Exploring the Potential for Large Language Models to Demonstrate Rational Probabilistic Beliefs

Gabriel Freedman, Francesca Toni
Department of Computing, Imperial College London, UK
{gif22, ft}@imperial.ac.uk

Abstract

Advances in the general capabilities of large language models (LLMs) have led to their use for information retrieval, and as components in automated decision systems. A faithful representation of probabilistic reasoning in these models may be essential to ensure trustworthy, explainable and effective performance in these tasks. Despite previous work suggesting that LLMs can perform complex reasoning and well-calibrated uncertainty quantification, we find that current versions of this class of model lack the ability to provide rational and coherent representations of probabilistic beliefs. To demonstrate this, we introduce a novel dataset of claims with indeterminate truth values and apply a number of well-established techniques for uncertainty quantification to measure the ability of LLM's to adhere to fundamental properties of probabilistic reasoning.

1 Introduction

In order for an agent to be an effective probabilistic reasoner, it must not violate the axioms of probability (Bas 2019). If an agent violates any of these axioms, it implies that it lacks the capacity to perform robust probabilistic reasoning, including uncertainty quantification. Three of the most fundamental properties of probabilistic reasoning, derived as corollaries from Kolmogorov's three axioms of probability theory (Kolmogorov 1963), are as follows. These properties clarify essential relationships between events, their complements, and their probabilities (for examples of each see Figure 1):

- **Complementarity**: Consider any event A within the sample space Ω, where Ω represents the set of all possible outcomes. Let A^c denote the complement of A, defined as the set of all outcomes in Ω that do not belong to A. The property of complementarity states that:

  $$P(A) + P(A^c) = 1.$$

  This result arises directly from the axioms of normalisation (the probability of the entire sample space is 1) and finite additivity (the probability of the union of mutually exclusive events is the sum of their probabilities).

- **Monotonicity (Specialisation)**: For two events A and A' such that $A' \subset A$ (i.e. A' is more specific than A), the probability of A' is less than or equal to the probability of A. Formally:

  $$P(A') < P(A).$$

  This reflects the principle that the probability of a more specific or refined event (a subset of another event) cannot exceed the probability of the broader event it is contained within.

- **Monotonicity (Generalisation)**: For two events A and A' such that $A \subset A'$ (i.e. A' is more general than A), the probability of A is less than or equal to the probability of A'. Formally:

  $$P(A) < P(A').$$

  This is the inverse of the specialisation relation.

Generative LLMs have demonstrated impressive performance in many reasoning tasks - including tasks which they have not been specifically trained for (Brown et al. 2020; Bubeck et al. 2023). This has led to the incorporation of LLMs into automated decision systems (ADSs) (Zhang et al. 2023; Ouyang and Li 2023; Wang et al. 2023). In order for ADSs to be trustworthy and contestable (Henin and Métayer 2021; Lyons, Velloso, and Miller 2021), they should be accompanied by a faithful representation of their reasoning. In the majority of real-world settings, in order for this to be an effective representation, it would need to include probabilistic uncertainty estimates.

Unlike some previous work, we are attempting to measure something distinct from 'subjective' uncertainty estimates (Geng et al. 2023), which aim to assess the uncertainty intrinsic to the model. This is measured by comparing the uncertainty estimate with the veracity of the model's output. Instead, we are interested in what may be called 'objective' uncertainty, as described in (Rayo 2019): *"The objective probability of A at time t is the subjective probability that a perfectly rational agent would assign to A, if she had perfect information about the world at times $\leq$ t and no information about the world at times $>$t."*. Thus, this is concerning the probability of a state of the world, regardless of the knowledge, or lack thereof, of the agent assigning the probability. This class of statements are at the core of academic disciplines and event forecasting.

Copyright © 2025 by the authors.
This open access article is published under the Creative Commons Attribution-NonCommercial 4.0 International License.

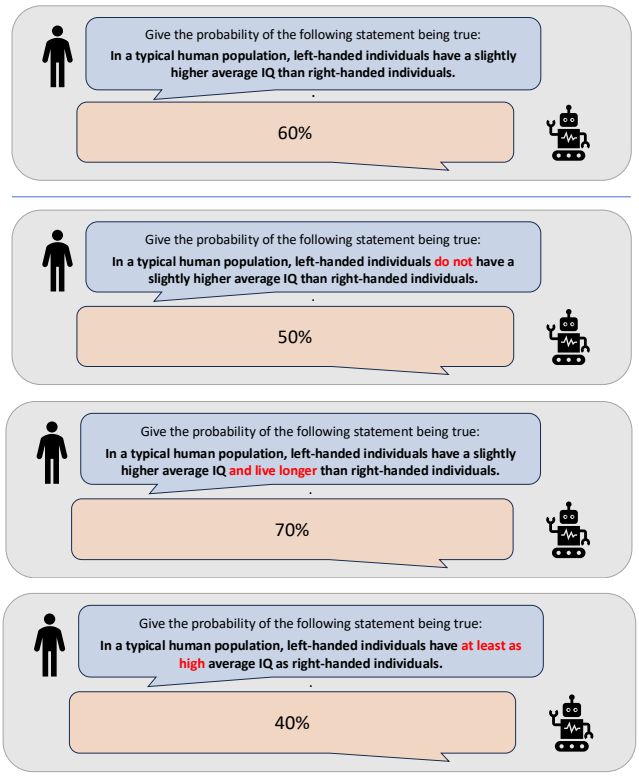

Figure 1: Examples of LLMs violating three principles of probabilistic reasoning. The image above the line shows the original claim with the corresponding probability assignment by an LLM, and below the line the images demonstrate the violation of the principles of: *complementarity* (top), *specialisation* (middle) and *generalisation* (bottom).

An example of a decision which requires 'objective' uncertainty estimations is determining the diets of the first inhabitants of America. The main sources of uncertainty in this inference come from imprecision in the measurement of carbon isotope levels in bone samples, as well as in auxiliary evidence about the climate (Booker and Ross 2011). Having a consistent and rational model of the uncertainties involved in this scenario allows for the effective integration of any new evidence or theories that come to light, and the possibility of overhauling existing conclusions.

In this paper we demonstrate that the current generation of state-of-the-art LLMs, both open and closed source, frequently violate the basic principles of probabilistic reasoning, even when applying techniques for enhancing the ability of LLMs to perform uncertainty quantification. This undermines a corpus of existing work that argues that LLMs exhibit sophisticated decision-making capabilities (Bubeck et al. 2023), including the ability to effectively quantify uncertainty (Lin, Trivedi, and Sun 2023; Tian et al. 2023; Hou et al. 2023).

In Section 2 we detail the construction of a novel dataset we use for evaluating LLMs' probabilistic beliefs. In Section 3 we describe our method for evaluating LLMs' adherence to the principles of complementarity and monotonicity and in Section 4 we outline our methodologies for measuring uncertainty. In Section 5 we report the results of our experiments using these methods. In Section 6 we briefly highlight relevant related work, and in the final Section we discuss the implications of our results and propose possible directions for future work. Code and data can be found at https://github.com/GIFRN/Rational-Probabilistic-Beliefs.

2 Novel Dataset

A contribution of this work is the construction of a novel, synthetically generated dataset of 517 claims with indeterminate truth values. Each claim is accompanied by its complement, as well as a more specific and more general version of the claim, amounting to 2068 samples in total. We name this the Rational Probabilistic Belief (RPB) dataset.

The initial claims were generated by prompting GPT-4o to generate claims that had an $n\%$ probability of being true, with n ranging between 10% and 90% in 10% intervals. To promote variety in the types of claims, we generated three variations of the basic prompt: asking the LLM to pose as an expert, and asking for historical claims and scientific claims. 200 claims were generated using each version of the prompt, resulting in 800 total.

In order to further ensure sufficient diversity of claims, we filtered out similar claims. We did this with simple string matching, removing any claim that overlapped with any other claim in the dataset by at least 90%. While this naive method does not capture semantically similar claims that are expressed differently, we note that the exact formulation of a statement sometimes has an impact on the uncertainty estimations of the LLMs, therefore making it desirable to have some samples where this is the case. We then manually remove any claim which is not making a valid statement that can be reasonably assigned a probability value. This leaves us with 517 base claims.

In order to generate the complementary, specialised and generalised versions of the claim, we also rely on GPT-4o. We use few-shot prompting with manually generated examples for each type of modification. We make sure the few-shot examples, especially for the specialised and generalised adaptions, are well chosen. This is because there is no set way to increase/decrease the specificity of a claim, and in order for it to create an appropriately challenging task, the newly generated claims must not be too distant (in a probabilistic sense) from the original.

For instance, if we consider the example in Figure 1, a logically valid specialisation of the original claim is: 'In a typical human population, left-handed individuals have a slightly higher average IQ and are more than twice as tall than right-handed individuals'. Formally, this is a subset of the original claim and is therefore a valid example of specialisation as we define it. However, the extra constraint of left-handed individuals being 'more than twice as tall' as right-handed individuals is so blatantly improbable that a system capable of any degree of moderately well-calibrated uncertainty quantification would assign a near-zero probability to it, thus undermining the informativeness of the task.

Therefore, to ensure the generated examples conform to these expectations, and fulfil the criteria for each of the respective properties, we again manually check each generated

sample. We adjust any samples that were found to violate either of these requirements.

3 Task Description

We carry out experiments to illustrate the degree to which LLMs adhere to the principles of complementarity and monotonicity. For the former, we compare the base claim with a 'specialised' claim which is a subset of the base claim, and a 'generalised' claim which is a superset. For the latter, we compare the models' uncertainty in a 'base' claim, with the uncertainty in the negated version of the claim (i.e. the complement of the claim). Following is a description of the tasks we use for evaluation.

3.1 Monotonicity

For measuring a model's ability to demonstrate probabilistic beliefs that adhere to the property of monotonicity, we compare the base claim with the specialised and generalised version of the claim (as described in Section 2) respectively. Therefore, to assess conformity to monotonicity, we measure whether the following inequalities hold:

P(original statement) $>$ *P(specialised statement).*

P(original statement) $<$ *P(generalised statement).*

We note that we define the inequalities as strict in both cases. While, theoretically, one can generate more/less specific versions of a claim that remain equally as likely, the practical instantiation of our dataset means that the 'difference' between the modified and original claim are always non-zero.

For instance, once again referring to the example in Figure 1, the difference between the original claim and the generalisation is something like: 'In a typical human population, left-handed individuals have the same IQ as right-handed individuals'. This statement clearly has a greater than zero probability, meaning that the conjunction of it with the original claim must result in an increased probability.

3.2 Complementarity

In order to evaluate adherence to complementarity, we assess the deviation of a model's evaluation of a claim's uncertainty, with that of the negated version of the same claim. Then, complementarity amounts to obtaining adherence to the following equation:

P(original statement) + *P(negated statement)* = 1.

A deviation from the value of 1 (i.e. either greater or less than) is considered a failure of the model to exhibit rational probabilistic beliefs, permitting some room for noise below a reasonable threshold.

4 Uncertainty Quantification

In order to quantify the models' performance on the tasks outlined in Section 3, we use three different methods to elicit probability estimates. We also experiment with a fourth for the larger, closed-source models only, due to technical constraints we discuss in more detail in the relevant section. Below, we describe each approach.

4.1 Direct Prompting

The first uncertainty quantification method we assess is a simple prompting based approach, wherein a prompt, including the target claim, is input into a LLM and the probability is taken directly from the generated output. Due to the high variation in performance depending on the formulation of the prompt (Zhuo et al. 2024), we form a validation set from 468 samples (117 of each variety of claim) in our dataset. Using this validation set we assess the performance of four distinct prompt variations.

We take a principled approach to candidate prompt selection. The first candidate is taken directly from existing literature (Freedman et al. 2024). The other three candidates were generated by GPT-o1 (using the web interface), having been provided a detailed task description. The best performing candidate was the one taken from the literature:

> **You are an analyst evaluating the validity of statements.**
> For the statement:
> **Statement:** "`{statement}`"
> Please give your confidence that the statement is correct, accurate, and truthful. Your response should be between **0% and 100%**, with **0%** indicating the statement is definitely invalid, **100%** indicating it is definitely valid, and values in between indicating varying levels of uncertainty. Your estimates should be well-calibrated, so feel free to err on the side of caution and output moderate probabilities if you are not completely sure in your assessment.
> **Only respond with your confidence score, no other text.**
> Please respond in the following form:
> **Likelihood:** The predicted likelihood that the considered statement is valid

When running the experiments the 'statement' variable is replaced by the target claim being evaluated. The probability estimate is extracted from the generated output using simple pattern matching. In order to negate the sampling bias introduced by the indeterminacy of LLM outputs, we generate 5 outputs for each sample, and take the mean of the scores as the final probability.

4.2 Chain-of-Thought

The chain-of-thought (Wei et al. 2022) approach we adopt is essentially a simple extension of the direct prompting method. While some instantiations of chain-of-thought use few-shot examples of how a problem can be broken down into discrete steps, this is more applicable to formal settings such as for mathematical problems. Instead, we simply append the sentence: 'Let's think step by step', to the end of the prompt, as introduced in Kojima et al. 2022.

Similarly to the direct prompting approach, we use the mean of 5 outputs as the final result.

4.3 Argumentative Large Language Models

Our third approach is taken from the literature on claim verification. Freedman et al. (2024) develop a method that uses outputs from LLMs to create formal arguments, called argumentative LLMs (ArgLLMs). Each argument is assigned

Model	Direct Prompting		Chain-of-Thought		Argumentative LLM	
	Spec.	Gen.	Spec.	Gen.	Spec.	Gen.
Llama-3	0.595	0.725	0.565	0.712	0.582	0.750
Llama-3.1	0.590	0.748	0.615	0.785	0.598	0.745
Mixtral	0.548	0.738	0.515	0.730	0.514	0.655
Gemma	0.452	0.515	0.462	0.545	0.500	0.507
Gemma-2	0.618	0.788	0.598	0.775	0.550	0.770
Gpt-4o-mini	0.627	0.771	0.610	0.767	**0.692**	**0.835**
Gpt-4o*	0.660	0.740	0.660	0.790	0.620	0.820
Average	**0.584**	0.718	0.576	**0.729**	0.582	0.726

Table 1: Accuracy for the **monotonicity** task across all evaluated models for two sets of 400 claim pairs (1200 samples total). Values represent performance on **specialisation** (Spec.) and **generalisation** (Gen.) tasks. Each column refers to different uncertainty quantification techniques described in Section 4. Best performances highlighted in bold.

*Reduced sample size evaluation (100 claim pairs per task).

Model	Direct Prompting			Chain-of-Thought			Argumentative LLM		
	>5%	>10%	>15%	>5%	>10%	>15%	>5%	>10%	>15%
Llama-3	0.965	0.942	0.905	0.960	0.932	0.892	0.858	0.720	0.615
Llama-3.1	0.850	0.765	0.675	0.908	0.822	0.740	0.840	0.735	0.652
Mixtral	0.870	0.775	0.700	0.892	0.788	0.728	0.890	0.775	0.719
Gemma	0.818	0.662	0.502	0.822	0.623	0.468	0.922	0.842	0.748
Gemma-2	0.922	0.832	0.788	0.922	0.845	0.775	0.850	0.720	0.613
Gpt-4o-mini	0.766	**0.529**	0.446	0.739	**0.529**	**0.441**	0.770	0.572	0.458
Gpt-4o*	0.820	0.590	0.460	0.800	0.560	0.460	**0.720**	0.610	0.510
Average	0.856	0.721	0.640	0.863	0.727	0.646	**0.833**	**0.711**	**0.599**

Table 2: Performance on the **complementarity** task across all evaluated models for two sets of 400 claim pairs (1200 samples total). Values represent the proportion of predictions deviating by more than 5%, 10%, and 15% from ground truth confidence scores (i.e. |1 - *P(original statement)* + *P(negated statement)*| * 100). Bold values indicate best performance per threshold column; note the smallest amount of deviation indicates the best performance.

*Reduced sample size evaluation (100 claim pairs).

a confidence score, also by LLMs, and then, using reasoning techniques from the Computational Argumentation literature, a final score is computed.

We follow a specific instantiation of ArgLLMs outlined in Freedman et al. (2024). This involves generating one argument supporting and one attacking the 'topic' claim, and then an additional layer of supporting and attacking arguments for both of these arguments. An 'argument strength' is also generated for each of the arguments, as well as the topic claim, meaning that six arguments and seven argument strengths are output for each claim.

For more specific details of the system, including the argument generation, argument strength attribution and final uncertainty calculation processes, please refer to Freedman et al. (2024). Due to the large number of generations required for for each sample, and our computational constraints, we are limited to producing one output per claim.

4.4 Top-K Logit Sampling

For our final method, we leverage the ability to obtain raw logit outputs from the models. These can be turned into a probability distribution over the entire token space by applying a softmax to the logits. In our case we limit this to the top 5 most likely next tokens. Since the tokens themselves are numerical (we use the same prompt as in Section 4.1), we are able to calculate a weighted average of the token value by the probability distribution.

Formally, let $\mathcal{V}$ denote the full vocabulary, comprising all possible tokens. For each token $i \in \mathcal{V}$, let z_i be the raw logit, and define the softmax probability:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{|\mathcal{V}|} \exp(z_j)}.$$

Let $S \subseteq \mathcal{V}$ be the set of k tokens with the highest probabilities p_i. Suppose each token in S corresponds to a numerical value t_i. Then the final numeric output is:

$$\hat{y} = \sum_{i \in S} p_i \, t_i.$$

Model	Specification	Generalisation	Complementarity		
			>5%	>10%	>15%
GPT-4o-mini	0.712	0.825	**0.729**	**0.549**	**0.441**
GPT-4o*	**0.760**	**0.830**	0.750	0.580	0.520

Table 3: Results for Top-K Logit Sampling method described in Section 4.4.

*Reduced sample size evaluation (100 claim pairs).

For instance an example output is:

$$t_1\text{: } 95 \quad p_1\text{: } 0.803$$
$$t_2\text{: } 90 \quad p_2\text{: } 0.179$$
$$t_3\text{: } 98 \quad p_3\text{: } 0.010$$
$$t_4\text{: } 85 \quad p_4\text{: } 0.007$$
$$t_5\text{: } 80 \quad p_5\text{: } 0.001$$

The weighted average of these outputs is then calculated as 94.

A shortcoming of this method is that for the smaller, open-source models we experiment with, we were not able to effectively apply this method. This is because they did not consistently conform to the formatting instructions requiring them to only return numerical outputs. Thus, the models did not return exclusively numerical token candidates, invalidating the weighted average calculation. We give a full breakdown of all models we experiment with in the next section.

5 Experiments

In this section we provide details about the experimental setup, as well as a discussion of the results.

5.1 Models

We use a number of open and closed-source models for our experiments, to ensure our findings are robust across different model sizes and architectures. We use seven main models: Mixtral (Mixtral-8x7B-Instruct-v0.1) (Jiang et al. 2024), Gemma (gemma-7b-it) (Mesnard et al. 2024), Gemma 2 (gemma-2-9b-it) (Riviere et al. 2024), Llama 3 (Meta-Llama-3-8B-Instruct) (Dubey et al. 2024), LLama 3.1 (Meta-Llama-3.1-8B-Instruct), GPT-4o mini (gpt-4o-mini-2024-07-18) (OpenAI 2024a) and GPT-4o (gpt-4o-2024-08-06) (OpenAI 2024b).

For the former six models we use the entire test set of 400 claims and their three variations each (1600 samples in total), and for GPT-4o, due to the cost associated with using closed-source models, we use a subset of 100 claims and their variations (400 samples in total). The model hyperparameters we use are the same across all models: temperature 0.7, top-p 0.95 and repetition penalty 1.0. In order to reduce the computational cost of running the open-source models we quantize all of them to 4-bits (Dettmers et al. 2023).

5.2 Results

The results for all the main experiments, excluding prompt selection, are reported in Tables 1 and 2. Additional results for the Top-K Logit Sampling method (Section 4.4), are reported for a reduced set of models in Table 3. Results are represented graphically in Figures 2 and 3.

Discussion Both the results reported in Table 1 and Table 2 suggest that the larger, closed-source models (specifically GPT-4o-mini and GPT-4o) perform better at both tasks. This appears to confirm the well-established relationship between model size and performance (Liang et al. 2022).

Despite the consistently stronger performance of the larger models, neither GPT-4o-mini nor GPT-4o showed robust adherence to all of the tested properties. We still observe that both can produce contradictory or incoherent outputs in terms of basic probability axioms. In particular, the violation rates for complementarity, while lower than those of other models, are still non-trivial. This suggests that even state-of-the-art LLMs, although generally more capable, may not maintain fully consistent probabilistic representations across diverse claim types and domains.

Examining the different uncertainty quantification techniques reveals additional insights. Chain-of-thought prompting offers some improvement over direct prompting, especially for smaller models, as it encourages more explicit reasoning steps that may guide the model to produce more stable probability assessments. However, for larger models, we find only marginal gains with chain-of-thought.

The ArgLLM technique achieves similar or slightly better performance than chain-of-thought in some scenarios, particularly for closed-source systems. By introducing additional structure via conflicting viewpoints and supporting/attacking sub-arguments, this approach may force the model to reconcile its own inconsistencies to some degree. However, it remains computationally expensive, as it entails generating multiple layers of arguments and confidence assignments for every claim.

The Top-K Logit Sampling method, examined only for GPT-4o-mini and GPT-4o in our experiments, showed modest improvements in monotonicity for GPT-4o, but not for GPT-4o-mini. While extracting probabilities at the token level mitigates some of the uncertainty introduced by sampling, it is limited by whether the model consistently follows the numerical response format. Our attempts to replicate this approach with the open-source models were largely unsuccessful, demonstrating the fragility of logit-based methods when generation style cannot be strictly controlled.

Moreover, when we analyse Figure 2, we see that despite having a smaller proportion of monotonically unsound probability estimations, when the larger models are incorrect

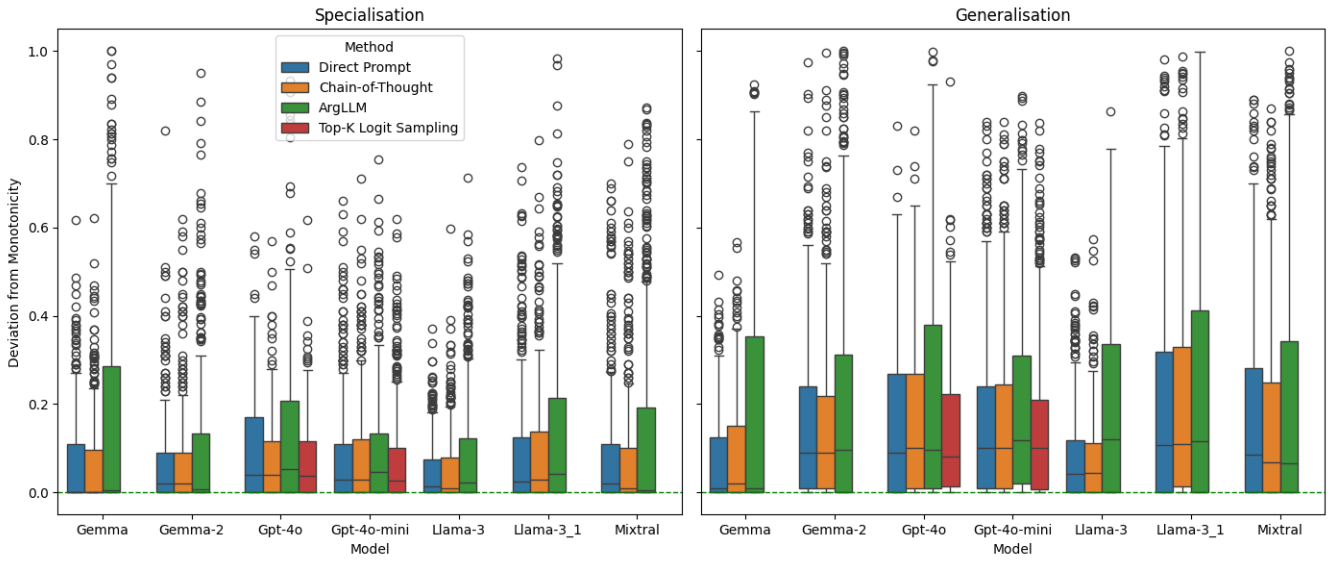

Figure 2: Adherence to **monotonicity** by model and uncertainty quantification methodology. The left panel is the Specialisation task, and the right Generalisation. Both tasks are described in detail in Section 3.1. The y-axis represents the magnitude of deviation from correctly monotonic probability estimations, and the x-axis is the model type.

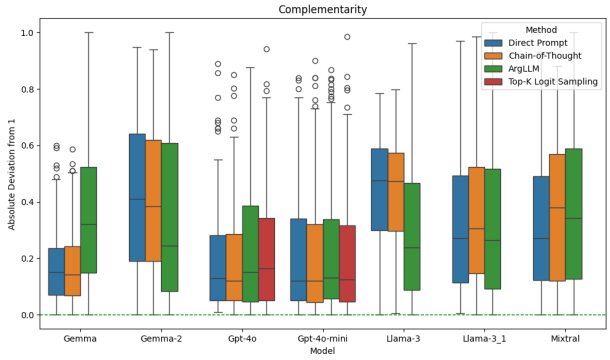

Figure 3: Adherence to **complementarity** by model and uncertainty quantification methodology. Detailed task description is provided in Section 3.2.

they are so by a greater magnitude. This can seen clearly, for example, by comparing the mean deviation on the generalisation task for GPT-4o and GPT-4o-mini, with that of Gemma, which is lower across all methods. Interestingly, Figure 3, while demonstrating the generally superior performance of the larger models, also appears to show Gemma as performing almost as well as the larger models, when using the direct prompting and chain-of-thought methods.

6 Related Work

Our findings build on a body of work demonstrating weaknesses in the ability of LLMs to adhere to basic logical principles. The reversal curse (Berglund et al. 2023) shows that in cases where a model has learnt "A is B", it has often not learnt "B is A". We demonstrate a similar effect but with probabilistic beliefs.

Similarly, Fluri, Paleka, and Tramèr (2023) demonstrate that LLMs succumb to various logical inconsistencies. They also show that LLMs can exhibit non-monotonic estimates, but restrict their analysis to the task of forecasting.

Wong et al. (2023) propose a method for integrating LLMs with a probabilistic logic engine. They also argue the necessity for LLMs to be able to reason probabilistically, and suggest that combining them with a symbolic module is the best way to achieve this.

Kuhn, Gal, and Farquhar (2022) also note the insufficiency of using direct prompting-based methods to ascertain the uncertainty of LLM outputs, instead developing a sampling-based method. However, this technique addresses subjective model uncertainty, and it is not clear whether there is an effective method to adapt it for representing objective uncertainty.

7 Conclusion and Future Work

In this paper we demonstrate that state-of-the-art LLMs fail at basic probabilistic reasoning. We observe that larger models demonstrate improved performance relative to their smaller counterparts. Nevertheless, the current scale of their failure is too significant to assume that further scaling will completely eradicate the problem.

A neurosymbolic approach, such as the one presented in Wong et al. (2023), bypasses the need for LLMs to reason probabilistically. Instead, they may rely on symbolic modules to handle any probabilistic inferences they may need to make. It is possible that a similar approach, which uses an appropriate symbolic knowledge representation, can provide an effective and robust solution to the problems we highlight in this paper.

Acknowledgments

This research was partially supported by ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme, and by UKRI through the CDT in Safe and Trusted Artificial Intelligence (Grant No. EP/S023356/1).

References

Bas, E. 2019. Basic concepts, axioms and operations in probability. *Basics of Probability and Stochastic Processes*.

Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Cooper Stickland, A.; Korbak, T.; and Evans, O. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv e-prints* arXiv–2309.

Booker, J. M., and Ross, T. J. 2011. An evolution of uncertainty assessment and quantification. *Sci. Iran.* 18:669–676.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. *Neural Information Processing Systems* abs/2005.14165.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J. A.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.-F.; Lundberg, S. M.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv* abs/2303.12712.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *CoRR* abs/2407.21783.

Fluri, L.; Paleka, D.; and Tramèr, F. 2023. Evaluating superhuman models with consistency checks. *arXiv e-prints* arXiv–2306.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2024. Argumentative large language models for explainable and contestable decision-making.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2023. A survey of language model confidence estimation and calibration. *ArXiv* abs/2311.08298.

Henin, C., and Métayer, D. L. 2021. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37:1397 – 1410.

Hou, B.; Liu, Y.; Qian, K.; Andreas, J.; Chang, S.; and Zhang, Y. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *ArXiv* abs/2311.08718.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc.

Kolmogorov, A. N. 1963. The theory of probability.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2022. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *NeurIPS ML Safety Workshop*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *ArXiv* abs/2305.19187.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction* 5:1 – 25.

Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR* abs/2403.08295.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini/`.

OpenAI. 2024b. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Ouyang, S., and Li, L. 2023. Autoplan: Automatic planning of interactive decision-making tasks with large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Rayo, A. 2019. Probability, subjective and objective. In *24-118-paradox-and-infinity-spring-2019*. MIT OpenCourseWare.

Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *CoRR* abs/2408.00118.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Wang, Z.; Liu, Z.; Zhang, Y.; Zhong, A.; Fan, L.; Wu, L.; and Wen, Q. 2023. Rcagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. *ArXiv* abs/2310.16340.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-

thought prompting elicits reasoning in large language models. In *NeurIPS*.

Wong, L. S.; Grand, G.; Lew, A. K.; Goodman, N. D.; Mansinghka, V. K.; Andreas, J.; and Tenenbaum, J. B. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv* abs/2306.12672.

Zhang, H.; Li, J.; Wang, Y.; and Songi, Y. 2023. Integrating automated knowledge extraction with large language models for explainable medical decision-making. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1710–1717.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1950–1976.