How Can Large Language Models Be More Reliable?

Yael Moros-Daval

Universitat Politècnica de València
ymordav@inf.upv.es

Introduction

Millions of users now rely on AI systems powered by Large Language Models (LLMs), which have become ubiquitous in many fields (Kasneci et al. 2023; Thirunavukarasu et al. 2023). Because these models are prone to errors, users must oversee their outputs and calibrate their expectations to ensure dependable performance. As LLMs grow, it is crucial to examine how their reliability has evolved. From the earliest LLMs (Raffel et al. 2020), researchers have both "scaled up" these models, by increasing parameter counts, expanding training corpora, and extending training durations, and "shaped up" their behavior via human-centered techniques such as instruction fine-tuning (Chung et al. 2024) or reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022).

It might seem obvious that as LLMs become more powerful and better aligned through these methods, their outputs would also grow more consistently accurate in ways that users can predict and adapt to (Schellaert et al. 2023). For example, early models failed simple arithmetic like "20 + 183", and these errors were so predictable that no one trusted them as calculators. However, newer models, enhanced in both scale and alignment, appear capable of handling additions even with 50-digit numbers, leading users to treat them as reliable calculation tools (for instance, when converting units) (Tsigaris and Teixeira da Silva 2023). However, these models make unexpected mistakes such as incorrectly answering "Add 3913 and 92". We lack a clear explanation for why a model will confidently deliver a wrong answer for a 100-digit addition instead of simply admitting its limitation. Ironically, this "never evasive" stance has been encouraged by developers who aim to create models that rarely refuse a user request (Bai et al. 2022).

Past Work

In (Zhou et al. 2024) we study this phenomenon by analysing three key reliability dimensions: difficulty concordance (whether model errors align with human-perceived difficulty), task avoidance (the tendency to refuse or hedge on hard questions), and prompting stability (sensitivity to paraphrases of the same question). We evaluate three model families, each in raw and shaped variants, across five benchmarks covering tasks like simple arithmetic or geographic queries.

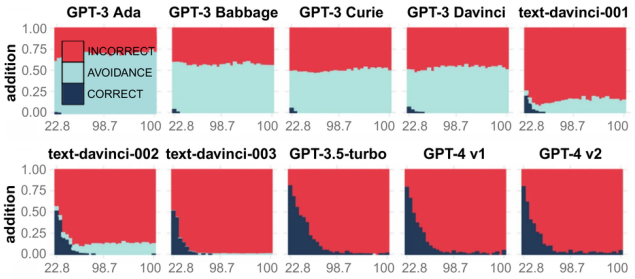

Figure 1: Performance of GPT models for a simple arithmetic task with increasing difficulty. The values are split by correct, avoidant and incorrect results. The x axis represents difficulty, which is split into 30 equal-sized bins, for which the ranges must be taken as indicative of different distributions of perceived human difficulty across benchmarks.

The results reveal a troubling pattern: although shaped-up models like GPT-4 show greater stability to prompt variations and higher correctness on average, they exhibit pronounced "difficulty discordance", continuing to err on easy instances while mastering harder ones, as we can observe in Figure 1. Moreover, as models scale and are more strongly aligned to always answer, task avoidance drops dramatically. GPT-4 almost never refuses, it gives plausible but incorrect responses instead. Task avoidance also fails to rise on harder items, so users have no reliable "safe operating area" where models either perform correctly or explicitly refuse.

On-going Work

Taking into account the limitations found in the study, my research thesis now focuses on the following question:

How can we make LLMs more reliable?

LLMs can achieve greater reliability by recognizing and refraining from answering questions for which they lack sufficient knowledge, rather than attempting to generate a response that may be incorrect. We are working on implementing such abstention as a two-step process. First, we must quantify the model's confidence in its own predictions, a

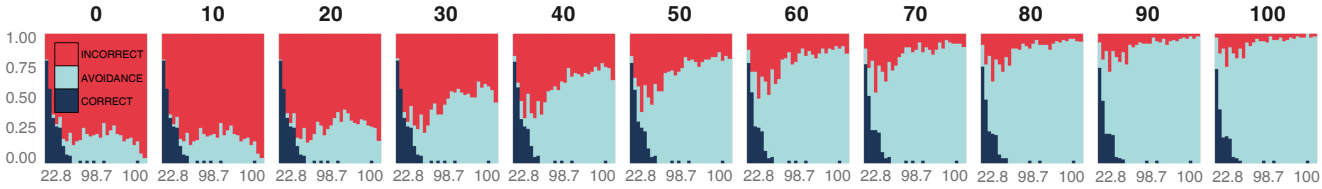

Figure 2: Performance of a LLaMa model for a simple arithmetic task with increasing difficulty. In each plot a different confidence threshold is applied (indicated above).

task addressed by the field of uncertainty estimation. There are two main approaches: black-box methods such as (Xiong et al. 2024), which treat the model as an opaque system and derive confidence scores from its outputs alone, and white-box methods e.g. (Kuhn, Gal, and Farquhar 2023; Duan et al. 2024), which leverage internal model states or probabilities.

Second, once we have a calibrated measure of confidence, we can establish rejection rules that dictate when the model should decline to answer. For instance, if we set an 80% confidence threshold, the model flags any answer below that level and simply replies, "I don't know" rather than risk a wrong response. In Figure 2, we see how the performance of a model varies with different confidence thresholds.

Crucially, this work goes beyond simply maximizing overall accuracy, it also considers question difficulty when evaluating rejection strategies. Traditional uncertainty-based rejection approaches have focused exclusively on boosting aggregate correctness, without distinguishing whether the model was abstaining on questions that users would typically find trivial or challenging. By incorporating a measure of question difficulty, we aim to define the model's "safe operating area". In practice, this means that users can trust the system to handle straightforward queries reliably, receiving either a correct response or an "I don't know", before encountering the limits of the model's competence. Such an approach not only improves accuracy but also enhances user trust, since it guarantees transparent handling of uncertainty and provides clear expectations about when the model will defer.

Acknowledgments

We acknowledge support from CIACIF/2023/276 funded by Generalitat Valenciana and European Social Found, and Spanish grant PID2024-162030OB-100 (ROBIN).

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70): 1–53.

Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; and Xu, K. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5050–5063.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Schellaert, W.; Martınez-Plumed, F.; Vold, K.; Burden, J.; Casares, P. A.; Loe, B. S.; Reichart, R.; O hÉigeartaigh, S.; Korhonen, A.; and Hernández-Orallo, J. 2023. Your Prompt is My Command: On Assessing the Human-Centred Generality of Multimodal Models. *Journal of Artificial Intelligence Research*, 77: 85–122.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Tsigaris, P.; and Teixeira da Silva, J. A. 2023. Can ChatGPT be trusted to provide reliable estimates? *Accountability in Research*, 1–3.

Xiong, M.; Hu, Z.; Lu, X.; LI, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Zhou, L.; Schellaert, W.; Martínez-Plumed, F.; Moros-Daval, Y.; Ferri, C.; and Hernández-Orallo, J. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032): 61–68.