# A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models

**Tiwalayo Eisape,**[†] **MH Tessler,**[‡] **Ishita Dasgupta,**[‡] **Fei Sha,**[§]
**Sjoerd van Steenkiste,**[§,*] **Tal Linzen**[§,*]
Massachusetts Institute of Technology[†], Google DeepMind[‡], Google Research[§]

## Abstract

A central component of rational behavior is logical inference: the process of determining which conclusions follow from a set of premises. Psychologists have documented several ways in which humans' inferences deviate from the rules of logic. Do language models, which are trained on text generated by humans, replicate such human biases, or are they able to overcome them? Focusing on the case of syllogisms—inferences from two simple premises—we show that, within the PaLM 2 family of transformer language models, larger models are more logical than smaller ones, and also more logical than humans. At the same time, even the largest models make systematic errors, some of which mirror human reasoning biases: they show sensitivity to the (irrelevant) ordering of the variables in the syllogism, and draw confident but incorrect inferences from particular syllogisms (syllogistic fallacies). Overall, we find that language models often mimic the human biases included in their training data, but are able to overcome them in some cases.

## 1 Introduction

The capacity to reason deductively—that is, to determine which inferences, if any, follow from a given set of premises—is central to rational thought (Newell and Simon, 1972; Laird et al., 1987; Fodor and Pylyshyn, 1988; Griffiths et al., 2010). Despite the importance of this capacity, human reasoning often displays systematic biases (Gigerenzer and Gaissmaier, 2011; Marcus, 2009; Kahneman, 2013; McClelland et al., 2010). In recent years, language models (LMs) trained with self-supervised objectives have been reported to display a range of capabilities, including the ability to reason (Brown et al., 2020; Chowdhery et al., 2022; Bubeck et al., 2023). Does LMs' logical reasoning follow the rules of logic to a greater extent than humans'? To the extent that LMs' reasoning deviates from normative logic, are their biases similar to humans' (Binz and Schulz, 2023; Dasgupta et al., 2022)?

In this work, we address these questions with a detailed study of a particularly simple case—inferences from pairs of premises, or *syllogisms*, such as the following:

> If **all bakers are artists**,
> and **some bakers are chemists**,
> ---
> then: **some artists are chemists**.

In a syllogism, each premise relates two terms with one of four quantifiers (traditionally known as "moods"): *all*, *some*, *none* and *some are not*. Only one term is shared between the premises (*bakers* in the example above). Inference is required to determine if there is a necessary relationship between the two remaining terms (here, *artists* and *chemists*) when the premises in question are true.

When human participants in experiments are asked to make syllogistic inferences, their responses often deviate from the rules of logic; in fact, for some syllogisms the vast majority of participants draw incorrect inferences (Khemlani and Johnson-Laird, 2012). This could pose a challenge to language models (LMs), as they learn from corpora consisting primarily of human-generated texts—texts which, in turn, reflect human beliefs and inferences. Is there sufficient signal in the training corpus to steer LMs away from (often incorrect)

[†]Work done when TE was a student researcher at Google Research. [*]SVS and TL are joint senior authors. Author contributions: TE, SVS, and TL co-led the project. TE conducted the experiments and analysis. TE, MHT, ID, FS, SVS, and TL helped with project framing, analysis and suggesting experiments. ID and SVS offered technical guidance and help with engineering. TE wrote the paper with help from MHT, ID, FS, SVS, and TL. Correspondence: eisape@mit.edu, svansteenkiste@google.com, linzen@google.com.



*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444
June 16-21, 2024 

| | | | |
|---|---|---|---|
| **A**: | **All** artists are bakers | **I**: | **Some** artists are bakers |
| **E**: | **No** artists are bakers | **O**: | **Some** artists **are not** bakers |

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| A-B | B-A | A-B | B-A |
| B-C | C-B | C-B | B-C |

Table 1: Syllogism moods (left) and variable orderings (right).

human inferences and toward a behavior consistent with the normative rules of logic—the behavior that is desirable for most applications?

We address this question in a detailed comparison between the PaLM 2 family of transformer LMs (Google, 2023) and studies from cognitive psychology, as well as a replication with the Llama 2 family of transformer LMs (Touvron et al., 2023). We report the following results:

1. LMs draw correct inferences more often than humans, and larger LMs tend to be more accurate than smaller ones, but the accuracy of even the best performing LM is only about 75%, and scale does not consistently lead to accuracy gains (Section 4.1).

2. LM errors are systematic, with very low accuracy on particular syllogism types (Section 4.1); the syllogisms that LMs struggle with are a subset of those that humans find difficult (Section 4.2).

3. Like humans, LMs are sensitive to the ordering of terms in the premises of a syllogism even when it is logically irrelevant (Section 4.2; this pattern is known as the "figural effect" in cognitive psychology; Johnson-Laird and Steedman 1978).

4. LMs show many of the same *syllogistic fallacies*, characterized by high confidence and low accuracy, as humans. Larger LMs are somewhat more susceptible to these fallacies than smaller ones (Section 4.2; Khemlani and Johnson-Laird 2017).

5. Using the Mental Models theory from cognitive psychology, we find quantitative evidence that larger LMs reason more deliberatively than smaller ones (Section 5; Khemlani and Johnson-Laird 2022).

Overall, we find that PaLM 2 LMs replicate many of the human biases discovered in psychology studies, consistent with the fact that LMs are trained on human-generated text. For some syllogisms, however, sufficiently large models overcome those biases and achieve dramatically better accuracy than humans, although their overall accuracy is still far from the perfect logical reasoner.

## 2 Background and Related Work

### 2.1 Syllogisms

Syllogisms are logical arguments consisting of two *premises* relating three variables, A, B and C (e.g., *artists*, *bakers*, and *chemists* in the example from the introduction). Each premise relates just two of the variables, through one of four quantifiers, often referred to as "moods" (Table 1, left). The variables in each of the premises can be ordered in either of the two directions—e.g., *all artists are bakers* vs. *all bakers are artists*—and so there are four possible pairs of orderings (Table 1, right). These orderings are traditionally referred to as "figures", but we will use the more transparent term "variable ordering". Taking the cross product of these building blocks yields 64 possible syllogisms: two premises, each of which can take one of four quantifiers and one of two possible orderings.

Though the premises only relate A and B, or B and C—never A and C—27 of the 64 syllogisms imply a quantified relationship between A and C (e.g., *some A are C*). In the remaining 37 syllogisms, no relation between A and C can be deduced; in human experiments, the expected response to these syllogisms is "nothing follows" (see Figure 10 in Appendix A for the full set of valid conclusions for each syllogism).

### 2.2 Human Syllogistic Reasoning

Psychologists, going back to the early 20th century, have found that the conclusions that humans draw from the premises of a syllogism often deviate from logical norms (for a review, see Khemlani and Johnson-Laird 2012). These errors are systematic: some syllogisms are much harder than others, and the incorrect conclusions that participants tend to draw are consistent across participants. For example, from the two premises (1) *no artists are bakers* and (2) *all bakers are chemists*, the vast majority of

participants incorrectly conclude that it is the case that *no artists are chemists*. We analyse such cases in detail in Section 4.2.

In addition to these specific, highly challenging syllogisms, several broader reasoning biases have been documented. When given a syllogistic argument where the variables in the premises are ordered "A-B, B-C", participants show a bias towards conclusions with an A-C ordering, even though reordering the variables in the premises does not affect the conclusions licensed by the syllogism (Johnson-Laird and Steedman, 1978). Participants are also more likely to produce a conclusion when it is true in the real world, independently of whether it follows from the premises ("content effects", Evans et al. 1983).

A number of theories have been proposed to explain human syllogistic reasoning. An influential account that we focus on in this work is the Mental Models Theory (Johnson-Laird and Byrne, 1991). This theory posits that human reasoners construct mental models populated by a small number of entities that instantiate the premises; e.g., to instantiate the premise *all artists are bakers*, a reasoner might construct a world with three specific artists, all of whom are bakers. These worlds are constructed based on a number of fallible heuristics, and human reasoning errors arise when those heuristics produce incorrect conclusions (see Section 5).

### 2.3 Language Models and Reasoning

LMs trained with self-supervised objectives on large text corpora have been instrumental in achieving high performance on a range of tasks. Some of the tasks in which LMs have shown promise have been referred to as reasoning tasks, including commonsense reasoning, natural language inference, or question answering (e.g., Chowdhery et al. 2022). In this work, we focus more specifically on deductive logical reasoning: drawing conclusions that *must*, rather than are likely to, be true given the premises, and where the inference is based only on the premises, and does not rely on world knowledge. Unlike work on datasets collected from textbooks or through crowdsourcing, we perform a well-controlled analysis of a simple logical task for which there is a wealth of human data.

Several studies have benchmarked LMs on logical reasoning tasks (Han et al., 2022; BIG-bench collaboration, 2022; Wu et al., 2023a; Betz et al., 2020; Saparov and He, 2022; Saparov et al., 2023; Ye et al., 2023) and examined LM reasoning biases

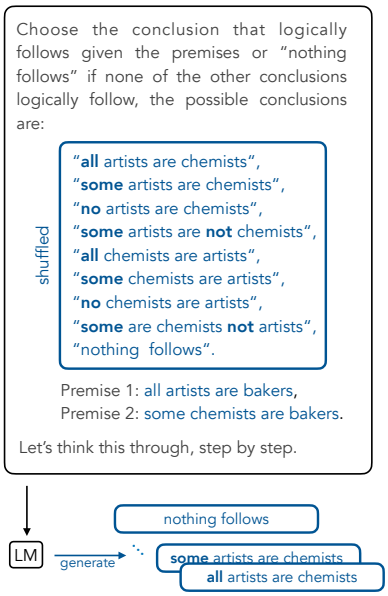



Figure 1: The zero-shot chain-of-thought prompt we use to assess LM syllogistic reasoning. The different parts of the prompt are grouped together for illustration purposes only; see also Figure 11 in the Appendix for a purely textual representation of the prompt.

(Dasgupta et al., 2022; Razeghi et al., 2022; Wu et al., 2023b; McCoy et al., 2023). Saparov and He (2022) take a similarly controlled experimental approach to ours (see also Saparov et al. 2023), but they analyze LMs' performance on formal logic rather than problems phrased in natural language as we do, and do not compare their results to humans. The closest study to ours is Dasgupta et al. (2022), which demonstrates content effects in a number of logical reasoning domains, including syllogisms. We extend their approach to study other aspects of syllogistic reasoning.

## 3 Methods

### 3.1 Data

The human behavioral data we use is drawn from Ragni et al. (2019), an online experiment where 139 participants responded once to each of the syllogisms. In each trial, a participant was presented with a syllogism and was instructed to choose among nine options: the eight possible conclusions and "nothing follows". The experimental trials were preceded by a brief training phase where participants were familiarized with the task.

Following Ragni et al. (2019), we generate syllogisms by replacing the abstract terms (A, B, C) in each syllogism with one of 30 content triples

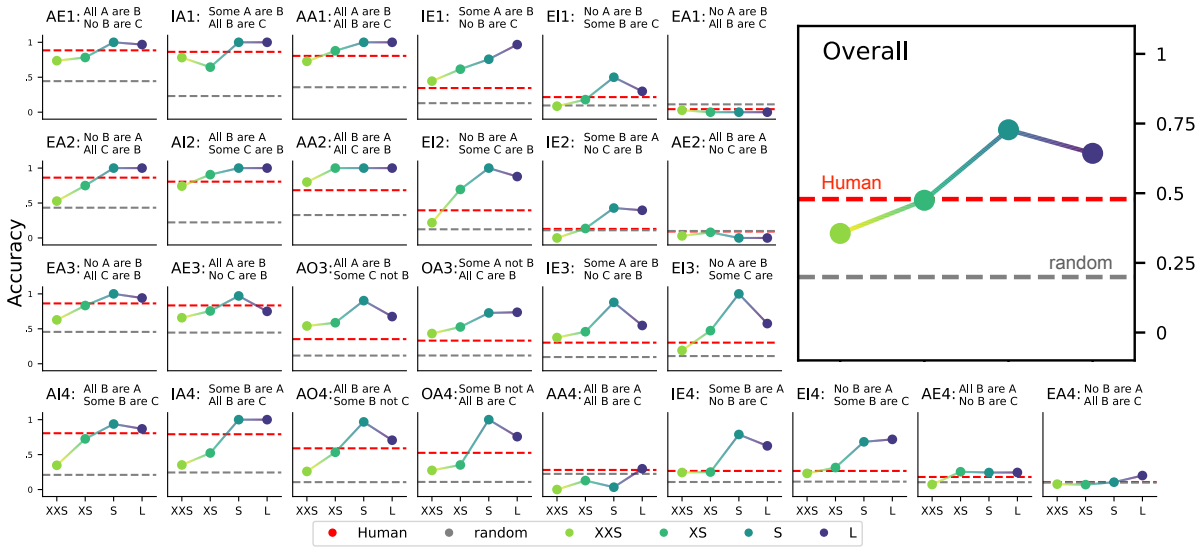



Figure 2: Accuracy of PaLM 2 models, humans (red), and random guessing (grey). Random guessing accuracy differs by syllogism as some syllogisms have more than one valid conclusion. Syllogisms are partitioned into variable ordering (by row) and ordered by decreasing human accuracy from left to right. The top right inset shows the average accuracy across all syllogisms. Syllogisms are identified with the letters of the moods of the premises (Table 1, left) and the number associated with their variable ordering (Table 1, right).

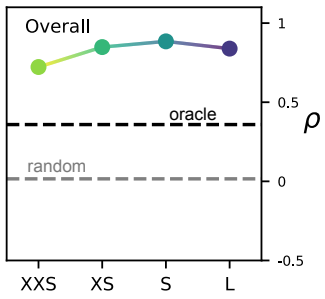



Figure 3: Correlation between PaLM 2 models' predictions and human predictions. The oracle here is a logically correct reasoner that samples a response at random from all valid responses; the correlation of such an oracle with humans is relatively low as it does not mimic human errors.

chosen such that there is no obvious semantic association between the terms (e.g., one of the triplets included *hunters*, *analysts* and *swimmers*; see Appendix A for the full list). This resulted in $64 \times 30 = 1920$ unique data points.

### 3.2 Models and Inference

Most of our analyses focus on the PaLM 2 family of LMs, which are publicly available in four sizes (XXS, XS, S, and L; Google 2023). These are transformer LMs trained on a large corpus of multilingual web documents, books, code, mathematics and conversations. We also repeat all of our analyses for the 7B-, 13B- and 70B-parameter versions of the Llama 2 family of transformer models (Touvron et al., 2023). Since, unlike for PaLM 2, we were unable to explore the different hyperparameters of our evaluation method for these models, we regard these results as preliminary and summarize them separately from the PaLM 2 results (Section 4.3 and Appendix E). All of the models we use are pretrained only, without additional fine-tuning to match human preferences.

Following the emerging standard practice for eliciting reasoning from LMs, we use zero-shot "chain-of-thought" prompting, where the model is instructed to "think step by step" (Kojima et al., 2022; Wei et al., 2022). We speculate that the more explicit reasoning process triggered by such prompts may more closely resemble the behavior of human participants in experiments; for an analysis of alternative prompting strategies that we explored before settling on this one, see Appendix B.1. The prompt we use is illustrated in Figure 1. We randomize the order of the conclusions in the prompt to control for LMs' sensitivity to answer ordering (Pezeshkpour and Hruschka, 2023).

For each of the 1920 reasoning problems, we estimated the distribution over conclusions for each LM with a rejection sampling approach. Samples were rejected if no conclusion was identified via uncased exact string match, and we took the LMs'

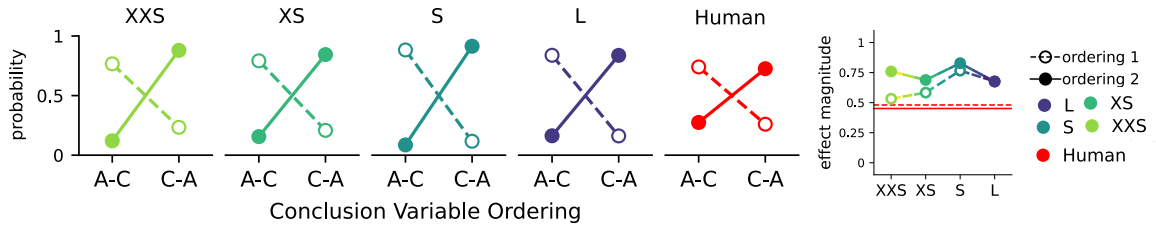



Figure 4: Variable ordering effects in PaLM 2 models and humans. **Left:** The marginal probabilities of A-C and C-A ordered conclusions. **Right:** The magnitude of the variable ordering effect (the absolute value of the difference between the C-A probability and the A-C probability).

response to be the conclusion with the highest probability in this distribution. Each distribution was estimated with 30 such proposals generated with a temperature of 0.5 and a maximum decoding length of 75 tokens. See Appendix B for further details and an exploration of the impact of different prompts and decoding parameters.

## 4 Results[1]

### 4.1 Do LMs Reason Accurately?

We first examine the PaLM 2 LMs' behavior on each of the 64 syllogism types separately. In practice the LMs rarely produced the output "nothing follows", which is the correct conclusion for 37 of the syllogisms. We return to this behavior briefly in Section 4.2, but in most of the following analyses, we restrict ourselves to the 27 syllogisms that license conclusions other than "nothing follows" (see Figure 2 for the full pattern of results on each of those 27 syllogisms). We compute the LMs' accuracy for each syllogism by dividing the number of logically valid conclusions produced by the LM by the total number of responses; note that some syllogisms have more than one valid conclusion (as many as four) and so the random baseline in Figure 2 varies by syllogism.

When averaged across all syllogisms, LM accuracy generally improves with scale, with the two largest models exceeding human accuracy. The relationship between scale and accuracy is not unambiguous, however: the largest model has somewhat lower accuracy than the second-largest one. There is considerable variation across syllogisms; for multiple syllogisms, accuracy is very low for all model sizes and can even decrease as model size increases (this is the case, for example, for EA1: *no A are B, all B are C*).

[1]An earlier version of this work, which was shared on arXiv, reported slightly different results for the experiments discussed due to a sampling bug.

### 4.2 Do LMs Reason Like Humans?

Human accuracy averaged across all syllogisms is roughly 50% (Figure 2; red dashed line); as such, high LM accuracy on this dataset does not necessarily imply humanlike reasoning. A comparison by syllogism type reveals that the syllogisms that PaLM 2 models struggle with are syllogisms that humans also find challenging, but the inverse is not true: multiple syllogisms that are hard for humans are solved correctly by larger models. For example, for the syllogism IE4 (*some B are A, no B are C*), human accuracy is barely above chance, but PaLM 2 Small and PaLM 2 Large are substantially more accurate.

**Comparing the distribution over responses.** So far we have focus on the proportion of correct responses. There are eight possible conclusions; is the distribution over all responses, including incorrect ones, similar across humans and PaLM 2 models? To compute the probability distribution over conclusions for each syllogism, we aggregate response counts for each syllogism and normalize them into a probability distribution as in Khemlani and Johnson-Laird (2016). We then correlate the probability estimates from humans with the estimates from PaLM 2 models across the entire dataset (Figure 3; for a by-syllogism breakdown, see Figure 14 in Appendix C). The correlation is fairly high across models, and is highest for PaLM 2 Small.

PaLM 2 Small and Large display both a high correlation with human responses and a higher-than-human accuracy. This suggests that the miscalibration to human data that models accrue due to higher accuracy is offset by a better fit to humans elsewhere in the dataset. The next analyses test this hypothesis, zooming in on two specific biases.

**Variable ordering effects.** Humans' syllogistic inferences are sensitive to variable ordering, even

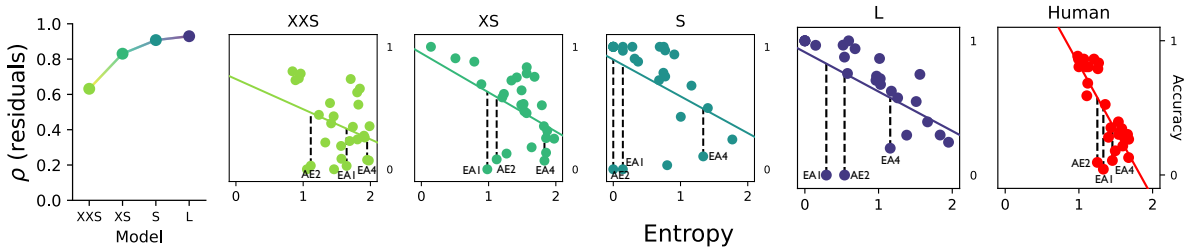



Figure 5: **Right:** Each syllogism plotted by accuracy (y-axis) and entropy (x-axis) and the regression line relating the two. Dashed lines black lines show the residuals for each of the top three human syllogistic fallacies. **Left:** The result of correlating PaLM 2's regression residuals with residuals estimated from human data.

when the ordering is logically irrelevant (Johnson-Laird and Steedman, 1978). Specifically, humans produce more conclusions with an A-C variable ordering when reasoning in response to a syllogism presented in ordering 1 (A-B, B-C); and they produce more conclusions with a C-A ordering when presented with a syllogism in ordering 2 (B-A, C-B). We aggregate the human and LM responses across all (A-B, B-C) syllogisms and across all (B-A, C-B) syllogisms separately and normalize the aggregated response counts. All four PaLM 2 models show an ordering effect in the same direction as humans (Figure 4, left). We compute the magnitude of the effect as $|P\,(\text{A-C}) - P\,(\text{C-A})\,|$, where $P\,(\text{A-C})$ is the probability placed on conclusions with the order A-C. All models display a moderately larger bias than humans (Figure 4, right). We do not find a clear trend in the magnitude of the bias as model size increases; if anything, the largest model shows a slightly weaker bias than the second-largest one.

**Syllogistic fallacies.** In general, humans are well-calibrated syllogistic reasoners—their accuracy is inversely correlated with the entropy of their responses (Figure 5; see also Khemlani and Johnson-Laird 2012). In other words, for most syllogisms where humans give incorrect answers, the particular incorrect answers they give vary substantially across individuals and trials. However, there are exceptions to this tendency: in some cases, humans confidently and consistently choose a particular incorrect answer (that is, low entropy coincides with low accuracy). For example, given the syllogism *no artists are bakers, all bakers are chemists*, humans overwhelmingly respond with the logically invalid conclusion *no artists are chemists*; the correct conclusion, *some chemists are not artists*, is

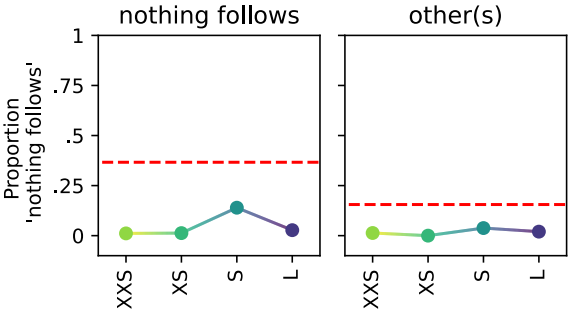



Figure 6: The proportion of "nothing follows" responses from humans and PaLM 2 models on the 37 syllogisms whose only valid conclusion is "nothing follows" (left) and the syllogisms that license conclusions other than "nothing follows" (right).

produced only 3% of the time, and the distribution over responses elicited from humans for this syllogism has one of the lowest entropies in the Ragni et al. (2019) dataset. We refer to such cases as *syllogistic fallacies* (Newsome and Johnson-Laird, 2006; Khemlani and Johnson-Laird, 2017).

To identify potential fallacies in LMs, we fit a regression line relating entropy (in nats) and accuracy, and then compute the distance from this line (the residual) for each syllogism (Figure 5, right; for alternaitve calibration measures, see Guo et al. 2017). The top three human syllogistic fallacies, defined as the top three outliers when plotting accuracy against entropy, are also outliers for the PaLM 2 models. We also correlate the residuals for all 27 syllogisms across humans and LMs, and find that larger models display stronger correlations (Figure 5, left).

**LMs avoid responding "nothing follows".** An important divergence from human behavior is that LMs rarely produce the response "nothing follows", even for the 37 syllogisms for which this is the

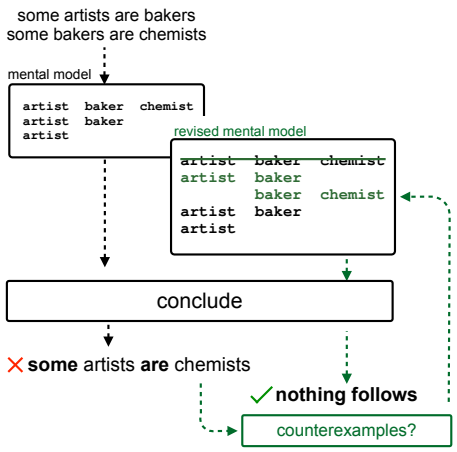



Figure 7: Schematic of mReasoner deducing an incorrect conclusion before finding counterexamples ("System 2" processes shown in green) and updating to the correct conclusion, "nothing follows".

correct conclusion. Humans are also reluctant to conclude "nothing follows" (Ragni et al., 2019), but the LMs' aversion to this response is much stronger than humans'—we observe accuracies close to 0% (Figure 6). This issue is particularly severe with the zero-shot chain-of-throught prompting method we use; in Appendix B.3, we describe an evaluation paradigm that can be used to elicit that conclusion, and leave further analysis of this behavior to future work.

### 4.3 Llama 2 Results

Llama 2 models' overall accuracy, when aggregated across all syllogisms, was similar to human accuracy, with a modest increase in accuracy as scale increases (Figure 18 in the Appendix). However, the breakdown by syllogism shows that this pattern masks substantial differences between humans and Llama 2: unlike PaLM 2 models, Llama 2 models struggle with some syllogisms that humans find easy, such as *some A are B, all B are C*. Llama 2 models do, however, display a human-like variable ordering effect (Figure 20 in the Appendix). We refer the reader to Appendix E for plots of the results and additional analyses of Llama 2 models.

## 5 Interpreting Language Models Using Mental Models Theory

We next analyze the behavior of PaLM 2 models using the Mental Models theory of human logical reasoning (Johnson-Laird, 1983), which has been developed over decades to account for human data. The theory takes humans to be resource-limited and simulation-based reasoners (Craik 1967; Lake et al. 2017; Lieder and Griffiths 2019; Johnson-Laird 1983, i.a.), with a potentially high degree of variability across individuals. The implementation we use—mReasoner[2] (Khemlani and Johnson-Laird, 2022)—captures these aspects of human reasoning with a small set of interpretable hyperparameters that enable it to construct, refine, and draw conclusions from internal mental models of the situations described in a syllogism.

Mental models consist of sets of entities instantiating the premises, where an entity is represented by a conjunction of logical properties. For example, Figure 7 illustrates a mental model corresponding to the syllogism *some artists are bakers, some bakers are chemists*. This model consists of just three entities, the first of whom is an artist who is also a baker and a chemist, the second is an artist and a baker who may or may not be a chemist (this uncertainty is represented in the figure with a blank space), and so on. The reasoner constructs and maintains its mental model with a set of actions parameterized by four hyperparameters:

- **LEN** ($\lambda \in [1, \infty)$): The number of entities generated by the reasoner is sampled from a Poisson distribution with a mean of **LEN**.
- **BROAD** ($\epsilon \in [0, 1]$): Determines the set of individuals that mReasoner samples from. There are two possible sets: a broader set of all individuals consistent with the premises, and a smaller, canonical set of individuals consistent with the premises. The canonical sets were determined from human experiments (Khemlani and Johnson-Laird 2022; for an example, see Figure 15 in Appendix D).
- **SYSTM2** ($\sigma \in [0, 1]$): The reasoner's propensity to reconsider its conclusion and search for counterexamples. Search is conducted by adding an entity to the model, moving a property from one entity to another, or decomposing one entity into two (these strategies are illustrated in Figure 16 in Appendix D).
- **WEAKEN** ($\omega \in [0, 1]$): Determines the model's reaction to finding a counterexample. The reasoner's options in this case are either to respond "nothing follows" or to weaken its response (i.e., amending erroneous global conclusions such as

[2] https://github.com/skhemlani/mReasoner

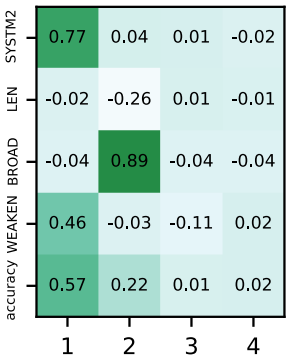



Figure 8: Correlations between the four principal components resulting from our analysis and mReasoner's original parameters (top four rows) as well as accuracy (bottom row).

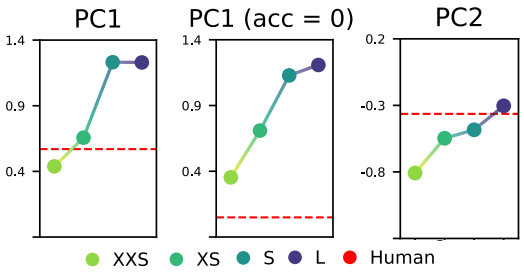



Figure 9: **Left and right:** Projecting PaLM 2 models onto the first two principal components of the feature space resulting from the behavior of simulations using mReasoner. **Center:** Projecting PaLM 2 models onto the same space when only their errors are taken into account.

*all A are C* to weaker particular conclusions such as *some A are C*). When **WEAKEN** is higher, mReasoner is more likely to weaken its response and less likely to answer "nothing follows".

Figure 7 illustrates how mReasoner might process the syllogism *some artists are bakers, some bakers are chemists*. First, it constructs a mental model, with length governed by **LEN** and content governed by **BROAD**, consisting of the entities mentioned above: an artist-baker-chemist, an artist-baker, and an artist. The conclusion *some artists are chemists* is consistent with this particular model (i.e., the first entity is both an artist and a chemist). This conclusion is not true in every model that is consistent with the premises, and as such it is not logically valid; however, if the reasoner does not trigger a System 2 process, it will (incorrectly) take this conclusion as valid and return it. Alternatively, with probability **SYSTM2** mReasoner will scrutinize the conclusion by amending its model in an attempt to find a counterexample. In this case, mReasoner successfully finds a counterexample by breaking the first entity into two entities that are still consistent with the premises but are not consistent with *some artists are chemists*; consequently, mReasoner corrects its answer to "nothing follows".

**Mapping LM predictions onto cognitively meaningful dimensions.** Syllogistic reasoning behavior is high-dimensional; in the set of syllogisms and conclusions we consider, there are 27 syllogisms and eight possible responses to each, for a total of 216 dimensions. We instantiate 1296 mReasoner models, one for each point in a parameter grid, and analyze the 923 of them that finished simulations before timing out (for the details of the parameter grid, see Table 3 in Appendix D.2). We evaluate each instance on each syllogism and represent the instance as a vector in a 216-dimensional space. Finally, we use PCA to identify the top four principal components in this space.

**Characterizing the space of reasoning behaviors described by mReasoner.** Although mReasoner is characterized by four parameters, we find a single principal component (PC 1) that captures 77% of the variance in the model's behavior. This component loads heavily on **SYSTM2** and, to a lesser degree, on **WEAKEN** (Figure 8). Following the terminology of Khemlani and Johnson-Laird (2016), we view this dimension as representing *deliberative* reasoning. Similarly, PC 2 loads heavily on **BROAD**. This dimension, however, describes much less of the behavioral variance of mReasoner.

**LMs show signatures of deliberative reasoning.** We project the 216-dimensional vectors describing the human data as well as the behavior of each of our LMs into the PC space. This allows us to interpret the LMs' behavior, in particular as model size increases, in terms of reasoning strategies (Figure 9). We find that larger LMs behave more like mReasoner instantiations with high **SYSTM2** and **WEAKEN** values, as indicated by the fact that their first principal component is higher; in the terminology of Khemlani and Johnson-Laird (2016), they show a stronger behavioral signature of deliberative reasoning.

**Deliberative reasoning is partly dissociable from accuracy.** PC 1 is strongly correlated not only with **SYSTM2**, but also with accuracy. Can the

changes in coordinates assigned to PaLM 2 be explained simply by differences in accuracy? To test this, we repeat our analysis, this time setting the probabilities of the correct answers to 0 for all mReasoner instantiations, LMs and renormalizing (Figure 9, center). In this control analysis, the accuracy of all models is 0% (by design), but larger models still display more deliberative reasoning. Here the deliberative component (PC 1) has zero correlation with accuracy but a correlation of 0.6 with **SYSTM2**; correlations with all other parameters are below 0.15. This indicates that even the models' errors become more consistent with deliberative reasoning.

## 6 Discussion

**Human-like reasoning or accurate reasoning?** Because of humans' systematic reasoning errors, syllogistic reasoning is a particularly clear demonstration of the tension between the two central aims of artificial intelligence: human-likeness and accuracy. We hypothesize that for most applications, accuracy is more important than human-likeness; one notable exception is cognitive modeling, where the goal is to better understand human reasoning by developing models that reason like humans. We consider this application to be an important direction for future work.

**Why are LMs more accurate than humans?** LMs learn from human-generated text, which is likely to reflect human beliefs and biases; it is natural to hypothesize that the language modeling objective would incentivize LMs to replicate those biases. We find only partial support for this hypothesis. While the largest model's responses are indeed slightly more correlated with human responses than the smaller ones, for some syllogisms where humans reason very poorly, the models overcome human biases and reason correctly. One possible explanation for this finding is that the data that PaLM 2 models were trained on includes not only natural language text, but also source code (Chowdhery et al., 2022), which may teach models to reason more effectively. The effect of the composition of the LM's training corpus can be tested in a controlled comparison in the future.

**Cognitive science for LM interpretation.** We have used cognitive science to shed light on LM reasoning in two ways. First, we used the biases documented in the cognitive psychology literature as hypotheses for the biases that LMs might acquire. This approach is motivated by the hypothesis that because LMs are trained on human-generated texts, which reflect human biases and beliefs, they will be incentivized to replicate those biases to improve perplexity. We found partial support for this hypothesis: larger LMs were more calibrated to human responses in some cases, in particular in our analysis of the correlation between accuracy and entropy (Section 4.2).

The second and more novel way in which we use cognitive science is in interpreting LM behavior using a computational cognitive model developed to explain human reasoning. Under the assumption that LM reasoning follows the same heuristic strategies as humans do (Section 5)—an assumption which, again, is informed by the fact that LMs learn from text generated by humans—we concluded from this analysis that LMs become more deliberative as their size increases.

## 7 Conclusion

Do LMs learn to reason correctly from self-supervised learning alone, even though much of their training data was produced by humans, whose reasoning often deviates from normative logic? We have addressed this question through a detailed examination of the syllogistic reasoning behavior of the PaLM 2 family of LMs. We find that the largest LMs make significantly fewer mistakes than humans but still display systematic errors (Section 4.1), and that while their mistakes are only partly aligned with human errors, LMs are susceptible to several qualitative reasoning biases shown by humans (Section 4.2).

## Acknowledgments

We thank Andrew Lampinen for helpful discussion and Sangeet Khemlani for open-sourcing MReasoner. TE is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

## Ethical Considerations and Limitations

Part of this work's motivation is to extend the understanding of similarities and differences between humans and current LMs, and we hope our work will have broader positive impacts, such as facilitating cognitively informed and ethical model development. Our results are limited, however, with challenges in directly comparing LM and human

behavior, and we comment on specific limitations below.

**Eliciting LM reasoning.** The space of possible ways to evaluate LMs on paradigms from human experiments is fairly large. One can generate from the model (Aina and Linzen, 2021), as we did; elicit meta-level judgements (Hu and Levy, 2023; Beguš et al., 2023); or simply compare the probabilities assigned by the LM to possible continuations (Linzen et al., 2016; Dasgupta et al., 2022). Evaluations can be done in a zero-shot way, as we did, or in a few-shot way, which may better approximate the training phase used in some human reasoning experiments, such as Ragni et al. (2019); for discussion, see Lampinen (2022). Finally, generative approaches can rely on a large set of possible prompts, and can be used with or without "chain-of-thought" statements encouraging the model to reveal its reasoning process (Kojima et al., 2022). Following preliminary experiments (Appendix B), we focused on zero-shot chain-of-thought; a more systematic evaluation of the different elicitation approaches would be an important direction for future work.

**The focus on Mental Models Theory.** In Section 5, we used a particular cognitive model, the Mental Models Theory, to interpret LM reasoning behavior. This is not the only possible mechanism that might underlie LM reasoning. Other accounts of human reasoning have argued that people do, in fact, apply normative logic rules (Rips, 1994), perform probabilistic inference with constrained resources (Chater and Oaksford, 1999), or combine probabilistic, heuristic and pragmatic reasoning (Tessler et al., 2022); and it is possible that LMs reason in a way that does not match any of these theories. We leave a systematic comparison of the fit of each of these theories to LM reasoning for future work.

## References


Laura Aina and Tal Linzen. 2021. The language model understood the prompt was ambiguous: Probing syntactic uncertainty through generation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 42–57, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gašper Beguš, Thomas Lu, and Zili Wang. 2023. Basic syntax from speech: Spontaneous concatenation in unsupervised deep neural networks.

Gregor Betz, Christian Voigt, and Kyle Richardson. 2020. Critical Thinking for Language Models.

BIG-bench collaboration. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Nick Chater and Mike Oaksford. 1999. The probability heuristics model of syllogistic reasoning. *Cognitive psychology*, 38(2):191–258.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Kenneth Craik. 1967. *The Nature of Explanation*, 1 edition. Cambridge University Press.

Ishita Dasgupta, Andrew K Lampinen, Stephanie C Y Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning.

J S Evans, J L Barston, and P Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Mem. Cognit.*, 11(3):295–306.

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

Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annu. Rev. Psychol.*, 62:451–482.

Google. 2023. PaLM 2 Technical Report.

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. 2010. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.*, 14(8):357–364.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: Natural Language Reasoning with First-Order Logic.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations.

Philip N Johnson-Laird and Mark Steedman. 1978. The psychology of syllogisms. *Cogn. Psychol.*, 10(1):64–99.

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA.

Philip Nicholas Johnson-Laird and Ruth MJ Byrne. 1991. *Deduction*. Lawrence Erlbaum Associates, Inc.

Daniel Kahneman. 2013. *Thinking, Fast and Slow*, 1 edition. Farrar, Straus and Giroux.

Sangeet Khemlani and P N Johnson-Laird. 2012. Theories of the syllogism: A meta-analysis. *Psychol. Bull.*, 138(3):427–457.

Sangeet Khemlani and P N Johnson-Laird. 2016. How people differ in syllogistic reasoning. https://modeltheory.org/papers/2016syllogisms-indvl-diffs.pdf. Accessed: 2023-6-22.

Sangeet Khemlani and P N Johnson-Laird. 2022. Reasoning about properties: A computational theory. *Psychol. Rev.*, 129(2):289–312.

Sangeet S Khemlani and P N Johnson-Laird. 2017. Illusions in reasoning. *Minds Mach.*, 27(1):11–35.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

John E Laird, Allen Newell, and Paul S Rosenbloom. 1987. SOAR: An architecture for general intelligence. *Artif. Intell.*, 33(1):1–64.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behav. Brain Sci.*, 40:e253.

Andrew Kyle Lampinen. 2022. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans.

Falk Lieder and Thomas L Griffiths. 2019. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.*, 43:e1.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Gary Marcus. 2009. *Kluge: The Haphazard Evolution of the Human Mind*, first edition edition. Mariner Books.

James L McClelland, Matthew M Botvinick, David C Noelle, David C Plaut, Timothy T Rogers, Mark S Seidenberg, and Linda B Smith. 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.*, 14(8):348–356.

Thomas R McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve.

Allen Newell and Herbert A Simon. 1972. Human problem solving. 920.

Mary R Newsome and P N Johnson-Laird. 2006. How falsity dispels fallacies. *Think. Reason.*, 12(2):214–234.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions.

Marco Ragni, Hannah Dames, Daniel Brand, and Nicolas Riesterer. 2019. When does a reasoner respond: Nothing follows? In *CogSci*, pages 2640–2546.

Yasaman Razeghi, Robert L Logan, IV, Matt Gardner, and Sameer Singh. 2022. Impact of Pretraining Term Frequencies on Few-Shot Reasoning.

Lance J Rips. 1994. *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Abulhair Saparov and He He. 2022. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples.

Michael Henry Tessler, Joshua B Tenenbaum, and Noah D Goodman. 2022. Logic, probability, and pragmatics in syllogistic reasoning. *Top. Cogn. Sci.*

Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *J. R. Stat. Soc. Series B Stat. Methodol.*, 61(3):611–622.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models.

Yongkang Wu, Meng Han, Yutao Zhu, Lei Li, Xinyu Zhang, Ruofei Lai, Xiaoguang Li, Yuanhang Ren, Zhicheng Dou, and Zhao Cao. 2023a. Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2347–2367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023b. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks.

Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. 2023. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14773, Stroudsburg, PA, USA. Association for Computational Linguistics.

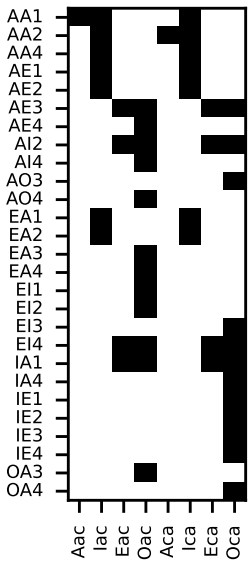



Figure 10: Valid conclusions for each syllogism. Conclusion identifiers show the conclusion mood (see Table 1) followed by 'ac' if the first variable in the conclusion is A and the second is C and 'ca' in the opposite case.

## A Further Details on Syllogism Dataset

Table 2 displays the full list of the content triples used in our experiments. The words in each triple were chosen to have minimal semantic associations with each other.

## B Prompting and Evaluation

Before settling on the generative chain-of-thought evaluation strategy that we focus on in this paper (described in detail in Section B.1), we explored two additional strategies for eliciting and scoring syllogistic inferences from LMs. First, we explored a multiple-choice approach, where, following the prompt, we computed the mutual information between the prompt and each of the nine possible conclusions (eight valid conclusions plus "nothing follows"; Section B.2); and second, we explored a simplified binary discrimination approach, where, following the prompt and a particular conclusion, we computed the mutual information between the prompt and each of the strings "valid" and "invalid" (Section B.3). Of these three methods, chain-of-thought prompting achieved the highest accuracy generally and had qualitatively similar performance across a range of hyperparameters, so we use it in the main text. That being said, the binary discrimination approach has the highest correlation with humans and is the only method that consistently

| | | |
|---|---|---|
| actuaries, sculptors, writers | assistants, poets, scientists | athletes, assistants, chefs |
| chemists, drivers, dancers | chemists, workers, painters | clerks, butchers, athletes |
| dancers, bankers, riders | doctors, riders, investors | drivers, porters, chemists |
| farmers, surfers, writers | gamblers, cleaners, models | golfers, cyclists, assistants |
| hunters, analysts, swimmers | joggers, actors, carpenters | linguists, cooks, models |
| linguists, skaters, singers | managers, clerks, butchers | miners, tellers, poets |
| models, tailors, florists | nurses, scholars, buyers | planners, sailors, engineers |
| riders, agents, waiters | riders, novelists, linguists | runners, opticians, clerks |
| scientists, novelists, florists | skaters, barbers, cooks | students, cashiers, doctors |
| students, hikers, designers | surfers, painters, porters | therapists, hikers, opticians |

Table 2: The 30 content word triples we use to construct syllogisms (e.g., for the first entry in the table, the variables A, B and C in the syllogism are replaced with *actuaries*, *sculptors* and *writers*, respectively). The words in each triple were chosen to be minimally semantically associated with each other.

provides the response "nothing follows" when appropriate, and as such is a promising method to explore in future work. The remainder of this appendix provides additional details about the different elicitation methods and the variations on those methods that we explored. All of the empirical results in this appendix are based on PaLM 2.

### B.1 Generative Evaluation with a Zero-Shot Chain-of-Thought Prompt

The zero-shot chain-of-thought approach is illustrated in Figure 1. We first describe the inference task to the model: "Choose the conclusion that necessarily follows from the premises or "nothing follows" if none of the other conclusions logically follow, ". We then define the conclusion space, with the string "the possible conclusions are: " followed by the list of all possible conclusions, including "nothing follows"; the possible conclusions are provided in a randomized order. Next, we provide the two premises for the syllogism being queried in the format: "Premise 1: `PREMISE1`, Premise 2: `PREMISE2`, ". Finally, we add the string "Let's think this through, step by step", which is intended to instruct the LM to produce a reasoning trace. We then generate from the LM, and determine for each of the conclusions whether they appear in the text generated by the LM. The conclusion that was detected most often, across content triples and samples, is taken to be the answer produced by the model.

#### B.1.1 Robustness to Prompt and Decoding Hyperparameters

The analyses presented in the main text are based on a decoding process in which we sequentially generate 75 tokens from the LM, with a temperature of 0.5, and take 30 such samples for each combination of syllogism type and content triple. Due to compute limitations, we are unable to conduct a systematic exploration of different variations on these hyperparameters for all model sizes; as such, we focus on PaLM 2 XS. As in the main text, we only report accuracy for the 27 syllogisms that have valid conclusions, and exclude the syllogisms for which "nothing follows" is the correct response.

**Prompts.** In addition to the prompt we used in the main text, which we refer to as `stepxstep`, we consider three variations on this prompt (Figure 11):

1. `logically`: The same as `stepxstep`, except the zero-shot reasoning trigger "Let's think this through, step by step" is replaced by "Think logically" (like `stepxstep`, this prompt is inspired by a prompt from Kojima et al. 2022).

2. `empty`: This prompt does not include any zero-shot reasoning trigger (that is, "Let's think this through, step by step" is replaced with the empty string).

3. `alt`: We created this prompt in an attempt to mitigate the LMs' reluctace to produce "nothing follows"; here the possibility of a "nothing follows" response is highlighted closer to the end of the prompt and in a more verbose way. This prompt also encourages the model to use the exact wording included in the prompt, and replaces "Let's think this through, step by step" with the slight variation "Let's think step by step".

**stepxstep**

```
Choose the conclusion that necessarily follows from the premises
or "nothing follows" if none of the other conclusions logically
follow,
the possible conclusions are:
"all artists are chemists",
"some artists are chemists",
"no artists are chemists",
"some artists are not chemists",
"all chemists are artists",
"some chemists are artists",
"no chemists are artists",
"some are chemists not artists",
"nothing follows".
Premise 1: all artists are bakers,
Premise 2: some chemists are bakers.
Let's think this through, step by step.
```

**logically**

```
Choose the conclusion that necessarily follows from the premises
or "nothing follows" if none of the other conclusions logically
follow,
the possible conclusions are:
"all artists are chemists",
"some artists are chemists",
"no artists are chemists",
"some artists are not chemists",
"all chemists are artists",
"some chemists are artists",
"no chemists are artists",
"some are chemists not artists",
"nothing follows".
Premise 1: all artists are bakers,
Premise 2: some chemists are bakers.
Think logically.
```

**empty**

```
Choose the conclusion that necessarily follows from the premises
or "nothing follows" if none of the other conclusions logically
follow,
the possible conclusions are:
"all artists are chemists",
"some artists are chemists",
"no artists are chemists",
"some artists are not chemists",
"all chemists are artists",
"some chemists are artists",
"no chemists are artists",
"some are chemists not artists",
"nothing follows".
Premise 1: all artists are bakers,
Premise 2: some chemists are bakers.
```

**alt**

```
Output the conclusion or conclusions that are logically true given the
premises.
The possible conclusions are as follows (your output should use this
exact wording): "all artists are chemists",
"some artists are chemists",
"no artists are chemists",
"some artists are not chemists",
"all chemists are artists",
"some chemists are artists",
"no chemists are artists",
"some are chemists not artists",
"nothing follows".
Premise 1: all artists are bakers,
Premise 2: some chemists are bakers.
In some cases, none of these conclusions will be logically valid,
output the words 'nothing follows' in this case.
Let's think step by step.
```

Figure 11: Variations on the prompt we used for the generative elicitation method; the prompt used in the main text is `stepxstep`.

In the experiments varying the prompt, we hold the decoding temperature at 0.5 and the maximum number of decoded tokens at 75. We find that the prompt variants show broadly similar patterns (Figure 12), though `stepxstep` achieves moderately higher accuracy than the other prompts.

**Decoding hyperparameters** Next, we hold the `stepxstep` prompt used in the main paper constant, and independently vary decoding length and temperature. First, we use the temperatures $\{0.25, 0.5, 0.75\}$, holding the decoding length at 75. Second, we vary the number of tokens decoded between 50, 75 and 100, keeping the temperature at 0.5. Here, we observe a slight increase in accuracy as the number of decoded tokens increases, which is expected (Figure 12).

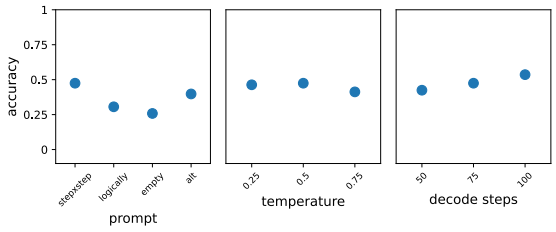



Figure 12: Accuracy for the chain-of-thought prompting method, with different prompts, temperatures and number of decoding steps.

## B.2 Multiple-Choice Discriminative Evaluation

In this approach to evaluating LM reasoning, we replace the generative evaluation with a discriminative scoring of each of the possible conclusions. The prompt is very similar: we remove the zero-shot chain-of-thought trigger from `stepxstep` and replace it with "The conclusion that necessarily follows is: ", then feed the prompt to the models and score each of the conclusions. To normalize for the idiosyncratic features of each conclusion, such as its length and prior probability, we use the *mutual information* between the prompt (p) and the conclusion (c) as the score (Holtzman et al., 2021):

$$\mathrm{MI}(\mathrm{c};\mathrm{p}) = \log P(\mathrm{c}|\mathrm{p}) - \log P(\mathrm{c}|\text{“”}) \quad (1)$$

We then renormalise these scores to compute a distribution over the conclusions (indexed by $i$):

$$P\left(\mathrm{c}_i\right) = \frac{\exp\left(\mathrm{MI}\left(\mathrm{c}_i;\mathrm{p}\right)\right)}{\sum_j \exp\left(\mathrm{MI}\left(\mathrm{c}_j;\mathrm{p}\right)\right)}, \quad (2)$$

and take the conclusion with the highest $P\left(\text{conclusion}_i\right)$ to be the LM's prediction for a given combination of syllogism and content triple. Results obtained using this method are shown in Figure 13.

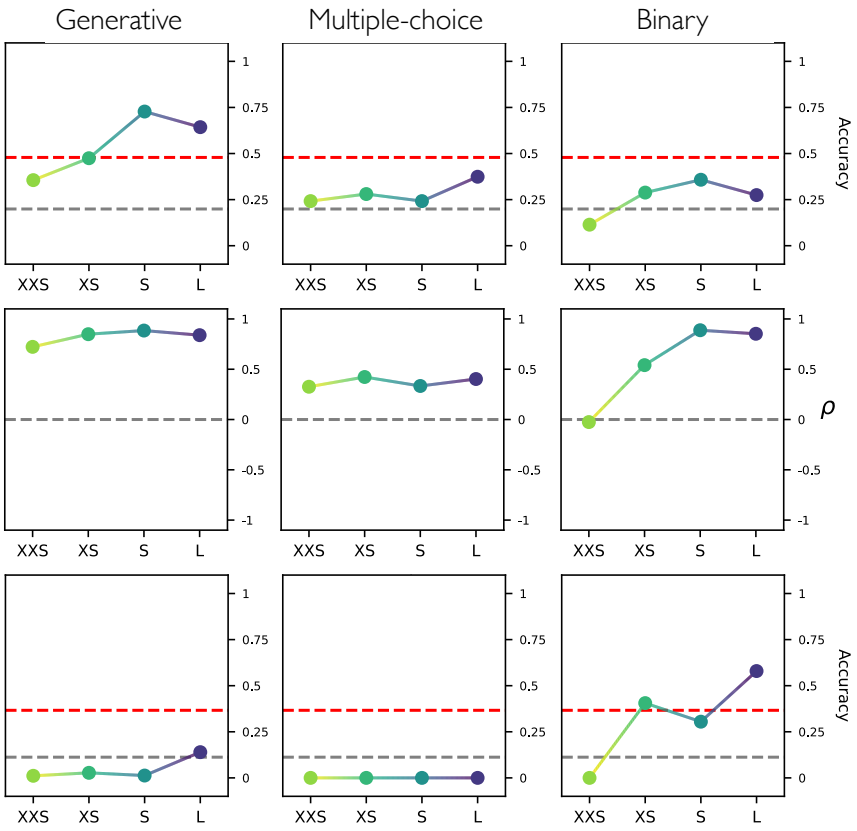



Figure 13: Comparison across reasoning elicitation methods with PaLM 2 models: the CoT generation method used in the main text (generative), as well as the binary and multiple-choice methods. We show accuracy among syllogisms with a valid conclusion (top), correlation with humans (middle), and accuracy among syllogisms where the correct response is "nothing follows" (bottom). The accuracy of the generative method is highest on the valid syllogisms, but the binary discrimination method achieves markedly higher accuracy on the "nothing follows" syllogisms. Both outperform the multiple-choice method substantially.

### B.3 Simplified Binary Evaluation

While the multiple-choice format is most similar to the paradigm used in human experiments, it poses a significantly harder task than simple binary discrimination (Dasgupta et al., 2022), which may be more sensitive. In the validity discrimination evaluation method, we present the LM with the prompt "Is this conclusion valid given the premises:" followed by the premises and a single conclusion (we refer to the concatenation of the prompt and conclusion$_i$ as prompt$_i$ below). We do this for all eight possible conclusions (omitting "nothing follows"). We, again, use the mutual information to score and compute the binary probability of "valid" as:

$$P(\text{“valid”}|\text{c}_i) = \frac{\exp\left(\text{MI}\left(\text{“valid”}; \text{p}_i\right)\right)}{\exp\left(\text{MI}\left(\text{“valid”}; \text{p}_i\right)\right) + \exp\left(\text{MI}\left(\text{“invalid”}; \text{p}_i\right)\right)}$$

We compute discrete conclusion decisions by normalizing $P(\text{“valid”})$ for each conclusion into a probability distribution:

$$P\left(\text{c}_i\right) = \frac{P\left(\text{“valid”}|\text{c}_i\right)}{\sum_j P\left(\text{“valid”}|\text{c}_j\right)}, \tag{3}$$

and taking the conclusion with the largest probability according to Equation 3 to be the LM's selected conclusion for a syllogism (the conclusion most likely to be valid according to the LM). In this approach, the LM's prediction is taken to be "nothing follows" if $P\left(\text{“valid”}|\text{conclusion}\right)$ does not exceed 50% for any of the conclusions. We note that this method is the only one that successfully elicits "nothing follows" conclusions for a substantial proportion of the syllogisms (Figure 13).

## C By-Syllogism Correlations with Human Responses

Figures 14 provides correlations between LMs and humans at the individual syllogism level. While larger LMs are generally more human-like, we observe a diversity of relationships between model scale and human-likeness, including cases such as IE2 where larger models are in fact less correlated with humans.

## D Mental Models Simulations: Additional Details

### D.1 Model details

This section provides additional details on mReasoner. Figure 15 shows an example of the "canonical sets" that mReasoner uses to heuristically sample entities, and Figure 16 illustrates the subroutines used to revise mental models.

| | | | | | | |
|---|---|---|---|---|---|---|
| **LEN** | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 |
| **BROAD** | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 0.9 |
| **SYSTM2** | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 0.9 |
| **WEAKEN** | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 0.9 |

Table 3: Parameter grid used to instantiate our mReasoner models.

### D.2 mReasoner instantiations

We instantiate one mReasoner model for every parameter vector in the grid shown in Table 3. This resulted in a total of 1,296 models. As the model's reasoning process is stochastic, we evaluate each model 100 times for each syllogism to estimate the distribution over responses. Due to resource constraints, we discarded models that did not finish these 100 iterations in 60 seconds, leaving us with 923 models spaced relatively evenly over the grid (i.e., this timeout criterion did not systemtically favor some hyperparameter values).

Each of the 923 models is represented by a 216-dimension vector, with eight possible conclusions for each of the 27 valid syllogisms ($27 \times 8 = 216$). We perform a standard PCA—the probabilistic PCA of Tipping and Bishop (1999) on the centered dataset, using scikit-learn (Pedregosa et al., 2011)—on these 923 vectors.

## E Llama 2: Additional Results and Plots

As we mentioned in Section 4.3, while the overall accuracy of Llama 2 models is similar to that of humans, this aggregate pattern masks large discrepancies with humans, and in particular poor accuracy on some syllogisms where humans rarely make mistakes, as well as high accuracy on syllogisms that humans struggle with. Consequently, these models exhibit a substantially lower correlation with human behavior across the board (Figure 21) than do PaLM 2 models (cf. Figure 14). Llama 2 models also demonstrate a slight decrease in correlation with human behavior as a function of model size in our analysis of the correlation between accuracy and entropy in syllogistic fallacies (Figure 19; cf. Figure 5 for PaLM 2 models).

We also repeated the mReasoner analysis (described in Section 5) to analyze Llama 2 models' behavior. Although the Llama 2 models, like

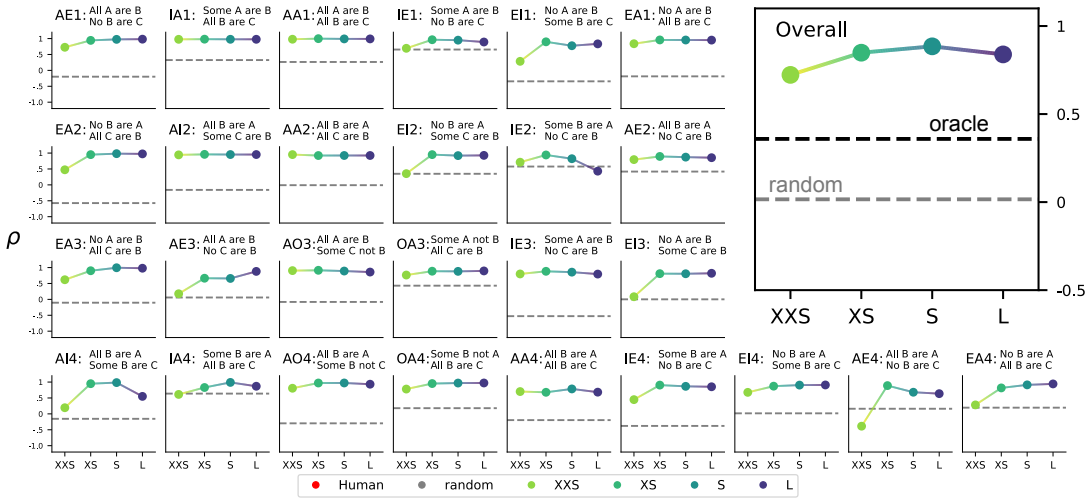



Figure 14: Correlation between PaLM 2 models' distribution over responses and the probabilities derived from normalizing human responses, broken down by syllogism. Syllogisms are partitioned into variable ordering type (by row) and ordered by decreasing human accuracy from left to right. Chance performance (dashed grey line) reflects random guessing. The top right inset shows correlation across the entire dataset.


Figure 15: The "canonical sets" used by mReasoner. The canonical set for a syllogism depends on the moods of the syllogism's premises. We show the possible individuals each premise contributes to a syllogism's canonical set here for the hypothetical content words *artists* and *bakers*.

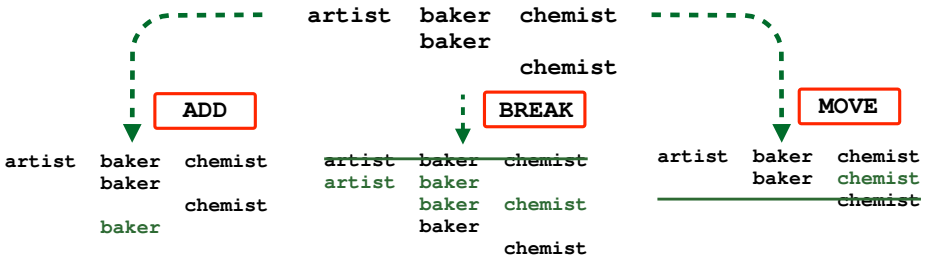
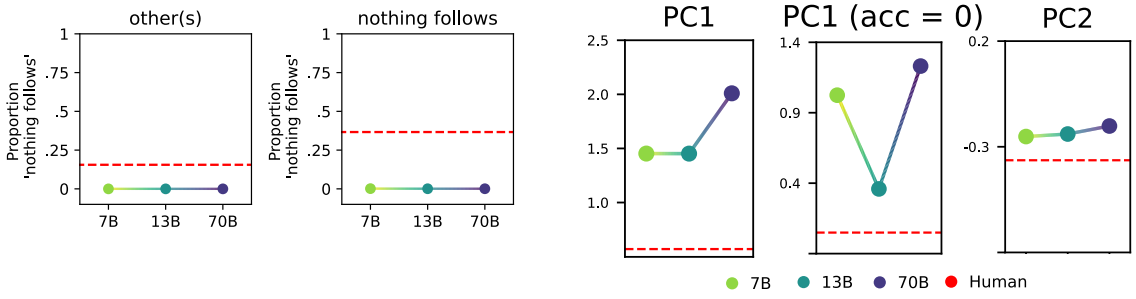



Figure 16: Subroutines used by mReasoner to revise mental models in order to check for counterexamples. We denote these subroutines as **ADD**, **BREAK**, and **MOVE**, following (Khemlani and Johnson-Laird, 2022). **ADD** adds one more entity to a mental model. **BREAK** decomposes an entity's properties into constituent entities with subsets of those properties. **MOVE** simply moves a property from one entity to another.

PaLM 2, exhibit increased signatures of deliberative reasoning compared to mReasoner when all of their predictions are considered (Figure 17b, left), when we control for accuracy by setting the probability of the correct answer to zero for all Llama 2 models, we find no significant correlation between model size and signatures of deliberative reasoning (Figure 17b, center; cf. Figure 9 for PaLM 2, where we do find such signatures even after controlling for accuracy).




(a) (b)

Figure 17: (a) Llama 2's behavior on the 37 syllogisms whose only valid conclusion is "nothing follows" (left) and the syllogisms that license conclusions other than "nothing follows" (right). (b) Llama 2's behavior when analyzed using the principal components of the mReasoner space (see Section 5). We find an increased signature of deliberative reasoning as a function of model size, but we no longer observe this effect when we control for accuracy (setting the probability of the correct answers to zero before projecting the models' behavior into this space).

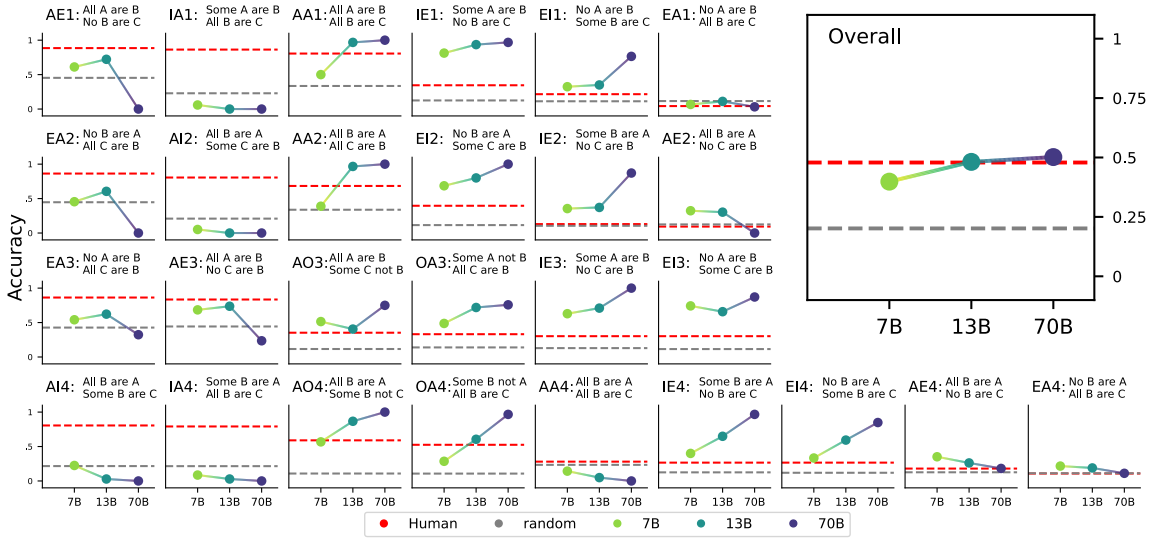



Figure 18: Accuracy of Llama 2 models, humans (red), and random guessing (grey). Random guessing accuracy differs by syllogism as some syllogisms have more than one valid conclusion. Syllogisms are partitioned into variable ordering (by row) and ordered by decreasing human accuracy from left to right. The top right inset shows the average accuracy across all syllogisms. Syllogisms are identified with the letters of the moods of the premises (Table 1, left) and the number associated with their variable ordering (Table 1, right).

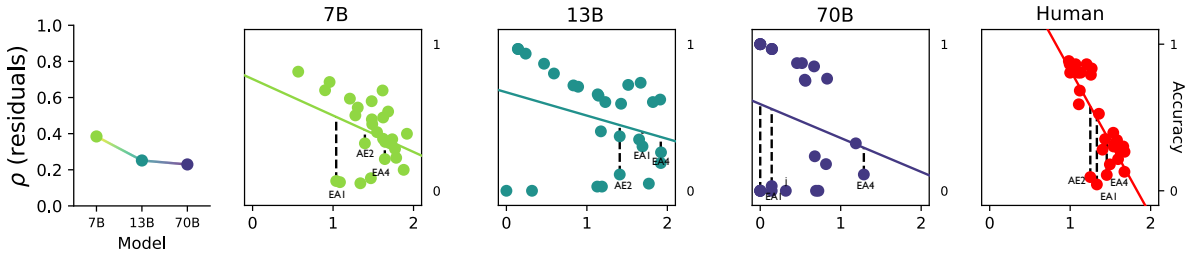



Figure 19: Analysis of Llama 2 models' handling of syllogistic fallacies. **Right:** Each syllogism plotted by accuracy (y-axis) and entropy (x-axis) and the regression line relating the two. Dashed lines black lines show the residuals for each of the top three human syllogistic fallacies. **Left:** The result of correlating Llama 2's residuals with residuals estimated from human data.

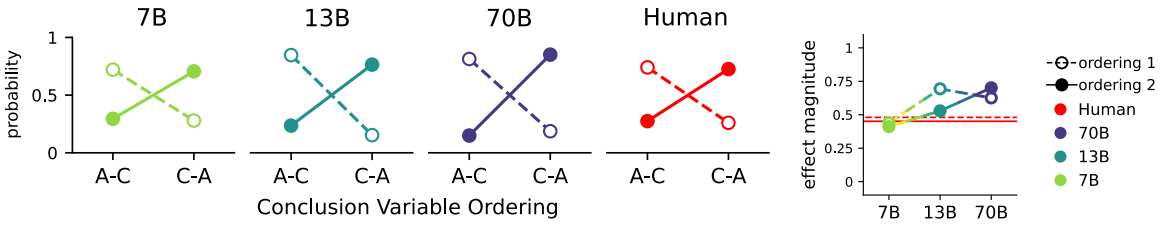



Figure 20: Variable ordering effects on Llama 2 responses. **Left:** The marginal probabilities of A-C and C-A ordered conclusions as estimated from human and LM responses. Humans and LMs both show variable ordering effects in the same direction. **Right:** The magnitude of the variable ordering effect (the absolute value of the difference between the probability of the C-A ordering and the probability of the A-C ordering).

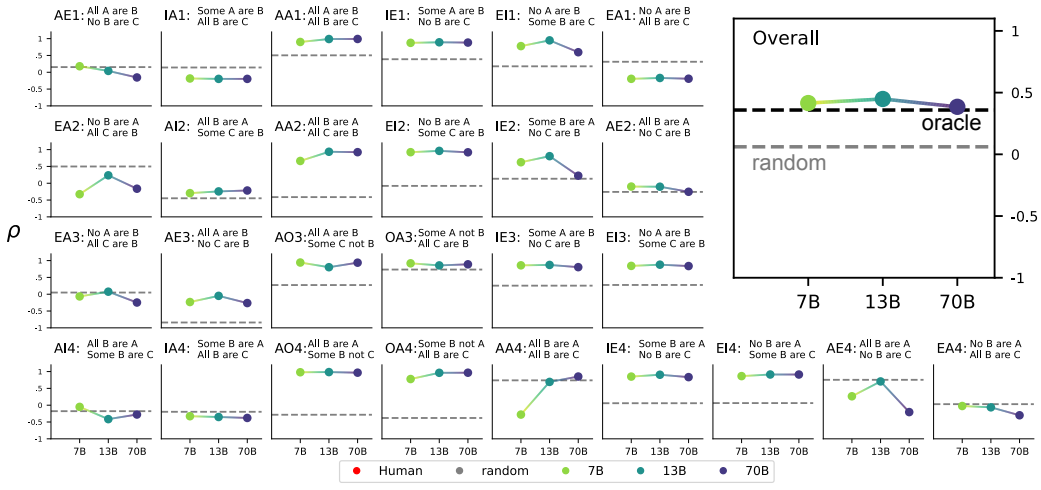



Figure 21: Correlation between the Llama 2 model's distribution over responses and the probabilities derived from normalizing human responses, broken down by syllogism. Syllogisms are partitioned into variable ordering type (by row) and ordered by decreasing human accuracy from left to right. Chance performance (dashed grey line) reflects random guessing. The top right inset shows correlation across the entire dataset.