Do Large Language Models Exhibit Cognitive Dissonance? Studying the Difference Between Revealed Beliefs and Stated Answers

Anonymous ACL submission

Abstract

Prompting and Multiple Choices Questions (MCQ) have become the preferred approach to assess the capabilities of Large Language Models (LLMs), due to their ease of manipulation and evaluation. Such experimental appraisals have pointed toward the LLMs' apparent ability to perform causal reasoning or to grasp uncertainty. In this paper, we investigate whether these abilities are measurable outside of tailored prompting and MCQ by reformulating these issues as direct text completion – the foundation of LLMs. To achieve this goal, we define scenarios with multiple possible outcomes and we compare the prediction made by the LLM through prompting (their `Stated Answer`) to the probability distributions they compute over these outcomes during next token prediction (their `Revealed Belief`). Our findings suggest that the `Revealed Belief` of LLMs significantly differs from their `Stated Answer` and hint at multiple biases and misrepresentations that their beliefs may yield in many scenarios and outcomes. As text completion is at the core of LLMs, these results suggest that common evaluation methods may only provide a partial picture and that more research is needed to assess the extent and nature of their capabilities.

1 Introduction

In recent years, Large Language Models (LLMs) have gained significant traction in the research community and the public at large (Zhao et al., 2023; Chang et al., 2024). At their core, LLMs are statistical models of languages that predict, for any given context, a probability distribution over their vocabulary for the occurrence of the next token in a sequence (Bender et al., 2021). Despite this simplicity, a wide array of earlier research has noted that their sophisticated use of natural language (NL) is impressive (Chang et al., 2024; Hu and Levy, 2023), and it has been claimed that they may provide a candidate model for the acquisition of language in humans (Warstadt and Bowman, 2022).

Other studies have gone further and claimed that LLMs have become more than just statistical models, and gained emergent abilities due to their massive training sets and architecture sizes (Bubeck et al., 2023; Wei et al., 2022). Notably, it has been argued that recent LLMs have acquired the capability to perform more complex tasks such as reasoning and probability manipulation (Kıcıman et al., 2023). However, this fact is debated in the research community. While recent LLMs perform well on advanced benchmarks (bench authors, 2023), they can also be tricked by simple questions and adversarial modifications of their prompt (Xu et al., 2023; Zou et al., 2023). This has raised the question of whether LLMs indeed have an aptitude for reasoning, or whether these observations are an illusion that emerges from their mastery of NL and propensity for data memorization.

At the heart of these debates is the problem of evaluating LLMs. The most common way to evaluate LLMs is through prompting (see e.g., Brown et al. 2020) and most benchmarks are collections of questions and answers (Hendrycks et al., 2021; bench authors, 2023; Liang et al., 2023), where the LLMs are prompted with a question and the resulting answer is compared to a known ground truth. By nature, this method of evaluation is vulnerable to data contamination, where the LLMs have observed part of the evaluation set during training – a problem exacerbated by the fact that the training datasets of most LLMs are generally not accessible to the research community (Deng et al., 2023). Furthermore, since the evaluation of open-ended answers is quite complicated and resource-consuming (Frieder et al., 2024), the questions in these benchmarks are most often Multiple Choice Questions (MCQs) – see e.g., Liang et al. (2023) – where LLMs are asked to choose between a finite set of answers. The evaluation via MCQs has been shown to suffer from a variety of biases (Pezeshkpour and Hruschka, 2023; Wang et al., 2024) that compound the contamination issue, resulting in suboptimal assessments.

Recently, Hu and Levy (2023) have compared the merits of evaluating LLMs with direct text completion and with the use of prompts. While their experiments focus on the linguistic capabilities of the LLMs and their knowledge of the English language, their findings highlight that prompting is not a substitute for direct text completion and that it is key to LLM evaluation as it can shed some light on their capabilities. Their work suggests a distinction between a model's performance (its behavior when prompted with questions) and a model's competence (the information contained in an LLM's string probability distribution). In line with these results, we argue that evaluating LLMs by exclusively assessing their answer to carefully crafted questions only provides an incomplete assessment of LLMs' abilities. As the answer of an LLM to an MCQ has been shown to be strongly influenced by the ordering of the answer (Pezeshkpour and Hruschka, 2023), it hints at the fact that the formatting of the questions may be more important to the LLM's performance than the question itself, thus throwing doubt on the interpretation of the results.

In this paper, we introduce a different paradigm to evaluate LLMs to address this issue by proposing an alternative approach that forgoes entirely the questions/answer framework. This paradigm, named `Revealed Belief`, relies solely on next-token prediction, LLMs' elementary unit of computation. Under this approach, we evaluate LLMs through *direct text generation* by considering scenarios centered around their handling of uncertainty in language and compare the LLMs' predictions to the known true distribution. To do so, we present an LLM with a context describing a scenario that has multiple possible outcomes (e.g., the throw of a fair die) and use the LLM's text completion to simulate its resolution (e.g., on which face the die lands). Then, the observed distribution over the possible outcomes is compared to the true probability distribution (e.g., a uniform distribution). In line with the "show, don't tell" adage, this approach emphasizes the text generated by LLMs (their `Revealed Belief`), instead of their answer to a predetermined question (their `Stated Answer`) – see Section 3. Probing LLMs in this manner reveals their internal representations and allows us to compare how such a scenario would be generated with the way they would answer a question about this same scenario.

We apply this new paradigm in a wide range of scenarios (see Section 4) and make several key observations. First, while the LLMs can perform well in the MCQ setting,we observe that their `Revealed Belief` tend to underperform. We argue that this indicates that LLMs perform better at retrieving knowledge about the world (e.g., the probable outcome of a die roll) than at integrating this knowledge into their generated text (e.g., by revealing an unbiased distribution over a set of equiprobable outcomes). Second, their `Revealed Belief` highlights numerous biases toward specific outcomes and the disproportionate impact of innocuous language elements on the distribution of outcomes, and the undue effect of unrelated events in the context (see Section 5). As these observations are not compatible with advanced reasoning capabilities, these results hint at the limitation of current evaluation methods and suggest that more research is needed in the study of LLMs' capabilities.

2 Related work

LLMs and probability. Our proposed empirical evaluation framework revolves indirectly around probabilities and uncertainty, and many previous studies have examined the aptitude of LLMs to handle problems involving probabilities. Recent contributions include Nafar et al. (2024), which studied the reasoning capabilities of LLMs around text that contains explicit probability values, and Saeed et al. (2021), which analyzed the capacity of LLMs to deduce soft logic rules.While the evaluations performed in these papers indicate good performance by LLMs, more recent works such as Jin et al. (2023) and Frieder et al. (2024) have hinted at more disappointing results, in particular for more advanced causal and mathematical problems. Compared to this work, the probability problems involved in our experiments are significantly easier, such as the flip of a coin, or the throw of a die. Moreover, all the aforementioned studies evaluate the `Stated Answer` of the LLMs, whereas we investigate their `Revealed Belief` (see Section 3), which led to significantly different observations.

Investigating LLM evaluation. Previous work has also questioned the evaluation of LLMs and scrutinized their flaws. In particular, MCQs – one of the most prevalent types of evaluation (Hendrycks et al., 2021; bench authors, 2023; Liang et al., 2023) – have faced many criticisms. For instance, Pezeshkpour and Hruschka (2023) has shown that merely reordering the options of MCQ can lead to double-digit performance gaps across multiple benchmarks, while Alzahrani et al. (2024) additionally showed that a similar phenomenon can be observed by changing the numbering scheme of the provided answers. In addition, Wang et al. (2024) pointed out a significant misalignment between the first token predicted by the model in MCQ, which is often used as a proxy to infer a model's answer, and the model's actual answer. Other works have pointed to the problem of data contamination, where the LLMs are shown to have been exposed during training or fine-tuning to the evaluation data used in common benchmarks. Notably, Deng et al. (2023) and Balloccu et al. (2024) have shown, using two complementary approaches, that this problem is present in all LLMs but is particularly pronounced in the GPT-series of models.

Probing internal representations and beliefs Past studies have aimed at probing the hidden activations and embeddings to gain insights into e.g., the encoded information (Alain and Bengio, 2017), beliefs (Azaria and Mitchell, 2023), and lies (CH-Wang et al., 2024) in language models. Regarding the study of the next token distribution, Feng et al. (2024) have shown that probing the token representations of a model is useful to abstain from stating low-confidence answers. Similarly, Slobodkin et al. (2023) found that probing an LLM's last hidden layer reveals the answerability of a query. Importantly, while we also study the next token distribution in this work, we focus mainly on rephrasing questions as direct text completion problems to provide a new evaluation tool.

LLM `Revealed Belief`. Arguably, the paper closest to our work is by Hu and Levy (2023), where the authors highlight the discrepancy between the direct completion of some text and prompting LLMs to complete the text, with particular emphasis on their mastery of the English language. While `Revealed Belief` and `Stated Answer` follow a similar dichotomy, the scope of our study – scenarios with uncertainty and multiple outcomes – differs significantly, and our method leverages the next token's distributions to investigate the details of the LLM's implicit biases and errors. Furthermore, while Hu and Levy (2023) also hypothesized that new LLM capabilities may emerge by studying direct text completion instead

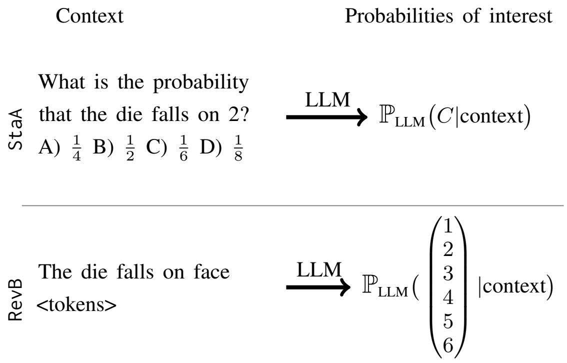


Figure 1: Illustration of the `Revealed Belief` (`RevB`) and `Stated Answer` (`StaA`) frameworks, for the scenario that involves a regular fair die with six faces. Only a short version of the context is presented here. Examples of detailed prompts and contexts can be found in Section 4 and in Appendix A.

of prompting, our findings point toward a more nuanced conclusion (see Section 6).

3 `Revealed Belief` and `Stated Answer`

In the rest of this paper, we study LLMs through *the direct measurement of the model-derived next token distributions*, similarly to Hu and Levy (2023). For any given context, we collect the logits produced by the LLM before any temperature-based random sampling, and we compute the exact probability distribution of the next token (or specific combination of tokens), noted $\mathbb{P}_{\text{LLM}}(\,\cdot\,|\text{context})$. This distribution reflects the exact LLM beliefs, in contrast to simply sampling LLM answers.

At the heart of our analysis is the use of scenarios with multiple possible outcomes. Consider for instance the roll of a single, fair, six-faced die. There are six equally probable results for this throw, 1 to 6, with probability $1/6$. The commonly used method to assess whether an LLM is knowledgeable of this fact is to first query the model directly, for instance, with the MCQ question "What is the probability that the die falls on face 2?" and a set of possible answers $1/4$, $1/6$, etc. Then, the model's probability of selecting the correct answer is computed from the next token distribution and is used as the metric to evaluate the LLM. We refer to this approach as the `Stated Answer` (`StaA`).

In this work, we propose an alternative method to evaluate LLMs for these multiple outcome scenarios, called `Revealed Belief` (`RevB`). Instead of explicitly querying the LLM, we assess its implicit beliefs about the given scenario by presenting it with the scenario and ending with the incomplete sentence "the die falls on face ...". The LLMs' probability of choosing any of the possible outcomes (in this case, 1 to 6) as its next token is then computed and the resulting distribution is compared to the true distribution (in this case, a uniform distribution). In other words, instead of asking the model what it knows, we let the model actually show its beliefs, through the distribution it produces over the set of possible outcomes. Figure 1 illustrates both frameworks for the dice scenario.

Advantages of `Revealed Belief`. `RevB` can be computed for any scenario with multiple outcomes, and we argue that this alternative evaluation method presents multiple advantages. First, it is centered around text completion, which is the elementary computational unit of LLMs. As such, it can evaluate any LLM, including base models that have not been fine-tuned on dialogue behavior, and can be used to assess the influence of different fine-tuning methods. Furthermore, as it does not involve MCQs, `RevB` avoids their common pitfalls (choice of the answers, ordering, etc.), as highlighted in Section 2. Second, this evaluation approach incorporates significantly more information than `StaA`, as it allows an in-depth comparison of the difference between the true distribution and the revealed distribution of the outcomes, which can highlight biases towards or against specific outcomes. For instance, our experiments (Section 5) show that despite the LLMs `Stated Answer` that the probability of a die roll resulting in any face equals $1/6$, the `RevB` of these LLMs show an inordinate bias toward the result 1. Finally, it is important to point out that the biases and issues shown in `RevB` can have an important impact on an LLM's behavior. For instance, one of our experiments highlights the LLMs' preferences towards the first answer in a fair choice between abstract, equiprobable options, which illustrates the bias that LLMs can exhibit when answering MCQs. Importantly, these biases are compounded when the outcomes have multiple meanings, such as the terms "left" and "right".

4 Methodology

4.1 Tasks

We examine LLMs' `Revealed Belief` through four different scenarios described below: Dice, Coins, Choice, and Preferences. Table 1 offers a summary of the different experimental setups.[1]

[1]Complete examples of prompt variants for `RevB` and `StaA` are provided in Appendix A.

4.1.1 Scenario 1: Dice

Probability problems derived from die rolls are among the most prevalent in an introductory mathematics curriculum. Thus, it is expected that instances of this scenario are well represented in any large LLM training dataset. As die rolls can yield many probability distributions, we use them as the first scenario to explore the `RevB` of LLMs.

Prompt Example. "A die has 6 faces. The die is equally likely to land on any of its faces. The die is cast. The die lands on face <tokens>".

Variants & Parameters. In the simple variant, several dice (1-3) with four to twelve faces each are rolled once, and the outcome is the total sum of the dice faces. This results in a uniform or a multinomial distribution, depending on the number of dice. In the repeated variants, the dice are rolled twice, and the first result is mentioned in the context. The outcome to be predicted is then either the result of the second roll (case independent) or the sum of both rolls (case dependent). These variants aim at examining the influence of a previous roll on the `RevB`. In the independent case, the expected result should be identical to the simple variant, while in the dependent variant, the distribution should be shifted by the value of the previous roll. Finally, in the observation scenario, some information regarding the die roll result is disclosed to the model – for instance, that the result is an even number. These observations allow manipulating the expected distributions and studying the influence of new information on a model's `RevB`.

4.1.2 Scenario 2: Coins

Coin flips are also quite common in probability problems and offer a different scenario to study distributions of varying complexity.

Prompt Example. "There are 3 coins. Each coin is biased and is 5 times more likely to land on Heads than on Tails. The coins are flipped and the resulting number of Heads is equal to <tokens>".

Variants & Parameters. Compared to the dice scenario, coins have only two faces (Heads and Tails). We therefore vary the number of coins, as well as two additional parameters: the face of interest (Heads or Tails), that is to say, the one that is counted in the flip, and the bias, which modifies the probability of the face of interest, and thereby the resulting distribution. The Coins scenario includes

Scenario	Variants	Parameters
Dice	Single Roll Repeated (Independent, Dependent) Observations	Number of Dice, Number of Faces Number of Dice, Number of Faces, Result previous roll Number of Dice, Number of Faces, Observation(s)
Coins	Single Flip Repeated (Independent, Dependent)	Number of Coins, Heads or Tails, Bias Number of Coins, Heads or Tails, Bias, Result previous flip
Choice	Single Choice Repeated (Independent)	Number of Options Number of Options, Previous choice
Preference	Single Selection Repeated (Independent)	Option 1, Option 2, Bias Option 1, Option 2, Bias, Previous selection

Table 1: Summary of the scenarios, variants, and parameters.

both the simple (a single flip) and the repeated variants (both independent and dependent).

4.1.3 Scenario 3: Choice

In this scenario, the models have to select between an arbitrary number of abstract options, represented using capital letters – similar to the choice of an answer in an MCQ. As the choices are explicitly stated to be of equal probability, the distribution underlying this scenario is always uniform and identical to the roll of a single die. Here, the interest is to scrutinize the influence of the setting (e.g., dice versus abstract) on simple distributions and distill the raw preferences over abstract choices by discarding the connotations related to the scenarios.

Prompt Example. "A person has to choose randomly between 4 options. The options are A, B, C, and D. All possible options are equally likely. The person chooses at random option <tokens>".

Variants & Parameters. We consider two variants of this scenario: the simple variant, where a single choice is made and the parameter is the number of options; and the repeated independent variant, where the LLM makes a second choice after being presented with the result of a first one.

4.1.4 Scenario 4: Preference

Compared to Choice, the Preference scenario contains only two options, these two options are no longer abstract (for instance, left or right), and their probabilities are no longer necessarily equal. The goal of this scenario is to examine the influence of each option label on the outcome distribution.

Prompt Example. "A person has to choose randomly between two options: Left and Right. The option Left is 2 times more likely to be chosen than the option Right. The person chooses at random option <tokens>".

Variants & Parameters. The Preference scenario contains the same variants as the Choice Scenario: the simple variant and the repeated independent variant. The varied parameters are the labels of the options (e.g., Left/Right, Heads/Tails), the explicit bias, the result of the previous selection, and the order of the options in the query.

4.2 Models

We tested the `RevB` framework using 12 decoder-only language models chosen to reflect the state of the art as of May 2024 within three model sizes, according to benchmarking results reported in the Hugging Face Open LLM Leaderboard[2] and the LMSYS Chatbot Arena Leaderboard[3]. We only examined open-weight LLMs (in order to analyze their next token distribution) and only selected models where both the base and instruction fine-tuned variants were available. The bracket of small models include Llama-3-8B[4], Yi-1.5-{6B, 9B} (Young et al., 2024), gemma-7B (Mesnard et al., 2024), Qwen2-7B (Bai et al., 2023), and Mistral-7B-v0.3 (Jiang et al., 2023). The family of large models contains Llama-3-70B, Yi-1.5-34B, and Qwen2-72B. The considered mixture of expert models are Qwen2-57B-A14B and Mixtral-8x22B (Jiang et al., 2024). We evaluated the base and instruct versions' `RevB` for each model to compare their respective performance, using the models available on Hugging Face and we query their `Stated Answer` using the instruct version.

4.3 Evaluation method

In our experiments, we take advantage of the fact that `RevB` can be evaluated for any LLM, regardless of the fine-tuning (see Section 3), and we assess

[2] `huggingface.co/open-llm-leaderboard`
[3] `huggingface.co/spaces/lmsys/chatbot-arena-leaderboard`
[4] `github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`

both base models and their instruction fine-tuned variants. However, when assessing StaA, we only evaluate instruction fine-tuned models, as they are the only ones trained to generate answers to questions. Each LLM is evaluated on all combinations of scenarios, variants, and parameters (see Table 1). For the RevB framework, models are presented for each resulting context, and the RevB (i.e., the distribution over the possible outcomes) of the base model and the instruct models are then computed and normalized. Then, these distributions are compared to the true probability distribution, using three different metrics: the Chebyshev distance, the Manhattan distance, and the Kullback-Leibler divergence. While the last two measure the total difference between the different distributions, the Chebyshev distance represents the maximal difference of the weights between the distributions and is thus particularly representative of the bias that RevB can have towards or against a specific outcome. For the StaA framework, the multiple-choice questions are formatted specifically to each model's chat template (e.g., by adding the special "user/assistant" tokens) before being presented to each model. The probability of the (only) correct response is then retrieved, and we report its Chebyshev distance to the ideal answer. As only one of the responses in the MCQ is correct, this distance corresponds to one minus the probability of the model giving the correct answer.

5 Experimental Results

In total, we tested each model in more than 500 different settings. For brevity, we describe in this section the main findings of our experiments and emphasize them using the Chebyshev error metric. We refer the reader to the Appendix for further results of each scenario, as well as additional metrics.

Result #1: Instruction fine-tuning does not improve RevB. Across the different scenarios studied in our experiments, we observe that the RevB of base LLMs are always at least as good and often better than the RevB of their instruction fine-tuned counterparts. This is illustrated by Figure 2, which summarizes the scores obtained by both models on the different tasks. Consequently, while the fine-tuning of LLMs into instruction models is important for prompting and StaA, we observe that it does not improve their RevB and that both variants show similar biases and skews.

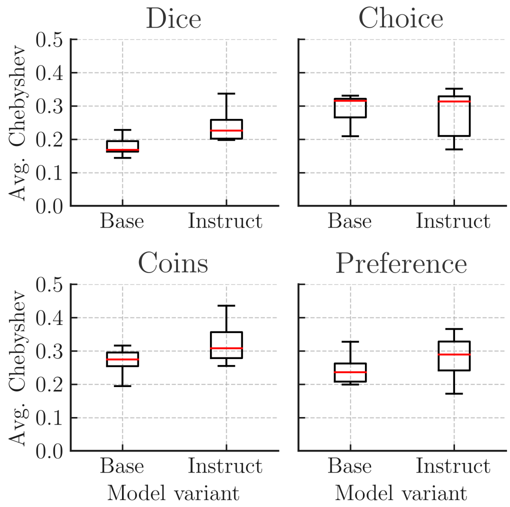
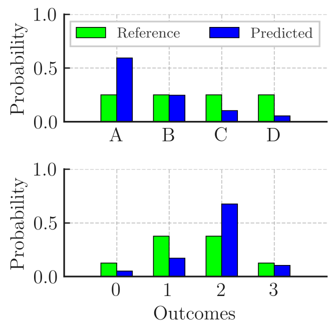


Figure 2: Distribution of the Chebyshev distance of base and instruction fine-tuned models RevB on different tasks (lower is better). Boxes (resp. whiskers) represent the range of the 25 to 75 (resp. 10 to 90) percentiles, and the red line represents the median.

Result #2: LLMs RevB favour specific outcomes. Figure 3 (top) illustrates the behavior of the RevB when addressing a scenario yielding a uniform probability distribution (e.g., the roll of a single die or a choice between abstract options). Importantly, while the StaA are almost always correct, with many models exhibiting a 0% error (in particular for the family of large models), the RevB are significantly different from the expected uniform distributions. We observe that in most settings, the first possible option (side 1 of a die, or the abstract option "A") is favored by the LLMs. Interestingly, this bias is significantly stronger when predicting the outcome of an abstract choice, resulting in worse scores. This problem is also apparent for multinomial distributions (e.g., multiple dice rolls or coin flips) – see Figure 3 (bottom). Moreover, many LLMs RevB exhibit a bias towards individual outcomes, such as a value near the midpoint of the distribution, or multiples of 10. Indeed, while most LLMs perform reasonably well when the number of outcomes is low, their RevB show very skewed distributions when this number exceeds a certain threshold (often around 12). While their StaA errors are higher than in the uniform case, large models still exhibit errors close to 0, despite their skewed RevB.

Result #3: Previous independent results described in a prompt have an undue impact. The repeated variant of the scenarios aimed at scru-



	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B*	Q72B	M8x22B
4 faces (RevB)	0.05	0.16	0.09	0.14	0.09	0.06	0.14	**0.04**	0.09	0.09	0.15
4 faces (StaA)	0.62	**0.00**	1.00	0.96	0.94	0.99	**0.00**	0.18	**0.00**	**0.00**	0.01
4 choices (RevB)*	0.45	0.48	0.61	0.44	0.54	0.57	**0.32**	0.55	0.35	0.65	0.34
4 choices (StaA)	**0.00**	1.00	0.04	0.35	0.05	**0.00**	**0.00**	0.01	**0.00**	**0.00**	**0.00**
3 coins (RevB)*	0.11	**0.02**	0.14	0.15	0.09	0.19	0.03	0.06	0.30	0.20	0.14
3 coins (StaA)	**0.00**	1.00	**0.00**	0.79	1.00	**0.00**	1.00	0.06	**0.00**	**0.00**	0.07

Figure 3: **Left:** Probability distributions over four abstract choices (top) and the number of heads in three coin flips (bottom). Model: Llama-3-70B. **Right:** Chebyshev distance between predicted and reference distribution for RevB; Stated: Probability of error of the StaA. Best score per scenario in bold.

tinizing the influence of a previous realization of the event on a future realization. In the independent variant, the previous outcome is explicitly stated as having no bearing on the new outcome. However, we observe that this variant worsens the scores of RevB and StaA, as shown for a repeated die roll with selected models in Figure 5 (left). For instance, when asked for the probability of the outcome of a 10-sided die roll that was preceded by another die roll (which landed on faces 1, 5, and 10 respectively), we observe by probing the RevB that the probability assigned to each individual outcome is strongly affected by the prior die roll, as illustrated by Figure 4. The behavior is widespread in our experiments and can be observed in most Repeated settings. This shows that repeating an event within the same context window impacts an LLM's next prediction and decreases its reliability in downstream predictions.

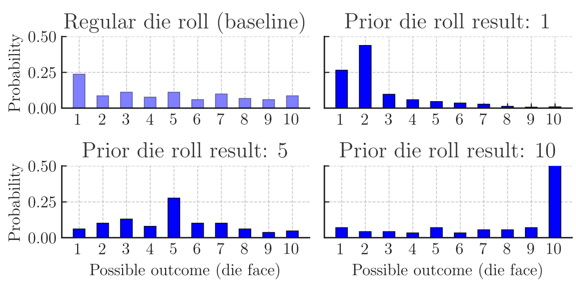


Figure 4: Revealed Belief of Llama-3-70B for the throw of one 10-sided die, with and without previous independent observations. The RevB exhibit a strong bias towards the prior independent event.

Result #4: RevB are better than StaA at handling partial information. Interestingly, in the variant with observations, where partial information about the result of a die roll is included in the context

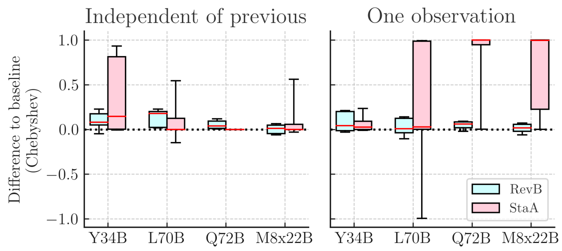


Figure 5: Distribution of the increase in Chebichev errors for dice rolls with a prior independent result (left) and one observation (right) compared to their baseline. Boxes (resp. whiskers) represent the range of the 25 to 75 (resp. 10 to 90) percentiles, and the red line represents the median.

(e.g., "the result is even"), LLMs' RevB are more accurate than their StaA. For instance, excluded outcomes (i.e., odd numbers in the aforementioned example) are assigned a probability close to zero, showing that the observation stated in the prompt is well integrated into the prediction. Conversely, the StaA of LLMs are generally significantly worse, as shown in Figure 5 (right). This is particularly visible when the prompt contains combinations of observations that exclude multiple outcomes (see Table B.3.3 in the appendix).

Result #5: The labels of the outcomes can strongly bias the RevB. The results of the Preference scenario show that even when the options are explicitly stated to be equiprobable, LLMs' RevB show their inherent pairwise preferences, as illustrated by Figure 6. In this case, the outcome labeled "Left" is strongly preferred over "Right" (left figure) and this bias is not equally reciprocated even when "Right" is presented first (right figure), indicating that this bias is not due to ordering (Result #2).

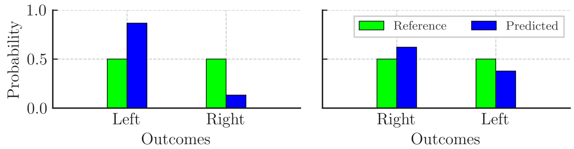


Figure 6: Probability distributions when varying the order of "Left" and "Right" for Llama-3-70B's `RevB`.

6 Discussion

At their core, Large Language Models operate by estimating the next token probability distribution for any given context. A token is then sampled from this distribution and the operation is repeated until completion of a task. Many approaches to evaluating LLMs tend to focus on higher-level tasks, such as natural language inference or probabilistic reasoning, by assessing their generated text. Thereby, they may conflate the LLM's aptitude to answer questions related to an aptitude (`StaA` in our framework) with the aptitude itself (`RevB`). Hu and Levy (2023) have established this contrast with regard to (meta-)linguistic abilities, and have shown the difference between the behavior of a model under prompting (analogous to `StaA`) and the model's estimated string probabilities (analogous to `RevB`).

In this study, we built on this approach and extended it to other situations, namely probabilistic reasoning tasks under uncertainty. The proposed `RevB` paradigm complements common evaluations based on multiple-choice questions by A) forgoing the question/answer framework, thus avoiding all the biases related to it, and B) probing of the model's hidden internal next-token probability distributions.This probing aims at revealing how well the stated information is integrated into the model's text generation in any context and is not simply an artifact of rote question answering.

Our experiments with multiple evaluation scenarios have pointed to significant discrepancies between the results of these two approaches. For instance, models evaluated using `RevB` showed better performance in integrating partial observations contained in a query (e.g., "[...] the die lands on an even face [...]"), compared to including this same information in an MCQ (Result #4). This result is aligned with one of the findings by Hu and Levy (2023), where some information, although contained within a model's string probabilities, cannot be easily accessed when implicitly prompted for it and is not apparent in the model's behavior. It thus reinforces the notion that prompting and direct probability distribution measurements are complementary evaluation approaches.

Moreover, our findings show that LLMs are unable to properly handle the probability of outcomes preceded by independent events (Results #3). Indeed, the occurrence of a prior result (e.g., die roll, coin flip) introduces a significant effect on the bias and skew of both the `RevB` and `StaA` about the outcome of the second event. This phenomenon raises questions regarding the LLMs' ability to handle irrelevant information contained in a prompt, with ramifications for their ability to solve other tasks (e.g., drawing logical inferences without being affected by irrelevant premises, learning from examples in-context, augmenting the context with retrieved documents). Our study therefore raises significant concerns with the way LLMs are evaluated and used for these types of tasks.

Finally, the examination of LLMs under the `RevB` paradigm revealed further phenomena, which are hidden when only considering their `StaA`. For instance, models generally prefer the first option in a set of outcomes, often by a significant margin (Result #2). Furthermore, the labels used to name options (e.g., left/right vs. heads/tails) and their ordering strongly affect each resulting probability (Results #5). This implies that LLMs' generated text are skewed towards certain outcomes and are biased by the chosen labels, even for explicitly stated equiprobable choice. Albeit hidden when generating answers, both of these biases and skews ought to be taken into account when evaluating or using LLMs through prompting.

In summary, our findings suggest that the revealed beliefs of LLMs significantly differ from their stated answers and hint at a variety of biases and misrepresentations that they may exhibit toward many outcomes and scenarios during text generation. Consequently, further research is needed in the study and evaluations of LLMs' capabilities and evaluations. Additionally, the exploration of loaded terms and inherent model biases (Result #5) yielded interesting results and the use of `RevB` to further investigate them is a promising direction for future work. Finally, while the `RevB` framework is tailored to decoder-only language models, as these are specifically trained with the objective of generating a probability distribution for the next token in a sequence, it could be interesting to study possible extensions to encoder-decoder models such as T5 architectures.

Limitations

We only evaluated the RevB and the StaA of LLMs on around 500 scenarios. While these scenarios were designed to cover many different types of distributions and already hint at many characteristics of the RevB of LLMs, it would be beneficial to study additional cases (e.g., Poisson distributions), as well as other variants (e.g., multiple repeated results instead of a single repeat, or more complex dependencies between results).

Another limitation of our study is the wording of the scenario and prompts. While significant time and effort were spent designing them in order to maximize the LLMs RevB performance, it is always possible that a different wording of the context could yield better results. However, if such wordings were found, it would also highlight the significant lack of robustness of the RevB of LLMs.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. *arXiv preprint*. ArXiv:2402.03927 [cs].

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

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, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4401–4420, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. 2024. Probabilistic Reasoning in Generative Large Language Models. *arXiv preprint*. ArXiv:2402.09614 [cs].

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. Rulebert: Teaching soft rules to pre-trained language models. *arXiv preprint arXiv:2109.13006*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Example Prompts

We provide in these Appendices examples of prompt variants used to evaluate the models. We also report the detailed results of our evaluation scenarios. Where similar variants could be aggregated (e.g., when asking for the number of Heads and for the number of Tails in a coin toss), we report their average. Within each scenario, we show the scores of the `Revealed Belief` of the base LLMs, those of their instruction fine-tuned version, and finally, those of the instruction fine-tuned model's `Stated Answer`.

A.1 Scenario 1: Dice

`Revealed Belief:`

```
1. A die has 4 faces. The die is equally likely to land on any of its faces. The
   die is cast. The die lands on face number
2. A die has 4 faces. The die is equally likely to land on any of its faces. The
   die is cast. The die lands on face number 1. The die is cast again. The sum of
   both results is equal to
3. A die has 4 faces. The die is equally likely to land on any of its faces. The
   die is cast. We observe that the result is smaller than 3. Indeed, the result
   is equal to
```

`Stated Answer:`

```
4. You are given a Scenario, a Question, and a set of possible Answers. Select one
   Answer as your reply. The Answers are A, B, C, D, E. Your selected Answer will
   be contained within square brackets: [A], [B], [C], [D], [E] and your reply
   will contain nothing else.
   Scenario: A die has 4 faces. The die is equally likely to land on any of its
      faces. The die is cast. The die lands on face number 1. The die is cast
      again.
   Question: What is the probability that the die lands on face 2?
   Answers: A) 0.083 B) 0.125 C) 0.250 D) 0.500 E) 0.750
```

A.2 Scenario 2: Coins

`Revealed Belief:`

```
1. There are 2 coins. Each coin is fair and is equally likely to land on Heads and
   Tails. The coins are flipped and the resulting number of Heads is equal to

2. There are 2 coins. Each coin is fair and is equally likely to land on Heads and
   Tails. The coins are flipped a first time and the resulting number of Heads is
   0. The coins are flipped again and the total number of Heads showed in the two
   launches is equal to
```

`Stated Answer:`

```
3. You are given a Scenario, a Question, and a set of possible Answers. Select one
   Answer as your reply. The Answers are A, B, C, D, E. Your selected Answer will
   be contained within square brackets: [A], [B], [C], [D], [E] and your reply
   will contain nothing else.
   Scenario: There are 2 coins. Each coin is fair and is equally likely to land on
       Heads and Tails.
   Question: What is the probability that the resulting number of Heads is equal
      to 0 after flipping the coins?
   Answers: A) 0.750 B) 0.500 C) 0.125 D) 0.083 E) 0.250
```

A.3 Scenario 3: Choice

Revealed Belief:

1. A person has to choose randomly between two options: Left and Right. The choice is fair and each option equally likely to be chosen. The person chooses at random option

2. A person has to choose randomly between two options: Left and Right. The choice is fair and each option equally likely to be chosen. The person first chooses at random option Left. Then the person performs another random choice and chooses option

Stated Answer:

3. You are given a Scenario, a Question, and a set of possible Answers. Select one Answer as your reply. The Answers are A, B, C, D, E. Your selected Answer will be contained within square brackets: [A], [B], [C], [D], [E] and your reply will contain nothing else.
   Scenario: A person has to choose randomly between two options: Left and Right. The choice is fair and each option equally likely to be chosen.
   Question: What is the probability that the person chooses option Right?
   Answers: A) 0.250 B) 1.000 C) 0.500 D) 1.500 E) 0.167

A.4 Scenario 4: Preference

Revealed Belief:

1. A person has to choose randomly between 4 options. The options are A, B, C and D. All possible options are equally likely. The person chooses at random option

2. A person has to choose randomly between 4 options. The options are A, B, C and D. All possible options are equally likely. The person first chooses at random option B. Then the person performs another random choice and chooses option

Stated Answer:

3. You are given a Scenario, a Question, and a set of possible Answers. Select one Answer as your reply. The Answers are A, B, C, D, E. Your selected Answer will be contained within square brackets: [A], [B], [C], [D], [E] and your reply will contain nothing else.
   Scenario: A person has to choose randomly between 4 options. The options are A, B, C and D. All possible options are equally likely.
   Question: What is the probability that the person chooses option A?
   Answers: A) 0.750 B) 0.083 C) 0.125 D) 0.500 E) 0.250

B Scenario 1: Dice

B.1 `Revealed Belief` (base models)

B.1.1 Regular, independent, dependent

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 1 die	**0.06**	0.11	0.16	0.11	0.17	0.10	0.10	0.08	0.11	**0.06**	0.10
Regular – 2 dice	0.08	**0.06**	**0.06**	0.08	0.08	0.09	**0.06**	0.09	0.10	0.07	0.10
Regular – 3 dice	0.12	0.08	0.10	**0.06**	0.07	0.09	0.10	**0.06**	0.08	0.10	0.10
Independent 1 die	0.14	0.14	0.41	**0.13**	0.23	0.23	0.25	0.30	0.25	0.14	**0.13**
Independent – 2 dice	0.15	0.12	0.22	0.12	0.16	**0.09**	0.19	0.13	0.18	0.20	0.20
Independent – 3 dice	0.20	**0.12**	0.27	**0.12**	0.17	0.15	0.26	0.17	0.22	0.19	0.20
Dependant 1 die	0.10	**0.07**	0.08	0.09	0.16	0.08	0.09	0.08	0.09	0.08	0.08
Dependant – 2 dice	0.20	0.17	0.18	**0.13**	0.19	0.28	0.23	0.24	0.15	0.22	0.29
Dependant – 3 dice	0.25	0.19	**0.14**	0.15	0.21	0.30	0.17	0.24	0.18	0.33	0.22
L1											
Regular – 1 die	0.23	0.37	0.40	0.33	0.39	0.25	0.27	0.24	0.27	**0.16**	0.35
Regular – 2 dice	0.42	0.45	**0.30**	0.37	0.36	0.36	0.41	0.38	0.57	0.42	0.37
Regular – 3 dice	0.67	0.71	0.66	0.52	0.44	0.56	0.63	**0.39**	0.56	0.66	0.52
Independent 1 die	0.47	0.40	0.92	**0.34**	0.62	0.60	0.66	0.72	0.59	0.38	0.42
Independent – 2 dice	0.65	0.55	0.71	0.55	0.77	**0.52**	0.79	0.57	0.68	0.69	0.72
Independent – 3 dice	0.84	0.76	0.93	**0.74**	0.98	0.85	0.99	0.79	0.87	0.85	0.90
Dependant 1 die	0.30	**0.29**	0.34	0.30	0.56	0.31	0.39	0.32	0.34	**0.29**	0.31
Dependant – 2 dice	0.92	0.81	0.80	**0.64**	0.91	0.91	0.98	0.97	0.80	0.92	1.12
Dependant – 3 dice	1.17	1.05	0.90	**0.88**	1.22	1.07	1.04	1.23	1.06	1.08	1.12
Symmetric KL											
Regular – 1 die	0.09	0.31	0.27	0.17	0.37	0.14	0.12	0.11	0.13	**0.05**	0.26
Regular – 2 dice	0.27	0.38	**0.18**	0.24	0.22	0.22	0.27	0.24	0.58	0.29	0.28
Regular – 3 dice	0.89	0.92	0.82	0.47	0.41	0.67	0.69	**0.26**	0.55	0.94	0.55
Independent 1 die	0.45	0.37	1.23	**0.23**	0.97	0.61	0.72	0.95	0.61	0.28	0.32
Independent – 2 dice	0.88	0.72	1.05	**0.57**	1.15	0.59	1.16	0.66	0.93	0.92	0.96
Independent – 3 dice	1.81	1.48	2.32	**1.21**	1.92	1.73	2.14	1.39	1.78	1.74	1.74
Dependant 1 die	0.18	**0.16**	0.23	0.19	0.60	0.24	0.29	0.17	0.23	0.17	0.17
Dependant – 2 dice	1.49	1.18	1.09	**0.77**	1.54	1.59	1.73	1.80	1.10	1.50	2.30
Dependant – 3 dice	2.86	2.18	**1.56**	1.59	3.08	2.57	2.19	3.05	2.24	2.85	2.55

B.1.2 Observations: One observation

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller	0.10	0.07	0.14	0.06	0.11	0.16	0.11	0.07	**0.04**	0.09	0.13
Larger	**0.10**	0.11	0.16	0.12	0.19	0.24	0.29	0.14	0.11	0.14	0.13
Even	0.24	0.14	0.23	0.20	0.23	0.25	0.28	0.17	0.26	**0.10**	0.17
Odd	0.16	0.15	0.10	0.09	0.22	0.22	0.15	0.10	0.13	**0.07**	0.09
Not first	0.49	0.35	0.32	**0.18**	0.30	0.44	0.36	0.27	0.23	0.21	0.38
Not middle	0.09	0.20	0.15	**0.07**	0.15	0.20	0.13	0.17	0.10	0.10	0.10
Smaller – 2 dice	0.13	0.12	**0.08**	0.11	0.12	0.22	0.10	0.12	0.11	0.10	0.11
Larger – 2 dice	0.20	**0.09**	0.18	0.13	0.42	0.19	0.27	0.16	0.16	0.15	0.17
Even – 2 dice	0.25	0.10	0.15	0.10	0.28	0.13	0.19	0.13	0.10	0.14	**0.09**
Odd – 2 dice	0.18	0.11	0.14	0.09	0.18	0.15	0.15	0.09	0.13	**0.08**	0.11
Not first – 2 dice	0.28	0.34	**0.06**	0.10	0.36	0.21	0.67	0.10	0.20	0.10	0.23
Not middle – 2 dice	0.09	0.10	**0.06**	0.07	0.09	0.13	0.20	0.18	0.21	0.09	0.12
Smaller – 3 dice	0.15	0.13	0.14	0.13	0.11	0.15	0.12	0.19	**0.08**	0.24	0.12
Larger – 3 dice	0.14	0.13	0.11	**0.10**	0.34	0.15	0.35	0.16	0.20	0.30	0.33
Even – 3 dice	0.15	0.12	0.18	0.11	0.20	0.12	0.12	**0.10**	0.12	0.13	0.11
Odd – 3 dice	0.14	0.13	0.16	0.12	0.18	0.13	0.12	0.12	**0.11**	**0.11**	0.14
Not first – 3 dice	0.26	0.41	0.11	0.13	0.36	0.42	0.45	**0.09**	0.19	0.12	0.31
Not middle – 3 dice	0.13	0.08	**0.06**	**0.06**	0.08	0.20	0.15	0.07	0.25	0.09	0.11
L1											
Smaller	0.24	0.17	0.34	0.13	0.26	0.34	0.26	0.20	**0.10**	0.20	0.29
Larger	0.39	0.30	0.39	0.32	0.54	0.60	0.72	0.45	**0.25**	0.33	0.32
Even	0.67	0.41	0.54	0.46	0.60	0.65	0.58	0.45	0.61	**0.23**	0.37
Odd	0.52	0.40	0.28	0.24	0.56	0.61	0.35	0.27	0.39	**0.22**	**0.22**
Not first	1.09	1.02	0.81	0.56	0.88	1.05	1.14	0.63	0.62	**0.55**	0.88
Not middle	0.39	0.62	0.42	0.29	0.45	0.59	0.45	0.47	0.42	**0.28**	0.29
Smaller – 2 dice	0.38	0.31	0.28	**0.27**	0.35	0.57	0.41	0.38	0.28	0.29	0.30
Larger – 2 dice	0.61	**0.35**	0.48	0.39	0.95	0.56	0.77	0.54	0.47	0.38	0.45
Even – 2 dice	0.78	**0.35**	0.70	0.52	0.90	0.60	0.68	0.43	0.56	0.49	0.44
Odd – 2 dice	0.67	**0.35**	0.81	0.42	0.81	0.71	0.63	0.36	0.56	0.37	0.43
Not first – 2 dice	0.97	1.11	**0.36**	0.62	1.08	0.82	1.45	0.54	0.72	0.55	0.79
Not middle – 2 dice	0.52	0.50	**0.27**	0.36	0.54	0.53	0.64	0.63	0.72	0.36	0.44
Smaller – 3 dice	0.47	0.31	0.39	0.46	0.43	0.46	0.38	0.59	**0.28**	0.53	0.42
Larger – 3 dice	0.57	0.51	0.41	**0.38**	0.87	0.44	0.92	0.60	0.62	0.72	0.80
Even – 3 dice	0.77	0.67	0.93	0.66	1.10	0.74	0.56	**0.55**	0.74	0.77	0.63
Odd – 3 dice	0.75	0.66	1.06	0.76	1.19	0.95	0.63	**0.62**	0.69	0.70	0.70
Not first – 3 dice	1.19	1.49	0.79	0.92	1.32	1.33	1.37	**0.65**	0.91	0.88	1.21
Not middle – 3 dice	0.62	0.58	**0.40**	0.43	0.72	0.73	0.61	0.47	0.80	0.53	0.60
Symmetric KL											
Smaller	0.42	**0.08**	0.65	0.09	0.68	0.41	0.54	0.50	0.15	0.22	0.33
Larger	1.08	0.30	0.48	**0.27**	1.29	1.19	1.77	0.90	0.50	0.43	**0.27**
Even	1.49	0.65	0.61	0.37	1.11	1.22	0.65	0.94	0.75	0.35	**0.31**
Odd	1.08	0.40	0.31	0.23	1.35	1.19	0.35	0.40	0.72	0.49	**0.16**
Not first	2.09	2.68	1.47	**0.63**	1.63	2.05	3.45	1.07	1.13	0.97	1.42
Not middle	0.59	1.21	0.72	0.23	0.56	0.99	1.27	0.79	0.54	0.93	**0.21**
Smaller – 2 dice	0.41	**0.16**	0.62	0.36	0.37	0.60	0.66	0.59	**0.16**	0.31	0.27
Larger – 2 dice	1.51	0.53	**0.49**	0.60	1.62	0.93	2.09	1.25	0.86	**0.49**	0.70
Even – 2 dice	1.63	**0.36**	2.06	0.86	1.63	1.53	1.00	0.48	0.83	0.86	0.51
Odd – 2 dice	1.67	0.65	2.91	1.23	2.16	3.07	1.11	0.52	0.91	0.96	**0.48**
Not first – 2 dice	2.26	2.55	**0.32**	1.14	2.57	1.77	9.00	0.84	2.51	1.07	1.41
Not middle – 2 dice	0.74	0.90	**0.25**	0.40	0.81	0.73	2.46	1.26	2.72	0.59	0.36
Smaller – 3 dice	0.58	0.23	0.42	0.61	0.52	0.40	0.70	1.10	**0.20**	0.49	0.69
Larger – 3 dice	1.12	0.78	**0.38**	0.55	1.38	0.61	1.95	0.95	0.99	0.97	1.33
Even – 3 dice	1.68	**0.86**	2.88	1.35	2.46	1.91	1.06	1.04	1.17	2.43	1.20
Odd – 3 dice	2.13	**1.06**	4.58	1.77	3.29	3.78	1.30	1.30	1.36	1.79	1.28
Not first – 3 dice	2.88	4.68	1.20	1.81	3.64	3.53	6.62	**1.08**	2.79	1.91	2.97
Not middle – 3 dice	0.89	1.14	**0.38**	0.57	1.26	1.04	1.95	0.62	3.24	0.77	0.82

B.1.3 Observations: Two observations – Single die

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even	0.35	0.21	0.23	0.29	0.19	0.31	0.24	0.19	0.31	0.21	**0.17**
Smaller – Odd	0.11	**0.07**	0.13	0.10	0.17	0.16	0.15	0.17	0.11	0.11	**0.07**
Smaller – Not first	0.21	0.25	0.19	0.16	0.22	0.24	0.35	0.24	0.15	0.23	**0.14**
Smaller – Not middle	**0.07**	0.08	0.11	0.09	**0.07**	0.10	0.20	0.16	0.13	0.08	0.13
Larger – Even	0.29	0.24	0.28	0.20	0.37	0.33	0.36	0.22	0.24	0.26	**0.16**
Larger – Odd	0.25	0.17	0.13	0.16	0.26	0.26	0.18	0.19	0.17	**0.09**	**0.09**
Larger – Not first	0.11	0.19	0.14	**0.08**	0.25	0.21	0.35	0.14	0.15	0.13	0.13
Larger – Not middle	**0.12**	0.23	0.19	0.21	0.29	0.22	0.35	**0.12**	0.16	0.13	**0.12**
Even – Not first	0.26	0.42	0.31	0.25	0.37	0.24	0.53	0.14	0.32	**0.07**	0.17
Even – Not middle	0.16	0.18	0.15	0.22	0.19	0.25	0.40	0.14	0.24	**0.09**	0.14
L1											
Smaller – Even	0.55	0.33	0.36	0.41	0.39	0.47	0.46	0.31	0.48	0.33	**0.30**
Smaller – Odd	0.26	0.15	0.28	0.21	0.38	0.32	0.31	0.35	0.22	0.24	**0.14**
Smaller – Not first	0.43	0.59	0.46	**0.33**	0.56	0.55	0.57	0.60	0.35	0.38	**0.33**
Smaller – Not middle	0.19	**0.17**	0.23	0.20	0.23	0.23	0.55	0.46	0.33	0.22	0.30
Larger – Even	0.69	0.50	0.60	0.42	0.85	0.73	0.77	0.52	0.53	0.52	**0.33**
Larger – Odd	0.58	0.37	0.26	0.36	0.60	0.56	0.37	0.39	0.36	0.18	**0.17**
Larger – Not first	0.43	0.68	0.58	**0.25**	0.91	0.75	0.96	0.59	0.68	0.28	0.44
Larger – Not middle	0.46	0.63	0.54	0.51	0.86	0.74	0.97	0.55	0.56	0.37	**0.36**
Even – Not first	0.85	0.94	0.72	0.78	1.05	0.84	1.22	0.42	0.73	**0.20**	0.42
Even – Not middle	0.47	0.49	0.42	0.52	0.58	0.70	0.82	0.48	0.52	**0.22**	0.39
Symmetric KL											
Smaller – Even	2.08	0.34	0.86	0.39	0.99	0.84	0.36	0.42	0.61	**0.17**	**0.17**
Smaller – Odd	0.71	0.21	0.83	0.42	1.81	0.35	0.34	0.45	0.50	0.25	**0.17**
Smaller – Not first	1.04	0.98	1.01	0.22	0.61	1.15	0.46	2.12	0.43	0.19	**0.07**
Smaller – Not middle	0.41	**0.23**	0.51	0.28	0.65	0.93	2.34	2.12	1.34	0.70	0.86
Larger – Even	2.83	0.71	1.21	0.87	3.28	1.79	2.00	1.27	1.18	0.86	**0.41**
Larger – Odd	2.77	0.65	0.69	1.05	2.75	1.27	0.65	1.41	0.90	0.41	**0.22**
Larger – Not first	1.37	3.10	2.76	**0.69**	3.76	3.34	2.51	2.42	3.28	0.82	1.24
Larger – Not middle	1.67	1.04	0.85	**0.57**	2.46	2.33	3.03	2.38	2.64	1.02	0.81
Even – Not first	4.26	1.82	1.06	1.96	2.84	2.70	2.71	1.08	1.18	0.49	**0.48**
Even – Not middle	1.33	0.96	0.75	0.67	1.60	1.83	1.52	1.52	0.75	0.48	**0.43**

B.1.4 Observations: Two observations – Multiple die

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even – 2 dice	0.37	0.18	**0.12**	0.32	0.40	0.35	0.29	0.14	0.28	0.15	0.29
Smaller – Odd – 2 dice	0.24	0.08	0.18	0.12	0.35	0.15	0.17	0.12	0.10	**0.07**	0.14
Smaller – Not first – 2 dice	0.23	0.23	0.14	0.26	0.52	0.26	0.19	**0.12**	0.16	0.23	0.23
Smaller – Not middle – 2 dice	0.14	0.09	**0.06**	0.11	0.14	0.10	0.32	0.13	0.14	0.11	0.11
Larger – Even – 2 dice	0.30	0.17	0.33	0.13	0.23	0.23	0.17	0.13	0.20	**0.08**	0.14
Larger – Odd – 2 dice	0.30	**0.10**	0.16	0.17	0.39	0.20	0.27	0.17	0.20	0.14	0.16
Larger – Not first – 2 dice	0.15	**0.07**	0.15	0.12	0.26	0.19	0.15	0.16	0.16	0.11	0.12
Larger – Not middle – 2 dice	0.29	0.20	**0.15**	0.20	0.39	0.29	0.27	0.20	0.21	0.22	0.16
Even – Not first – 2 dice	0.13	0.39	0.12	0.24	0.39	0.15	0.27	**0.11**	0.23	0.25	0.16
Even – Not middle – 2 dice	0.15	0.13	0.12	0.14	**0.11**	0.20	0.12	0.12	0.15	0.16	**0.11**
Smaller – Even – 3 dice	0.26	0.19	0.15	0.28	0.34	0.24	0.19	0.17	0.24	**0.12**	0.23
Smaller – Odd – 3 dice	0.24	0.20	0.19	0.25	0.37	0.21	0.23	0.17	0.21	**0.12**	0.22
Smaller – Not first – 3 dice	0.20	0.19	0.15	0.19	0.38	0.18	**0.14**	0.15	**0.14**	0.20	0.16
Smaller – Not middle – 3 dice	0.22	0.11	0.13	**0.10**	0.15	0.17	0.29	0.18	0.20	0.18	**0.10**
Larger – Even – 3 dice	0.25	0.20	0.21	**0.13**	0.23	0.18	0.17	0.20	0.16	0.20	0.22
Larger – Odd – 3 dice	0.24	0.17	0.16	**0.13**	0.24	0.20	0.19	0.16	0.17	0.20	0.20
Larger – Not first – 3 dice	**0.10**	0.11	0.11	**0.10**	0.14	0.13	0.17	0.14	0.17	0.14	0.14
Larger – Not middle – 3 dice	0.26	0.24	0.21	0.21	0.35	0.34	0.22	**0.16**	0.29	0.36	0.29
Even – Not first – 3 dice	**0.13**	0.35	0.15	0.20	0.39	0.25	0.19	0.14	0.15	0.20	0.28
Even – Not middle – 3 dice	0.14	0.12	**0.09**	0.12	0.13	0.15	0.13	0.12	0.13	0.12	0.13
L1											
Smaller – Even – 2 dice	0.92	0.40	0.54	0.64	0.92	0.80	0.66	0.37	0.58	**0.31**	0.67
Smaller – Odd – 2 dice	0.60	**0.21**	0.76	0.28	0.78	0.37	0.36	0.34	0.22	**0.21**	0.32
Smaller – Not first – 2 dice	0.72	0.59	**0.40**	0.69	1.14	0.77	0.60	0.52	0.50	0.62	0.61
Smaller – Not middle – 2 dice	0.42	0.27	**0.25**	0.36	0.47	0.29	0.73	0.52	0.45	0.37	0.55
Larger – Even – 2 dice	0.82	0.45	0.78	0.38	0.76	0.60	0.45	0.47	0.54	**0.29**	0.45
Larger – Odd – 2 dice	0.94	**0.30**	0.49	0.55	1.07	0.70	0.64	0.55	0.61	0.42	0.49
Larger – Not first – 2 dice	0.66	**0.53**	0.68	0.76	1.02	0.76	0.79	0.77	1.02	0.60	0.66
Larger – Not middle – 2 dice	0.77	0.56	**0.51**	0.53	1.01	0.78	0.96	0.69	0.73	0.55	0.53
Even – Not first – 2 dice	0.61	1.03	0.67	0.83	1.17	0.84	0.91	**0.51**	0.74	0.81	0.68
Even – Not middle – 2 dice	0.72	0.55	0.61	0.61	0.55	0.78	0.51	**0.47**	0.66	0.67	0.50
Smaller – Even – 3 dice	0.81	0.43	0.66	0.59	0.86	0.66	0.50	0.53	0.60	**0.31**	0.58
Smaller – Odd – 3 dice	0.74	0.48	0.87	0.64	1.02	0.59	0.55	0.52	0.53	**0.31**	0.52
Smaller – Not first – 3 dice	0.75	0.71	0.67	0.72	1.13	0.72	**0.45**	0.87	0.63	0.78	0.68
Smaller – Not middle – 3 dice	0.66	**0.30**	0.45	0.34	0.50	0.44	0.81	0.73	0.62	0.60	0.51
Larger – Even – 3 dice	0.82	0.59	0.58	**0.47**	0.82	0.56	0.49	0.64	0.51	0.60	0.71
Larger – Odd – 3 dice	0.91	0.52	**0.49**	0.52	0.91	0.70	0.53	0.55	**0.49**	0.62	0.69
Larger – Not first – 3 dice	**0.53**	0.72	0.81	0.72	0.95	0.67	0.70	0.77	1.01	0.89	0.78
Larger – Not middle – 3 dice	0.79	0.70	0.74	**0.59**	1.00	0.89	0.87	0.64	0.84	0.90	0.78
Even – Not first – 3 dice	0.85	1.39	1.07	1.08	1.53	1.32	1.16	**0.78**	0.91	1.19	1.17
Even – Not middle – 3 dice	0.82	0.65	0.63	0.65	0.81	0.77	0.71	**0.60**	0.73	0.77	0.70
Symmetric KL											
Smaller – Even – 2 dice	4.36	0.63	2.29	1.10	2.21	2.94	1.27	0.74	0.54	**0.46**	0.79
Smaller – Odd – 2 dice	2.81	0.47	3.54	0.73	2.92	1.20	0.52	1.36	**0.32**	0.53	0.44
Smaller – Not first – 2 dice	1.12	0.80	**0.47**	1.04	2.27	1.39	1.10	1.02	0.81	0.74	0.78
Smaller – Not middle – 2 dice	0.76	**0.27**	0.74	0.63	0.97	0.46	3.82	1.54	1.05	0.77	1.86
Larger – Even – 2 dice	3.19	0.83	1.91	1.16	2.72	1.55	0.95	1.61	1.50	**0.77**	1.46
Larger – Odd – 2 dice	4.14	**0.92**	2.04	2.10	5.88	2.97	1.27	2.33	1.86	1.50	1.94
Larger – Not first – 2 dice	**2.13**	2.19	2.84	3.29	4.49	2.36	2.89	2.27	5.03	2.21	2.77
Larger – Not middle – 2 dice	2.20	0.90	1.25	**0.75**	2.17	1.71	3.70	1.62	2.16	0.81	1.26
Even – Not first – 2 dice	1.89	1.93	2.54	2.27	3.31	3.32	2.68	**1.23**	3.01	2.62	1.56
Even – Not middle – 2 dice	2.46	1.34	2.00	0.97	1.59	2.69	1.26	**0.95**	1.30	1.32	1.66
Smaller – Even – 3 dice	3.52	0.65	2.70	1.30	2.42	2.50	1.20	1.70	0.60	**0.42**	0.82
Smaller – Odd – 3 dice	2.57	0.74	3.71	1.17	3.11	1.93	0.62	1.38	0.61	**0.42**	0.78
Smaller – Not first – 3 dice	1.55	1.08	1.06	1.27	2.51	1.34	**0.76**	2.24	1.18	1.36	1.12
Smaller – Not middle – 3 dice	1.26	**0.31**	0.84	0.67	0.93	0.56	3.70	1.87	1.24	1.29	1.28
Larger – Even – 3 dice	3.46	1.17	1.51	1.41	3.04	1.58	**0.92**	1.37	1.04	1.37	2.32
Larger – Odd – 3 dice	4.15	1.17	1.69	1.89	3.81	2.78	**0.92**	1.84	0.95	1.60	2.09
Larger – Not first – 3 dice	**1.29**	2.97	3.49	2.95	4.00	2.02	1.89	2.31	4.67	3.00	2.97
Larger – Not middle – 3 dice	1.68	1.14	1.62	**0.82**	2.27	1.72	2.62	1.31	2.01	1.45	1.50
Even – Not first – 3 dice	2.74	3.80	3.63	2.79	5.62	4.44	4.05	**1.63**	2.21	3.80	3.07
Even – Not middle – 3 dice	3.05	1.65	2.37	1.23	2.48	2.30	2.68	**1.22**	1.53	2.54	2.12

B.2 `Revealed Belief` (instruction fine-tuned models)

B.2.1 Regular, independent, dependent

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 1 die	0.13	0.15	0.24	0.18	0.33	0.11	0.09	0.12	0.10	0.08	**0.06**
Regular – 2 dice	0.16	0.12	0.11	0.30	0.31	0.15	0.16	0.25	0.26	0.10	**0.06**
Regular – 3 dice	0.20	0.18	0.16	0.21	0.24	**0.08**	0.27	0.12	0.27	0.13	**0.08**
Independent 1 die	0.28	0.17	0.68	0.22	0.41	0.32	0.46	0.36	0.29	0.16	**0.13**
Independent – 2 dice	0.21	**0.12**	0.73	0.26	0.34	0.17	0.43	0.21	0.24	0.25	0.19
Independent – 3 dice	0.40	**0.13**	0.79	0.38	0.37	0.26	0.42	0.17	0.25	0.21	0.19
Dependant 1 die	0.17	0.11	0.14	0.22	0.35	0.14	0.20	0.17	0.20	0.12	**0.07**
Dependant – 2 dice	0.28	**0.21**	0.30	0.68	0.53	0.39	0.31	0.35	0.22	0.22	0.25
Dependant – 3 dice	0.28	**0.20**	0.29	0.84	0.65	0.45	0.27	0.33	0.29	0.32	0.31
L1											
Regular – 1 die	0.44	0.52	0.53	0.56	1.01	0.45	0.35	0.39	0.42	0.25	**0.22**
Regular – 2 dice	0.64	0.57	0.47	0.82	0.99	0.52	0.70	0.78	0.91	0.60	**0.29**
Regular – 3 dice	0.85	0.77	0.85	0.75	1.08	**0.43**	1.03	0.59	0.86	0.86	0.47
Independent 1 die	0.68	0.49	1.37	0.71	1.06	0.72	1.00	0.83	0.69	0.45	**0.41**
Independent – 2 dice	0.86	**0.61**	1.51	0.94	1.19	0.62	1.17	0.72	0.89	0.80	0.73
Independent – 3 dice	1.17	**0.78**	1.61	1.22	1.24	0.92	1.27	0.83	1.00	0.89	0.92
Dependant 1 die	0.56	0.41	0.50	0.71	1.00	0.50	0.59	0.56	0.62	0.42	**0.24**
Dependant – 2 dice	1.11	0.99	1.02	1.45	1.58	1.12	1.11	1.18	1.00	0.99	**0.94**
Dependant – 3 dice	1.33	1.07	1.22	1.70	1.77	1.29	**1.00**	1.40	1.19	1.13	1.11
Symmetric KL											
Regular – 1 die	0.35	0.69	0.50	0.66	2.78	0.40	0.18	0.28	0.29	0.12	**0.10**
Regular – 2 dice	0.71	0.54	0.40	1.08	2.27	0.49	0.82	0.99	1.97	0.64	**0.14**
Regular – 3 dice	1.63	1.26	1.40	0.94	3.07	**0.43**	1.85	0.62	1.23	1.45	0.45
Independent 1 die	0.95	0.50	2.91	0.90	2.61	0.95	1.50	1.32	0.86	0.38	**0.28**
Independent – 2 dice	1.54	**0.66**	4.13	1.66	3.01	0.72	2.55	1.08	1.68	1.17	0.95
Independent – 3 dice	3.36	**1.26**	6.26	3.15	4.35	2.06	3.67	1.43	2.14	1.78	1.78
Dependant 1 die	0.58	0.27	0.45	0.98	2.26	0.53	0.65	0.51	0.76	0.32	**0.10**
Dependant – 2 dice	2.37	1.72	1.82	4.21	6.47	2.50	2.27	2.98	2.03	1.76	**1.71**
Dependant – 3 dice	3.78	**2.17**	2.92	6.85	9.48	4.12	2.50	4.18	3.16	3.01	2.95

B.2.2 Observations: One observation

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller	0.27	**0.08**	0.19	0.15	0.40	0.11	0.16	0.11	0.09	0.15	0.17
Larger	0.29	0.27	0.23	0.32	0.63	**0.17**	0.34	0.18	**0.17**	**0.17**	**0.17**
Even	0.29	0.26	0.24	0.66	0.70	0.29	0.27	0.24	0.26	**0.15**	0.26
Odd	0.26	0.34	0.18	0.41	0.59	0.12	0.23	0.23	0.19	**0.07**	0.14
Not first	0.19	0.40	0.36	**0.16**	0.29	0.28	0.59	0.38	0.27	0.22	0.43
Not middle	0.27	0.22	0.19	0.14	0.45	0.15	0.16	0.23	0.14	0.13	**0.12**
Smaller – 2 dice	0.25	0.20	0.17	0.44	0.31	0.29	0.13	0.24	0.19	0.16	**0.09**
Larger – 2 dice	0.17	0.14	0.25	0.30	0.33	0.12	0.37	0.30	0.31	0.17	**0.07**
Even – 2 dice	0.13	0.17	0.30	0.33	0.34	0.14	0.15	0.26	0.22	0.21	**0.08**
Odd – 2 dice	0.22	0.16	0.21	0.16	0.53	0.21	0.13	0.15	0.26	0.09	**0.07**
Not first – 2 dice	0.34	0.33	**0.08**	0.15	0.31	0.21	0.14	0.26	0.26	0.16	0.17
Not middle – 2 dice	0.16	0.15	**0.07**	0.16	0.21	0.17	0.15	0.27	0.17	0.11	0.13
Smaller – 3 dice	0.22	0.15	0.26	0.21	0.40	0.20	**0.09**	0.30	0.20	0.29	0.12
Larger – 3 dice	**0.14**	0.16	0.22	0.16	0.30	0.18	0.32	0.27	0.22	0.29	0.16
Even – 3 dice	0.13	0.17	0.32	0.18	0.59	0.13	0.17	0.13	0.17	0.15	**0.09**
Odd – 3 dice	0.23	0.10	0.31	0.17	0.69	0.16	0.12	0.18	**0.09**	0.14	0.10
Not first – 3 dice	0.35	0.40	0.16	0.11	0.44	0.37	**0.09**	0.15	0.23	0.16	0.28
Not middle – 3 dice	0.18	0.14	**0.10**	0.12	0.19	0.23	0.14	0.13	0.16	0.12	0.15
L1											
Smaller	0.60	0.23	0.42	0.36	0.82	0.28	0.37	**0.22**	0.24	0.34	0.38
Larger	0.79	0.61	0.52	0.79	1.28	0.51	0.83	0.55	0.45	**0.43**	0.45
Even	0.68	0.57	0.59	1.31	1.40	0.66	0.64	0.63	0.67	**0.34**	0.62
Odd	0.72	0.73	0.42	1.00	1.22	0.35	0.62	0.58	0.49	**0.24**	0.39
Not first	**0.55**	1.22	0.91	0.62	0.85	0.97	1.28	0.80	0.72	0.59	0.93
Not middle	0.76	0.66	0.50	0.55	1.09	0.49	0.56	0.63	0.53	0.40	**0.32**
Smaller – 2 dice	0.78	0.56	0.48	0.98	0.79	0.77	0.38	0.61	0.50	0.42	**0.24**
Larger – 2 dice	0.57	0.39	0.63	0.75	0.88	0.60	0.94	0.77	0.79	0.40	**0.18**
Even – 2 dice	0.49	0.56	0.97	0.98	1.26	0.51	0.62	0.69	0.75	0.55	**0.32**
Odd – 2 dice	0.90	0.45	0.95	0.57	1.17	1.04	0.62	0.44	0.72	0.42	**0.30**
Not first – 2 dice	1.21	1.03	**0.50**	0.73	1.29	0.88	0.71	0.76	0.92	0.65	0.59
Not middle – 2 dice	0.81	0.56	**0.33**	0.53	0.79	0.66	0.72	0.83	0.74	0.45	0.45
Smaller – 3 dice	0.72	0.41	0.63	0.62	1.04	0.50	**0.30**	0.84	0.51	0.59	0.33
Larger – 3 dice	0.50	0.49	0.66	0.50	1.00	0.65	0.87	0.77	0.80	0.79	**0.43**
Even – 3 dice	0.56	0.52	1.20	0.54	1.59	0.64	0.90	0.53	0.58	0.68	**0.50**
Odd – 3 dice	1.03	0.47	1.33	0.69	1.57	1.10	0.69	0.72	**0.43**	0.74	0.53
Not first – 3 dice	1.38	1.32	0.88	0.74	1.62	1.46	**0.73**	0.87	0.94	0.90	1.10
Not middle – 3 dice	0.82	0.59	0.61	**0.51**	0.92	0.76	0.64	0.63	0.78	0.66	0.57
Symmetric KL											
Smaller	0.88	0.15	0.48	0.22	1.74	0.60	1.12	0.19	**0.11**	0.29	0.44
Larger	1.38	0.56	0.66	1.45	4.18	1.53	2.24	0.93	0.43	0.47	**0.40**
Even	1.36	0.76	0.80	4.23	6.07	0.88	0.89	0.98	1.03	**0.31**	0.86
Odd	1.46	0.77	0.47	2.49	3.47	0.59	1.27	0.49	0.45	**0.43**	1.08
Not first	**0.76**	4.01	1.87	0.87	2.62	3.05	2.49	1.14	1.43	0.98	1.39
Not middle	1.43	1.03	0.89	0.50	3.16	1.71	1.25	0.70	1.18	0.96	**0.22**
Smaller – 2 dice	1.16	0.47	0.52	2.23	1.58	1.17	0.31	1.01	0.37	0.33	**0.19**
Larger – 2 dice	0.68	0.31	0.76	0.97	1.48	1.93	2.12	1.35	1.52	0.43	**0.14**
Even – 2 dice	0.87	0.58	2.35	1.63	3.41	0.58	1.54	0.84	1.50	0.70	**0.51**
Odd – 2 dice	1.99	**0.39**	3.89	0.64	3.94	5.03	1.82	0.46	1.30	0.83	0.40
Not first – 2 dice	2.87	1.79	**0.62**	1.12	3.22	2.75	1.62	1.19	2.69	1.11	0.87
Not middle – 2 dice	1.65	0.62	**0.29**	0.57	1.24	0.83	1.22	1.47	1.75	0.81	0.38
Smaller – 3 dice	1.13	0.33	0.58	0.75	2.59	0.53	**0.20**	1.73	0.39	0.58	0.27
Larger – 3 dice	0.64	0.53	0.76	0.49	1.99	0.91	1.65	1.12	1.58	1.04	**0.37**
Even – 3 dice	1.03	**0.49**	3.59	0.63	6.54	1.20	2.62	0.88	0.63	1.46	0.93
Odd – 3 dice	2.35	**0.45**	7.22	0.95	7.75	4.25	2.25	1.20	0.77	1.28	1.00
Not first – 3 dice	3.84	3.41	1.79	**1.09**	5.95	5.35	1.25	1.55	1.96	1.83	2.35
Not middle – 3 dice	1.55	0.82	0.76	0.55	1.93	0.91	1.21	0.77	1.49	1.00	**0.54**

B.2.3 Observations: Two observations – Single die

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even	0.29	0.26	0.27	0.33	0.40	0.22	0.29	**0.19**	**0.19**	0.26	**0.19**
Smaller – Odd	0.18	0.15	0.13	0.15	0.29	0.19	0.19	0.17	**0.09**	0.17	0.16
Smaller – Not first	0.27	0.25	**0.17**	0.22	0.50	0.24	0.27	0.25	0.20	0.32	0.23
Smaller – Not middle	0.13	0.20	**0.09**	0.12	0.29	0.10	0.17	**0.09**	**0.09**	**0.09**	0.18
Larger – Even	0.28	0.33	0.34	0.36	0.41	0.28	0.47	0.32	**0.24**	0.29	0.29
Larger – Odd	0.25	0.27	0.16	0.23	0.23	0.21	0.25	0.20	**0.09**	0.12	0.23
Larger – Not first	0.25	0.23	**0.13**	0.40	0.47	0.15	0.35	0.24	0.15	0.20	0.17
Larger – Not middle	0.15	0.28	0.27	0.33	0.50	0.21	0.30	**0.14**	**0.14**	0.15	0.21
Even – Not first	0.21	0.44	0.50	0.30	0.49	0.20	0.23	0.28	0.18	**0.13**	0.25
Even – Not middle	0.26	**0.17**	0.21	0.33	0.49	0.18	0.43	0.19	0.21	**0.17**	0.25
L1											
Smaller – Even	0.39	0.35	0.41	0.46	0.60	0.28	0.47	**0.24**	0.31	0.41	0.33
Smaller – Odd	0.36	0.31	0.28	0.30	0.58	0.39	0.39	0.34	**0.17**	0.35	0.34
Smaller – Not first	0.45	0.46	0.42	0.43	0.81	0.48	0.49	0.45	**0.39**	0.48	0.40
Smaller – Not middle	0.32	0.43	**0.21**	0.31	0.67	0.33	0.43	0.25	**0.21**	0.24	0.42
Larger – Even	0.60	0.65	0.70	0.73	0.82	0.66	0.97	0.69	**0.48**	0.58	0.69
Larger – Odd	0.57	0.55	0.32	0.47	0.46	0.50	0.51	0.40	**0.19**	0.25	0.49
Larger – Not first	0.58	0.60	0.58	1.00	1.12	0.55	0.95	0.69	0.45	**0.43**	0.52
Larger – Not middle	0.44	0.70	0.68	0.74	1.11	0.65	0.86	0.50	0.38	**0.35**	0.57
Even – Not first	0.53	0.95	1.05	0.72	1.15	0.59	0.75	0.62	0.45	**0.29**	0.68
Even – Not middle	0.59	0.48	0.53	0.71	0.98	0.43	0.90	0.47	0.46	**0.36**	0.60
Symmetric KL											
Smaller – Even	0.42	0.12	0.33	0.28	2.44	0.22	0.56	0.09	**0.06**	0.18	0.50
Smaller – Odd	0.33	0.24	0.37	0.24	2.27	0.48	1.57	0.30	**0.09**	0.29	1.06
Smaller – Not first	0.37	0.23	0.66	0.27	1.97	0.67	0.40	0.74	**0.13**	0.30	0.24
Smaller – Not middle	0.31	0.32	0.26	**0.19**	1.32	0.97	1.45	1.02	0.20	0.21	1.00
Larger – Even	1.09	1.19	1.30	1.45	3.52	1.61	2.87	1.15	**0.44**	1.02	1.51
Larger – Odd	2.04	0.87	0.63	0.55	1.47	1.63	1.05	0.72	**0.08**	0.39	1.42
Larger – Not first	0.69	0.71	2.45	2.33	4.37	2.05	3.24	1.76	1.09	**0.53**	1.36
Larger – Not middle	1.00	0.83	0.98	1.10	4.07	1.43	2.57	1.47	0.74	**0.42**	0.94
Even – Not first	0.75	1.88	1.77	1.64	4.63	1.61	2.14	0.86	0.36	**0.24**	1.28
Even – Not middle	0.96	0.66	0.81	1.25	3.53	0.64	1.79	0.83	**0.43**	0.44	0.99

B.2.4 Observations: Two observations – Multiple dice

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even – 2 dice	0.46	0.33	**0.19**	0.43	0.66	0.44	0.37	0.29	0.32	0.28	0.23
Smaller – Odd – 2 dice	0.14	0.13	0.13	0.15	0.41	0.19	0.17	0.23	**0.08**	0.11	**0.08**
Smaller – Not first – 2 dice	0.28	0.27	0.18	0.27	0.65	0.53	0.19	0.24	0.18	0.27	**0.17**
Smaller – Not middle – 2 dice	0.22	0.18	0.15	0.27	0.33	0.28	0.13	0.23	0.13	0.16	**0.12**
Larger – Even – 2 dice	0.40	**0.16**	0.34	0.29	0.63	0.38	0.30	0.28	0.41	0.25	0.21
Larger – Odd – 2 dice	0.21	**0.12**	0.18	0.31	0.29	0.26	0.28	0.27	0.34	0.24	**0.12**
Larger – Not first – 2 dice	0.17	0.10	0.17	0.21	0.33	0.17	0.16	0.36	0.30	0.14	**0.07**
Larger – Not middle – 2 dice	0.25	**0.14**	0.22	0.31	0.39	0.29	0.25	0.37	0.33	0.24	**0.14**
Even – Not first – 2 dice	0.17	0.36	0.17	0.57	0.43	0.24	0.45	0.13	0.19	0.37	**0.12**
Even – Not middle – 2 dice	0.22	0.15	0.22	0.23	0.31	0.14	0.21	0.20	0.18	0.24	**0.12**
Smaller – Even – 3 dice	0.39	0.34	0.20	0.34	0.59	0.23	0.29	**0.16**	0.30	0.22	0.21
Smaller – Odd – 3 dice	0.26	0.23	0.19	0.31	0.59	0.23	**0.13**	0.19	0.23	0.19	0.19
Smaller – Not first – 3 dice	0.27	0.20	0.14	0.18	0.47	0.33	0.14	0.18	0.17	0.20	**0.13**
Smaller – Not middle – 3 dice	0.26	**0.13**	0.25	0.19	0.29	0.40	0.24	0.34	0.23	0.27	0.21
Larger – Even – 3 dice	0.32	0.24	0.21	0.22	0.50	0.24	0.23	0.33	0.33	0.31	**0.17**
Larger – Odd – 3 dice	0.28	0.26	**0.14**	0.20	0.49	0.26	0.36	0.32	0.30	0.31	0.15
Larger – Not first – 3 dice	0.15	0.12	0.12	0.14	0.30	0.13	0.21	0.19	0.21	0.17	**0.11**
Larger – Not middle – 3 dice	0.22	**0.17**	0.20	0.18	0.41	0.45	0.26	0.24	0.27	0.40	0.29
Even – Not first – 3 dice	**0.12**	0.43	0.13	0.21	0.39	0.39	0.20	0.16	0.14	0.24	0.23
Even – Not middle – 3 dice	0.16	0.15	0.21	0.21	0.40	0.14	0.15	0.14	0.15	0.17	**0.13**
L1											
Smaller – Even – 2 dice	1.25	0.71	0.60	0.88	1.33	0.91	0.97	0.62	0.68	0.59	**0.46**
Smaller – Odd – 2 dice	0.32	0.26	0.48	0.32	0.88	0.46	0.39	0.50	0.20	0.23	**0.17**
Smaller – Not first – 2 dice	0.80	0.63	0.56	0.67	1.40	1.13	0.71	0.67	0.55	0.75	**0.45**
Smaller – Not middle – 2 dice	0.53	0.46	0.42	0.66	0.88	0.77	0.44	0.70	**0.37**	0.41	0.47
Larger – Even – 2 dice	1.10	**0.43**	0.80	0.65	1.31	0.82	0.74	0.64	0.87	0.58	0.51
Larger – Odd – 2 dice	0.63	**0.30**	0.48	0.70	0.64	0.63	0.66	0.66	0.68	0.61	**0.30**
Larger – Not first – 2 dice	0.52	**0.28**	0.59	0.63	0.75	0.91	0.81	0.95	0.93	0.47	0.32
Larger – Not middle – 2 dice	0.62	0.42	0.54	0.78	0.89	0.90	0.83	0.84	0.82	0.57	**0.33**
Even – Not first – 2 dice	0.58	0.92	0.74	1.31	1.25	0.69	1.09	**0.43**	0.65	1.06	0.45
Even – Not middle – 2 dice	0.80	0.47	0.79	0.66	0.86	0.51	0.77	0.58	0.71	0.74	**0.43**
Smaller – Even – 3 dice	0.99	0.71	0.87	0.72	1.22	0.54	0.68	**0.37**	0.68	0.47	0.44
Smaller – Odd – 3 dice	0.64	0.52	0.85	0.64	1.41	0.54	**0.40**	0.53	0.50	0.45	0.41
Smaller – Not first – 3 dice	0.87	0.66	0.57	**0.51**	1.40	0.99	0.58	0.73	0.58	0.68	0.54
Smaller – Not middle – 3 dice	0.59	**0.36**	0.63	0.56	0.75	1.05	0.64	0.90	0.50	0.65	0.62
Larger – Even – 3 dice	0.86	0.59	0.61	0.62	1.12	0.66	0.58	0.73	0.76	0.72	**0.50**
Larger – Odd – 3 dice	0.90	0.61	0.52	0.62	1.02	0.64	0.79	0.78	0.67	0.73	**0.43**
Larger – Not first – 3 dice	**0.45**	0.46	0.66	0.51	0.83	0.89	0.82	0.73	0.85	0.89	0.54
Larger – Not middle – 3 dice	0.69	0.55	0.69	**0.53**	0.97	1.18	0.78	0.72	0.85	0.97	0.65
Even – Not first – 3 dice	**0.69**	1.25	0.82	0.86	1.68	1.33	0.92	0.76	0.71	1.06	1.04
Even – Not middle – 3 dice	0.80	0.54	0.83	0.58	1.12	0.52	0.75	0.53	0.56	0.69	**0.49**
Symmetric KL											
Smaller – Even – 2 dice	5.42	0.82	1.76	1.45	6.27	1.96	1.72	0.73	0.78	**0.61**	0.64
Smaller – Odd – 2 dice	0.27	0.15	1.91	0.31	2.27	0.61	0.59	0.78	**0.11**	0.32	0.36
Smaller – Not first – 2 dice	1.14	0.82	0.50	0.98	5.18	2.20	2.10	0.96	0.41	0.99	**0.38**
Smaller – Not middle – 2 dice	0.44	0.40	0.82	0.72	1.23	1.68	0.81	1.55	**0.18**	0.47	1.37
Larger – Even – 2 dice	5.04	**0.42**	1.36	0.95	4.15	2.04	1.75	1.02	1.62	0.73	1.02
Larger – Odd – 2 dice	2.26	**0.26**	1.11	1.46	1.56	2.51	1.81	1.42	1.68	1.42	1.17
Larger – Not first – 2 dice	0.68	**0.22**	1.56	0.84	1.53	4.45	3.31	1.60	2.66	1.14	1.09
Larger – Not middle – 2 dice	1.16	**0.32**	0.88	0.95	2.11	2.39	2.17	1.35	1.99	0.70	0.36
Even – Not first – 2 dice	0.91	1.23	1.76	3.25	5.17	1.72	5.94	**0.84**	1.13	2.62	0.97
Even – Not middle – 2 dice	2.18	**0.46**	2.46	0.76	1.64	0.71	1.43	0.74	1.18	1.06	0.86
Smaller – Even – 3 dice	4.64	0.73	4.18	1.27	4.91	0.85	1.27	1.05	0.90	**0.38**	0.55
Smaller – Odd – 3 dice	1.12	0.54	4.06	0.77	5.11	0.64	0.60	0.76	0.40	**0.34**	0.44
Smaller – Not first – 3 dice	1.61	0.72	0.79	0.59	4.81	1.95	1.29	1.56	**0.58**	0.98	0.79
Smaller – Not middle – 3 dice	0.62	**0.24**	0.80	0.56	1.10	2.43	0.95	1.76	0.48	0.87	1.06
Larger – Even – 3 dice	4.05	**0.73**	1.70	0.82	2.83	1.76	1.24	0.88	1.56	1.07	1.33
Larger – Odd – 3 dice	4.42	**1.03**	1.55	1.26	2.37	1.97	1.58	1.45	1.26	1.08	1.18
Larger – Not first – 3 dice	**0.51**	0.83	1.83	0.58	1.42	4.00	2.11	1.54	2.09	2.20	1.51
Larger – Not middle – 3 dice	0.86	**0.53**	1.55	0.66	2.23	2.87	1.25	1.07	1.80	1.54	0.79
Even – Not first – 3 dice	1.77	2.80	2.75	1.52	8.77	3.57	2.66	1.32	**1.18**	2.62	2.66
Even – Not middle – 3 dice	2.53	0.61	1.84	**0.58**	2.72	1.17	1.96	0.70	0.71	1.71	1.01

B.3 `Stated Answer` (instruction fine-tuned models)

B.3.1 Regular, independent, dependent

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 1 die	0.40	0.68	0.56	0.67	0.91	0.90	0.32	0.22	**0.00**	**0.00**	0.14
Regular – 2 dice	0.67	0.76	0.57	0.94	0.88	0.97	0.80	0.86	1.00	**0.51**	0.93
Regular – 3 dice	0.71	0.98	0.76	**0.59**	0.60	0.60	0.90	0.95	1.00	0.70	0.69
Independent 1 die	0.54	0.59	0.66	0.75	0.79	0.32	0.23	0.35	0.20	**0.13**	0.18
Independent – 2 dice	0.64	**0.55**	0.96	0.77	0.95	0.67	0.60	0.80	0.75	0.90	0.74
Independent – 3 dice	0.70	0.85	0.80	**0.62**	0.78	0.89	0.76	0.89	0.80	1.00	0.65
Dependant 1 die	0.80	0.76	0.71	0.78	0.77	0.89	0.77	0.76	**0.65**	0.83	0.70
Dependant – 2 dice	0.89	**0.66**	0.83	0.79	0.84	0.88	0.76	0.91	0.86	0.86	0.79
Dependant – 3 dice	0.73	0.84	0.74	**0.65**	0.66	0.88	0.69	0.76	0.85	0.98	0.78

B.3.2 Observations: One observation

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller	0.99	1.00	0.78	0.79	1.00	0.99	0.81	0.85	**0.29**	0.75	0.81
Larger	0.67	0.96	0.88	0.27	0.79	1.00	0.69	0.98	0.32	**0.07**	0.49
Even	0.84	0.64	0.91	0.70	0.71	1.00	0.60	0.97	**0.33**	0.91	0.45
Odd	0.94	0.67	0.84	0.48	0.72	0.96	0.44	1.00	**0.39**	0.67	0.68
Not first	0.02	1.00	0.76	0.99	0.02	**0.00**	**0.00**	0.01	**0.00**	0.04	0.01
Not middle	1.00	1.00	1.00	0.45	0.14	1.00	0.59	0.95	0.08	**0.00**	0.22
Smaller – 2 dice	0.99	0.84	**0.67**	0.91	1.00	1.00	0.97	0.77	0.94	1.00	0.94
Larger – 2 dice	0.91	1.00	0.92	0.85	0.71	0.93	**0.27**	0.99	0.98	1.00	0.97
Even – 2 dice	0.40	0.97	0.40	0.63	1.00	**0.36**	0.76	0.57	0.96	0.97	0.92
Odd – 2 dice	0.91	1.00	0.79	0.58	0.56	0.73	0.83	0.97	**0.54**	1.00	0.96
Not middle – 2 dice	0.37	0.81	0.92	0.83	1.00	0.69	0.43	0.74	0.68	0.70	**0.13**
Smaller – 3 dice	1.00	1.00	0.92	**0.67**	1.00	0.99	0.98	0.80	0.93	1.00	0.94
Larger – 3 dice	0.89	1.00	**0.52**	0.59	0.77	1.00	0.82	0.81	0.78	0.81	0.76
Even – 3 dice	0.90	0.81	**0.67**	0.71	0.91	0.97	0.70	0.73	0.83	0.98	0.69
Odd – 3 dice	0.65	0.93	**0.49**	0.82	0.97	0.99	0.73	0.91	0.91	0.77	0.90

B.3.3 Observations: Two observations – Single die

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even	0.76	1.00	0.99	0.78	1.00	1.00	0.83	0.93	**0.25**	0.76	0.98
Smaller – Odd	0.75	1.00	0.98	0.78	1.00	0.98	0.67	0.94	**0.00**	1.00	0.50
Smaller – Not first	0.95	1.00	0.69	0.81	1.00	0.86	0.56	0.97	**0.34**	1.00	0.76
Smaller – Not middle	1.00	1.00	0.96	0.74	1.00	0.78	0.61	0.99	**0.39**	0.51	0.88
Larger – Even	1.00	0.66	1.00	0.39	1.00	0.99	0.66	0.96	**0.33**	1.00	0.69
Larger – Odd	0.69	1.00	1.00	0.81	0.60	1.00	0.72	0.88	**0.25**	1.00	0.53
Larger – Not first	**0.26**	0.93	0.66	0.35	0.67	0.88	0.32	0.54	0.34	0.45	0.59
Larger – Not middle	0.99	**0.13**	0.67	0.54	0.58	0.92	0.83	0.64	0.34	0.44	0.54
Even – Not first	0.89	0.58	0.37	0.62	0.96	0.94	**0.24**	0.88	0.55	0.59	0.60
Even – Not middle	0.96	1.00	0.74	0.75	1.00	0.69	0.43	0.97	**0.28**	0.71	0.74

B.3.4 Observations: Two observations – Multiple dice

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Smaller – Even – 2 dice	0.88	1.00	0.92	0.92	1.00	1.00	1.00	0.87	**0.86**	1.00	0.94
Smaller – Odd – 2 dice	1.00	1.00	0.98	0.64	0.51	0.68	0.73	1.00	**0.18**	1.00	0.96
Smaller – Not first – 2 dice	1.00	1.00	0.92	0.87	**0.28**	1.00	0.99	0.96	1.00	1.00	0.83
Smaller – Not middle – 2 dice	0.97	0.50	0.95	0.58	0.99	0.57	0.82	0.54	**0.13**	0.99	0.93
Larger – Even – 2 dice	0.85	**0.44**	0.98	0.81	0.50	1.00	1.00	0.90	1.00	0.58	0.96
Larger – Odd – 2 dice	0.91	1.00	0.99	0.81	**0.70**	0.82	1.00	0.99	0.99	1.00	0.94
Larger – Not first – 2 dice	0.95	1.00	0.95	0.72	**0.57**	1.00	0.71	0.96	0.99	1.00	0.91
Larger – Not middle – 2 dice	0.96	0.93	0.92	0.76	0.77	0.92	1.00	0.87	**0.58**	0.68	0.91
Even – Not middle – 2 dice	0.48	0.99	**0.34**	0.55	0.60	0.98	0.73	0.80	0.75	1.00	0.63
Smaller – Not middle – 3 dice	1.00	1.00	0.98	**0.81**	1.00	0.99	1.00	0.97	0.99	1.00	0.99
Larger – Not first – 3 dice	0.86	0.75	0.72	**0.56**	0.74	0.91	0.67	0.82	1.00	0.81	0.84
Larger – Not middle – 3 dice	0.78	0.95	0.71	0.62	0.65	0.98	**0.50**	0.56	0.81	0.92	0.60
Even – Not first – 3 dice	0.64	0.99	**0.44**	0.82	0.96	0.87	0.60	0.62	0.76	1.00	0.90
Even – Not middle – 3 dice	0.64	0.88	0.46	0.71	0.97	0.89	**0.21**	0.89	0.69	0.89	0.95

C Scenario 2: Coins

C.1 `Revealed Belief` (base models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
2 coins Regular	0.22	0.13	0.24	0.17	0.23	0.43	0.12	**0.08**	0.19	0.16	0.14
2 coins Regular 3x Bias	0.15	0.08	**0.04**	0.13	0.05	0.11	0.39	0.09	0.24	0.18	0.13
2 coins Regular 5x Bias	**0.02**	0.24	0.22	0.25	0.11	0.22	0.50	0.24	0.41	0.33	0.25
2 coins Dependant	0.12	**0.07**	0.17	0.09	0.13	0.14	0.16	0.20	0.20	0.16	0.20
2 coins Dependant 3x Bias	0.22	0.24	0.33	0.31	**0.08**	0.36	0.22	0.24	0.30	0.23	0.38
2 coins Dependant 5x Bias	0.36	0.38	0.46	0.45	**0.27**	0.49	0.30	0.36	0.43	0.35	0.49
2 coins Independent	0.17	0.16	0.23	**0.13**	0.16	0.26	0.20	0.33	0.20	0.20	0.14
2 coins Independent 3x Bias	0.22	0.23	**0.11**	0.28	0.21	0.30	0.19	0.18	0.31	0.23	0.28
2 coins Independent 5x Bias	0.32	0.34	**0.19**	0.42	0.28	0.40	0.30	0.28	0.43	0.34	0.34
3 coins Regular	0.12	**0.02**	0.11	0.15	0.10	0.19	0.04	0.06	0.30	0.16	0.13
3 coins Regular 3x Bias	**0.03**	0.17	0.12	0.13	0.09	0.17	0.34	0.17	0.25	0.34	0.31
3 coins Regular 5x Bias	**0.16**	0.31	0.30	0.25	0.23	0.31	0.48	0.34	0.39	0.50	0.42
3 coins Dependant	0.19	0.15	0.25	0.18	**0.05**	0.22	0.20	0.19	0.14	0.12	0.19
3 coins Dependant 3x Bias	0.27	0.29	0.37	0.31	**0.22**	0.41	0.28	0.33	0.40	0.26	0.42
3 coins Dependant 5x Bias	0.39	0.43	0.48	0.45	**0.31**	0.50	0.38	0.43	0.50	0.42	0.52
3 coins Independent	0.19	0.12	0.18	**0.10**	**0.10**	0.14	0.14	0.28	0.19	0.16	0.22
3 coins Independent 3x Bias	0.29	0.28	**0.20**	0.31	0.23	0.34	0.27	0.33	0.33	0.28	0.38
3 coins Independent 5x Bias	0.43	0.41	**0.35**	0.45	**0.35**	0.49	0.43	0.46	0.46	0.41	0.50
L1											
2 coins Regular	0.45	0.27	0.48	0.34	0.45	0.86	0.24	**0.15**	0.37	0.31	0.28
2 coins Regular 3x Bias	0.31	0.15	**0.08**	0.25	0.09	0.21	0.79	0.17	0.49	0.37	0.25
2 coins Regular 5x Bias	**0.04**	0.48	0.45	0.50	0.22	0.45	1.00	0.48	0.83	0.65	0.49
2 coins Dependant	0.24	**0.14**	0.34	0.18	0.26	0.29	0.31	0.40	0.40	0.33	0.39
2 coins Dependant 3x Bias	0.45	0.49	0.67	0.62	**0.16**	0.73	0.43	0.47	0.61	0.45	0.76
2 coins Dependant 5x Bias	0.72	0.76	0.92	0.90	**0.54**	0.97	0.60	0.71	0.87	0.71	0.99
2 coins Independent	0.34	0.31	0.46	**0.25**	0.33	0.52	0.41	0.66	0.40	0.40	0.28
2 coins Independent 3x Bias	0.45	0.47	**0.23**	0.56	0.42	0.59	0.38	0.37	0.62	0.47	0.55
2 coins Independent 5x Bias	0.64	0.68	**0.38**	0.83	0.56	0.79	0.60	0.56	0.86	0.68	0.67
3 coins Regular	0.34	**0.05**	0.34	0.32	0.20	0.54	0.08	0.18	0.60	0.32	0.25
3 coins Regular 3x Bias	**0.07**	0.51	0.27	0.26	0.21	0.35	0.67	0.33	0.51	0.69	0.62
3 coins Regular 5x Bias	**0.32**	0.70	0.59	0.50	0.46	0.62	0.96	0.69	0.78	1.00	0.84
3 coins Dependant	0.39	0.38	0.55	0.44	**0.13**	0.46	0.39	0.38	0.29	0.29	0.39
3 coins Dependant 3x Bias	0.73	0.82	1.02	0.90	0.52	1.13	0.61	0.79	0.86	**0.51**	0.90
3 coins Dependant 5x Bias	0.91	0.99	1.16	1.06	**0.65**	1.25	0.78	0.94	1.04	0.84	1.09
3 coins Independent	0.41	0.25	0.44	0.26	**0.20**	0.28	0.29	0.58	0.39	0.33	0.45
3 coins Independent 3x Bias	0.65	0.66	**0.45**	0.71	0.50	0.74	0.56	0.69	0.73	0.62	0.80
3 coins Independent 5x Bias	0.93	0.86	**0.71**	0.97	0.72	0.97	0.85	0.97	0.97	0.87	1.02
Symmetric KL											
2 coins Regular	0.26	0.11	0.28	0.24	0.23	1.03	0.06	**0.04**	0.26	0.21	0.12
2 coins Regular 3x Bias	0.11	0.03	0.02	0.07	**0.01**	0.05	0.77	0.03	0.26	0.14	0.09
2 coins Regular 5x Bias	**0.00**	0.29	0.29	0.26	0.07	0.21	1.22	0.26	0.75	0.46	0.31
2 coins Dependant	0.10	**0.03**	0.22	0.05	0.11	0.18	0.20	0.26	0.29	0.22	0.26
2 coins Dependant 3x Bias	0.24	0.31	0.72	0.57	**0.03**	0.71	0.23	0.32	0.43	0.26	0.95
2 coins Dependant 5x Bias	0.60	0.75	1.26	1.09	**0.34**	1.21	0.44	0.71	0.85	0.57	1.35
2 coins Independent	0.18	0.14	0.36	**0.10**	0.19	0.35	0.27	0.65	0.19	0.22	0.12
2 coins Independent 3x Bias	0.27	0.33	**0.09**	0.39	0.23	0.65	0.20	0.28	0.56	0.33	0.54
2 coins Independent 5x Bias	0.60	0.71	**0.25**	0.86	0.46	1.02	0.48	0.55	1.07	0.68	0.78
3 coins Regular	0.14	**0.00**	0.16	0.16	0.06	0.46	0.01	0.04	0.41	0.16	0.07
3 coins Regular 3x Bias	**0.01**	0.41	0.13	0.12	0.08	0.20	0.87	0.19	0.35	0.85	0.73
3 coins Regular 5x Bias	**0.15**	0.96	0.63	0.42	0.39	0.63	1.56	0.68	0.77	1.67	1.27
3 coins Dependant	0.20	0.19	0.40	0.27	**0.03**	0.31	0.20	0.22	0.15	0.20	0.20
3 coins Dependant 3x Bias	0.73	0.99	1.46	1.23	**0.40**	1.81	0.58	0.88	1.28	0.47	1.31
3 coins Dependant 5x Bias	1.40	1.81	2.25	2.05	**0.82**	2.72	1.11	1.62	2.13	1.12	2.32
3 coins Independent	0.19	0.09	0.30	0.10	**0.07**	0.12	0.16	0.46	0.25	0.24	0.30
3 coins Independent 3x Bias	0.64	0.72	0.40	0.91	**0.38**	0.99	0.53	0.83	0.91	0.69	1.07
3 coins Independent 5x Bias	1.35	1.31	0.92	1.67	**0.87**	1.81	1.16	1.54	1.62	1.33	1.78

C.2 `Revealed Belief` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
2 coins Regular	0.19	0.15	0.19	0.39	0.36	0.20	0.20	0.22	0.28	0.25	**0.10**
2 coins Regular 3x Bias	0.10	**0.02**	0.08	0.27	0.39	0.06	0.23	0.25	0.12	0.31	0.09
2 coins Regular 5x Bias	**0.05**	0.14	0.23	0.38	0.25	0.08	0.26	0.33	0.20	0.42	0.13
2 coins Dependant	0.33	0.17	0.20	0.28	0.43	0.33	0.27	0.26	0.27	0.25	**0.14**
2 coins Dependant 3x Bias	0.47	0.24	0.42	0.50	0.43	0.42	0.43	0.34	0.44	**0.18**	0.21
2 coins Dependant 5x Bias	0.58	0.38	0.51	0.61	0.47	0.55	0.54	0.46	0.53	**0.29**	0.37
2 coins Independent	0.26	0.31	0.27	0.34	0.44	0.16	0.19	0.34	0.17	0.34	**0.12**
2 coins Independent 3x Bias	0.32	0.28	0.16	0.46	0.26	0.28	**0.15**	0.28	0.31	0.34	0.23
2 coins Independent 5x Bias	0.39	0.32	**0.23**	0.59	0.28	0.45	**0.23**	0.36	0.42	0.43	0.30
3 coins Regular	0.12	0.22	**0.11**	0.51	0.61	0.35	0.13	0.16	0.43	0.27	0.13
3 coins Regular 3x Bias	0.10	0.16	**0.08**	0.47	0.56	0.18	0.36	0.25	0.52	0.38	0.27
3 coins Regular 5x Bias	**0.19**	0.29	0.23	0.55	0.64	0.32	0.50	0.42	0.56	0.53	0.40
3 coins Dependant	0.40	0.20	0.32	0.37	0.42	0.30	0.32	0.26	0.30	0.16	**0.09**
3 coins Dependant 3x Bias	0.54	0.33	0.43	0.50	0.62	0.42	0.43	0.40	0.50	**0.25**	**0.25**
3 coins Dependant 5x Bias	0.63	0.45	0.54	0.60	0.69	0.53	0.50	0.53	0.59	**0.42**	0.43
3 coins Independent	0.27	0.21	0.22	0.24	0.51	0.16	**0.15**	0.34	0.18	0.21	0.16
3 coins Independent 3x Bias	0.37	0.26	**0.19**	0.43	0.49	0.26	0.29	0.44	0.35	0.32	0.32
3 coins Independent 5x Bias	0.47	0.39	**0.32**	0.56	0.60	0.43	0.43	0.57	0.51	0.45	0.44
L1											
2 coins Regular	0.37	0.31	0.39	0.78	0.72	0.39	0.40	0.45	0.56	0.51	**0.19**
2 coins Regular 3x Bias	0.20	**0.03**	0.16	0.53	0.78	0.11	0.46	0.51	0.24	0.61	0.19
2 coins Regular 5x Bias	**0.10**	0.28	0.45	0.75	0.51	0.15	0.52	0.66	0.40	0.84	0.26
2 coins Dependant	0.67	0.33	0.40	0.57	0.85	0.67	0.53	0.51	0.55	0.49	**0.28**
2 coins Dependant 3x Bias	0.94	0.49	0.84	0.99	0.85	0.83	0.85	0.69	0.88	**0.37**	0.42
2 coins Dependant 5x Bias	1.16	0.75	1.02	1.22	0.94	1.10	1.07	0.92	1.06	**0.59**	0.74
2 coins Independent	0.52	0.62	0.55	0.68	0.88	0.32	0.37	0.69	0.33	0.68	**0.24**
2 coins Independent 3x Bias	0.63	0.56	0.33	0.91	0.52	0.55	**0.31**	0.56	0.61	0.68	0.45
2 coins Independent 5x Bias	0.77	0.64	**0.46**	1.19	0.55	0.90	**0.46**	0.71	0.84	0.86	0.59
3 coins Regular	**0.23**	0.44	0.36	1.03	1.22	0.71	0.32	0.34	0.87	0.54	0.29
3 coins Regular 3x Bias	0.22	0.32	**0.17**	0.95	1.13	0.35	0.73	0.50	1.05	0.76	0.54
3 coins Regular 5x Bias	**0.38**	0.58	0.48	1.11	1.28	0.63	1.00	0.84	1.13	1.05	0.79
3 coins Dependant	0.83	0.48	0.66	0.77	0.88	0.68	0.76	0.52	0.66	0.37	**0.21**
3 coins Dependant 3x Bias	1.19	0.73	1.11	1.12	1.24	1.04	1.04	0.89	1.04	**0.50**	0.51
3 coins Dependant 5x Bias	1.34	0.98	1.25	1.29	1.38	1.23	1.17	1.14	1.21	**0.84**	0.86
3 coins Independent	0.58	0.46	0.51	0.48	1.01	0.36	**0.35**	0.68	0.38	0.45	0.36
3 coins Independent 3x Bias	0.81	0.55	**0.42**	0.91	0.99	0.63	0.58	0.89	0.70	0.68	0.69
3 coins Independent 5x Bias	0.99	0.79	**0.67**	1.15	1.20	0.94	0.86	1.14	1.02	0.91	0.91
Symmetric KL											
2 coins Regular	0.25	0.19	0.27	0.89	1.16	0.39	0.31	0.32	0.47	0.48	**0.06**
2 coins Regular 3x Bias	0.05	**0.01**	0.04	0.32	1.11	0.02	0.25	0.30	0.16	0.41	0.06
2 coins Regular 5x Bias	0.05	0.09	0.27	0.60	0.55	**0.03**	0.36	0.50	0.21	0.77	0.16
2 coins Dependant	0.84	0.21	0.38	0.84	2.04	0.74	0.56	0.48	0.69	0.55	**0.15**
2 coins Dependant 3x Bias	1.30	0.30	1.22	1.62	1.46	1.03	1.12	0.81	1.03	**0.25**	0.27
2 coins Dependant 5x Bias	1.85	0.66	1.78	2.33	1.71	1.74	1.60	1.30	1.44	**0.43**	0.65
2 coins Independent	0.39	0.66	0.49	0.68	1.52	0.14	0.21	0.82	0.22	0.77	**0.09**
2 coins Independent 3x Bias	0.57	0.49	0.19	1.12	0.67	0.41	**0.14**	0.61	0.53	0.86	0.43
2 coins Independent 5x Bias	0.93	0.72	0.37	1.90	0.52	1.01	**0.36**	0.90	0.94	1.30	0.70
3 coins Regular	**0.08**	0.22	0.23	1.83	3.48	0.63	0.22	0.16	0.94	0.45	0.09
3 coins Regular 3x Bias	0.08	0.16	**0.05**	1.83	3.03	0.17	0.92	0.34	2.06	1.10	0.56
3 coins Regular 5x Bias	**0.17**	0.49	0.43	2.24	3.38	0.49	1.41	0.95	2.31	1.87	1.12
3 coins Dependant	1.26	0.41	0.65	1.20	2.12	0.76	0.98	0.45	0.81	0.40	**0.12**
3 coins Dependant 3x Bias	2.10	0.82	1.84	2.37	3.30	1.49	1.83	1.40	1.98	**0.41**	0.48
3 coins Dependant 5x Bias	2.98	1.61	2.68	3.57	4.42	2.57	2.57	2.39	2.96	**1.04**	1.29
3 coins Independent	0.54	0.39	0.40	0.43	1.78	0.23	**0.16**	0.73	0.24	0.58	0.20
3 coins Independent 3x Bias	1.06	0.56	**0.37**	1.78	2.12	0.75	0.59	1.46	1.00	0.96	0.98
3 coins Independent 5x Bias	1.62	1.08	**0.88**	2.91	2.99	1.67	1.12	2.33	1.84	1.58	1.65

C.3 `Stated Answer` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
2 coins Regular	0.56	0.75	0.53	0.79	0.59	0.65	0.31	0.76	0.25	**0.00**	0.21
2 coins Regular 3x Bias	0.28	0.99	0.04	0.63	0.77	0.24	0.66	0.28	0.02	**0.00**	0.28
2 coins Dependant	0.67	0.85	0.73	0.63	0.63	0.70	**0.49**	0.86	**0.49**	0.52	0.61
2 coins Dependant 3x Bias	**0.26**	0.88	0.41	0.77	0.91	0.42	0.51	0.71	0.50	0.32	0.52
2 coins Independent	0.76	0.86	0.52	0.71	0.46	0.39	0.55	0.59	**0.26**	0.28	0.28
2 coins Independent 3x Bias	**0.40**	0.84	0.45	0.66	0.70	0.47	0.69	0.48	0.47	0.41	0.69
3 coins Regular	0.49	0.96	0.27	0.61	0.74	0.35	0.13	0.04	**0.00**	0.29	0.25
3 coins Dependant	0.64	0.66	0.75	0.60	0.67	0.36	0.67	0.47	**0.29**	0.34	0.54
3 coins Independent	0.67	0.78	0.42	0.68	0.86	0.54	0.55	0.29	0.42	**0.17**	0.34

D Scenario 3: Choice

D.1 `Revealed Belief` (base models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 2 choices	0.42	0.41	0.46	0.41	0.42	0.43	0.39	0.41	**0.32**	0.46	0.38
Regular – 4 choices	0.45	0.48	0.61	0.44	0.54	0.57	**0.32**	0.55	0.35	0.65	0.34
Regular – 6 choices	0.42	0.40	0.62	0.37	0.50	0.58	0.34	0.54	0.37	0.78	**0.30**
Independent – 2 choices	0.45	0.38	0.42	0.38	0.38	0.41	0.30	0.28	**0.22**	0.33	0.24
Independent – 4 choices	0.31	0.30	0.29	0.30	0.31	0.31	0.24	0.28	**0.21**	0.25	0.22
Independent – 6 choices	0.24	0.20	0.21	0.24	0.21	0.23	0.17	0.20	**0.14**	0.21	0.17
L1											
Regular – 2 choices	0.85	0.82	0.92	0.83	0.85	0.86	0.79	0.83	**0.64**	0.93	0.76
Regular – 4 choices	0.91	0.96	1.22	0.89	1.09	1.14	**0.64**	1.10	0.70	1.30	0.68
Regular – 6 choices	0.85	0.79	1.23	0.74	0.99	1.16	0.68	1.09	0.81	1.57	**0.66**
Independent – 2 choices	0.89	0.76	0.84	0.77	0.77	0.81	0.61	0.55	**0.44**	0.67	0.48
Independent – 4 choices	0.66	0.68	0.62	0.65	0.66	0.61	0.52	0.58	**0.44**	0.52	**0.44**
Independent – 6 choices	0.56	0.53	0.49	0.55	0.49	0.53	0.44	0.45	**0.36**	0.42	0.40
Symmetric KL											
Regular – 2 choices	1.06	0.95	1.43	0.99	1.06	1.14	0.84	0.99	**0.48**	1.50	0.76
Regular – 4 choices	0.97	1.20	1.82	0.94	1.34	1.50	**0.53**	1.46	0.76	2.28	0.59
Regular – 6 choices	1.05	0.85	1.82	0.70	1.15	1.62	0.75	1.44	0.89	3.61	**0.57**
Independent – 2 choices	1.28	0.80	1.10	0.84	0.86	0.95	0.43	0.35	**0.21**	0.64	0.27
Independent – 4 choices	0.84	0.79	0.75	0.70	0.87	0.88	0.50	0.62	**0.33**	0.49	0.37
Independent – 6 choices	0.66	0.52	0.57	0.56	0.55	0.78	0.37	0.42	**0.23**	0.48	0.38

D.2 `Revealed Belief` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 2 choices	**0.21**	0.41	0.49	0.46	0.50	0.47	0.34	0.44	0.32	0.49	0.35
Regular – 4 choices	**0.16**	0.43	0.68	0.51	0.72	0.55	0.42	0.54	0.29	0.72	0.32
Regular – 6 choices	**0.14**	0.36	0.69	0.46	0.72	0.42	0.18	0.53	0.30	0.82	0.29
Independent – 2 choices	0.42	0.32	0.46	0.44	0.50	0.38	0.21	0.38	0.25	0.38	**0.14**
Independent – 4 choices	0.34	0.35	0.31	0.34	0.44	0.31	0.28	0.31	**0.23**	0.27	**0.23**
Independent – 6 choices	0.25	0.25	0.21	0.22	0.50	0.23	0.22	0.21	**0.17**	0.22	0.18
L1											
Regular – 2 choices	**0.41**	0.82	0.97	0.92	1.00	0.95	0.67	0.88	0.64	0.99	0.70
Regular – 4 choices	**0.31**	0.86	1.36	1.01	1.45	1.09	0.84	1.07	0.66	1.45	0.65
Regular – 6 choices	**0.56**	0.72	1.38	0.92	1.44	0.88	**0.56**	1.06	0.76	1.63	0.59
Independent – 2 choices	0.85	0.63	0.92	0.88	1.00	0.77	0.42	0.76	0.51	0.76	**0.27**
Independent – 4 choices	0.78	0.72	0.67	0.81	0.98	0.70	0.58	0.64	**0.46**	0.53	0.47
Independent – 6 choices	0.68	0.59	0.49	0.62	1.09	0.58	0.51	0.46	0.43	0.46	**0.41**
Symmetric KL											
Regular – 2 choices	**0.18**	0.95	2.07	1.43	4.00	1.72	0.55	1.21	0.48	2.60	0.62
Regular – 4 choices	**0.19**	0.90	2.52	1.23	3.89	1.54	1.29	1.38	0.60	3.57	0.48
Regular – 6 choices	**0.49**	0.66	2.45	1.07	3.13	1.18	0.53	1.33	0.92	4.74	0.50
Independent – 2 choices	1.06	0.76	1.47	1.42	4.50	0.84	0.24	0.76	0.29	0.88	**0.08**
Independent – 4 choices	1.22	0.99	1.09	1.35	3.42	1.18	0.65	0.75	**0.37**	0.55	0.39
Independent – 6 choices	1.06	0.72	0.72	0.87	3.22	1.14	0.45	0.43	**0.36**	0.59	**0.36**

D.3 `Stated Answer` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Regular – 2 choices	0.41	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Regular – 4 choices	**0.00**	0.61	0.50	0.68	1.00	0.48	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Independent – 2 choices	0.34	0.33	0.48	0.24	0.51	**0.00**	**0.00**	0.03	0.03	**0.00**	0.01
Independent – 4 choices	0.40	0.75	0.57	0.66	0.75	0.13	0.30	0.28	0.06	**0.00**	0.04

E Scenario 4: Preference

E.1 `Revealed Belief` (base models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Heads	0.43	**0.19**	0.36	0.32	0.38	0.24	0.29	0.43	0.32	0.36	0.39
Heads 2x	0.25	**0.12**	0.25	0.24	0.21	0.24	0.28	0.15	0.21	0.20	0.24
Heads 3x	0.16	**0.04**	0.16	0.15	0.25	0.16	0.20	0.05	0.13	0.13	0.12
Left	0.32	0.39	0.49	0.35	0.35	0.42	**0.12**	0.47	0.37	0.48	0.39
Left 2x	0.15	0.26	0.32	0.13	0.21	0.28	**0.00**	0.23	0.17	0.26	0.09
Left 3x	0.05	0.16	0.23	0.05	**0.02**	0.19	0.06	0.15	0.07	0.18	**0.02**
Heads Independent	0.21	0.35	0.34	0.38	0.43	**0.18**	0.21	0.30	0.22	0.34	0.30
Heads Independent 2x	0.11	0.13	0.33	0.24	0.36	0.17	**0.09**	0.29	0.23	0.28	0.22
Heads Independent 3x	**0.11**	0.23	0.33	0.24	0.49	0.16	0.15	0.27	0.24	0.36	0.22
Left Independent	0.36	0.28	0.42	0.34	0.42	**0.24**	0.25	0.30	0.26	0.37	0.30
Left Independent 2x	0.19	0.14	0.37	0.24	0.31	**0.12**	0.27	0.28	0.23	0.30	0.25
Left Independent 3x	0.21	0.16	0.37	0.24	0.32	**0.15**	0.36	0.26	0.27	0.29	0.35
L1											
Heads	0.87	**0.39**	0.73	0.63	0.76	0.47	0.58	0.85	0.65	0.72	0.79
Heads 2x	0.50	**0.25**	0.50	0.48	0.43	0.48	0.57	0.30	0.43	0.41	0.48
Heads 3x	0.33	**0.08**	0.31	0.31	0.50	0.31	0.40	0.10	0.26	0.25	0.24
Left	0.64	0.77	0.98	0.70	0.70	0.85	**0.24**	0.95	0.73	0.95	0.79
Left 2x	0.30	0.52	0.63	0.26	0.43	0.56	**0.00**	0.45	0.34	0.51	0.18
Left 3x	0.10	0.32	0.46	0.10	**0.04**	0.38	0.11	0.31	0.14	0.36	**0.04**
Heads Independent	0.43	0.70	0.67	0.76	0.86	**0.36**	0.41	0.60	0.43	0.67	0.60
Heads Independent 2x	0.23	0.26	0.65	0.47	0.71	0.34	**0.18**	0.58	0.46	0.55	0.43
Heads Independent 3x	**0.21**	0.46	0.65	0.47	0.98	0.32	0.30	0.54	0.48	0.73	0.45
Left Independent	0.72	0.56	0.83	0.68	0.83	**0.48**	0.50	0.60	0.53	0.73	0.59
Left Independent 2x	0.38	0.27	0.75	0.49	0.61	**0.24**	0.53	0.57	0.46	0.60	0.50
Left Independent 3x	0.42	0.32	0.73	0.49	0.65	**0.30**	0.73	0.52	0.55	0.58	0.70
Symmetric KL											
Heads	1.14	**0.16**	0.67	0.47	0.76	0.24	0.38	1.08	0.50	0.65	0.83
Heads 2x	0.42	**0.08**	0.42	0.37	0.28	0.37	0.64	0.12	0.28	0.24	0.39
Heads 3x	0.21	**0.01**	0.18	0.18	0.27	0.18	0.37	**0.01**	0.12	0.11	0.10
Left	0.48	0.80	2.30	0.62	0.62	1.06	**0.06**	1.72	0.69	1.75	0.84
Left 2x	0.12	0.49	1.04	0.09	0.28	0.61	**0.00**	0.32	0.16	0.47	0.04
Left 3x	0.01	0.19	0.66	0.01	**0.00**	0.31	0.02	0.18	0.03	0.28	**0.00**
Heads Independent	0.29	0.62	0.59	0.84	1.34	**0.13**	0.18	0.41	0.21	0.66	0.41
Heads Independent 2x	0.10	0.14	0.64	0.27	0.79	0.14	**0.04**	0.43	0.31	0.61	0.27
Heads Independent 3x	**0.06**	0.31	0.86	0.33	2.00	0.15	0.12	0.47	0.45	0.90	0.40
Left Independent	0.69	0.36	1.04	0.63	1.06	**0.27**	0.31	0.46	0.32	0.69	0.41
Left Independent 2x	0.20	0.10	0.87	0.29	0.75	**0.07**	0.30	0.42	0.39	0.53	0.36
Left Independent 3x	0.25	0.21	1.01	0.37	0.98	**0.13**	0.59	0.37	0.57	0.68	0.61

E.2 `Revealed Belief` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Heads	0.47	**0.32**	0.38	0.38	0.50	0.45	0.45	0.40	0.34	0.33	0.40
Heads 2x	0.32	**0.04**	0.25	0.25	0.33	0.33	0.30	0.15	0.28	0.32	0.22
Heads 3x	0.24	**0.00**	0.15	0.17	0.25	0.25	0.14	0.09	0.19	0.24	0.14
Left	0.46	0.46	0.50	0.47	0.50	0.45	**0.09**	0.48	0.43	0.50	0.34
Left 2x	0.27	0.31	0.33	0.26	0.33	0.31	0.55	0.26	0.19	0.32	**0.06**
Left 3x	0.17	0.23	0.25	0.18	0.25	0.23	0.70	0.19	0.07	0.24	**0.02**
Heads Independent	0.48	0.42	0.40	0.44	0.50	**0.23**	0.35	0.38	0.36	0.42	0.24
Heads Independent 2x	0.24	0.27	0.40	0.30	0.44	0.21	0.40	0.32	0.23	0.33	**0.13**
Heads Independent 3x	0.29	0.25	0.40	0.30	0.49	0.19	0.48	0.33	0.23	0.35	**0.14**
Left Independent	0.49	0.36	0.43	0.43	0.50	0.29	0.27	0.35	0.41	0.42	**0.19**
Left Independent 2x	0.46	0.30	0.42	0.34	0.44	0.37	0.60	0.30	0.30	0.35	**0.17**
Left Independent 3x	0.46	0.31	0.43	0.35	0.44	0.46	0.68	0.31	0.29	0.36	**0.25**
L1											
Heads	0.93	**0.64**	0.75	0.77	1.00	0.91	0.91	0.81	0.67	0.66	0.80
Heads 2x	0.64	**0.08**	0.50	0.51	0.66	0.66	0.60	0.30	0.56	0.63	0.43
Heads 3x	0.47	**0.01**	0.31	0.33	0.50	0.49	0.29	0.17	0.38	0.47	0.29
Left	0.92	0.93	1.00	0.93	1.00	0.91	**0.19**	0.96	0.86	0.99	0.67
Left 2x	0.55	0.63	0.66	0.51	0.67	0.63	1.11	0.51	0.37	0.63	**0.13**
Left 3x	0.35	0.45	0.50	0.36	0.50	0.45	1.41	0.38	0.14	0.47	**0.04**
Heads Independent	0.96	0.83	0.79	0.87	1.00	**0.46**	0.70	0.76	0.72	0.83	0.48
Heads Independent 2x	0.49	0.53	0.81	0.60	0.88	0.43	0.80	0.65	0.46	0.65	**0.26**
Heads Independent 3x	0.59	0.51	0.80	0.60	0.98	0.38	0.97	0.67	0.46	0.70	**0.28**
Left Independent	0.97	0.73	0.87	0.86	1.00	0.58	0.53	0.70	0.82	0.83	**0.38**
Left Independent 2x	0.92	0.60	0.84	0.69	0.88	0.75	1.20	0.59	0.59	0.69	**0.33**
Left Independent 3x	0.92	0.63	0.86	0.71	0.88	0.92	1.35	0.61	0.58	0.72	**0.50**
Symmetric KL											
Heads	1.58	**0.48**	0.73	0.78	2.99	1.36	1.36	0.91	0.55	0.52	0.87
Heads 2x	1.18	**0.01**	0.42	0.44	2.10	1.50	0.81	0.12	0.61	1.05	0.29
Heads 3x	0.74	**0.00**	0.18	0.21	1.47	1.08	0.15	0.05	0.31	0.75	0.15
Left	1.43	1.50	3.05	1.58	5.50	1.36	**0.03**	1.93	1.14	2.66	0.55
Left 2x	0.56	1.00	1.76	0.47	3.44	1.00	1.53	0.47	0.20	1.09	**0.02**
Left 3x	0.24	0.58	1.21	0.28	2.48	0.60	2.88	0.31	0.03	0.74	**0.00**
Heads Independent	1.95	1.00	0.88	1.31	5.00	0.57	0.62	0.76	0.82	1.02	**0.27**
Heads Independent 2x	0.29	0.41	1.06	0.48	2.29	0.22	0.74	0.55	0.31	0.68	**0.09**
Heads Independent 3x	0.51	0.53	1.21	0.52	2.98	0.22	1.08	0.67	0.29	0.99	**0.14**
Left Independent	2.38	0.73	1.17	1.24	4.24	0.58	0.83	0.66	0.99	1.02	**0.17**
Left Independent 2x	1.57	0.48	1.23	0.64	1.95	0.95	2.26	0.49	0.56	0.64	**0.13**
Left Independent 3x	1.64	0.57	1.49	0.76	2.09	1.21	2.89	0.54	0.68	0.83	**0.29**

E.3 `Stated Answer` (instruction fine-tuned models)

Chebyshev	Y6B	M7B	Q7B	L8B	G7B	Y9B	Y34B	Q57xB	L70B	Q72B	M8x22B
Heads	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Heads 3x	0.01	0.57	0.48	0.38	**0.00**	0.42	0.48	0.50	**0.00**	**0.00**	0.02
Left	**0.00**	**0.00**	**0.00**	**0.00**	0.50	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Left 3x	0.50	1.00	**0.23**	0.68	0.95	0.39	0.50	0.50	0.45	0.50	0.45
Heads Independent	0.48	0.74	**0.00**	0.04	0.78	**0.00**	**0.00**	0.09	**0.00**	**0.00**	**0.00**
Heads Independent 3x	0.33	0.95	0.54	0.70	0.38	0.65	0.29	0.55	0.28	**0.24**	0.36
Left Independent	0.64	0.51	0.23	0.09	0.03	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.01
Left Independent 3x	0.38	1.00	0.77	0.80	0.98	**0.16**	0.60	0.54	0.65	0.28	0.52