Large language models are not probabilistic reasoners

Anonymous ACL submission

Abstract

Advances in the general capabilities of large language models (LLMs) have led to the possibility of incorporating them into automated decision systems. A faithful representation of probabilistic reasoning in these models can be essential to ensure the reasoning of the automated decision systems incorporating them is trustworthy and explainable. 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 consistent and coherent probability estimates. We then suggest possible directions that future research can take to alleviate this weakness.

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. Two of the most fundamental properties of probabilistic reasoning (both corollaries of Kolmogorov's original three axioms (Kolmogorov, 1963)) are:

- *Consistency*. A probability measure P on a sample space Ω assigns to every event A a unique probability $P(A)$, where $0 \leq P(A) \leq 1$.
- *Complementarity*. For any event A, with complement A^c, $P(A) + P(A^c) = 1$.

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 et al., 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 attempt to assess the uncertainty intrinsic to the model. It 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: *"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."* (Rayo, 2019). 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.

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 LLMs, including the GPT-3.5 and GPT-4 (Achiam et al., 2023) family of models, frequently violate the basic principles of probabilistic reasoning. 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 et al., 2023; Tian et al., 2023; Hou et al., 2023). In Section 2 we describe our method for evaluating LLMs' adherence to the principles of consistency and complementarity. In Section 3 we report the results of our experiments using this method. In Section 4 we briefly highlight relevant related work, and in the final Section we discuss the implications of our results and propose a number of possible directions for future work.

2 Method

We carry out experiments to illustrate the degree to which LLMs adhere to the principles of consistency and

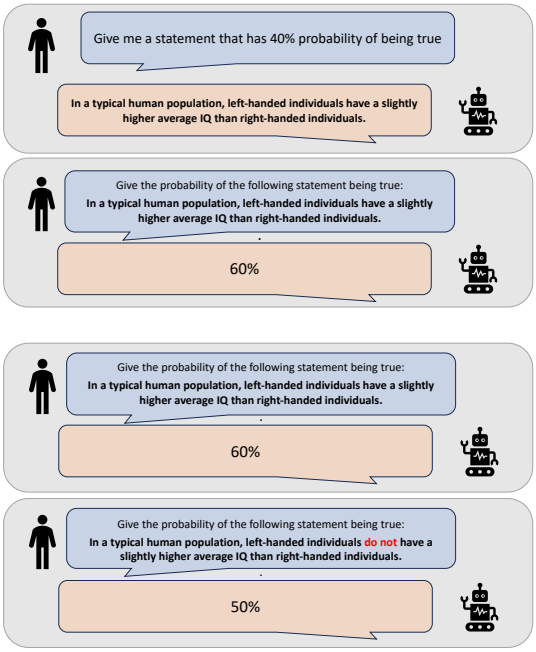

Figure 1: Examples of LLMs violating the principles of *consistency* (top) and *complementarity* (bottom).

complementarity. For the former we use two separate methods to elicit the models' degree of confidence about particular statements. We then measure the discrepancy between the values produced by the different methods. For the latter, we compare the models' degree of confidence in a statement, with the degree of confidence in the negated statement (the complement of the statement). We provide details below.

2.1 Consistency

We develop a method for recursive uncertainty estimation. First, we prompt the model to produce a statement which has a particular likelihood of being true. We use a prompt in the following format (referred to as the *statement generation* prompt):

```
Please provide a statement that would be assigned X% probability of being true by experts with access to all the necessary evidence. Do not make it a prediction about the future. Do not provide any justification. Do not mention the probability of the statement. Only respond with the statement and nothing else.
```

We then feed the statements back into the same model (in a separate context window), and prompt it to produce an uncertainty estimate (referred to as the *probability assignment* prompt):

```
Please state the percentage probability of being true that experts with access to all the necessary evidence would assign to the following statement (only return a single numerical percentage and no other explanation or commentary): <statement>.
```

In order to demonstrate inconsistency, we measure the discrepancy between the percentage used as input in the first (statement generation) prompt, and the percentage output by the model after the second (probability assignment) prompt.

In order to reduce the bias introduced by the particular format of the pair of prompts, we carry out the experiments with three additional adapted (pairs of) prompts. Two specify that the statement must be 'historical' and 'scientific' respectively, while the third replaces

```
would be assigned X% probability of being true by experts with access to all the necessary evidence
```

with

```
should be assigned X% probability of being true
```

We refer to the original pair of prompts as *expert*, to the second and third as *history* and *science*, respectively, and to the last one as *base*.

2.2 Complementarity

In order to evaluate adherence to complementarity, we first negate the statement produced by the statement generation prompt. We then use the same probability assignment prompt as previously to elicit a confidence score for the negated statement. Then, complementarity amounts to obtaining:

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

3 Experiments

3.1 Consistency

We evaluate four LLMs (GPT-4-0613, GPT-4-preview-0125, GPT-3.5-turbo-0125 and Mixtral-8x7B-Instruct-v01 (Jiang et al., 2024)).[1] For each LLM, we use values of X in the statement generation prompt ranging from 0 to 100 in increments of 10. For each one of the four prompt types and for each LLM, we produce 20 samples at each percentage value, resulting in 220 samples.Thus, overall, we produce 880 samples per LLM.

The results are overviewed in Table 1. Here, the 'Average difference' measures the cumulative magnitude of the difference between the probability specified in the statement generation prompt, and the probability output by the model in response to the probability assignment prompt. Also, the 'Proportion difference $\geq 15\%$' measures the frequency of significantly diverging samples, i.e. when the input and output percentages differ by more than 15% (in absolute terms). This is an important metric as the cases when a model's probabilistic estimates are dramatically inconsistent pose a greater threat in terms of possible downstream impact.

Note that we run the experiments three times, using temperatures of 1, 0.5 and 0, respectively, for the statement generation prompt. We choose the first two values

[1]For all experiments, for running inferences on Mixtral we used a Tesla A100 for a total of 20 GPU hours. For GPT models we spent a total of $50 in API credits.

Model	Prompts	Average difference	Proportion difference $\geq 15\%$
GPT-4	expert	18.6% / 20.2% / 18.5%	0.450 / 0.556 / 0.500
	history	22.4% / 20.1% / 23.1%	0.609 / 0.543 / 0.636
	science	18.9% / 18.6% / 13.1%	0.519 / 0.532 / 0.375
	base	19.5% / 23.3% / 13.9%	0.518 / 0.561 / 0.444
	Average	20.05%	**0.534**
GPT-4-turbo	expert	19.5% / 19.9% / 17.6%	0.554 / 0.582 / 0.636
	history	18.2% / 17.7% / 9.3%	0.538 / 0.573 / 0.286
	science	17.4% / 19.5% / 11.7%	0.522 / 0.619 / 0.300
	base	19.9% / 18.9% / 11.7%	0.537 / 0.459 / 0.300
	Average	**18.57%**	0.540
GPT-3.5-turbo	expert	22.7% / 23.6% / 7.0%	0.545 / 0.543 / 0.200
	history	26.1% / 21.7% / 18.5%	0.593 / 0.492 / 0.429
	science	38.1% / 39.0% / 27.5%	0.704 / 0.619 / 0.500
	base	22.8% / 18.1% / 22.8%	0.582 / 0.389 / 0.429
	Average	26.15%	0.550
Mixtral	expert	25.0% / 21.3% / 24.7%	0.532 / 0.554 / 0.714
	history	23.8% / 22.7% / 12.0%	0.596 / 0.622 / 0.4
	science	25.2% / 26.7% / 24.7%	0.596 / 0.611 / 0.714
	base	29.8% / 27.0% / 16.7%	0.630 / 0.591 / 0.667
	Average	24.91%	0.593

Table 1: The average difference and proportion of differences greater than 15% between the percentage values input in the statement generation prompt and output after the probability assignment prompt in each pair, as we describe in Section 2.1. The results are given when setting the temperature of the models for generating the statements to 1 / 0.5 / 0 (in the first two cases with 20 samples and in the latter case with a single sample, per percentage value). The reported averages are weighted by number of samples for each temperature and are taken across all four prompt types and three temperatures.

to ensure that there is enough diversity in the outputs. In the case where the temperature is set to 0, we only produce a single sample for each percentage value as the output does not differ.

Furthermore, when the same statement is produced multiple times at a different probability (e.g. 'Humans will discover evidence of microbial life on Mars' is produced twice as a statement with 50% and 60% probability, respectively) we assign to the statement a range of 'ground-truth' probabilities (e.g. [50%,60%]), and calculate the proportion difference by measuring distances (e.g. the difference between [50%,60%] and i) 55% is 0, ii) 40% is 10, and iii) 80% is 20).

For the probability assignment prompt (second in each pair), we run the experiments in Table 1 using a temperature of 0, so that the output is as deterministic, and thus reproducible, as possible.

We also run two further experiments where we increase the temperature for assigning probabilities (second prompt). This allows us to sample from the outputs. We set the temperature to 0.5 and take 10 samples of the probability for each statement. For this experiment we use the best performing model (on average difference in the consistency experiment), GPT-4-turbo, with the 'expert' prompt. We run the experiment with the samples produced at temperature 1 and 0.5. For the temperature 1 samples, we get an average difference of 18.3%, and a proportion of samples with difference $\geq 15\%$ of 0.547. This is compared to 19.5% and 0.554 when using one probability produced with temperature set to 0. Likewise, for the temperature 0.5 samples, we observe a marginal improvement of 17.7% and 0.532, compared to 19.9% and 0.582. This suggests our approach of setting temperature to 0 offers a reliable approximation of a more thorough sampling-based approach.

Discussion The results in Table 1 demonstrate the systemic inconsistency in LLM outputs. GPT-4 and GPT-4-turbo perform the best on both metrics (average difference and proportion difference $\geq 15\%$), with the latter exhibiting an average difference of around 18.5% between confidence extraction methods. The smaller models both average roughly 25%. This confirms the well-established relationship between model size and performance (Liang et al., 2022).

Figure 2 gives a breakdown of the inconsistencies by input percentage value (`X` in the statement generation prompt). We observe that all models perform almost perfectly at 0% input probabilities and all models apart from Mixtral perform almost perfectly at 100% input probabilities. This suggests that the LLMs we evaluate have learnt robust 'concepts' of necessity and impossibility. This is an important aspect of reasoning and may offer an explanation for the impressive performance these models have demonstrated on binary reasoning

tasks (Bubeck et al., 2023; Liang et al., 2022). However, as demonstrated, this does not necessarily generalise to the ability to reason with other input probabilities, which is necessary for more complicated reasoning tasks.

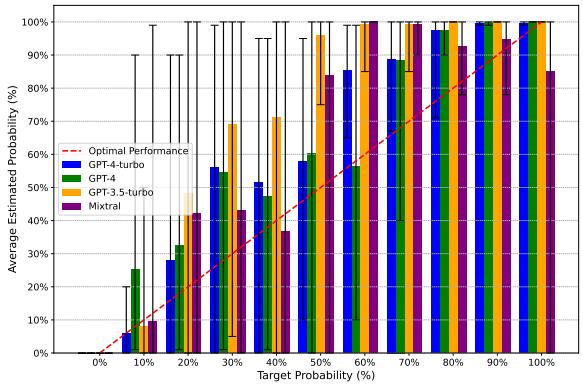

Figure 2: Breakdown of model inconsistency by input percentage value (x-axis gives the input probability value; y-axis gives the output probability value).

3.2 Complementarity

To measure adherence to complementarity we measure the deviation of *P(original statement)* + *P(negated statement)* from the expected value of 1. We evaluate the adherence of the two models which demonstrate the smallest proportion of significant inconsistencies and lowest average inconsistency: GPT-4 (GPT-4-0613) and GPT-4-turbo (GPT-4-preview-0125). To do so, we use a subset of the samples produced in the consistency experiment in order to limit cost (this subset amounts to under 200 samples obtained using the initial ('expert') variant of the prompts, after removing duplicates).

We automate the negation process by using GPT-4-turbo. However, we verify that it has correctly negated the statements before performing the evaluation. The results are overviewed in Table 2.

Discrepancy	**Proportion of Samples**	
	GPT-4-turbo	GPT-4
$\delta \leq 1\%$	0.30	0.54
$5\% \geq \delta > 1\%$	0.23	0.14
$10\% \geq \delta > 5\%$	0.1	0.08
$\delta > 10\%$	0.37	0.23

	GPT-4-turbo	GPT-4
Mean Discrepancy	17.71%	9.21%
Standard Deviation	26.58%	18.147%

Table 2: Summary of discrepancies between probabilities assigned to statements and their negations. Discrepancy: $\delta = |100\% - (P(\text{original statement}) + P(\text{negated statement}))|$

Discussion The results indicate that for GPT-4 over a fifth and for GPT-4-turbo over a third of the pairs of (original and negated) statements deviate by more than 10% from the rational value of 100%. Furthermore, the mean deviation is approximately 9% and 17%. This indicates these models have a severely limited capacity for probabilistically modelling negation, which is a relatively simple concept. Interestingly, this is a reversal of the respective performances of GPT-4 and GPT-4-turbo we observe in the consistency experiments.

4 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 probablistic beliefs.

Similarly, Fluri et al. (2023) demonstrate that LLMs succumb to various logical inconsistencies. They also show that LLMs are non-monotonic when forecasting, which is an additional violation of probabilistic reasoning to the ones we have demonstrated.

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 et al. (2022) also note the insufficiency of using direct prompting-based methods to ascertain the uncertainty of LLM outputs. They devise a sampling-based method which uses the relative frequency of semantic clusters as a way to measure model uncertainty. This technique addresses subjective model uncertainty, and it is not clear whether there is an effective method to adapt it for representing objective uncertainty.

5 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 extent of their failure is too significant to extrapolate that further scaling will eradicate the problem.

Evidence that neural models can assign effective quality scores to code (Deepmind, 2023) might provide a blueprint for how a similar approach can be developed for probability attribution. However, the utility of this method does not guarantee that it will still violate the principles of probabilistic reasoning.

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

Limitations

In our experiments, we make extensive use of LLMs. This hurts the reproducibility of some of the experiments we have run, as the outputs of these models are non-deterministic. However, we have included the raw outputs of all the experiments we carried out, and provided details of the important hyperparamater settings in the body of the paper.

We made sure to include one open source model (Mixtral), but unfortunately, at this time, the best performing models are closed-source. This means that once again, reproducibility of our experiments is harmed, as there is an associated monetary cost with doing so. However, as with the previous consideration, we have tried to remedy this by including all the raw outputs of our experiments with our submission.

While we did attempt to vary the models we evaluate, we ended up using a single vendor for the majority of the experiments. Ideally we would have been able to evaluate a far greater number of models, and use a greater number of samples per model. However, monetary and computational constraints limited us in this respect.

Ethical Considerations

The use of LLMs has an associated cost, either financial or in access to compute, as well as an environmental cost. This adds an extra barrier to use and research, compared with other domains in computer science. Furthermore, closed-source models place even greater restriction on use.

Our research is examining fundamental capacities of these models. The reality of the current landscape is that closed-source models are currently the best performing in this class of model. Therefore, any thorough analysis of their limitations is contingent on conducting experiments with the closed-source versions.

For running inferences on Mixtral we used a Tesla A100 for a total of 20 GPU hours. For GPT models we spent a total of $50 in API credits.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wi-

ethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

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

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv e-prints*, pages arXiv–2309.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Neural Information Processing Systems*, abs/2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Google Deepmind. 2023. Alphacode 2 technical report.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. 2023. Evaluating superhuman models with consistency checks. *arXiv e-prints*, pages arXiv–2306.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *ArXiv*, abs/2311.08298.

Clément Henin and Daniel Le Métayer. 2021. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, 37:1397 – 1410.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *ArXiv*, abs/2311.08718.

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*.

Andrei Nikolaevich Kolmogorov. 1963. The theory of probability.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *NeurIPS ML Safety Workshop*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *ArXiv*, abs/2305.19187.

Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction*, 5:1 – 25.

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

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

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Zefan Wang, Zichuan Liu, Yingying Zhang, Aoxiao Zhong, Lunting Fan, Lingfei Wu, and Qingsong Wen. 2023. Rcagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. *ArXiv*, abs/2310.16340.

Li Siang Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv*, abs/2306.12672.

Haodi Zhang, Jiahong Li, Yichi Wang, and Yuanfeng Songi. 2023. Integrating automated knowledge extraction with large language models for explainable medical decision-making. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717.