
MULTIPLE CHOICE QUESTIONS: REASONING MAKES LARGE LANGUAGE MODELS (LLMs) MORE SELF-CONFIDENT EVEN WHEN THEY ARE WRONG

Tairan Fu

College of Mechanical and Electrical Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, China
ftr258@nuaa.edu.cn

Javier Conde, María Grandury, Pedro Reviriego

ETSI de Telecomunicación
Universidad Politécnica de Madrid
Madrid, Spain
{javier.conde.diaz, pedro.reviriego}@upm.es

Gonzalo Martínez

Universidad Carlos III de Madrid
Madrid, Spain
gonzmart@pa.uc3m.es

María Grandury

SomosNLP/Universidad Politécnica de Madrid
Madrid, Spain
mariagrandury@gmail.com

ABSTRACT

One of the most widely used methods to evaluate LLMs are Multiple Choice Question (MCQ) tests. MCQ benchmarks enable the testing of LLM knowledge on almost any topic at scale as the results can be processed automatically. To help the LLM answer, a few examples called few shots can be included in the prompt. Moreover, the LLM can be asked to answer the question directly with the selected option or to first provide the reasoning and then the selected answer, which is known as chain of thought. In addition to checking whether the selected answer is correct, the evaluation can look at the LLM-estimated probability of its response as an indication of the confidence of the LLM in the response. In this paper, we study how the LLM confidence in its answer depends on whether the model has been asked to answer directly or to provide the reasoning before answering. The results of the evaluation of questions on a wide range of topics in seven different models show that LLMs are more confident in their answers when they provide reasoning before the answer. This occurs regardless of whether the selected answer is correct. Our hypothesis is that this behavior is due to the reasoning that modifies the probability of the selected answer, as the LLM predicts the answer based on the input question and the reasoning that supports the selection made. Therefore, LLM estimated probabilities seem to have intrinsic limitations that should be understood in order to use them in evaluation procedures. Interestingly, the same behavior has been observed in humans, for whom explaining an answer increases confidence in its correctness.

Keywords LLMs · Evaluation · Confidence · Multiple Choice Questions

1 Introduction

The evaluation of Large Language Models (LLMs) is challenging, as their answers are in natural language and they have to be evaluated on their performance on a large number of topics and tasks [1].

A potential approach is human evaluation, so people evaluate LLM responses. However, this does not scale to tens of thousands of questions for each model, with new models appearing every day. To address this issue, initiatives such as Chatbot Arena [2] resort to the community to assess human preferences. However, questions, answers, and participants are not controlled, so the results provide a comparative ranking of models, but not a detailed analysis of their specific capabilities. A more scalable alternative would be to use an LLM to evaluate other LLMs [3]. This method has limitations, as the LLM that is judging may have biases towards its own content or toward long responses [4, 5], and

someone has to evaluate this LLM. Today, the most widely used method to evaluate LLMs is to run different benchmarks that are mostly made up of multiple-choice questions. This enables automation of the process and evaluation of specific tasks, for example, mathematics [6], reasoning [7], or knowledge of many different topics [8], [9].

The results of a Multiple Choice Question (MCQ) test are typically measured in terms of the percentage of correct answers using a given number of examples in the prompt to help the model [1]. This accuracy metric does not provide any insight into the confidence of the LLM in its responses, which is an important feature of the LLM [10],[11],[12]. However, as in order to select each new token, an LLM computes estimates of the probability that each token in its dictionary is the next token, these probabilities can be used to develop confidence estimates [13]. For example, if there are four possible options to answer a question, A, B, C, D , and the LLM estimated probabilities for each of them are $0.5, 0.3, 0.2, 0.1$, the model does not have much confidence in its response.

The LLM responses depend not only on the question but also on the text produced by the LLM before selecting an option. In fact, it is well known that in many cases LLMs have better results when they are asked to think and decompose the problem into several steps, a technique known as Chain of Thoughts (CoT) [14]. This can be done in MCQ tests by asking the LLM to first provide the reasoning and then select an option on the prompt. An interesting question is whether this reasoning has any impact on the confidence of the LLM in its choice. In more detail, are the LLM estimated probabilities for its selected option different when the model reasons from those of when the model answers directly? If there is a difference, does it apply to choices that are correct, wrong, or to both? Is the use of logprobs therefore a recommended measure to evaluate the confidence level of LLMs? In which cases is it recommended?

In this paper, we study how the self-confidence of LLMs in their MCQ responses varies when the models answer directly or when they first provide the reasoning and then the selected answer. The main findings of our evaluation show that

1. Models tend to be more confident when they reason before answering.
2. This increase in self-confidence occurs both when the model response is correct and when it is incorrect.
3. The increase occurs for all the models tested.
4. The increase occurs for practically all categories/topics of the questions, but it is larger for topics that require reasoning.
5. All tested models experience similar trends across all dimensions.

These effects are discussed and linked to the operation of LLMs and also to human cognition to try to understand their causes and implications.

The rest of the paper is organized as follows, in Section 2 the methodology used in the evaluation is described discussing the procedure, models, and benchmarks used. The results are presented in section 3 and discussed in section 4. The paper ends with the conclusion in Section 5.

2 Methodology

2.1 Procedure

In our evaluation, we consider two different prompts when asking the question to the model. In the first, the model is asked to answer directly:

“Please respond with only the letter of the solution, in the format {‘sol’: ‘solution’}. Do not respond with any other information. Here is an example:

Input: A car travels 60 kilometers per hour for 2 hours and then 80 kilometers per hour for 3 hours. What is the average speed of the car for the entire trip? a) 70 km/h, b) 72 km/h, c) 75 km/h, d) 74 km/h

Output: {‘sol’: ‘b’}”

In the second prompt, as per CoT, the model is asked to provide step-by-step reasoning before selecting an option:

“Please think step by step before answering, considering at least three steps. Once you have the solution, end the response only with the letter of the solution, in the format {‘sol’: ‘solution’}. Here is an example:

Input: A car travels 60 kilometers per hour for 2 hours and then 80 kilometers per hour for 3 hours. What is the average speed of the car for the entire trip? a) 70 km/h, b) 72 km/h, c) 75 km/h, d) 74 km/h

*Output: First, I need to calculate the total distance traveled. For the first part of the trip, the car travels at 60 km/h for 2 hours, so the distance is $60 * 2 = 120$ kilometers. Next, for the second part of the trip, the car travels at 80 km/h for 3 hours, so the distance is $80 * 3 = 240$ kilometers. The total distance traveled is $120 + 240 = 360$ kilometers. Now, I*

need to calculate the total time spent. The total time is $2 + 3 = 5$ hours. To find the average speed, I divide the total distance by the total time: $360 \text{ kilometers} \div 5 \text{ hours} = 72 \text{ km/h}$. Therefore, the correct answer is {'sol': 'b'}.

The responses are parsed to extract the option selected by the model and its estimated probability.

2.2 LLMs

In order to ensure that the results are representative of the current LLMs, we select open and proprietary models from different companies and sizes. In more detail, we evaluate the following LLMs.

- Two models from Meta: LLama3.1-8B and LLama3.2-11B [15].
- One model from Mistral: Mistral-7B [16].
- One model from Google: Gemma-2-9B [17].
- One model from 01.AI: Yi-1.5-9B [18].
- Two models from OpenAI: GPT-4o-mini and GPT-4o [19].

2.3 Tests

The benchmark selected for our experiments is the Massive Multitask Language Understanding (MMLU) [8] as it covers a wide range of topics and we are interested in evaluating if the self-confidence in the results depends on the nature of the question. The dataset has 57 categories and more than 15,000 questions in total.

3 Evaluation results

The 57 categories of MMLU questions were run on the selected models with the direct and CoT prompts described in the previous section¹. First, we look at the aggregated results in terms of accuracy. The accuracies with both prompts are shown in Figure 1 for the different models. It can be seen that accuracy increases when the models reason before selecting the option, as reported in the literature [20].

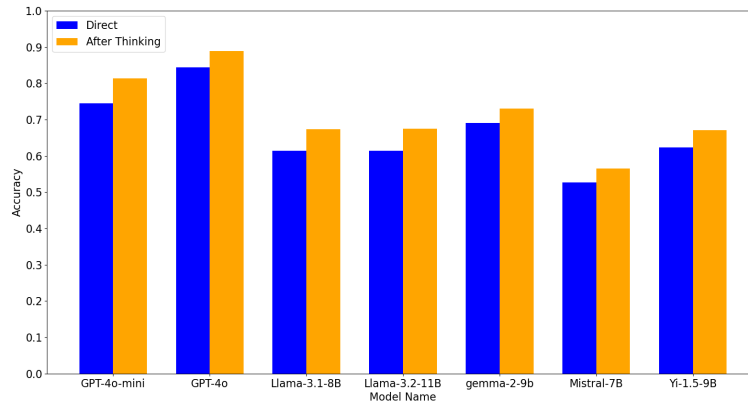


Figure 1: Accuracy Comparison Across Models on MMLU Categories with Direct and CoT Prompts

The next step is to examine the confidence that LLMs have in their responses. This is illustrated in Figure 2 which shows the average probability of the selected option for both direct and CoT prompts. It can be observed that all LLMs increase their confidence in the selected option. This could be a side effect of the CoT prompt achieving better accuracy and thus having more correct answers of which the models could be more confident. To see if that is the case, the analysis is now done independently for correct and incorrect answers in Figures 3 and 4. It can be seen that, indeed, models are more confident when the correct option is selected, but that the increase in self-confidence occurs for both correct and incorrect answers. In fact, the increment in the models' self-confidence is larger when the answer is incorrect. Therefore, the results cannot be attributed to better accuracy when using the CoT prompt.

¹The results and scripts used are available at https://github.com/aMa2210/LLM_MCQ_LogProbs

Thinking Makes Large Language Models (LLM) More Self-Confident

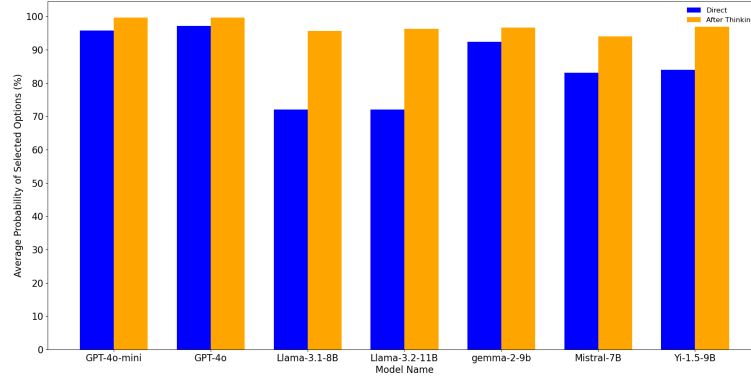


Figure 2: Average Probabilities of Selected Option Across Models on MMLU with Direct and CoT Prompts

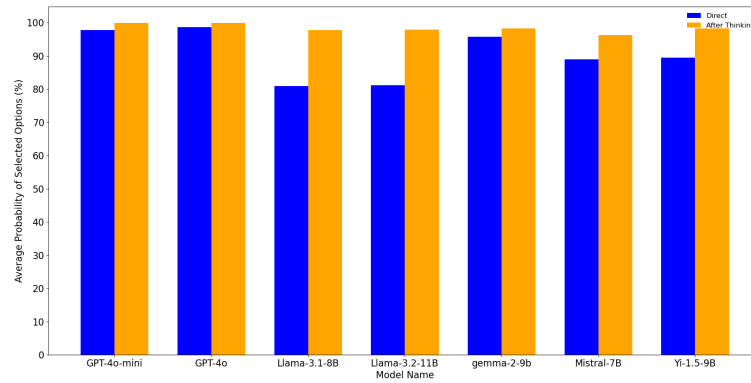


Figure 3: Average Probabilities of Correctly Selected Option Across Models on MMLU with Direct and CoT Prompts

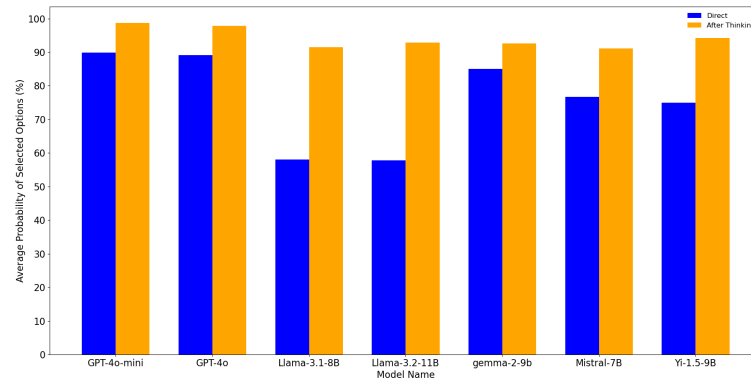


Figure 4: Average Probabilities of Incorrectly Selected Option Across Models on MMLU with Direct and CoT Prompts

To better understand this increase in the models' self-confidence we first analyze the distribution of the probabilities in Figures 5 and 6 for correct and incorrect answers. In both cases, there is a clear effect of values concentrating closer to one with the CoT prompt, which is consistent with the results obtained for the average and reported in previous figures.

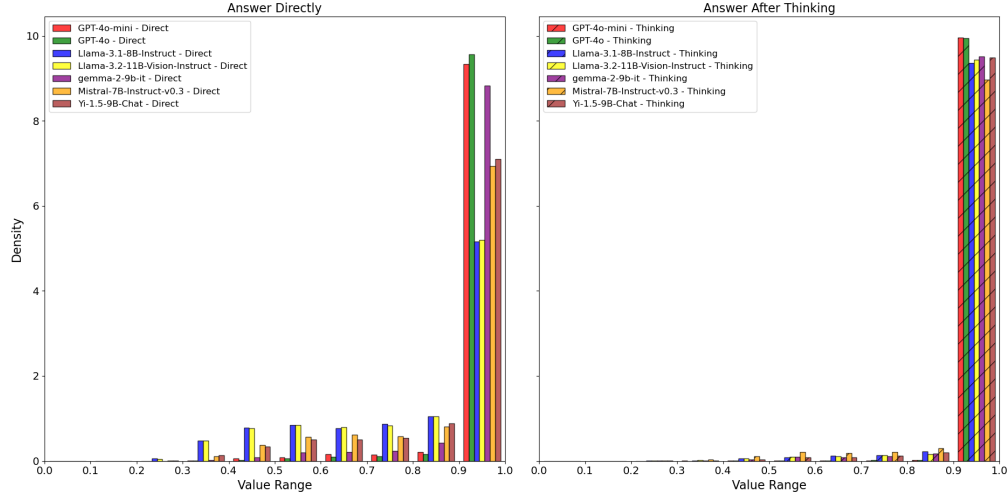


Figure 5: Probability Distribution of Correctly Selected Option Across Models in MMLU

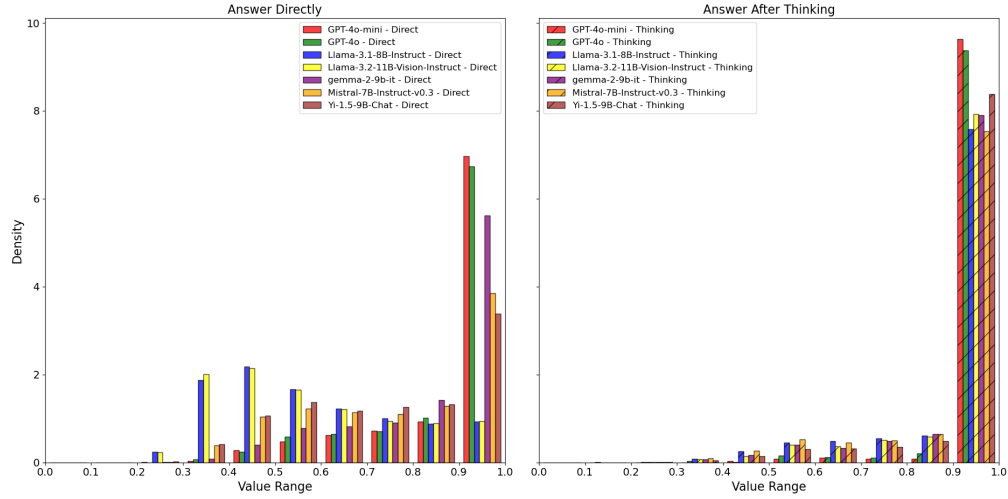


Figure 6: Probability Distribution of Incorrectly Selected Option Across Models in MMLU

It is of interest to see if this effect is consistent across the different categories in MMLU or if it only applies to a few, for example, those in which reasoning helps more. To visualize this, the increment in self-confidence is plotted versus the gain in accuracy as shown in Figure 7. The MMLU subjects are displayed in growing order of probability of the selected answer when incorrectly averaged over the seven models². It can be observed that an increase in self-confidence occurs for all categories in practically all models. The subjects with larger gains are mostly related to science except for global facts that is in the top ten. There also seems to be some correlation between increased accuracy and increased self-confidence.

²The exact computation first normalizes the increment by the mean on the models and computes the average over the seven models excluding the lowest and highest values.

Thinking Makes Large Language Models (LLM) More Self-Confident

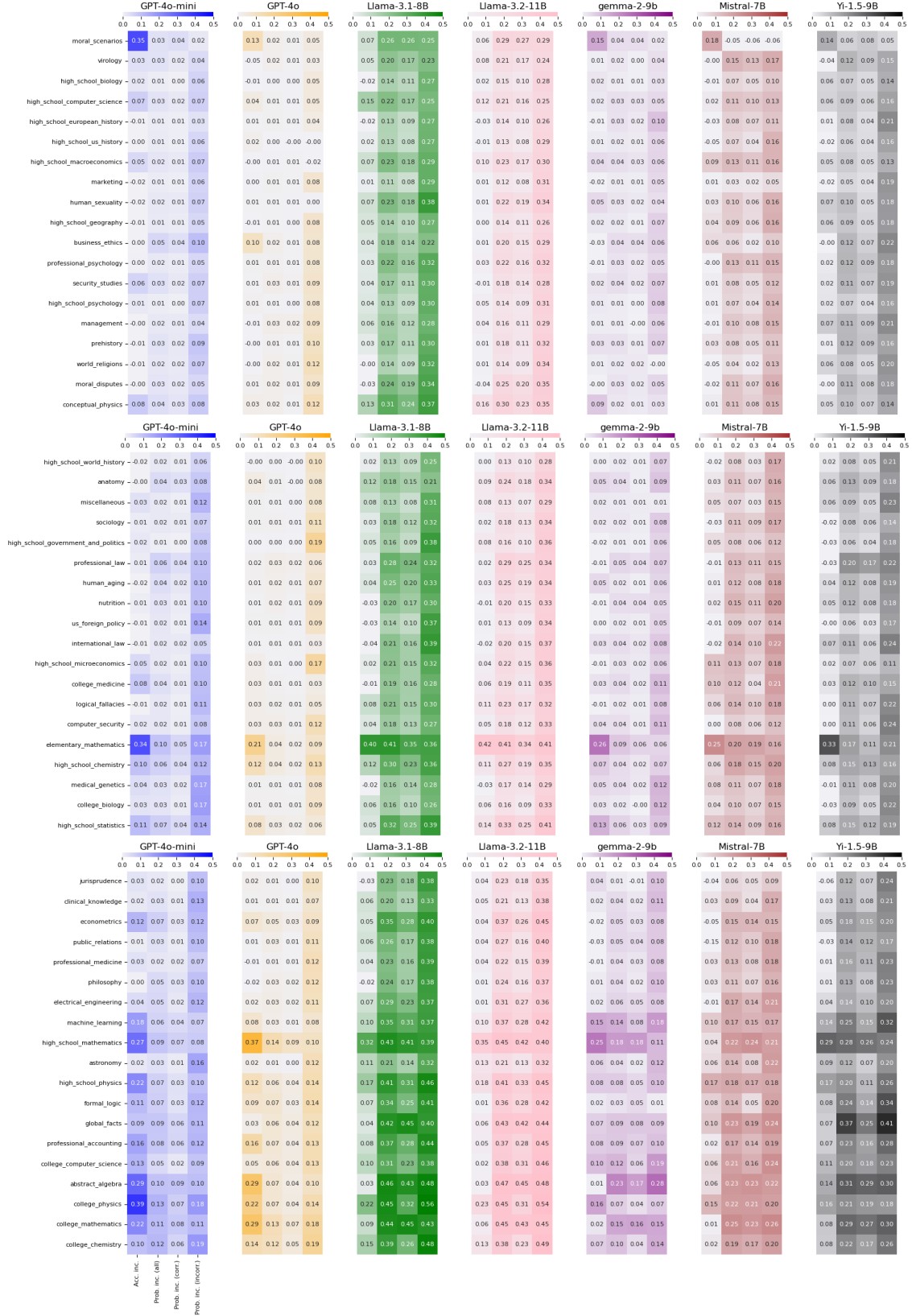


Figure 7: Increments in accuracy, in the probability of the selected option, in the probability of the selected option for correct answers and in the probability of the selected option for incorrect answers for the different subjects in MMLU across models.

Finally, we consider whether the increase in probability after thinking is different when the selected option changes. The results are summarized in Figure 8 and it can be seen that the increase is larger when the thinking process changes the selected option from incorrect to correct. This was observed in most subjects within MMLU.

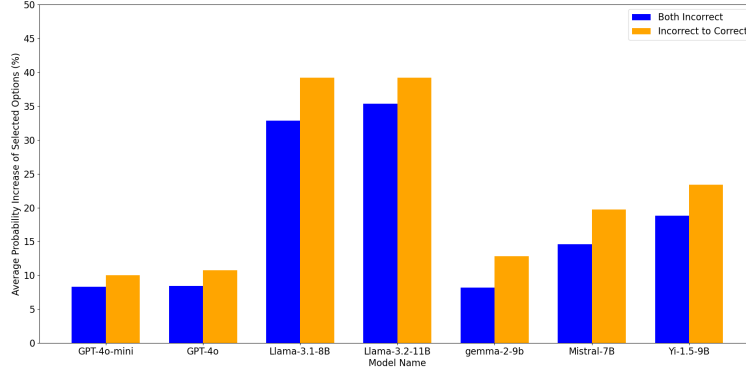


Figure 8: Average increase in the probability of the selected option when options for both prompts are incorrect and when the direct answer is incorrect and the thinking answer is correct

4 Discussion

The results in Figure 6 show that when LLMs generate incorrect answers, the frequency of the "wrong and confident" scenario significantly exceeds that of the "wrong and not confident" scenario, particularly when LLMs are required to reason, which is consistent with human behavior when answering MCQ [21]. Furthermore, as shown in Figure 7, for questions such as those related to history or moral disputes, which require minimal reasoning, the impact of reasoning on the accuracy of LLMs is negligible, and in many cases, accuracy actually decreases. However, simultaneously, the confidence of the model in providing incorrect answers increases significantly. This suggests that LLMs generating more reasoning information may actually be harmful. Experiments on MCQ in medical exams [22] have shown that non-analytical reasoning, which relies on intuition to quickly answer questions, led to the correct answer even more effectively than analytical reasoning. This is primarily because non-analytical reasoning can more efficiently utilize the test taker's prior experience with similar questions, whereas reasoning processes may cause this experience to become ineffective, thus leading to incorrect choices. This provides a potential explanation for our findings. For questions involving common sense, LLMs undoubtedly possess vast amounts of experience during their training process. When reasoning is required before answering, the influence of this experience is diminished, causing the model to potentially rely on erroneous reasoning based on faulty premises, resulting in incorrect answers. Although this effect is not as pronounced as in humans, it suggests that, for certain categories of questions, allowing LLMs to rely on intuition might be a more reliable approach. In such cases, applying logprobs to evaluate the model's performance may not be appropriate.

The increase in LLM self-confidence when it provides reasoning before answering can be related to the auto-regressive nature of these models that predict the next token based on the previous ones. This means that if the reasoning is convincing and supports the selection of a given option, the model would tend to assign it a larger probability as the next token. In fact, this behavior has been consistently observed in humans, when they explain the answer, their confidence in their response increases, as stated in [23], "explaining is believing". The same observation applies to the limited correlation between confidence and accuracy, with incorrect answers having many times higher confidence values, which has also been reported in studies with humans [24]. Therefore, it seems that more studies are needed to see if the confidence of the models follows the same patterns as in humans. If that is the case, it could provide insight into how LLMs work.

Finally, the results show how the confidence of the model is, as in humans, highly dependent on several factors and therefore should be used with caution as a tool to evaluate LLM performance. More research is needed to understand when confidence is a valid performance indicator. Existing studies on human behavior with regard to confidence provide valuable information that should be used in these research efforts.

5 Conclusion

This paper has studied how the self-confidence of LLMs in their answers to multiple choice questions depends on whether the models answer directly with the selected options or if they provide first step-by-step reasoning and then select an option. The results for the 57 subjects of the MMLU benchmark and in seven different LLMs show that the estimated probability of the selected response increases when the models provide reasoning before answering. This occurs regardless of whether the option selected by the LLM is correct. In fact, the increase in self-confidence is larger when the selected option is incorrect. These results are consistent with human studies on confidence, and suggest that further research is needed to understand when and how LLM confidence estimates can be used for evaluation.

Acknowledgments

This work was supported by the FUN4DATE (PID2022-136684OB-C22) project funded by the Spanish Agencia Estatal de Investigacion (AEI) 10.13039/501100011033 and by the Chips Act Joint Undertaking project SMARTY (Grant no. 101140087).

References

- [1] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023.
- [2] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- [3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.
- [4] Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.
- [5] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- [6] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.
- [7] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019.
- [8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [9] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.
- [10] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, 2024.
- [11] Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. Efficient and effective uncertainty quantification for LLMs. In *Neurips Safe Generative AI Workshop 2024*, 2024.
- [12] Yudi Pawitan and Chris Holmes. Confidence in the reasoning of large language models. *arXiv preprint arXiv:2412.15296*, 2024.
- [13] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*, 2024.
- [14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

- [15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [16] Albert Q. Jiang et al. Mistral 7b, 2023.
- [17] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv e-prints*, pages arXiv–2408, 2024.
- [18] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.
- [19] OpenAI. Gpt-4 technical report, 2023.
- [20] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, 2024.
- [21] Donald A Curtis, Samuel L Lind, Christy K Boscardin, and Mark Dellenges. Does student confidence on multiple-choice question assessments provide useful information? *Medical education*, 47(6):578–584, 2013.
- [22] Steven J Durning, Ting Dong, Anthony R Artino, Cees van der Vleuten, Eric Holmboe, and Lambert Schuwirth. Dual processing theory and experts’ reasoning: exploring thinking on national multiple-choice questions. *Perspectives on medical Education*, 4:168–175, 2015.
- [23] Derek J Koehler. Explanation, imagination, and confidence in judgment. *Psychological bulletin*, 110(3):499, 1991.
- [24] Jody M Shynkaruk and Valerie A Thompson. Confidence and accuracy in deductive reasoning. *Memory & cognition*, 34(3):619–632, 2006.