence is related to content, correctness, and human response times

Language models do not produce a single answer; rather, they produce a probability distribution over the possible answers. This distribution can provide further insight into their processing. For example, the probability assigned to the top answers, relative to the others, can be interpreted as a kind of confidence measure. By this measure, language models are often somewhat calibrated, in the sense that the probability they assign to the top answer approximates the probability that their top answer is correct (e.g. 63). Furthermore, human Response Times (RTs) relate to many similar variables, such as confidence, surprisal, or task difficulty; thus, many prior works have

related language model confidence to human response or reaction times for linguistic stimuli (e.g. 64, 65). In this section, we correspondingly analyze how the language model confidence relates to the task content and logic, the correctness of answers, and the human response times.

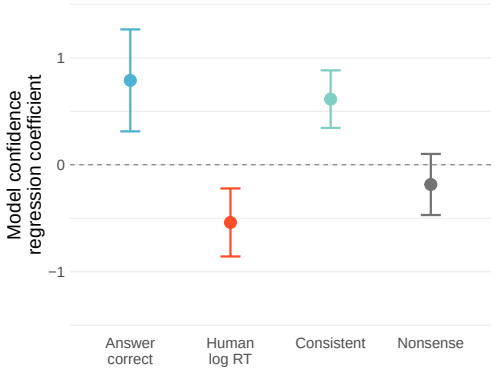
We summarize these results in Fig. 7. We measure model confidence as the difference in prior-corrected log-probability between the top answer and the second highest—thus, if the model is almost undecided between several answers, this confidence measure will be low, while if the model is placing almost all its probability mass on a single answer, the confidence measure will be high. In mixed-effects regressions predicting model confidence from task variables and average human RTs on the same problem, we find a variety of interesting effects. First, language models tend to be more confident on correct answers (that is, they are somewhat calibrated). Task variables also affect confidence; models are generally less confident when the conclusion violates beliefs, and more confident for the realistic rules on the Wason task.Furthermore, even when controlling for task variables and accuracy, there is a statistically-significant negative association with human response times on the NLI and syllogisms tasks (respectively $t(655) = -3.39$, $p < 0.001$; and $t(353) = -2.03$, $p < 0.05$; Appx. C.4)—that is, models tend to show more confidence on problems where humans likewise respond more rapidly. We visualize this relationship in Fig. 8.

Analyzing components of the Wason responses

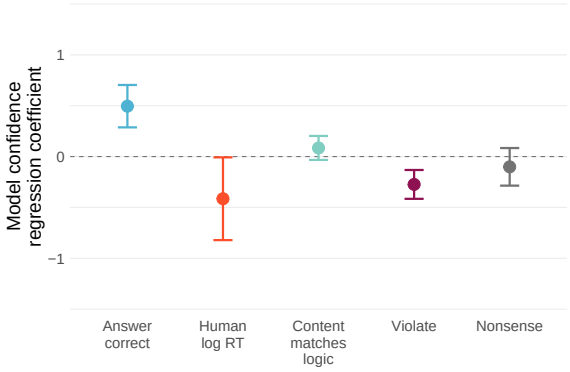
Because each answer to the Wason problems involves selecting a pair of cards, we further analyzed the individual cards chosen. The card options presented each problem are designed so that two cards respectively match and violate the antecedent, and similarly for the consequent. The correct answer is to choose one card for the antecedent and one for the consequent; more precisely, the card for which the antecedent is true (AT), and the card for which the consequent is false (CF). In Fig. 9 we examine human and model choices; we quantitatively analyze these

choices using a multinomial logistic regression model in Appx. C.3.2.

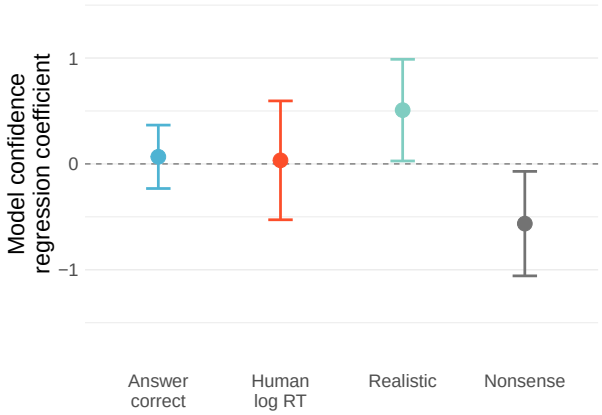
Even in conditions when performance is close to chance, behavior is generally not random. As in prior work, humans do not consistently choose the correct answer (AT, CF). Instead, humans tend to exhibit the matching bias; that is, they tend to choose each of the two cards that match each component of the rule (AT, CT). However, in the Realistic condition, slow humans answer correctly somewhat more frequently. Humans also exhibit errors besides the matching bias; including an increased rate of choosing the two cards corresponding to a single component of the rule — either both antecedent cards, or both consequent cards. Language models tend to give more correct responses than humans, and to show facilitation in the realistic rules compared to arbitrary ones. Relative to humans, language models show fewer matching errors, fewer errors of choosing two cards from the same rule component, but more errors of choosing the antecedent false options. These differences in error patterns may indicate differences between the response processes engaged by the models and humans. (Note, however, that while the models accuracies do not change too substantially with alternate scoring methods, the particular errors the models make are somewhat sensitive to scoring method — without the DC-PMI correction the model errors more closely approximate the human ones in some cases; Appx. B.2.3.)



(a) Natural language inference.

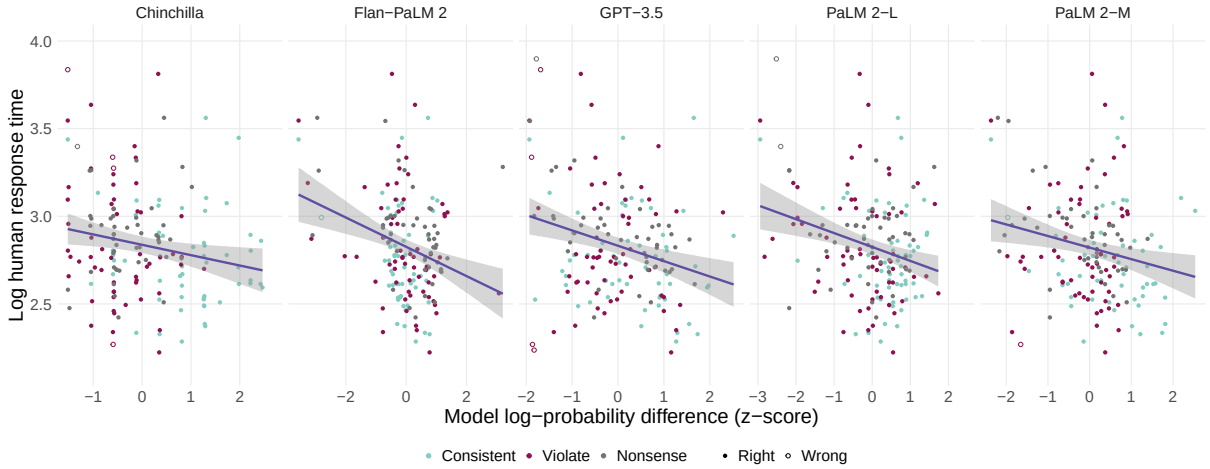


(b) Syllogisms.

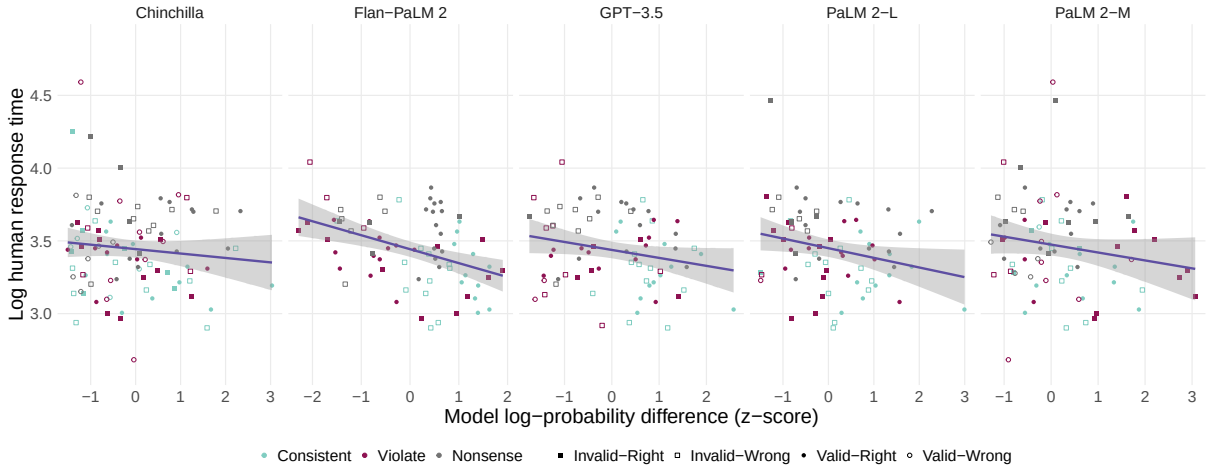


(c) Wason selection task.

Figure 7: Language model confidence—as measured by the difference in(prior-corrected) log-probability between the chosen answer and the next most probable—is associated with correct answers, task variables, and human average response times. (a-b) On the NLI and syllogism tasks, models are generally more confident in correct answers and belief-consistent conditions, less confident in belief-violating conditions, and less confident on problems that humans take longer to answer. (c) On the Wason task, effects are weaker. Human RT and correct answers are not associated with confidence; however, the models do show more confidence on Realistic problems, and less on Nonsense ones. (Effects are calculated from a mixed-effects regression predicting the difference in log-probability between the top and second-highest answer, z-scored within each model, and controlling for all other significant predictors. Errorbars are parametric 95%-CIs. Note that human RT is calculated across all human subjects for the Wason task, not just slow subjects.)



(a) Natural language inference raw results.



(b) Syllogisms raw results.

Figure 8: Human response times are generally negatively related to model confidence (measured as the difference in log-probabilities between the correct answer and the incorrect answer). That is, on problems for which the model displays greater confidence, humans tend to respond more quickly. This relationship holds on both (a) the NLI tasks, and (b) the syllogism tasks. (Points show average response times for individual problems, broken down by whether the humans or models answered correctly or not; see Appx. C.4 for details.)

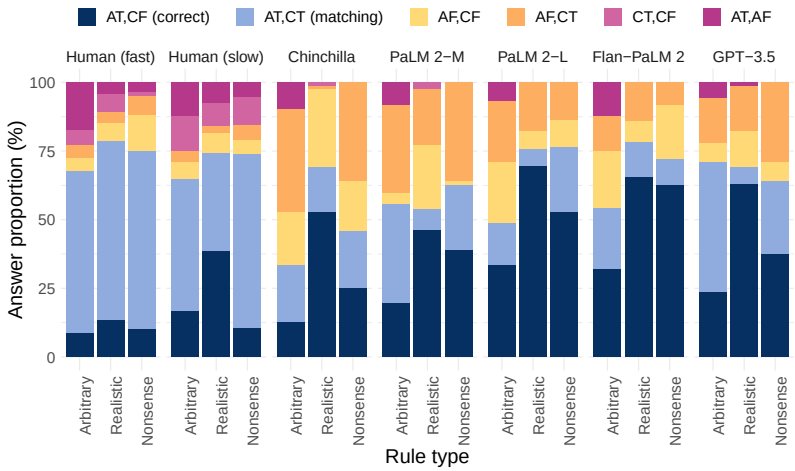


Figure 9: Answer patterns for the Wason tasks, broken down into components: the pairings of individual cards that each participant chose (AT = Antecedent True, CF = Consequent False, etc.). Behavior is not random, even when performance is near chance. As above, humans do not consistently choose the correct answer (AT, CF; dark blue); instead, humans more frequently exhibit the matching bias (AT, CT; light blue). Humans also show other errors, however, including a surprisingly high rate of choosing two cards corresponding to a single rule component (dark/-light pink). Language models answer correctly more often than humans, but intriguingly choose options with the antecedent false and a consequent card (yellow/orange) more frequently. (Note that all participants and language models were forced to choose exactly two cards.)

Discussion

Humans are imperfect reasoners. We reason most effectively about entities and situations that are consistent with our understanding of the world. Even in these familiar cases, we often make mistakes. Our experiments show that language models mirror these patterns of behavior. Language models likewise perform imperfectly on logical reasoning tasks, but this performance depends on content and context. Most notably, such models often fail in situations where humans fail — when stimuli become too abstract or conflict with prior expectations about the world.

Beyond these simple parallels in accuracy across different conditions and items, we also observed more subtle parallels in language model confidence. The model's confidence tends to be higher for correct answers, and for cases where prior expectations about the content are consistent with the logical structure. Even when controlling for these effects, model confidence is related to human response times. Thus, language models reflect human content effects on reasoning at multiple levels. Furthermore, these core results are generally robust across different language models with different training and tuning paradigms, different prompts, etc., suggesting that they are a fairly general phenomenon of predictive models that learn from human-generated text.

Prior research on language model reasoning. Since Brown et al. (7) showed that large language models could perform moderately well on some reasoning tasks, there has been a growing interest in language model reasoning (66). Typical methods focus on prompting for sequential reasoning (9, 67, 10), altering task framing (68, 69) or iteratively sampling answers (70).

In response, some researchers have questioned whether these language model abilities qualify as "reasoning". The fact that language models sometimes rely on "simple heuristics" (15), or reason more accurately about frequently-occurring numbers (14), 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" (2; cf. 34). In this work, we emphasize how *both* humans and language models rely on content when answering reasoning problems — using simple heuristics in some contexts, and answering more accurately about frequently-occurring situations (71, 28). 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 (72, 73, 74). The dual system idea has more recently become popular (75, 76), including in machine learning (e.g. 77). 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. 78). These calls to action usually advocate for an explicit duality; with a neural network based system providing the system 1 and a system with more explicit symbolic or otherwise structured system being the system 2.

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 and task difficulty. 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 (9, 67, 10).

These observations suggest the possibility that the unitary language model may have implicitly learned a context-dependent control mechanism that arbitrates between conflicting re-

sponses (such as more intuitive answers vs. logically correct ones). This perspective suggests several possible directions for future research. First, it would be interesting to seek out mechanistic evidence of such as conflict-arbitration process within language models. Furthermore, it suggests that augmenting language models with a second system might not be necessary to achieve relatively reliable performance. Instead, it might be sufficient to further develop the control mechanisms within these models by altering their context and training, as we discuss below.

From a human cognitive neuroscience perspective, these issues are more complex. The idea of context-dependent arbitration between conflicting responses has been influential in the literature on human cognitive control (79, 80), and has been implicated in humans reasoning successfully in tasks that require following novel, arbitrary reasoning procedures or over-riding pre-existing response tendencies (81, 82). However, these control processes are generally believed to principally reside in frontal regions outside the language areas, or in a network of control-related brain areas that interface with the language regions and other domain-specific brain areas but is not housed therein (81). Thus, understanding the full detail of human cognition in such language-based logical tasks may require incorporating a control network into the architecture more explicitly. Nevertheless, our results with language models suggest that this controller could be more intertwined with the statistical inference system than it would be in a classic dual-systems model; moreover, that the controller does not need to be implemented as a classical symbol system to achieve human-competitive logical reasoning performance.

Neural mechanisms of human reasoning. Deep learning models are increasingly used as models of neural processing in biological systems (e.g. 83, 84), 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 (85, 86),

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" (87; see also 88, 89). Language models also predict low-level behavioral phenomena; e.g. surprisal predicts reading time (64, 90). 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. This is particularly exciting because neural models for these phenomena in humans are currently lacking or incomplete. Indeed, even prior high-level explanations of these phenomena have often focused on only a single task, such as explaining *only* the Wason task content effects with appeals to evolved social-reasoning mechanisms (25). Our results suggest that there could be a more general explanation.

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. 74, 52). Others interpret them as a rational adaptation to reasoning under constraints such as limited memory or time (e.g. 91, 92, 93) — where content effects actually support fast and effective reasoning in commonly encountered tasks (71, 28). Our results show that content effects can emerge from simply training a large transformer to imitate language produced by human culture, without explicitly incorporating any human-specific internal mechanisms.

This observation suggests two possible origins for these content effects. First, the content

effects could be directly learned from the humans that generated the data used to train the language models. Under this hypothesis, poor logical inferences about nonsense or belief-violating premises come from *copying* the incorrect inferences made by humans about these premises. Since humans also learn substantially from other humans and the cultures in which we are immersed, it is plausible that both humans and language models could acquire some of these reasoning patterns by imitation.

The other possibility is that (like humans) the model's exposure to the world reflects semantic truths and beliefs and that language models and humans both converge on these content biases that reflect this semantic content for more task-oriented reasons: because it helps humans to draw more accurate inferences in the situations they encounter (which are mostly familiar and believable), and helps language models to more accurately predict the (mostly believable) text that they encounter. In either case, humans and models acquire surprisingly similar patterns of behavior, from seemingly very different architectures, experiences, and training objectives. A promising direction for future enquiry would be to causally manipulate features of language model's training objective and experience, to explore which features contribute to the emergence of content biases in language models. These investigations could offer insights into the origins of human patterns of reasoning, and into what data we should use to train language models.

Why might model response patterns differ from human ones? The language model response patterns do not perfectly match all aspects of the human data. For example, on the Wason task several models outperform humans on the Nonsense condition, and the error patterns on the Wason tasks are somewhat different than those observed in humans (although human error patterns also vary across populations; 57, 58). Similarly, not all models show the significant interaction between believability and validity on the syllogism tasks that humans do (20),

although it is present in most models (and the human interaction similarly may not appear in all cases; 53). Various factors could contribute to differences between model and human behaviors.

First, while we attempted to align our evaluation of humans and models as closely as possible (cf. 69), it is difficult to do so perfectly. In some cases, such as the Wason task, differences in the form of the answer are unavoidable — humans had to select answers individually by clicking on cards to select them, and then clicking continue, while models had to jointly output both answers in text, without a chance to revise their answer before continuing. Moreover, 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. 94). Similarly, presenting negative examples can help humans learn, but is generally detrimental to model performance (e.g. 95) — presumably because the model infers that the task is to sometimes output wrong answers, while humans might understand the communicative intent behind the use of negative examples. Thus, while we tried to match instructions between humans and models, it is possible that idiosyncratic details of our task framing may have caused the model to infer the task incorrectly. To minimize this risk, we tried various different prompting strategies, and where we varied these details we generally observed similar overall effects. Nevertheless, it is possible that some aspect of the problem instructions or framing contributes to the response patterns.

More fundamentally, language models do not directly experience the situations to which language refers (96); 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 (97, 98). Finally, language models experience language passively, while humans experience language as an active, conventional system for social communication (e.g. 99); active

participation may be key to understanding meaning as humans do (36, 100). 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 humans 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 (101, 102, 103, 57, 58, 104). However, causal evidence is scarce, because years of education are difficult to experimentally manipulate; thus the association may be partly due to selection effects, e.g. continuing in formal education might be more likely in individuals with stronger prior abilities. Nevertheless, the association with formal education raises an intriguing question: could language models learn to reason more reliably with targeted formal education?

Several recent results suggest that this may indeed be a promising direction. Pretraining on synthetic logical reasoning tasks can improve model performance on reasoning and mathematics problems (105, 106). In some cases language models can either be prompted or can learn to verify, correct, or debias their own outputs (107, 108, 109, 63). Finally, language model reasoning can be bootstrapped through iterated fine-tuning on successful instances (110). 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 environmental and educational factors (36, 111), rather than a core ability that must be built in to an intelligent system.

Limitations In addition to the limitations noted above — such as the challenges of perfectly aligning comparisons between humans and language models — there are several other limita-

tions to our work. First, our human participants exhibited relatively low performance on the Wason task. However, as noted above, there are well-known individual differences in these effects that are associated with factors like depth of mathematical education. We were unfortunately unable to examine these effects in our data, but in future work it would be interesting to explicitly explore how educational factors affect performance on the more challenging Wason conditions, as well as more general patterns like the relationship between model confidence and human response time. Furthermore while our experiments suggest that content effects in reasoning can emerge from predictive learning on naturalistic data, they do not ascertain precisely which aspects of the large language model training datasets contribute to this learning. Other research has used controlled training data distributions to systematically investigate the origin of language model capabilities (112, 113); it would be an interesting future direction to apply analogous methods to investigate the origin of content effects.

Materials and Methods

Creating datasets While many of these tasks have been extensively studied in cognitive science, the stimuli used in cognitive experiments are often online in articles and course materials, and thus may be present in the training data of large language models, which could compromise results (e.g. 114, 115). To reduce these concerns, we generate new datasets, by following the design 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. 1). 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"). Note that for the Wason task, this change alters the kinds of inferences that need to be made; while for the basic Wason task, matching each card to the antecedent or consequent is nontrivial (e.g. realizing that "shoes" is a plural word, not singular), it is difficult to match these inferences with Nonsense words that have no prior associations. As shown in the examples, we use a format where the cards either have more or less of a nonsense attribute, which makes the inferences perhaps more direct than other conditions (although models perform similarly on the basic inferences across conditions; Appx. 21).

In Appx. B.1 we validate the semantic content of our datasets, by showing that participants find the propositions and rules from our Consistent and Realistic stimuli much more plausible than those from other conditions.

We attempted to create these datasets in a way that could be presented to the humans and language models in precisely the same manner (for example, prefacing the problems with the same instructions for both the humans and the models).[1]

[1]N.B. this required adapting some of the problem formats and prompts compared to an earlier preprint of this paper that did not evaluate humans; see Appx. A.1.4.

Models & evaluation We evaluate several different families of language models. First, we evaluate several base LMs that are trained only on language modeling: including Chinchilla (45) a large model (with 70 billion parameters) trained on causal language modeling, and PaLM 2-M and -L (47), which are trained on a mixture of language modeling and infilling objectives (116). We also evaluate two instruction-tuned models: Flan-PaLM 2 (an instruction-tuned version of Palm 2-L), and GPT-3.5-turbo-instruct (46), which we generally refer to as GPT-3.5 for brevity.[2] We observe broadly similar content effects across all types of models, suggesting that these effects are not too strongly affected by the particular training objective, or by standard instruction-tuning.

For each task, we present the model with brief instructions that approximate the relevant portions of the human instructions. We then present the question, which ends with "Answer:" and assess the model by evaluating the likelihood of continuing this prompt with each of a set of possible answers. We apply the DC-PMI correction proposed by Holtzman et al. (117) — 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". This approach can also be interpreted as correcting for the prior over utterances. For the NLI task, however, the direct answer format means that the DC-PMI correction would therefore control for the very bias we are trying to measure. Thus, for the NLI task we simply choose the answer that receives the maximum likelihood among the set of possible answers. We also report syllogism and Wason results with maximum likelihood scoring in Appx. B.2.2; while overall accuracy changes (usually decreases, but with some exceptions), the direction of content effects is generally preserved

[2]We fortuitously performed this evaluation during the short window of time in which scoring was available on GPT-3.5-turbo-instruct.

under alternative scoring methods.

Human experiments The human experiments were conducted in 2023 using an online crowdsourcing platform, and recruiting only participants from the UK who spoke English as a first language, and who had over a 95% approval rate. We did not further restrict participation. We offered pay of £2.50 for our task. Our intent was to pay at a rate exceeding £15/h, and we exceeded this target, as most participants completed the task in less than 10 minutes.

The human participants were first presented with a consent form detailing the experiment and their ability to withdraw at any time. If they consented to participate, they then proceeded to an instructions page. After the instructions they were presented with one question from each of our three tasks, one at a time. Each participant saw the tasks in a randomized order, and with randomized conditions. Subsequently, the participants were presented with three rating questions, rating the believability of a rule from one of the Wason tasks (not the one they had completed), and their degree of agreement with a concluding proposition from a syllogism task, and a concluding proposition from the NLI task. In each case, the ratings were provided on a continuous scale from 0 to 100 (with 50% indicated as neither agree nor disagree). On each question and rating, the participants had to answer in less than a time limit of 5 minutes (to ensure they were not abandoning the task entirely). This time limit was reset on the next question. See Appx. A.3 for further detail on the experimental methods.

We first collected a dataset of responses from 625 participants. After observing the low accuracy in the Wason tasks, we collected an additional dataset from 360 participants in which we offered an additional performance bonus of £0.50 for answering the Wason question correctly, to motivate subjects. In this replication, we collected data only on the Realistic and Abstract Wason conditions. In our main analyses, we collapse across these two subsets, but we present the results for each experiment separately in Appx. 23.

Due to infrastructure restrictions in the framework used to create the human tasks, we assigned participants to conditions and items randomly rather than with precise balancing. Furthermore, a few participants timed out on some questions, and there were a handful of instances of data not saving properly due to server issues. Thus, the exact number of participants for which we have data varies slightly from task to task and item to item.

Statistical analyses Our main analyses are quantified with mixed-effects logistic regression models that include task condition variables as predictors, and control for random effects of items, and, where applicable, models. The key results of these models are reported in the main text. The full model specifications and full results are provided in Appx. C.

References

1. P. C. Wason, Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20**, 273 - 281 (1968).

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

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

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

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

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

7. 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).

8. 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).

9. 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* (2021).

10. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916* (2022).

11. 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* (2022).

12. R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004* (2023).

13. 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).

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

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

16. K. Valmeekam, A. Olmo, S. Sreedharan, S. Kambhampati, Large language models still can't plan (a benchmark for llms on planning and reasoning about change) (2022).

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

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

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

20. J. Evans, J. L. Barston, P. Pollard, On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition* **11**, 295–306 (1983).

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

22. K. C. Klauer, J. Musch, B. Naumer, On belief bias in syllogistic reasoning. *Psychological review* **107**, 852 (2000).

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

24. P. W. Cheng, K. J. Holyoak, Pragmatic reasoning schemas. *Cognitive psychology* **17**, 391–416 (1985).

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

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

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

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

29. D. J. Benjamin, Errors in probabilistic reasoning and judgment biases, *Tech. rep.*, National Bureau of Economic Research (2018).

30. C. R. Peterson, Z. Ulehla, Uncertainty, inference difficulty, and probability learning. *Journal of Experimental Psychology* **67**, 523–530 (1964).

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

32. P. C. Wason, P. N. Johnson-Laird, *Psychology of reasoning: Structure and content*, vol. 86 (Harvard University Press, 1972).

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

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

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

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

37. 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).

38. S. Bhatia, R. Richie, W. Zou, Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences* **29**, 31–36 (2019).

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

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

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

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

43. Z. Jiang, J. Araki, H. Ding, 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).

44. A. Søgaard, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 8240–8245.

45. 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).

46. Gpt-3.5, `https://platform.openai.com/docs/models/gpt-3-5`. Retrieved September 19th, 2023.

47. R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

48. B. MacCartney, C. D. Manning, *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (2007), pp. 193–200.

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

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

51. R. Smith, *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed. (Metaphysics Research Lab, Stanford University, 2020), Fall 2020 edn.

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

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

54. D. Trippas, M. F. Verde, S. J. Handley, Using forced choice to test belief bias in syllogistic reasoning. *Cognition* **133**, 586–600 (2014).

55. M. H. Tessler, Understanding belief bias by measuring prior beliefs for a bayesian model of syllogistic reasoning. *Proceedings of ESSLLI* pp. 225–237 (2015).

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

57. M. Inglis, A. Simpson, Mathematicians and the selection task. *International Group for the Psychology of Mathematics Education* (2004).

58. C. Cresswell, C. P. Speelman, Does mathematics training lead to better logical thinking and reasoning? a cross-sectional assessment from students to professors. *PLOS ONE* **15**, 1-21 (2020).

59. J. S. B. Evans, J. Curtis-Holmes, Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning* **11**, 382–389 (2005).

60. J. S. B. Evans, S. Newstead, J. Allen, P. Pollard, Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology* **6**, 263–285 (1994).

61. W. A. Wickelgren, Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica* **41**, 67–85 (1977).

62. S. L. Wise, X. Kong, Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education* **18**, 163–183 (2005).

63. 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, J. Kaplan, Language models (mostly) know what they know (2022).

64. A. Goodkind, K. Bicknell, *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (2018), pp. 10–18.

65. R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, R. Levy, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), pp. 32–42.

66. M. Binz, E. Schulz, Using cognitive psychology to understand gpt-3 (2022).

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

68. D. Khashabi, C. Baral, Y. Choi, H. Hajishirzi, *Findings of the Association for Computational Linguistics: ACL 2022* (2022), pp. 589–612.

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

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

71. H. Mercier, D. Sperber, *The Enigma of Reason* (Harvard University Press, 2017).

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

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

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

75. D. Kahneman, *Thinking, fast and slow* (Macmillan, 2011).

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

77. Y. Bengio, The consciousness prior. *arXiv preprint arXiv:1709.08568* (2017).

78. M. Nye, M. Tessler, J. Tenenbaum, 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 (2021).

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

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

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

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

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

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

85. R. Cao, D. Yamins, Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490* (2021).

86. R. Cao, D. Yamins, Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489* (2021).

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

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

89. 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**, 369–380 (2022).

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

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

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

93. H. A. Simon, *Utility and probability* (Springer, 1990), pp. 15–18.

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

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

96. J. L. McClelland, F. Hill, M. Rudolph, J. Baldridge, 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**, 25966–25974 (2020).

97. M. W. Alibali, R. Boncoddo, A. B. Hostetter, Gesture in reasoning: An embodied perspective. *The Routledge handbook of embodied cognition* p. 150 (2014).

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

99. H. H. Clark, *Using language* (Cambridge university press, 1996).

100. D. Schlangen, Norm participation grounds language. *arXiv preprint arXiv:2206.02885* (2022).

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

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

103. N. Attridge, A. Aberdein, M. Inglis, Does studying logic improve logical reasoning? (2016).

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

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

106. Y. Wu, M. N. Rabe, W. Li, J. Ba, R. B. Grosse, C. Szegedy, *International Conference on Machine Learning* (PMLR, 2021), pp. 11251–11262.

107. T. Schick, S. Udupa, 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 (2021).

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

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

110. E. Zelikman, Y. Wu, N. D. Goodman, Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465* (2022).

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

112. S. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. McClelland, F. Hill, Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems* **35**, 18878–18891 (2022).

113. B. Prystawski, N. D. Goodman, Why think step-by-step? reasoning emerges from the locality of experience. *arXiv preprint arXiv:2304.03843* (2023).

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

115. J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021), pp. 1286–1305.

116. Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, S. Zheng, *et al.*, *The Eleventh International Conference on Learning Representations* (2022).

117. A. Holtzman, P. West, V. Shwartz, Y. Choi, L. Zettlemoyer, Surface form competition:

Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315* (2021).

118. A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, *International Conference on Learning Representations* (2019).

119. A. Hosseini, S. Reddy, D. Bahdanau, R. D. Hjelm, A. Sordoni, A. Courville, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Online, 2021), pp. 1301–1312.

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

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

Funding: This work was not supported by external funding.

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

Competing Interests The authors declare that they have no competing financial interests.

Data and materials availability: Additional data and materials are available upon request.

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

A Supplemental methods

A.1 Datasets

If seas are bigger than puddles, then puddles are...
Choice: smaller than seas
Choice: bigger than seas
Answer: **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 10: Examples of the three logical reasoning tasks we evaluate, as they were presented to the models: (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.

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 13) 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 (118) 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".

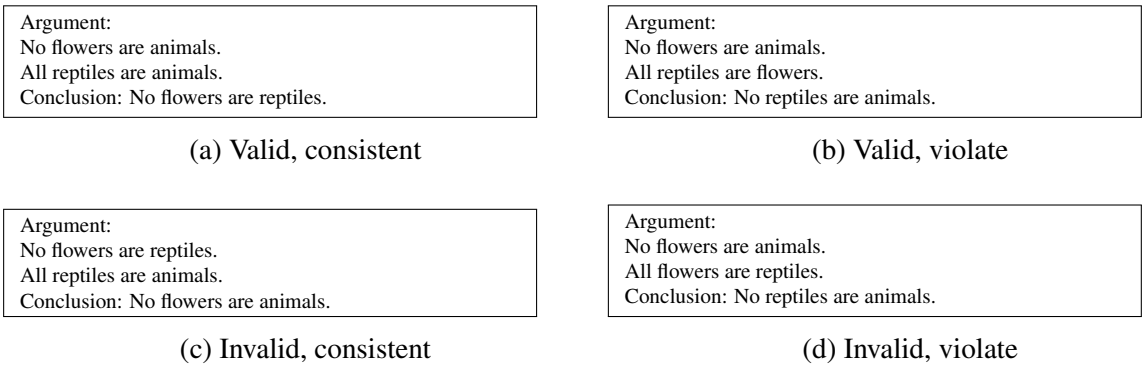
We generated a natural inference task derived from these comparison sentences as follows. We began with the *consistent* version, by taking the the raw output from the LM, "puddles are smaller than seas" as the hypothesis and formulating a premise "seas are bigger than puddles" such that the generated 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 the task, we simply replace the nouns ('seas' and 'puddles') with nonsense words.



(a) Valid, consistent

(b) Valid, violate

(c) Invalid, consistent

(d) Invalid, violate

Figure 11: Example syllogism cluster, showing 2×2 design of valid (top row), invalid (bottom row), and consistent (left column) and violate (right column) arguments.

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.

A.1.2 Syllogisms data generation

We generated a new set of problems for syllogistic reasoning. Following the approach of Evans et al. (20), 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. 10b 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. 11 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

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

language models and humans (cf. 119, 120). We then sampled an identical set of nonsense arguments by simply replacing the entities in realistic arguments with nonsense words.

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 also tried some minor variations in preliminary experiments (such as changing the prompt or prefixing the conclusion with "Therefore:" or omitting the prefix before the conclusion), but observed qualitatively similar results so we omit them here.

A.1.3 Wason data generation

As above, we generated a new dataset of Wason problems to avoid potential for dataset contamination (see Fig. 10c 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 independently; it does not see all answers at once.)

See Figs. 13 and 14 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. 10c for an example. We demonstrate in Appx. B.3 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 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 (119).

Finally, we created two types of control rules based on the realistic rules, which we present here. 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. For example, one shuffled-realistic rule is "If they are doctors, then they must have a parachute."

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 the violate rule "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 (viz. requiring the item that is *not* a parachute to skydive), 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.1.4 Differences from an earlier preprint of this paper:

Readers of an earlier preprint of this paper (`https://arxiv.org/abs/2207.07051v1`) may notice some differences in task format and performance of Chinchilla, especially on the NLI task. These differences are due to our attempts to adapt the tasks in order to present them to human participants. In order to align comparisons between humans and the language models (cf. 69), we then ported the human-oriented changes back into the format used for language model evaluation.

For example, in the original paper we did not show the model the two possible choices for the NLI task; we simply evaluated the model's likelihood of each continuation. However, because we presented the tasks multiple choice to the humans in multiple choice format, we showed them the two possible answers. Thus, in the current version of the paper we also included the two answer choices in the prompt when evaluating the language models, followed

```
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 13: Realistic Wason rules and instances used.

```
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 14: Arbitrary Wason rules and instances used.

by "Answer:", and only then evaluate the model (see Fig. 10a). Likewise, in the original version of the paper we did not provide instructions before the tasks; in this version we attempted to match the relevant portions of the human instructions.

These changes mean that the results in the current version of the paper cannot be directly compared to the results in the earlier version.

A.2 Evaluation

DC-PMI correction: We use the DC-PMI correction (117) for the syllogisms and Wason tasks; i.e., we choose an answer from the set of possible answers ($\mathcal{A}$) as follows:

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

Where the baseline prompt is the task instruction prompt, followed by "Answer:" and $p(x \,|\, y)$ denotes the model's evaluated likelihood of continuation x after prompt y.

Instruction prompt: We prefixed each question with a two-part instruction prompt that attempted to match the human generic and task-specific instructions (see below). We began each of these prompts with the performance relevant generic instructions that preceded our human experiment:

```
In this task, you will have to answer a series of questions. You will have to choose the best answer to complete a sentence,
↪  paragraph, or question. Please answer them to the best of your ability.\n\n
```

After two linebreaks, a task-specific instruction was appended:

NLI:

```
Please choose the best completion for the following sentence:\n
```

Syllogisms:

```
Please assume that the first two sentences in the argument are true. Determine whether the argument is valid, that is, whether
↪  the conclusion follows from the first two sentences:\n'
```

Wason:

```
Please answer the following question carefully:\n
```

Finally, the question was appended to this prompt.

A.3 Human experiments

The exact text seen by the participants before each question was as follows:

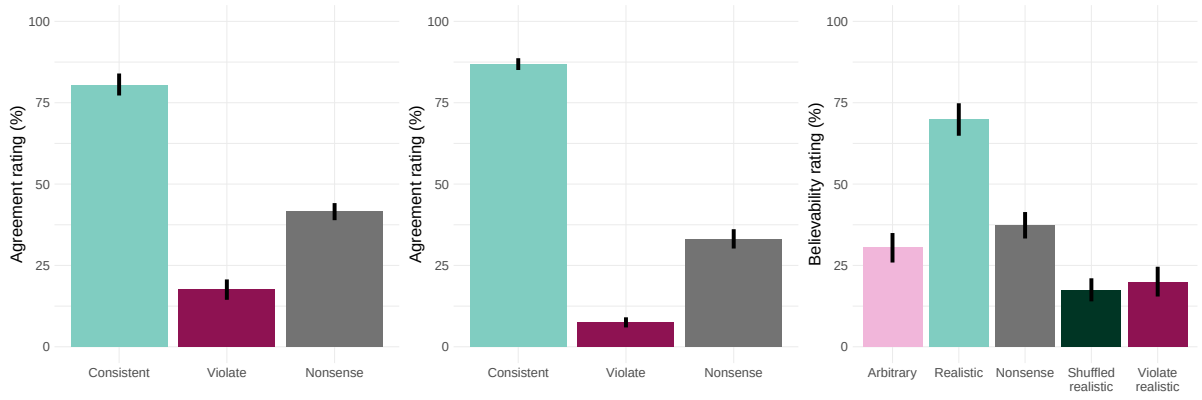
```
NLI_DEFAULT_PREFACE = (
    "Please choose the best completion for the following sentence:")
SYLLOGISMS_DEFAULT_PREFACE = (
    "Please assume that the first two sentences in the argument are true. "
    "Determine whether the argument is valid, that is, whether the
        conclusion "
    "follows from the first two sentences:")
WASON_DEFAULT_PREFACE = (
    "Please answer the following question carefully:"
)
WASON_BONUS_PREFACE = (
    "Please answer the following question carefully; <font color='#bb0044'>
        we "
    "will pay you an additional performance bonus of 0.5 GBP if you answer
        "
    "this question correctly</font>:"
)
PRIOR_AGREEMENT_PREFACE = (
    "Please rate how much you agree with the following statement, on a
        scale "
    "from 0% (disagree completely) to 50% (neither agree nor disagree) to
        100% "
```

```
    "(agree completely)."
)
WASON_BELIEVABLE_PREFACE = (
    "Please rate how believable the following rule is, on a scale from 0% "
    "(completely unbelievable) to 50% (neither believable nor unbelievable)
        to "
    "100% (completely believable)."
)
```

B Supplemental analyses

B.1 Believability of the propositions and rules

In order to assess the validity of our new datasets, we collected believability ratings from each subject, after they had completed the three tasks tasks, on one stimulus from each task type (not the version they had seen). Specifically, we asked the participants how believable a Wason rule was, and how much they agreed or disagreed with a proposition. In Fig. 15 we show that participants found Consistent and Realistic stimuli much more believable than those in other conditions.



(a) NLI (b) Syllogisms (c) Wason

Figure 15: Our datasets align with human beliefs. When participants were asked how much they believed propositions or rules from our three tasks (a-c), they rated the Consistent or Realistic conditions as much more believable than the Violate ones, with Nonsense in between.

B.2 Robustnesss of the main language model results to raw-likelihood scoring and few-shot prompting

In this section, we show that the content effects we observe are robust to various manipulations of the evaluation context.

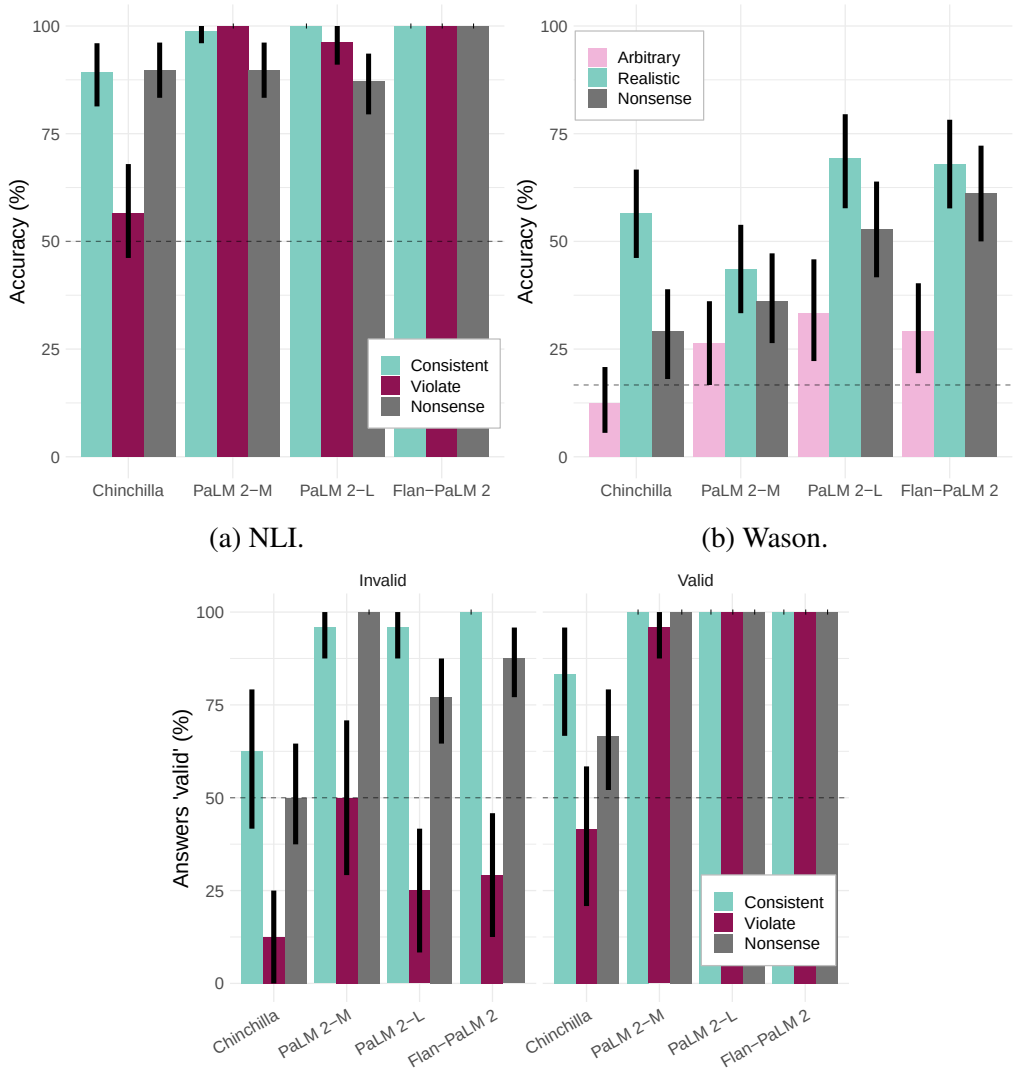
B.2.1 Removing instruction prompts

In the main text experiments, we provided models with an instruction prompt that roughly matched the human instructions (cf. 69). However, it is unclear how substantial a role this prompt played in performance, and human-likeness of the content effects. In Fig. 16 we show performance of a subset of the models when removing this instruction prompt; in most cases, results are similar, with a few notable exceptions. In particular Chinchilla shows much stronger content effects on the NLI tasks without instructions.

B.2.2 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 (117). However, it is also common to score language models using raw likelihood comparisons. Would we observe the same content effects in that case?

In Fig. 17 we show the results of raw-likelihood scoring. On the syllogisms tasks, this scoring method results in substantially more answer bias — several of the models say valid in response to every problem, regardless of the content or logical structure. Thus, performance is much worse overall. However, for the models that do show any variability with content, the content effects are broadly similar to those observed in the main text: the models are more likely to say an argument is valid if the conclusion is belief-consistent than if the conclusion violates beliefs. Furthermore, if the instruction prompt is removed, the bias is substantially reduced, and stronger content effects are revealed.



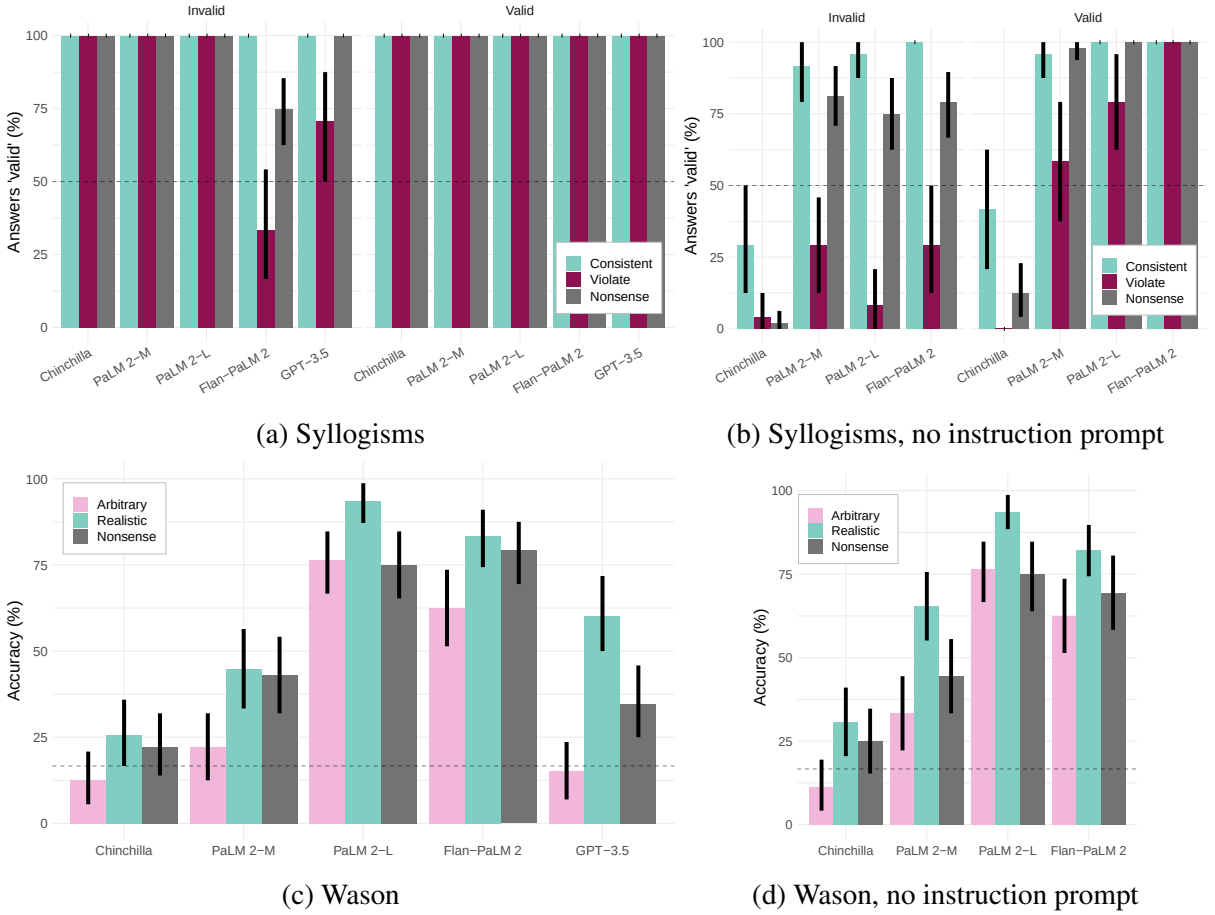
(a) NLI.

(b) Wason.

(c) Syllogisms.

Figure 16: Performance of a subset of the models when evaluated without an instruction prompt. Overall results and content effects are similar; however, in a few cases performance is noticeably impaired, particularly for Chinchilla on the Violate condition of NLI.

In the Wason tasks, the effects on accuracy are more complex. While some models perform worse without the prior correction (e.g. Chinchilla), others perform much better. In particular, PaLM-2 L achieves over 75% performance in every condition (including Arbitrary and Nonsense). However, all models that perform above chance show the same content effects observed in the main text: better performance on Realistic than Arbitrary rules. (In Appx. B.2.3 we also explore the effect of scoring with raw likelihoods on the individual card choices on the Wason task.)



(a) Syllogisms

(b) Syllogisms, no instruction prompt

(c) Wason

(d) Wason, no instruction prompt

Figure 17: Scoring using the raw answer likelihoods — rather than the Domain-Conditional PMI prior correction — for the Syllogisms and Wason tasks. (a) On the syllogisms tasks, removing the prior correction results in substantial answer biases for many models: much greater likelihood to say "valid" than "invalid." Overall performance is much worse due to this bias; indeed, several models answer "valid" for every argument in every condition. However, for those that do not — Flan-PaLM 2 and GPT-3.6 — the direction of the content effects is as in the main text: the models are more likely to answer "valid" if the conclusion is belief-consistent. (b) However, the answer bias on the syllogisms with raw-likelihood scoring seems to be strongly driven by the instruction prompt; without the prompt, the raw likelihoods yield less biased responses, and strong overall content effects. (c-d) On the Wason tasks, with or without the instruction prompt, removing the prior correction improves performance from some models, but hurts performance from others. Regardless, all models show the same pattern of content effects: facilitation in the Realistic rules compared to Arbitrary. (Compare to Figs. 4 and 5, respectively, which use DC-PMI scoring.))

B.2.3 Effects of scoring method and answer order on the Wason answer choices

In Fig. 18 we show the effect of scoring method (DC-PMI vs. raw likelihoods) and the order in which the cards were presented (antecedent cards first or consequent cards first) on the models' answer choices on the Wason task. Scoring method does affect the error distribution fairly substantially, even where accuracy is similar; answer order has smaller effects.

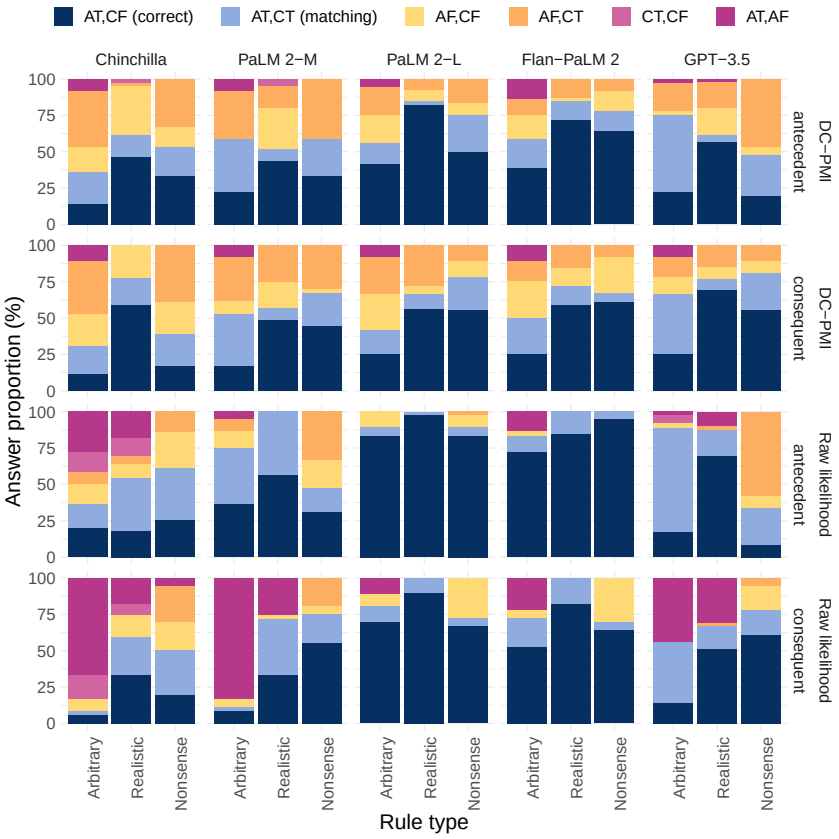


Figure 18: Effect of scoring method (DC-PMI in the top two rows vs. raw likelihoods in the bottom two) and ordering of the cards (antecedent cards first or consequent cards first; respectively in rows 1 and 3, and 2 and 4) on model choices. The DC-PMI prior correction does shift error patterns somewhat, and the models commit relatively more of the AT,AF answers with raw likelihood scoring, while with the DC-PMI scoring, the humans commit more of these errors than the models. The ordering of the cards does not have too substantial an effect, particularly with DC-PMI scoring. Generally, content effects — that is, the advantage of the Realistic rules over arbitrary ones — persists regardless of scoring method or order.

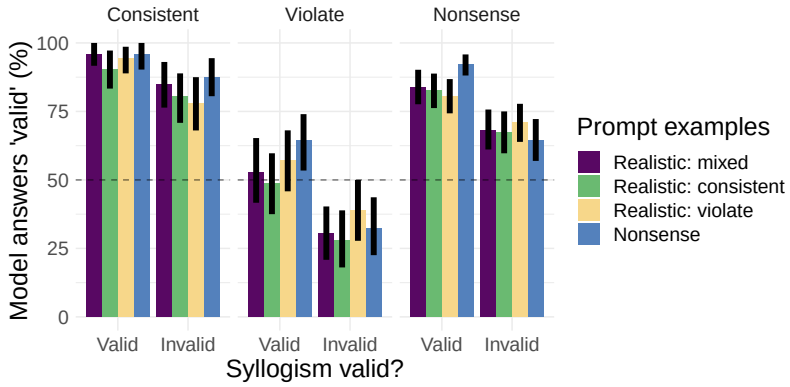


Figure 19: Chinchilla evaluated 5-shot on the syllogisms task, with different types of prompt examples. Content effects are very slightly reduced relative to the original experiments, but remain robust. The particular type of problems used in the prompt examples do not strongly affect performance. (The "Realistic: mixed" condition includes realistic examples from both the consistent and violate subsets.)

B.2.4 Few-shot prompting of Chinchilla

In all the main text experiments, we evaluated the models zero-shot, with only instructions. However, language model performance is generally improved by few-shot prompting (e.g. 7). We therefore evaluated whether few shot prompting with different kinds of prompt examples would alter the content effects we observed. (Note that, for computational reasons, we restrict these analyses to the Chinchilla model.) 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.

In Fig. 19 we show 5-shot prompting results for Chinchilla on the Syllogisms tasks. Content effects are slightly weaker than without the examples, but remain robust.

In Fig. 20 we show 5-shot prompting results for Chinchilla on the Wason selection tasks. Content effects are exaggerated with the 5-shot prompts, because the model improves noticeably at Realistic rules, but improves less (if at all) on Arbitrary ones. We also see a noticeable effect of the type of examples used in the prompt, with Realistic examples offering optimal benefits.

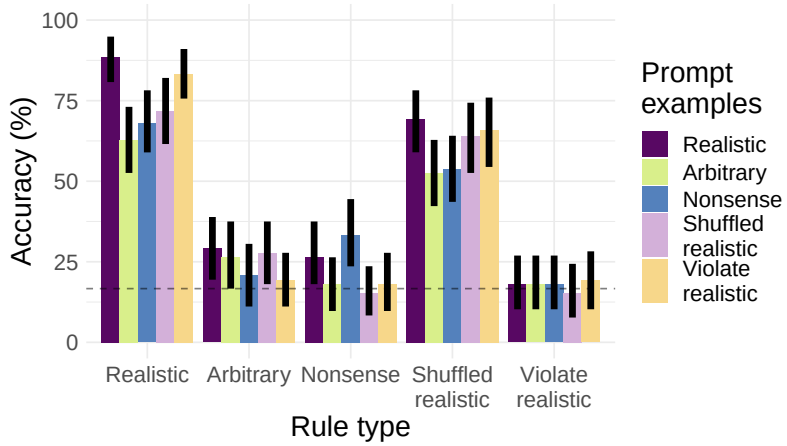


Figure 20: Chinchilla evaluated 5-shot on the Wason task, with different types of prompt examples. Again, content effects remain strong — or are even amplified — with few-shot prompts. Realistic prompt examples appear to be most beneficial overall, but especially for realistic and shuffled realistic probes, thus they actually enhance content effects. Other types of prompts are generally helpful in a more limited set of conditions; there may be an overall benefit to prompts matching probes.

B.3 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 of the rule "if the workers work as a doctor then they must have received an MD" we prompted Chinchilla 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'. To avoid order

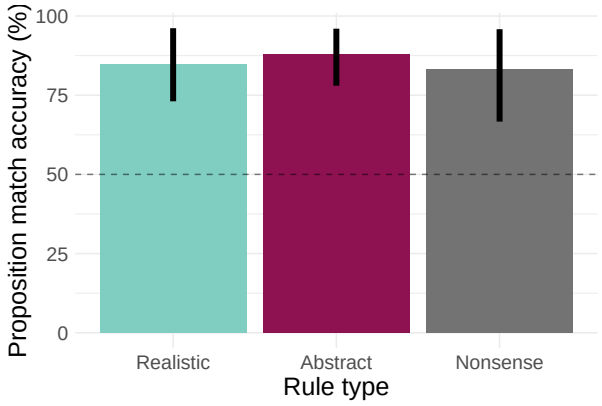


Figure 21: The component propositions (antecedents and consequents) of the Wason rules have similar difficulty across conditions. This plot shows Chinchilla's accuracy on forced choices of which instance matches a proposition, across conditions. (Note that the shuffled realistic rules use the same component propositions as the realistic rules.)

biases, 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. 21)—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.4 Additional recombined realistic conditions for the Wason tasks

The Wason task rules can be realistic or unrealistic in multiple ways. For example, the component propositions can be realistic even if the relationship between them is not. We therefore generate two variations on realistic rules:

Shuffled realistic rules, which combine realistic 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]").

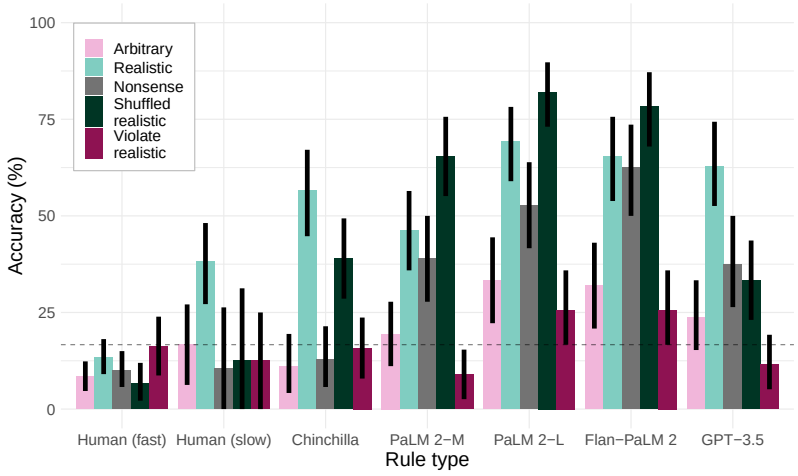


Figure 22: Evaluating models and humans on shuffled realistic and violate versions of the Wason rules. Humans

We also evaluated models and humans on these rules. For shuffled rules, results are well above chance. Surprisingly, one family of models (PaLM 2) even perform better at shuffled realistic than realistic rules. For violate rules, by contrast, performance is generally close to 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. Humans generally perform poorly on either rule variant.

B.5 Human performance on the Wason tasks, in our original sample & replication

As mentioned in the main text, after collecting our original sample on the Wason task, we recruited an additional set of participants to whom we offered a performance bonus on this task, in an attempt to increase performance. We present the results broken down by sample in Fig.

23. We performed mixed-effects logistic regressions (Table 1) to test for an improvement in performance in the sample with a performance bonus; this effect was marginally significant. However, performance remains low overall, and we do not observe a significant difference in the content effect.

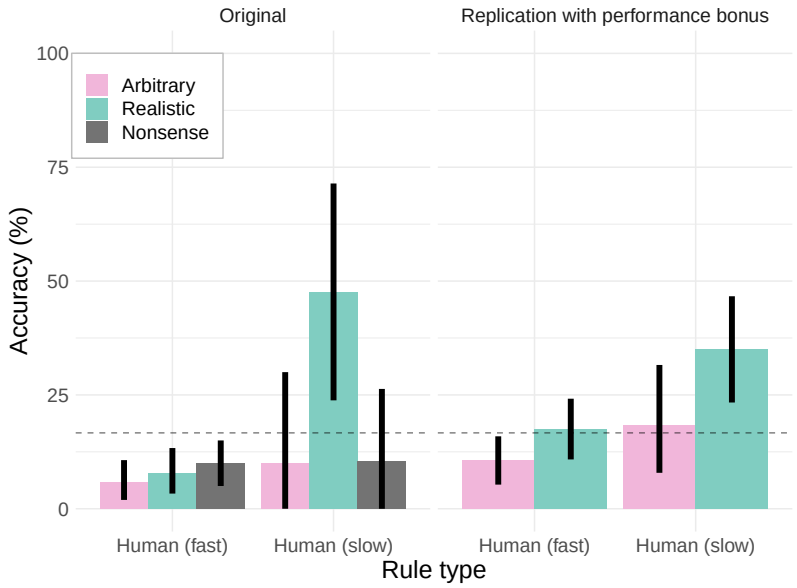


Figure 23: Breakdown of human results in our original experiment, and our replication (where we also added a performance bonus of 0.5 GPB for the Wason question). We observe a significant advantage for the slower humans in the Realistic condition in each case. The performance bonus does not seem to clearly improve performance.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + replication_experiment +
    (1 | wason_name)
   Data: wason_human_correct_df

     AIC      BIC   logLik deviance df.resid
   476.1    493.5   -234.0    468.1      570

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.7247 -0.4879 -0.3449 -0.2718  3.9974

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.1704   0.4127
Number of obs: 574, groups:  wason_name, 25

Fixed effects:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.6190     0.3029  -8.646  < 2e-16 ***
wason_conditionRealistic   0.8261     0.3066   2.694  0.00706 **
replication_experimentTRUE   0.5274     0.2695   1.957  0.05034 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Additive model.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition * replication_experiment +
    (1 | wason_name)
   Data: wason_human_correct_df

     AIC      BIC   logLik deviance df.resid
   477.7    499.5   -233.9    467.7      569

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.7198 -0.4788 -0.3544 -0.2523  4.3184

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.1762   0.4197
Number of obs: 574, groups:  wason_name, 25

Fixed effects:
                                                Estimate Std. Error z value
(Intercept)                                      -2.7723     0.4162  -6.662
wason_conditionRealistic                          1.0550     0.5078   2.078
replication_experimentTRUE                        0.7380     0.4606   1.602
wason_conditionRealistic:replication_experimentTRUE  -0.3262     0.5673  -0.575
                                                Pr(>|z|)
(Intercept)                                     2.71e-11 ***
wason_conditionRealistic                          0.0378 *
replication_experimentTRUE                        0.1091
wason_conditionRealistic:replication_experimentTRUE   0.5653
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Interaction model.

Table 1: Mixed-effects linear regressions for differences in human performance on the replication sample on the Wason task. We do observe a marginally-significant effect of the experiment in the additive model (top). However, we do not observe significant differences in the content effect in an interaction model (bottom).

B.6 Human response time distributions on the Wason tasks

In Fig. 24 we show the distribution of response times for humans in the Wason tasks. There is a mean difference in response times, with participants spending about 12 seconds longer on Realistic questions on average. This difference may be due to the time needed to read the extra sentences giving the realistic context, or to the participants engaging more deeply with the problems that seem more sensible. However, in Appx. C.3.1 we show that this difference alone does not explain the advantage of the Realistic conditions.

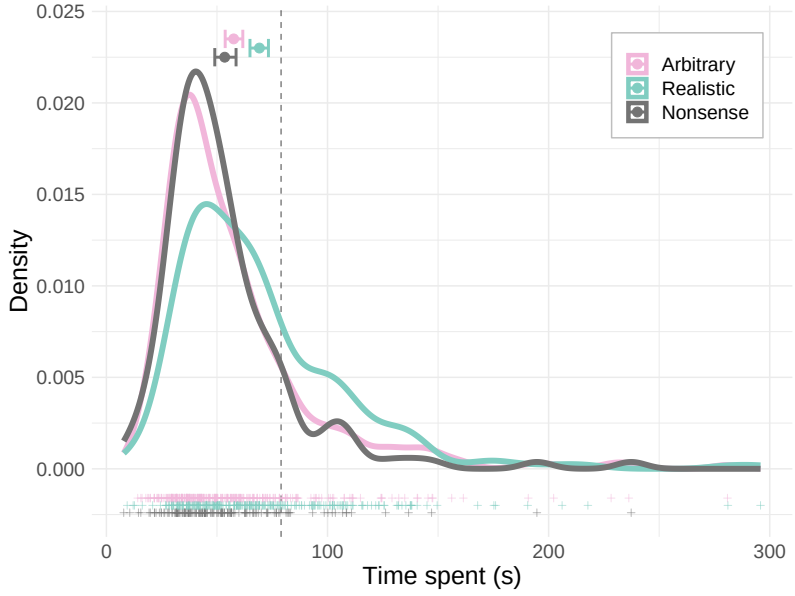


Figure 24: Human response time distributions on the Wason tasks. The Realistic condition results in significantly longer response times. The vertical dashed line indicates the cutoff for "slow" subject group; 85% of the subjects were faster than this in the original experiment.

B.6.1 Response time effects on NLI and syllogisms

Given the strong effect of response time on Wason task performance, we also analyzed the effects on the NLI and Syllogism tasks (Figs. 25 & 25; Table 2). In these tasks we do not see clear effects, though there are hints of an interesting potential interaction in the syllogisms task.

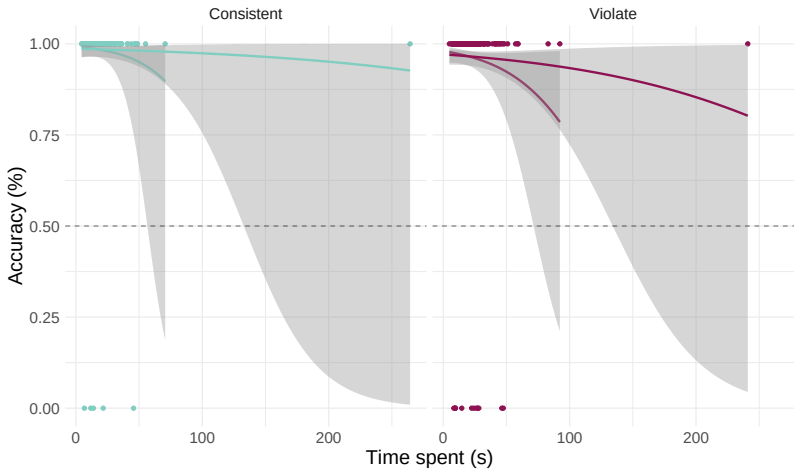


Figure 25: There is little effect of response time on accuracy in the NLI tasks.

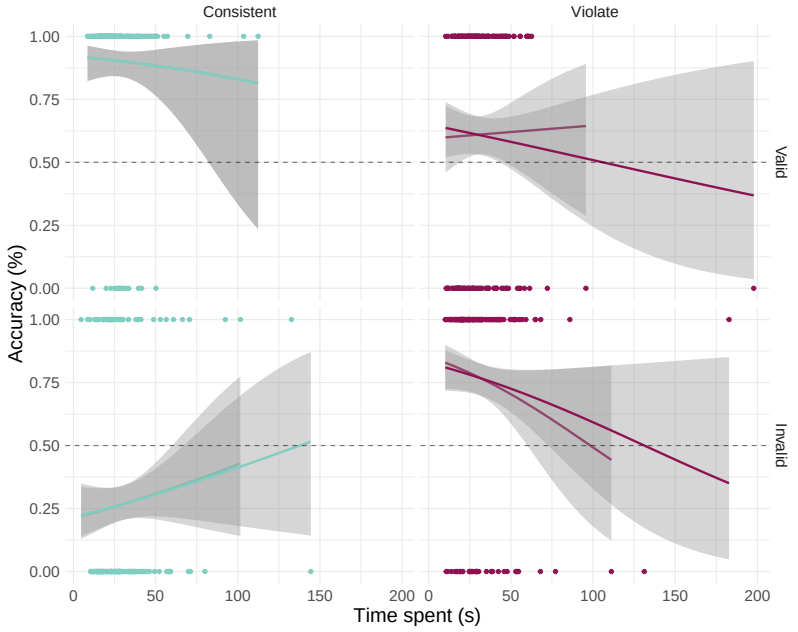


Figure 26: Effects of response time on accuracy on the syllogisms task.

B.7 Item-level effects

In this section, we perform item level analyses for each task.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent * (scale(log(rt)) +
    consistent_plottable) + (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == "Human")

     AIC      BIC   logLik deviance df.resid
   693.4    724.6   -339.7    679.4      631

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5854 -0.6137  0.3423  0.6457  2.1597

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0.08975  0.2996
Number of obs: 638, groups:  syllogism_name, 12

Fixed effects:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                      0.61895    0.19140   3.234  0.00122 **
logic_belief_consistent1                         3.16927    0.33673   9.412  < 2e-16 ***
scale(log(rt))                                  -0.08297    0.09858  -0.842  0.39998
consistent_plottableViolate                      0.20829    0.21271   0.979  0.32746
logic_belief_consistent1:scale(log(rt))         -0.35038    0.19413  -1.805  0.07109 .
logic_belief_consistent1:consistent_plottableViolate -2.40845    0.42012  -5.733 9.88e-09 ***
```

Table 2: Mixed-effects regression examining the continuous effect of RT on the Syllogism tasks. There is no main effect, but there is a marginally-significant interaction with the content effect, such that slower responses are more helpful on problems where content contradicts logic.

B.7.1 NLI

First, for the NLI task, we plot the item-level correlations in accuracy in Fig. 27. Surprisingly (given the close-to-ceiling performance), we find that Human success rates are significantly predictive of LM success rates, even when controlling for condition (Table 3).

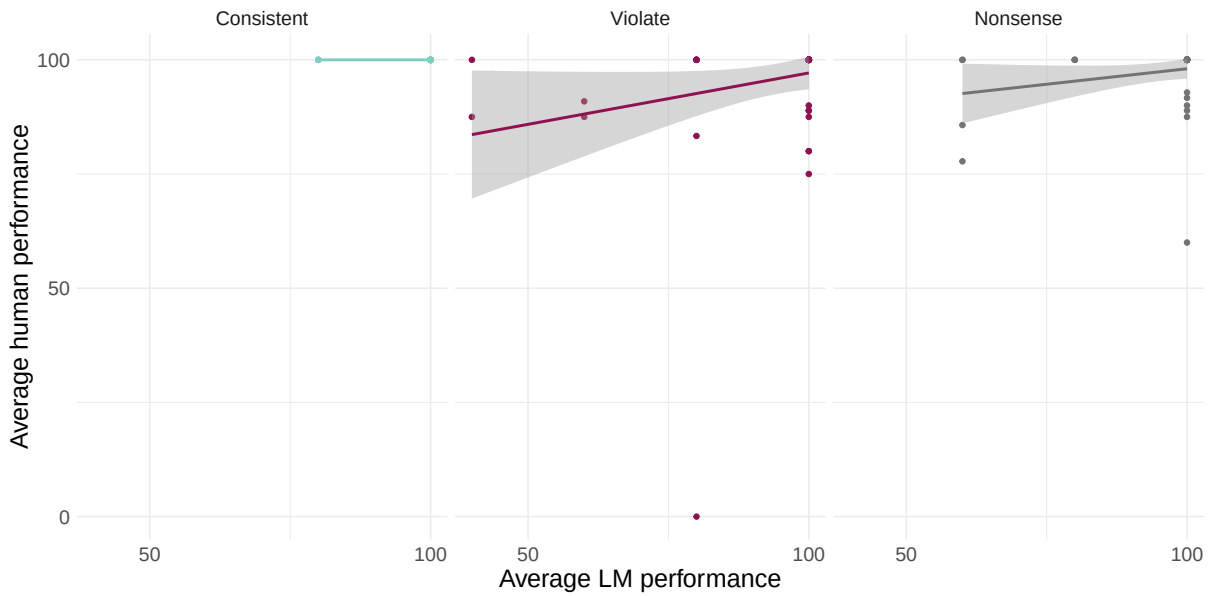


Figure 27: Association of human and average model accuracy on the NLI task.

```
Linear mixed model fit by REML ['lmerMod']
Formula: LM ~ Human + consistent_plottable + (1 | model)
   Data: nli_item_level_df

REML criterion at convergence: -404.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.2460  0.0661  0.1428  0.2724  1.6139

Random effects:
 Groups   Name        Variance  Std.Dev.
 model    (Intercept) 0.0008882 0.0298
 Residual             0.0350993 0.1873
Number of obs: 845, groups:  model, 5

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                   0.73890    0.07288  10.139
Human                         0.24681    0.07077   3.488
consistent_plottableViolate  -0.02879    0.01560  -1.846
consistent_plottableNonsense -0.02429    0.01649  -1.473
```

Table 3: Mixed-effects linear regression for item-level association of human and model accuracy on the NLI task, controlling for consistency.

B.7.2 Syllogisms

For the Syllogisms task, we plot the item-level correlations in accuracy in Fig. 27. We again find a significant relationship between human success rates and language model success ($t = 4.98$, $p < 0.001$ when controlling for task variables; Table 4).

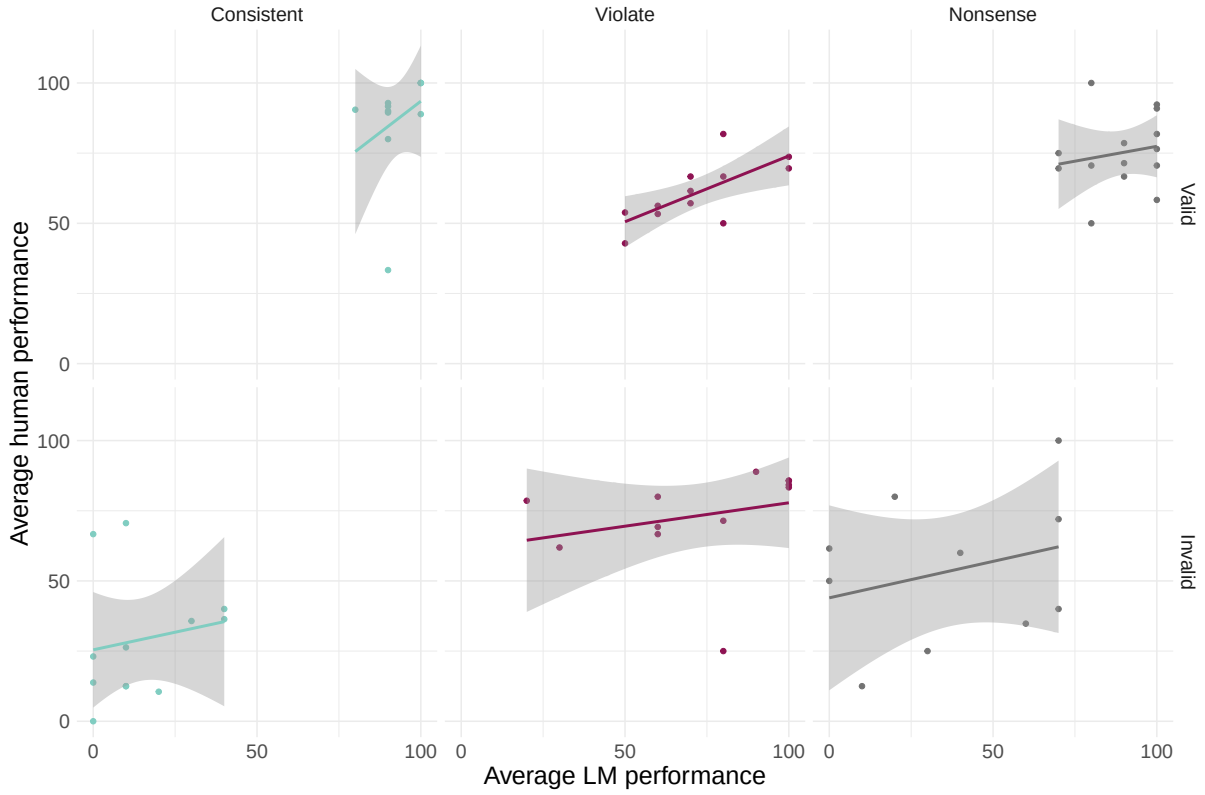


Figure 28: Association of human and average model accuracy on the Syllogisms task.

```
Linear mixed model fit by REML ['lmerMod']
Formula: LM ~ Human + logic_belief_consistent * consistent_plottable +
    (1 | model)
   Data: syl_item_level_df

REML criterion at convergence: 313.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5622 -0.5048  0.2668  0.6984  2.8012

Random effects:
 Groups   Name        Variance Std.Dev.
 model    (Intercept) 0.004978 0.07056
 Residual             0.131391 0.36248
Number of obs: 355, groups:  model, 5

Fixed effects:
                                                Estimate Std. Error t value
(Intercept)                                      0.23822    0.07538   3.160
Human                                            0.51089    0.10262   4.978
logic_belief_consistent                          0.24473    0.04505   5.432
consistent_plottableViolate                      0.14737    0.04827   3.053
consistent_plottableNonsense                     0.09873    0.04806   2.054
logic_belief_consistent:consistent_plottableViolate -0.27191    0.05281  -5.148
```

Table 4: Mixed-effects linear regression for item-level association of human and model accuracy on the Syllogisms task, controlling for content and logic.

B.7.3 Wason

For the Wason task, we plot the item-level correlations in accuracy in Fig. 27. Perhaps because human performance is low overall, we do not observe a significant relationship between human success rates and language model success (Table 5).

Due to the item-level effects observed in some of the main regressions, we also plot performance of each model or human group on each of the Wason rules in Fig. 30. Overall, the variability seems mostly as expected. However, there are some interesting patterns, including one arbitrary rule that most subject perform well on. That particular rule is:

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

It is not particularly apparent to us why this rule might be easier.

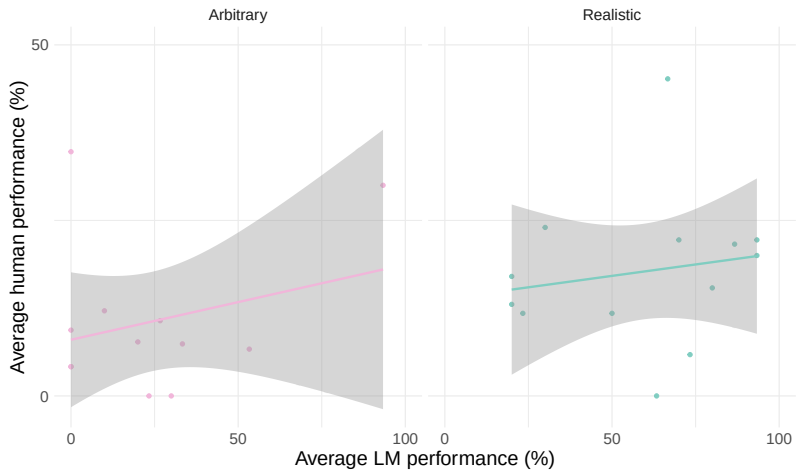


Figure 29: Association of human and average model accuracy on the Wason task. Note the vertical axis scale—human performance is low overall.

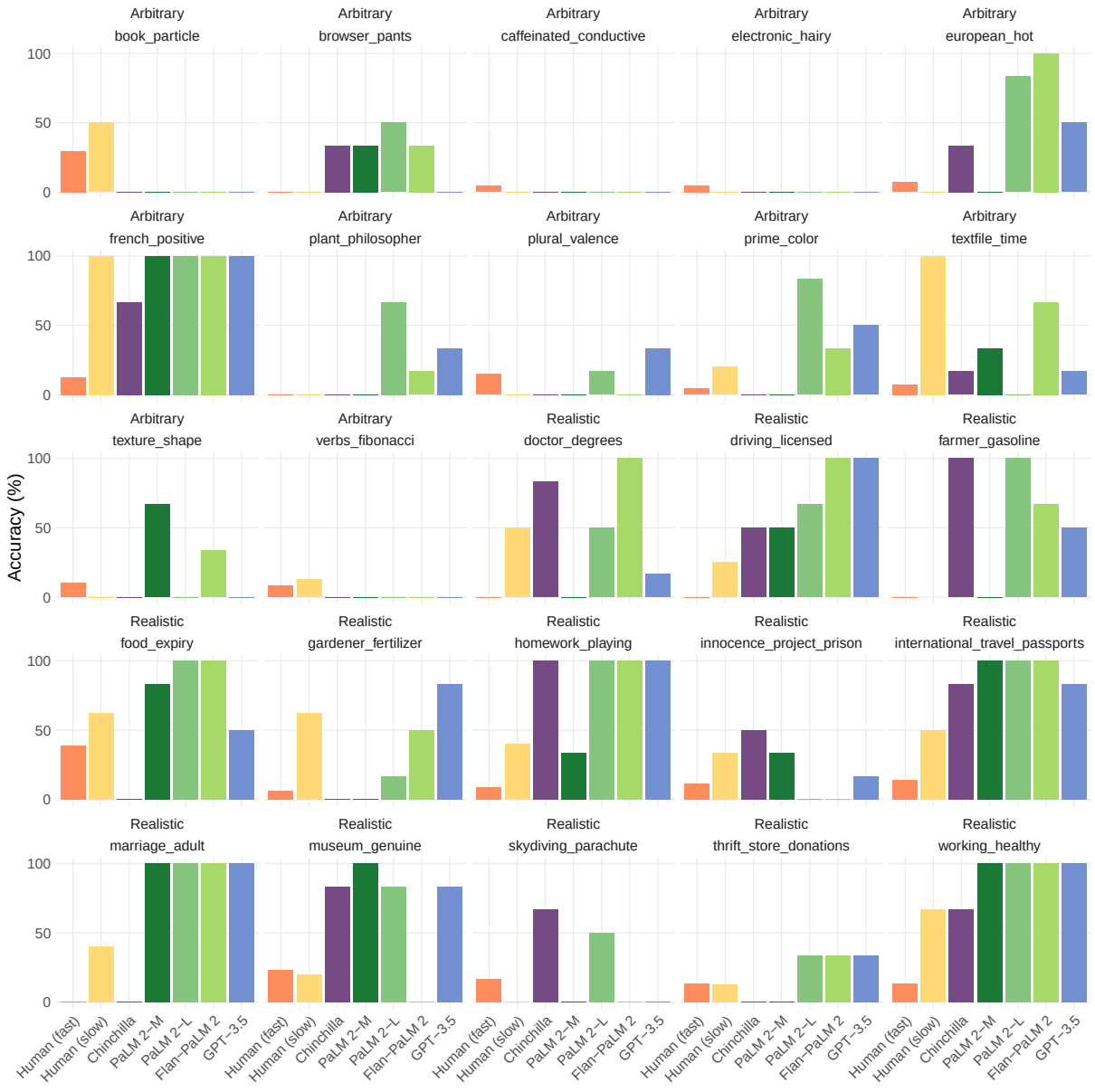


Figure 30: Accuracy of humans and each model on each rule for the Wason tasks. Note that due to sampling variability, the number of human participants who experienced each rule varies, particularly for the slow subjects. There are various suggestive patterns, including an arbitrary rule (`french_positive`) that models and slower humans perform quite well on, and realistic rules (like `skydiving_parachute`) that all perform surprisingly poorly on. (Note that the variation within a model comes from testing on multiple variations of each problem, with different card orders and card names; see Appx. A.1.)

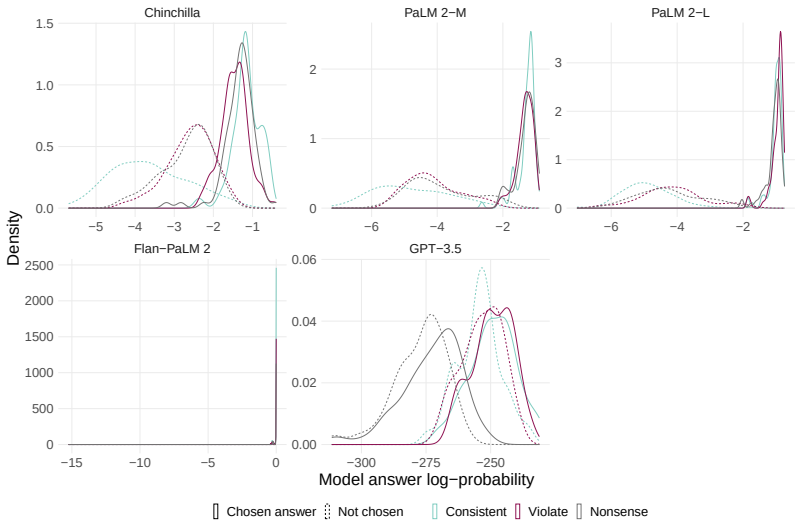
```
Linear mixed model fit by REML ['lmerMod']
Formula: LM ~ Human + wason_condition + (1 | model)
   Data: wason_item_level_df

REML criterion at convergence: 181.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7748 -0.7994 -0.1560  0.8760  1.9239

Random effects:
 Groups   Name        Variance Std.Dev.
 model    (Intercept) 0.006414 0.08009
 Residual             0.145352 0.38125
Number of obs: 185, groups:  model, 5

Fixed effects:
                       Estimate Std. Error t value
(Intercept)             0.20251    0.06860   2.952
Human                   0.36974    0.29860   1.238
wason_conditionRealistic  0.32436    0.07148   4.538
wason_conditionNonsense   0.19533    0.06967   2.804
```

Table 5: Mixed-effects linear regression for item-level association of human and model accuracy on the Syllogisms task, controlling for content and logic.

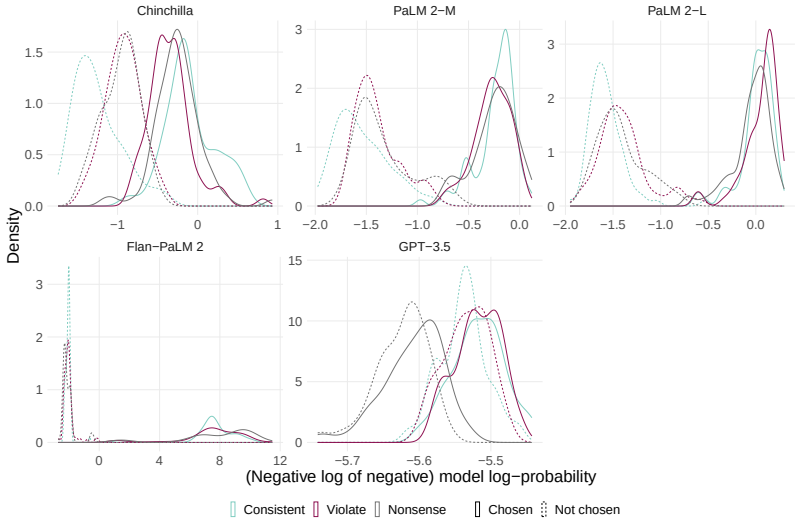
B.8 Model answer log-probability distributions

In this section we plot the log-probability distributions of the models on the different tasks (Figs. 31, 32, 33). There are a variety of interesting effects of task variables, and some striking differences among the models.

For example, the instruction-tuned models (Flan-PaLM 2 and GPT-3.5) have numerically much greater magnitude log-probabilities to the answers, especially GPT-3.5. This may be an artifact of the tuning process. Furthermore, the larger models tend to show clearer separation between the chosen answer and the others (e.g., comparing PaLM 2-L to -M).



(a) Raw log-probabilities.



(b) Log transformed.

Figure 31: Model log-probability distributions for the answer choices on the Natural Language Inference (NLI) task. We visualize these in two ways: (a) the raw log-probabilities, and (a) the negative log of the negative log-probabilities — this transform makes the distribution for Flan-PaLM 2 clearer. Across both plots, there is fairly clear separation between the distributions of chosen and unchosen answers for most models. There are various interesting effects of content on the log-probabilities, e.g. changes in the mean and variance of the distributions. There are also striking differences among the models, possibly hinting at the effects of different training processes.

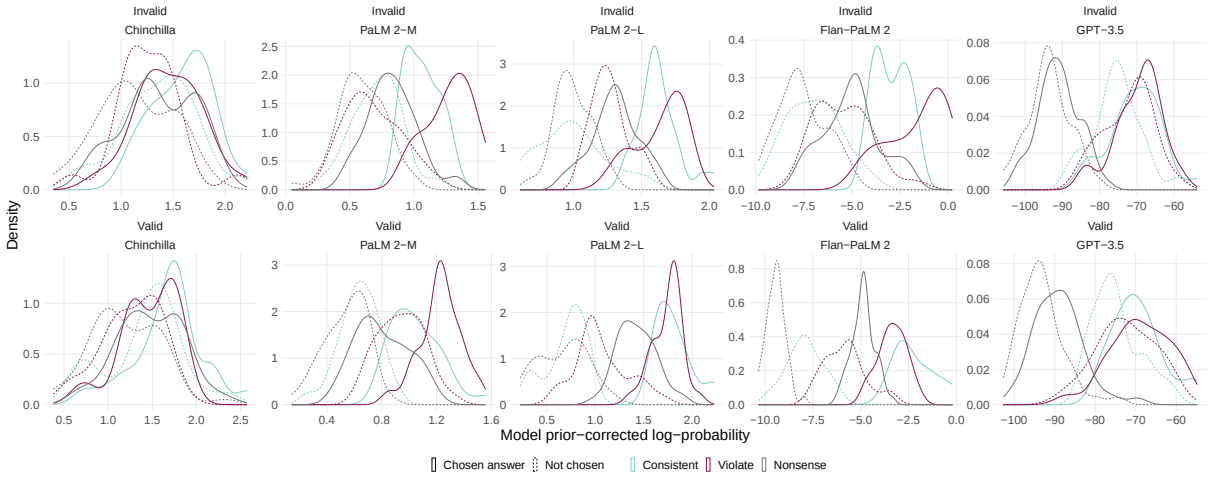


Figure 32: Model prior-corrected log-probability distributions for the answer choices on the syllogisms task. The degree of separation between the distributions depends on the model, validity, and content. Again, there are differences among the models. For example, larger models show more cleanly separated distributions (PaLM 2-L vs. -M), and the instruction tuned models (Flan-PaLM 2 and GPT-3.5) show much larger magnitude prior corrected log probabilities.

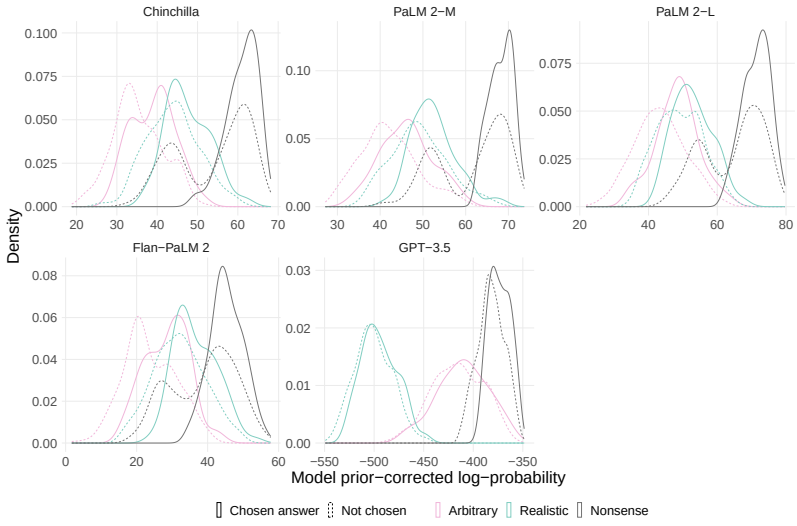
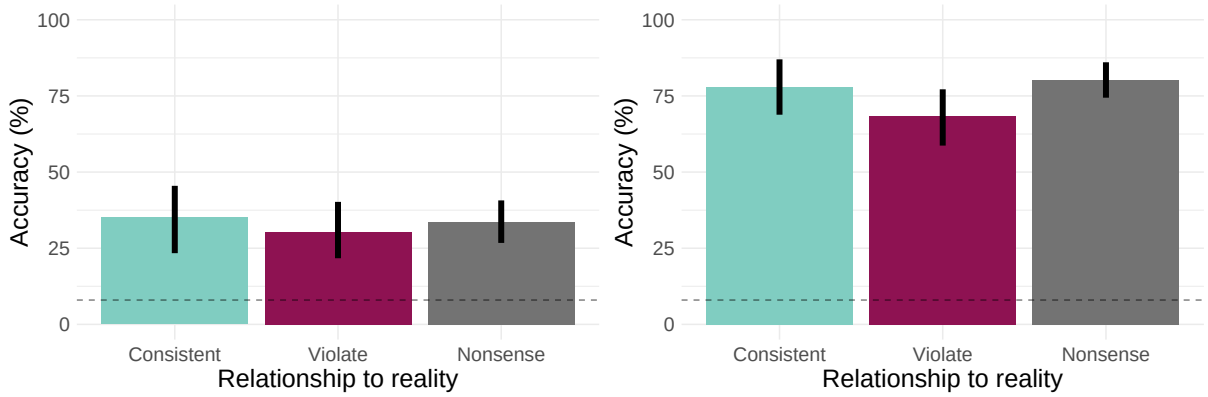


Figure 33: Model prior-corrected log-probability distributions for the answer choices on the Wason selection task. The degree of separation between the chosen and not-chosen answer distributions is generally lower than in other tasks, possibly reflecting the greater difficulty of the Wason task, or the greater problem-to-problem variability. By contrast, the separation by content is striking for some models, e.g. GPT-3.5.

B.9 Chinchilla can identify the valid conclusion of a syllogism from among all possible conclusions with high accuracy

In Fig. 34 we show the accuracy of Chinchilla 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. (54) that humans exhibit less bias when making a forced choice among two possible arguments (one valid and one invalid) rather than deciding if a single syllogism is valid or invalid.

Note that in this case scoring with the Domain-Conditional PMI (117)—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 34: Chinchilla's 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.

C Statistical analyses

In this section, we provide the full results for all statistical analyses reported in the main text. We generally report results from mixed-effects logistic regressions, controlling for the random effects of the different stimuli used.[3]

C.1 NLI

We report statistical analyses of content effects on the NLI tasks for humans and all models in Tables 6-11. We generally fit mixed effects logistic regressions, but the regressions for PaLM 2-L and Flan-PaLM 2 failed to converge due to ceiling effects. We therefore also report χ^2 tests of the difference in correct responses across conditions. In all cases, we do not find a significant content effect on this simple task.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ consistent_plottable + (1 | name)
   Data:
nli_joint_df %>% filter(subject == "Human", consistent_plottable !=
    "Nonsense")

     AIC      BIC   logLik deviance df.resid
   133.2    146.8    -63.6    127.2      677

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5119  0.0105  0.0106  0.0282  0.4931

Random effects:
 Groups Name        Variance Std.Dev.
 name   (Intercept) 25.82    5.081
Number of obs: 680, groups:  name, 122

Fixed effects:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                 9.082      2.072   4.384 1.17e-05 ***
consistent_plottableViolate   -2.051      1.608  -1.276    0.202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Mixed-effects logistic regression.

```
Chi-squared test for given probabilities
X-squared = 0.33937, df = 1, p-value = 0.5602
```

(b) χ^2 test.

Table 6: Statistical analyses of human performance on the NLI tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. There are no significant content effects.

[3]Unless otherwise noted, we conservatively approximate the degrees of freedom for all t-tests by treating all random effects as though they were fixed effects (i.e. by subtracting the number of levels of each random variable from the residual degrees of freedom), rather than using a variance-based approximation.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ consistent_plottable + (1 | name)
   Data:
nli_joint_df %>% filter(subject == "Chinchilla", consistent_plottable !=
    "Nonsense")

     AIC      BIC   logLik deviance df.resid
    46.6     55.8    -20.3     40.6      153

Scaled residuals:
      Min        1Q    Median        3Q       Max
-0.084635  0.000969  0.000969  0.001903  0.001903

Random effects:
 Groups Name        Variance Std.Dev.
 name   (Intercept) 2646     51.43
Number of obs: 156, groups:  name, 153

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  13.877      3.373   4.114  3.9e-05 ***
consistent_plottableViolate  -1.357      3.760  -0.361    0.718
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Mixed-effects logistic regression.

```
Chi-squared test for given probabilities
X-squared = 0.34266, df = 1, p-value = 0.5583
```

(b) χ^2 test.

Table 7: Statistical analyses of Chinchilla's performance on the NLI tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. There are no significant content effects.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ consistent_plottable + (1 | name)
   Data:
nli_joint_df %>% filter(subject == "PaLM 2-M", consistent_plottable !=
    "Nonsense")

     AIC      BIC   logLik deviance df.resid
    22.0     31.1     -8.0     16.0      153

Scaled residuals:
      Min        1Q    Median        3Q       Max
-0.074976  0.000648  0.000648  0.000804  0.000804

Random effects:
 Groups Name        Variance Std.Dev.
 name   (Intercept) 3553     59.61
Number of obs: 156, groups:  name, 153

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 14.2497     3.5260   4.041 5.31e-05 ***
consistent_plottableViolate  0.4323     5.5651   0.078    0.938
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)
cnsstnt_plV -0.629
```

(a) Mixed-effects logistic regression.

```
Chi-squared test for given probabilities
X-squared = 0.0066225, df = 1, p-value = 0.9351
```

(b) χ^2 test.

Table 8: Statistical analyses of PaLM 2-M's performance on the NLI tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. There are no significant content effects.

```
Chi-squared test for given probabilities
X-squared = 0, df = 1, p-value = 1
```

Table 9: Statistical analysis of PaLM 2-L's performance on the NLI tasks, using a χ^2 test, as the logistic regression failed to converge. There are no significant content effects.

```
Chi-squared test for given probabilities
X-squared = 0.0064516, df = 1, p-value = 0.936
```

Table 10: Statistical analysis of Flan-PaLM 2's performance on the NLI tasks, using a χ^2 test, as the logistic regression failed to converge. There are no significant content effects.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ consistent_plottable + (1 | name)
   Data:
nli_joint_df %>% filter(subject == "GPT-3.5", consistent_plottable !=
    "Nonsense")

     AIC      BIC   logLik deviance df.resid
    31.3     40.5    -12.7     25.3      153

Scaled residuals:
      Min        1Q    Median        3Q       Max
-0.078288  0.000867  0.000867  0.001145  0.001145

Random effects:
 Groups Name        Variance Std.Dev.
 name   (Intercept) 3151     56.13
Number of obs: 156, groups:  name, 153

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 14.0988     3.3443   4.216 2.49e-05 ***
consistent_plottableViolate  -0.5577     4.1579  -0.134    0.893
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Mixed-effects logistic regression.

```
Chi-squared test for given probabilities
X-squared = 0.027027, df = 1, p-value = 0.8694
```

(b) χ^2 test.

Table 11: Statistical analyses of GPT-3.5-turbo-instruct's performance on the NLI tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. There are no significant content effects.

C.2 Syllogisms

We report mixed effects logistic regressions for humans and all models in Tables 12-17. We analyze these results using a variable which corresponds to the main content effect (`logic_belief_consistent`), which is 1 when the logical answer matches the believability of the conclusion — i.e. when the argument is valid and the conclusion is believable, or the argument is invalid and the conclusion is unbelievable — and 0 when there is a mismatch. This measure corresponds to the difference score reported in Fig. 2b. We ran three nested models for humans and each language model — one regression only incorporating the content effect predictor (whether the logic matches the consistency), another adding consistency condition, and a third adding the interaction of the two.

```
response_correct ~ logic_belief_consistent + (1 | syllogism_name)
response_correct ~ logic_belief_consistent + consistent_plottable_f + (1 | syllogism_name)
response_correct ~ logic_belief_consistent * consistent_plottable_f + (1 | syllogism_name)
```

For humans and each language model, we report the best-fitting regression, measured by the BIC (and omitting models which failed to converge). However, since the interaction effect is theoretically interesting (e.g. 53), and several of the interaction models fail to converge, we also report two-way χ^2 tests of the interactions for each model. For PaLM 2-L all regressions failed to converge due to ceiling effects; thus we also report a χ^2 test of the content effect for this model only. All models show a significant content effect; all except Chinchilla and PaLM 2-M show a significant interaction.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent * consistent_plottable_f +
    (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == this_subject)

     AIC      BIC   logLik deviance df.resid
   692.8    715.1   -341.4    682.8      633

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7130 -0.6097  0.3328  0.6018  1.9316

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0.08992  0.2999
Number of obs: 638, groups:  syllogism_name, 12

Fixed effects:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    0.7182     0.1359   5.286 1.25e-07 ***
logic_belief_consistent1                       1.9502     0.2081   9.372  < 2e-16 ***
consistent_plottable_f1                        0.1863     0.2114   0.881    0.378
logic_belief_consistent1:consistent_plottable_f1  -2.4800     0.4172  -5.945 2.76e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Mixed-effects logistic regression.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 11.402, df = 1, p-value = 0.0007338
```

(b) χ^2 test of interaction.

Table 12: Statistical analyses of human performance on the Syllogism tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test of the interaction effect. There is a significant content effect and a significant interaction effect, that is, different sensitivity to logic in the Consistent compared to Violate conditions.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent + (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == this_subject)

     AIC      BIC   logLik deviance df.resid
   125.8    133.5    -59.9    119.8       93

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7321 -0.8819  0.5774  0.5774  1.1339

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0        0
Number of obs: 96, groups:  syllogism_name, 12

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.4236     0.2212   1.915  0.05549 .
logic_belief_consistent1   1.3499     0.4425   3.051  0.00228 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Mixed-effects logistic regression.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 6.0122e-31, df = 1, p-value = 1
```

(b) χ^2 test of interaction.

Table 13: Statistical analyses of Chinchilla's performance on the Syllogism tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test of the interaction effect. Chinchilla shows significant content effects, but no interaction with consistency.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent + (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == this_subject)

     AIC      BIC   logLik deviance df.resid
   102.0    109.6    -48.0     96.0       93

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4202 -0.6095  0.4132  0.4132  1.6408

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0        0
Number of obs: 96, groups:  syllogism_name, 12

Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.3886     0.2611   1.488    0.137
logic_belief_consistent1   2.7581     0.5222   5.281 1.28e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Mixed-effects logistic regression.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 0, df = 1, p-value = 1
```

(b) χ^2 test of interaction.

Table 14: Statistical analyses of PaLM 2-M's performance on the Syllogism tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. PaLM 2-M shows significant content effects, but no interaction with consistency.

```
Chi-squared test for given probabilities
X-squared = 6.3913, df = 1, p-value = 0.01147
```

(a) χ^2 test of content effect.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 14.318, df = 1, p-value = 0.0001544
```

(b) χ^2 test of interaction.

Table 15: Statistical analyses of PaLM 2-L's performance on the Syllogism tasks, using (a) a χ^2 test of the content effect as none of the regressions converged, and (b) a χ^2 test of the interaction. PaLM 2-L shows both significant content effects, and a significant interaction with consistency (as measured by the χ^2 test, as the regression with an interaction failed to converge).

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent + (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == this_subject)

     AIC      BIC   logLik deviance df.resid
   100.1    107.8    -47.0     94.1       93

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3166 -1.0000  0.3015  0.4761  1.0000

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0        0
Number of obs: 96, groups:  syllogism_name, 12

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.1989     0.2984   4.019 5.86e-05 ***
logic_belief_consistent1   2.3979     0.5967   4.019 5.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Mixed-effects logistic regression.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 17.913, df = 1, p-value = 2.312e-05
```

(b) χ^2 test of interaction.

Table 16: Statistical analyses of Flan-PaLM 2's performance on the Syllogism tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. Flan-PaLM 2 shows both significant content effects, and a significant interaction with consistency (as measured by the χ^2 test, as the regression with an interaction failed to converge).

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ logic_belief_consistent + (1 | syllogism_name)
   Data: syllogism_model_df %>% filter(subject == this_subject)

     AIC      BIC   logLik deviance df.resid
   131.8    139.5    -62.9    125.8       93

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4832 -0.9199  0.6742  0.6742  1.0871

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0        0
Number of obs: 96, groups:  syllogism_name, 12

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.3107     0.2127   1.461   0.1440
logic_belief_consistent1   0.9555     0.4253   2.247   0.0247 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Mixed-effects logistic regression.

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 25.507, df = 1, p-value = 4.408e-07
```

(b) χ^2 test.

Table 17: Statistical analyses of GPT-3.5-turbo-instruct's performance on the syl tasks, using (a) a mixed-effects logistic regression or (b) a χ^2 test. GPT-3.5 shows both significant content effects, and a significant interaction with consistency (as measured by the χ^2 test, as the regression with an interaction failed to converge).

C.3 Wason

We report mixed-effects logistic regressions for humans (both all humans, and the fast and slow groups individually) and all models in Tables 18-25. We observe a significant effect of content in most cases. However, the fast humans alone do not show a significant content effect. Furthermore, the content effects in PaLM 2-M and Flan-PaLM 2 are only marginally significant, due to high item level variance.

In Appx. C.3.1 we further analyze the human data whil incorporating response time in the regression.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data: wason_joint_df %>% filter(subject_no_rt == "Human", wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
     478      491     -236      472      571

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.7162 -0.4629 -0.3279 -0.2881  3.4711

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.2309   0.4805
Number of obs: 574, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.2951     0.2554  -8.988   <2e-16 ***
wason_conditionRealistic   0.8219     0.3235   2.541   0.0111 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 18: Statistical analysis of human performance (collapsing across fast and slow subjects) on the Wason tasks, using a logistic regression. There is a significant content effect.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   305.1    317.4   -149.6    299.1      442

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.5732 -0.3620 -0.3034 -0.2709  3.7157

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.273    0.5225
Number of obs: 445, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.4740     0.2948  -8.393   <2e-16 ***
wason_conditionRealistic   0.4570     0.3877   1.179    0.239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 19: Statistical analysis of human (fast subjects only) performance on the Wason tasks, using a logistic regression. We do not observe a significant content effect

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   156.8    165.4    -75.4    150.8      126

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.8934 -0.7468 -0.4277  1.1193  2.3380

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.1906   0.4365
Number of obs: 129, groups:  wason_name, 24

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.6534     0.4303  -3.842 0.000122 ***
wason_conditionRealistic   1.1518     0.5009   2.299 0.021495 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 20: Statistical analysis of human (slow) performance on the Wason tasks, using a logistic regression. There is a significant content effect.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   136.9    146.0    -65.5    130.9      147

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9739 -0.2756 -0.1256  0.4490  2.6782

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 6.068    2.463
Number of obs: 150, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.584      1.154  -3.106  0.00190 **
wason_conditionRealistic    3.576      1.354   2.640  0.00828 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 21: Statistical analysis of Chinchilla's performance on the Wason tasks, using a logistic regression. There is a significant content effect.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   115.5    124.6    -54.8    109.5      147

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.14109 -0.13581 -0.04492  0.15813  1.50757

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 30.12    5.488
Number of obs: 150, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -5.842      2.296  -2.544   0.0110 *
wason_conditionRealistic    5.122      2.777   1.844   0.0651 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 22: Statistical analysis of PaLM 2-M's performance on the Wason tasks, using a logistic regression. There is a marginally-significant content effect, due to high item-level variance.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   142.6    151.6    -68.3    136.6      147

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2846 -0.1708  0.1686  0.2894  2.2759

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 9.488    3.08
Number of obs: 150, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.921      1.182  -1.625   0.1043
wason_conditionRealistic    3.908      1.759   2.222   0.0263 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 23: Statistical analysis of PaLM 2-L's performance on the Wason tasks, using a logistic regression. There is a significant content effect.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   129.1    138.1    -61.6    123.1      147

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4634 -0.2275 -0.1330  0.1271  2.2738

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 18.14    4.259
Number of obs: 150, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.143      1.591  -1.347   0.1778
wason_conditionRealistic    4.539      2.551   1.779   0.0752 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 24: Statistical analysis of Flan-PaLM 2's performance on the Wason tasks, using a logistic regression. There is a marginally-significant content effect, due to high item-level variance.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + (1 | wason_name)
   Data:
wason_joint_df %>% filter(subject == this_subject, wason_condition %in%
    c("Arbitrary", "Realistic"))

     AIC      BIC   logLik deviance df.resid
   154.3    163.3    -74.1    148.3      147

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1094 -0.4257 -0.1980  0.3909  2.3489

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 5.008    2.238
Number of obs: 150, groups:  wason_name, 25

Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.105      0.867  -2.428  0.01517 *
wason_conditionRealistic    3.092      1.188   2.603  0.00923 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 25: Statistical analysis of GPT-3.5-turbo-instruct's performance on the Wason tasks, using a logistic regression. There is a significant content effect.

C.3.1 Human analyses incorporating response time

Here we present two regression analyses of the human results that incorporate the response time. In Table 26 we show a mixed-effects logistic regression controlling for log response time; the content effect remains significant. Thus, the content effects are not solely driven by the differences in response time noted above (Appx. 24).

However, it is also possible to conceive of the shift in response time as *a part of* the content effect. We can analyze the data this way by z-scoring response time within each condition; thus, the effect of the mean difference in response time will be included in the condition predictor. We present these results in Table 27. Both content and z-scored response time remain significant predictors of success.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + scale(log(rt)) + (1 | wason_name)
   Data: wason_human_correct_df

     AIC      BIC   logLik deviance df.resid
   459.5    476.9   -225.8    451.5      570

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0996 -0.4417 -0.3265 -0.2246  4.5293

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.2539   0.5039
Number of obs: 574, groups:  wason_name, 25

Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.3248     0.2655  -8.755  < 2e-16 ***
wason_conditionRealistic   0.6659     0.3350   1.988   0.0468 *
scale(log(rt))             0.5637     0.1269   4.442 8.89e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 26: Statistical analysis of human (both fast and slow) performance on the Wason tasks, using a logistic regression and also controlling for (log) response time. The content effect remains significant.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: response_correct ~ wason_condition + zscored_rt_by_condition +
    (1 | wason_name)
   Data: wason_human_correct_df

     AIC      BIC   logLik deviance df.resid
   459.5    476.9   -225.8    451.5      570

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1002 -0.4417 -0.3265 -0.2247  4.5269

Random effects:
 Groups     Name        Variance Std.Dev.
 wason_name (Intercept) 0.254    0.504
Number of obs: 574, groups:  wason_name, 25

Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.4329     0.2700  -9.011  < 2e-16 ***
wason_conditionRealistic   0.8793     0.3349   2.626  0.00865 **
zscored_rt_by_condition    0.5540     0.1247   4.443 8.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 27: Statistical analysis of human (both fast and slow) performance on the Wason tasks, using a logistic regression and also controlling for (log) response time, but z-scored *within* condition. Again the content effect is significant.

C.3.2 Multinomial regression of the response patterns on the Wason tasks

In Table 28 we present the results of a multinomial logistic regression predicting which of the six possible subsets of answers the humans and language models chose on the Wason task. This regression quantitatively supports the claim that the behavior is nonrandom, and more generally quantifies the qualitative observations of response patterns made in the main text.

A matrix: 5 × 9 of type dbl

	(Intercept)	Realist	Nonsense	humanslow	chinchi	palm2_m	palm2_l	flanpalm2	gpt3.5
AT,CF (correct)	-2.64	1.43	0.75	1.20	2.31	2.31	3.17	3.19	2.34
AT,AF	-1.24	-1.35	-2.16	0.32	0.08	-0.21	-0.06	0.53	-0.59
AF,CT	-2.65	0.06	0.32	0.02	2.75	2.81	2.72	2.29	2.29
AF,CF	-2.45	0.64	0.21	0.22	2.31	1.38	2.05	2.27	1.13
CT,CF	-2.47	0.30	-1.07	0.97	-1.16	-0.60	-12.95	-12.33	-13.45

(a) Coefficients.

A matrix: 5 × 9 of type dbl

	(Intercept)	Realist	Nonsense	humanslow	chinchi	palm2_m	palm2_l	flanpalm2	gpt3.5
AT,CF (correct)	0.18	0.17	0.17	0.25	0.24	0.23	0.25	0.25	0.22
AT,AF	0.16	0.29	0.44	0.35	0.44	0.46	0.51	0.41	0.49
AF,CT	0.22	0.20	0.19	0.50	0.28	0.27	0.30	0.33	0.28
AF,CF	0.20	0.20	0.21	0.38	0.27	0.30	0.31	0.29	0.31
CT,CF	0.26	0.33	0.56	0.35	1.03	0.75	0.00	306.59	0.00

(b) Standard errors.

Table 28: Results of a multinomial logistic regression predicting the answer choices (reference level is AT,CT — the matching bias) from participants and language models based on condition (reference level is the Arbitrary condition), and participant group (reference level is fast humans). The regression was performed with dummy coding, so coefficients represent the difference in log odds relative to the reference level in each case. We present both the (a) coefficients estimated by the regression and (a) their standard errors. There are a variety of noticeable effects, including the overall matching bias in the fast humans (the fact that the intercept coefficients are all negative), the basic content effect that Realitic problems are more likely to yield correct answers, and the finding that language models and slow humans tend to give correct answers more often than fast humans. Additionally, many qualitative patterns reported in Fig. 9 are statistically borne out by this analysis. Note that due to some models rarely giving some responses, certain coefficient estimates are unstable, particularly in the CT,CF row.

C.4 Response time and model log-probability differences

In this section we present the mixed-effects linear regressions comparing human response times and model log-probabilities on the NLI and syllogisms tasks, in Tables 29 and 30, respectively. In order to make these comparisons, we breakdown each problem into cases where both humans

and models got it correct, and cases where both got it wrong, and only compare log-probabilities and response times within these cases. This breakdown is necessary to control for accuracy in these models, as it is a significantly related to both response times and log-probabilities. Note, however, that this means that problems where a model answered correctly but humans never answered correctly, or vice versa, are omitted.

In both tasks, we see significant effects of the content on the model log-probability differences; even controlling for these we see significant relationships to the human response times, such that on items on which the humans respond more slowly, the models show smaller differences in log-probabilities.

```
Linear mixed model fit by REML ['lmerMod']
Formula:
zscored_logprob_diff ~ log(Human) + consistent_plottable + response_correct +
    (1 | model) + (1 | name)
   Data: nli_logprob_rt_corr_df

REML criterion at convergence: 2074.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7422 -0.6124  0.0099  0.6353  3.8258

Random effects:
 Groups   Name        Variance Std.Dev.
 name     (Intercept) 0.2772   0.5265
 model    (Intercept) 0.0000   0.0000
 Residual             0.5468   0.7394
Number of obs: 831, groups:  name, 171; model, 5

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                    0.7544     0.5353   1.409
log(Human)                    -0.5392     0.1590  -3.392
consistent_plottableConsistent   0.6136     0.1345   4.563
consistent_plottableNonsense    -0.1841     0.1427  -1.291
response_correctTRUE            0.7889     0.2383   3.311
```

Table 29: Statistical analysis of the relationship between human response times and language model log-probability differences on the NLI tasks, using a mixed-effects regression controlling for the task variables and answer correctness, as well as random effects of the item and LM. Note that the model log-probabilities are significantly affected by the content, even though the model accuracy is not.

```
Linear mixed model fit by REML ['lmerMod']
Formula: zscored_logprob_diff ~ log(Human) + logic_belief_consistent +
    consistent_plottable + response_correct + (1 | model) + (1 |
    syllogism_name)
   Data: syllogism_logprob_rt_corr_df_2

REML criterion at convergence: 1077

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.33241 -0.73009 -0.02953  0.61543  2.99660

Random effects:
 Groups         Name        Variance Std.Dev.
 syllogism_name (Intercept) 0.055391 0.23535
 model          (Intercept) 0.005615 0.07493
 Residual                   0.829587 0.91082
Number of obs: 394, groups:  syllogism_name, 36; model, 5

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                   1.18573    0.70847   1.674
log(Human)                   -0.41458    0.20342  -2.038
logic_belief_consistent       0.08477    0.05899   1.437
consistent_plottableviolate  -0.27369    0.07090  -3.860
consistent_plottablenonsense -0.10094    0.09245  -1.092
response_correctTRUE          0.49560    0.10421   4.756
```

Table 30: Statistical analysis of the relationship between human response times and language model log-probability differences on the Syllogisms tasks, using a mixed-effects regression controlling for the task variables and answer correctness, as well as random effects of the item and LM.