Why Accuracy Is Not Enough: A Practical Study of Hallucination Risks in Large Language Models

Krunal Panchal
Independent Scholar
Krunalcp@live.com

Abstract - Large Language Models (LLMs) have demonstrated great success on a variety of natural language tasks and are commonly measured by metrics that are accuracy-based. Nevertheless, a high accuracy level does not imply that the outputs of the models are accurate and true to fact. An increasing worry in practice is the hallucination, in which responses produced by LLMs seem to be confident and fluent yet include false or fake information. This conduct is very risky in areas like education, research, healthcare and decision support systems. It is a practical study of the risks of hallucinations in large language models, wherein the author argues that accuracy is not enough to measure the model safety and reliability. We examine standard sources of hallucination, the weaknesses of the traditional assessment measures, and real-world cases in which false results may result in damage. Instead of putting forward sophisticated algorithms, the work focuses on the clarity of definitions, practical observations, and lightweight evaluation considerations that can be easily accessed by a researcher and a student. The research will create awareness of the risks of hallucinations and promote the adoption of wider evaluation practices when implementing systems based on LLM.

I. INTRODUCTION

Massive Language Models (LLMs) have taken center stage in the current artificial intelligence applications. They are commonly applied in question answering applications, text summarization applications, code generation applications, and conversational assistant applications. The recent developments have demonstrated that these models can be extremely accurate on benchmark datasets and be useful on various tasks with minimal or no task-specific training. This has led to the accuracy being considered as the main measure of quality of models.

Although these successes have been made, accuracy is not the only factor that can fully describe the reliability of LLM outputs. Hallucination is one of the most urgent problems of these models, as the system comes up with the information that is fluent and persuasive yet incorrect and even altogether made-up. In contrast to mere mistakes, hallucinations may be hard to identify by the users as the answers seem sure and properly organized. This gives an illusion of trust and may have grave consequences when such outputs are not checked and applied.

The issue of hallucination points to the disconnect between the evaluation and use of LLMs in the real world. Conventional evaluation measures consider superficial correctness or other measures of similarity to reference text, however, without sufficient consideration of factual consistency, uncertainty or risk. As the trend towards implementing LLaMs in high-impact fields, including education, research support, healthcare records, and policy formulation, is on the rise, it is necessary to learn and address the risks of hallucinations. This paper presents an argument that the accuracy is not sufficient to gauge the safety and reliability of large language models. We introduce a practical investigation of the behavior of hallucinations analyzing its main causes, general types, and difficulties in evaluation.

This work is aimed at no longer presenting a new model or training technique, but to offer a clear and approachable analysis, which will help researchers, practitioners, and students better understand the reasons behind hallucinations and their importance. Occupying all the aforementioned practical observations and evaluation perspectives, this research intends to make a contribution to the current debate on trustful and responsible use of large language models.

II. BACKGROUND AND DEFINITIONS

2.1 Large Language Models

Large Language Models (LLM) are neural network systems that are trained on huge amounts of textual data to give a forecast on the next word in a sequence. After acquiring patterns in language, one can make coherent sentences, respond to questions, summarize documents and many other activities involving language using these models. Common examples of popular LLMs are being trained with transformer-based architectures, and optimized to generate high probabilities of likely word sequences. One of the main peculiarities of LLMs is that they are not aware of facts or have no clear understanding of them. They instead use statistical relations trained on training data. Although this method allows one to achieve great fluency and generalization, there are risks involved as well as the models produce the answers that seemingly seem correct, but are not informed by the verified information.

2.2 What Is Hallucination in Large Language Models?

Within the large language models context, hallucination defines the production of a fluent and plausible content that is factually

incorrect, misleading or even fabricated. Hallucinations are not mere errors in predictions as the model tends to introduce false information in a confident and structured manner that makes it hard to detect a mistake by the users.

Hallucinations may be subdivided into two broad categories:

- Factual hallucinations, in which the model alters or distorts unbiased facts, i.e., wrong dates, names, or references.
- Semantic hallucinations, when the content that is produced is not in the correct context or meaning to what was intended, although individual utterances may make sense.

These hallucinations do not just happen. They are direct consequences of how language is generated by the LLMs, on the basis of the likelihood, but not on the basis of fact checking.

2.3 Accuracy Versus Trustworthiness

Language models are usually evaluated based on accuracy by comparing the generated output and reference answers. Statistics like BLEU or ROUGE test the similarity between texts, but it does not test the validity or reliability of the information. Consequently, a model can achieve high marks on the accuracy measures even when hallucinated material is produced.

Trustworthiness, in its turn, is associated with the ability of users to trust model outputs in practice. Not only should a credible system give out accurate answers, but it must not give out unsure or wrong information as true. Hallucinations undermine trust since they obscure where right knowledge and speculation that is generated are at.

2.4 Why Hallucination Is a Practical Risk

The problem of hallucination is particularly troublesome when the application of the LLM is in the fields of education, research support, healthcare records, or decision support systems. Users in such environments can have the belief that there is reliability in the fluent responses and they may not even check the information independently.

As LLMs keep on being scaled in high volumes, the behavior of hallucinations is also necessary in order to be used responsibly. To solve this problem, it is necessary to reconsider the evaluation based on accuracy and use the more extended perspectives which embrace the notions of reliability, uncertainty, and possible harm.

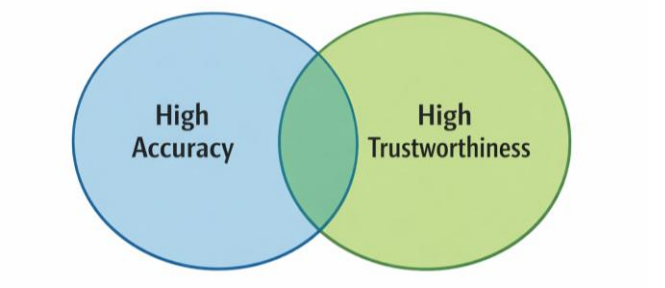

Figure 1: Conceptual Illustration of Accuracy vs. Hallucination

This statistic shows the disparity between accuracy-based assessment and hallucination risk of artificial intelligence in large language models. One model can score high in accuracy on benchmark tasks and still produce fluent errorful hallucinated outputs. The figure emphasizes the fact that accuracy is not itself an indicator of trustworthiness or reliability in the real world.

III. LARGE LANGUAGE MODELS CAN ALSO HAVE SOURCES OF HALLUCINATION

The phenomenon of hallucination in large language models does not happen by chance. Rather, it is a combination of a number of interacting elements concerning data, model design, and practices of evaluation. The interpretation of these sources is crucial in finding out why hallucinations will not go despite the models scoring high accuracy.

3.1 Training Data Limitations

Large language models are trained using large volumes of text, collected as a wide variety of sources, such as books, articles, websites, and forums. In as much as this diversity facilitates comprehensive coverage of the language, it also creates inconsistencies, old facts and unfinished information. Models acquire statistical trends based on this information and do not differentiate between validated facts and unreliable information.

Consequently, the model can come up with responses that seem right, but are founded on poor or false links when presented with questions that are similar to the ones observed during the training process. This issue is further intensified with niche issues, current events or areas that need an accurate factual foundation.

3.2 Generation based on likelihood.

LLMs create text through predicting the most probable word to come next using the context. This goal focuses on fluency and coherence, as opposed to truth. In case the model is not strictly sure or has insufficient information, this model can still give a confident answer since it is usually statistically preferred to give a plausible answer rather than to acknowledge uncertainty.

Such conduct is the reason why hallucinations are most of the time well organized and linguistically persuasive. The model maximizes the probability rather than factual accuracy, a mismatch between language generation and truth verification.

3.3 Inability to have Explicit World Models.

In contrast to the knowledge-based systems of the past, LLMs do not have a clear representation of facts or rules. They are not knowing that a statement is true, they are simply knowing that similar statements have been found in the training data. The model lacks any external grounding or verification mechanisms which make it impossible to tell the difference between what is well-known and what is invented.

This is a critical limitation where the use of LLMs is conducted in the decision-support system, where the user can mistakenly believe the generated answers are fact-checked or authoritative.

3.4 Evaluation and Feedback Gaps

Numerous hallucinations were not recognized due to the fact that such errors are not punished in the common assessment techniques. Benchmark datasets tend to reward shallow similarity and not factual consistency. Therefore, hallucination behavior can continue even between model updates without any impact on reported performance measures.

IV. PRACTICAL MEASUREMENT OF HALLUCINATION.

Hallucination is very difficult to measure due to its consideration of factual accuracy, contextual precision, and uncertainty which are not adequately measured using traditional accuracy measures. This part covers the real-life methods of detecting and assessing the behavior of hallucinations in LLCs.

4.1 Shortcomings of Traditional Metrics.

BLEU and ROUGE are common metrics that are used to assess language generation tasks. These measurements put generated text against reference outputs in terms of word overlap or similarity. Although they are useful in the quality of linguistics, they cannot be used to know whether the information created is factual.

A model can also score highly because it is able to give a text that is more or less similar to reference answers with the rest of it having erroneous or made up information. This is the drawback of the conventional measures of trustworthiness.

4.2 Human Evaluation

Human assessment is one of the surest ways of hallucination detection. Evaluators are reviewing whether the model outputs are true, appropriate and cautionary. Human assessment however is cumbersome, subjective and cannot be scaled, particularly when using large datasets.

In spite of these shortcomings, human judgment is frequently required to detect subtle hallucinations that are not readily detected by automated metrics.

4.3 Verification Through Factual Approaches.

New solutions are trying to compare model outputs with external sources like databases, search engine or verified documents. These techniques can identify discrepancies and unsupported statements by comparing generated statements and confirmed information. Even though these methods enhance detection, they are not infallible and might ave difficulty with vague queries or incomplete data feeds.

4.4 Toward Lightweight Hallucination Evaluation.

Hallucination evaluation that uses a combination of more than one metric should be used instead of using one measure. As demonstrated by the lightweight evaluation models that incorporate all these factors, a more realistic view of model reliability without too much complexity can be developed.

V. CONCLUSION AND FUTURE WORK

Large language models have performed remarkably well in their accuracy and fluency in many tasks. Nevertheless, as demonstrated in this paper, precision is not the only factor that can be used to evaluate the security and reliability of these systems. Hallucination is also a major concern especially when the LLMs are applied in high-impact or decision making contexts. This study can demonstrate that there is a need in having widespread assessment practices that transcend conventional accuracy measures by investigating the cause of hallucination and the drawbacks of current evaluation procedures. The practical examples and assessment considerations illustrated how the outputs provided by the fluent yet wrong system may compromise the trust of the user and result in the actual consequences. Future research is to be aimed at the development of the standardized evaluation frameworks that explicitly measure the risk and uncertainty of hallucinations. The consequences of hallucinations could also be minimized by integrating outside checking systems and enhancing transparency of model confidence. Finally, to be responsible in deploying large language models, it is important to acknowledge that they possess constraints and implement evaluation methods that place more emphasis on reliability rather than performance.

VI. REFERENCES

[1] T. Brown et al., “Language Models Are Few-Shot Learners,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[3] A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI Technical Report, 2019.
[4] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, 2021.
[5] A. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of ACL*, pp. 311–318, 2002.
[6] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *ACL Workshop on Text Summarization*, pp. 74–81, 2004.
[7] K. Ethayarajh et al., "Understanding and Evaluating Factuality in Abstractive Summarization," *Proceedings of ACL*, pp. 541–556, 2022.
[8] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.
[9] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed., Pearson, 2009.
[10] Y. Nie et al., "Combining Fact Extraction and Verification with Neural Semantic Matching Networks," *Proceedings of AAAI*, pp. 6859–6866, 2019.
[11] D. Hendrycks et al., "Measuring Massive Multitask Language Understanding," *International Conference on Learning Representations (ICLR)*, 2021.
[12] A. Amodei et al., "Concrete Problems in AI Safety," *arXiv preprint arXiv:1606.06565*, 2016.
[13] M. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Philosophy & Technology*, vol. 33, pp. 681–694, 2020.
[14] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of ACM SIGKDD*, pp. 1135–1144, 2016.
[15] S. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *Proceedings of ACL*, pp. 8342–8360, 2020.
[16] Y. Zou and L. Schiebinger, "AI Can Be Sexist and Racist — It's Time to Make It Fair," *Nature*, vol. 559, no. 7714, pp. 324–326, 2018.
[17] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to Sequence Learning with Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3104–3112, 2014.
[18] J. Liang et al., "Towards Understanding and Mitigating Hallucinations in Neural Machine Translation," *Proceedings of NAACL*, pp. 196–209, 2021.
[19] A. Maynez et al., "On Faithfulness and Factuality in Abstractive Summarization," *Proceedings of ACL*, pp. 1906–1919, 2020.
[20] M. He et al., "Evaluating the Factual Consistency of Abstractive Text Summarization," *Proceedings of EMNLP*, pp. 933–946, 2019.
[21] S. Zhang et al., "Fact-Checking Neural Models with External Knowledge," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3901–3914, 2022.
[22] J. Lehman et al., "Semantic Evaluation of Language Models," *Transactions of the ACL (TACL)*, vol. 10, pp. 123–138, 2022.
[23] C. Lin et al., "Trustworthy AI: From Principles to Practice," *IEEE Computer*, vol. 54, no. 9, pp. 36–45, 2021.