Real or Hallucinated? Improving the Truthfulness of LLMs

Salil Fernandes, Shruti Pareshbhai Gandhi, Xenus Gonsalves, Vedant Hareshbhai Patel, Reuben Suju Varghese
[salilfer, gandhisp, xgonsalv, patelved, rsvarghe] @usc.edu
Group 40
University of Southern California

1 Introduction

Our paper will focus on hallucinations that occur in current Large Language Models (LLMs), including OpenAI's ChatGPT, Google's Gemini, and Meta's LLaMA. These hallucinations often arise in responses to multi-step logic questions, current-events-based queries, highly specialized knowledge questions, or ambiguous questions with multiple interpretations.

Examples include:

- How many "r"s are there in "strawberry"?
- Cadmium Chloride is slightly soluble in this chemical, it is also called what?
- I get out on the top floor (third floor) at street level. How many stories is the building above the ground?

This raises a fundamental question: Should we trust LLMs blindly when they struggle with basic reasoning, such as counting letters in a word? Is hallucination inevitable?

2 Relevant Background Information

Hallucinations in Large Language Models (LLMs) refer to occurrences where the model produces inaccurate, deceptive, or invented information with great certainty. These mistakes frequently arise in multi-step reasoning, memory recall, spatial awareness, and language ambiguity. Tackling hallucinations is essential since they affect the trust, dependability, and safety of AI-generated material. To enhance our understanding and address these problems, we investigate different detection and mitigation methods to assess their efficiency in real-world situations.

2.1 Detection Methods

- **Entropy-Based Confidence Scores:** Identifies low-confidence tokens where hallucinations are probable.
- **Self-Consistency Checking:** Evaluates various answers to the same question; discrepancies indicate potential hallucinations.
- **Fact-Checking Models (FactScore, BERTScore, ROUGE/BLEU):** Assesses factual consistency with established information.
- **Retrieval-Augmented Generation (RAG) Comparison:** Evaluates the outputs of LLM against sourced factual information.
- **Human Evaluation:** Specialists evaluate factual precision in essential applications

2.2 Reduction Techniques

- **Retrieval-Augmented Generation (RAG):** Improves answers by incorporating external factual information.
- **Post-Hoc Fact Verification:** Utilizes fact-checking algorithms to confirm the accuracy of provided responses.
- **Confidence Filtering:** Rejects or alters responses with low confidence as determined by entropy scores.
- **Self-Consistency Voting:** Chooses the most reliable answer from various generated responses.

A challenge in detection is that evaluating by humans demands considerable workforce resources, which we do not have. Moreover, because of limited resources, we will exclusively test free-tier LLMs, since paid models are unavailable for our research.

Other methods, like prompt engineering (which employs structured prompts to minimize ambiguity) and fine-tuning with factually accurate data, are likewise successful in diminishing hallucinations. Nonetheless, these approaches fall beyond the limits of our research.

Finally, what sets our approach apart is that while prior research studies have addressed techniques for detecting and reducing hallucinations, many

have not thoroughly applied or evaluated these methods in real-world scenarios. Our research seeks to systematically assess various detection and mitigation strategies on publicly accessible LLMs, offering empirical insights into their efficacy.

3 Literature Review

3.1 Hallucination Causes

The origination of hallucinations in large language models occurs through their data acquisition methods and their training procedures and inference operations. The combination of wrong data sources and individual biases produces incorrect factual outputs alongside the situation in which models attempt to respond outside their fundamental knowledge base (Lin, 2022; Singhal, 2023). The use of inferior data processing methods including memory retrieval errors and knowledge database failures leads to hallucinations (Kang, 2023; Wang, 2020).

The problems in training methodology includes architectural flaws and suboptimal training objectives which lead to inaccurate factual information (Chen, 2023b; Schulman, 2023). Model accident happens more likely because of two factors: capability misalignment when models miss their purpose and belief misalignment due to reinforcement learning adjustments (Sharma, 2023; Chen, 2023c). The defective decoding methods found in inference processes generate hallucinations through two factors: random sample generation in the system and limited context-based processing (Miao, 2011; Min, 2023).

3.2 Hallucination Detection and Benchmark

Various detection techniques have been developed to identify hallucinations in LLM outputs. The detection of factual hallucinations uses three evaluation approaches including self-checking systems together with third-party knowledge verification and entropy analysis for measuring uncertainty (Maynez, 2020; Scialom, 2021). The detection of faithful text generation ensures that outputs fit within the provided input context while remaining free of added false statements (Cheng, 2023).

TruthfulQA serves as a benchmark to measure factual accuracy while HaluBench delivers various examples of hallucinations in its assessment framework (Li, 2023). HaluEval and FELM provide standardized measures for evaluating hallucinations in dialogue and summarization settings (Chen, 2023a; Gunasekar, 2023).

3.3 Hallucination Mitigation

The target areas for mitigation strategies include data, training and inference phases. The enhancement of factual datasets for data-related mitigation requires both biased information filtering and debiasing methods to prevent misinformation reinforcement (Abbas, 2023; Mitchell, 2022). Training models require three strategies that involve model adjustment for better hallucination control and external retrieval validation through knowledge integration (Ram, 2023; Pan, 2023). Self-consistency checks along with controlled generation constitute factually enhanced decoding techniques that maintain accurate responses at inference time (Zhang, 2023). The combination of faithfulness-enhanced decoding techniques leads to better consistency because it ensures logical coherence in generated text (O'Brien, 2023). Together these methods decrease hallucinations and enhance LLM reliability.

4 Work Plan

- **Dataset Preparation:** Collect human-annotated datasets covering several query categories and cross validating them to ensure that the LLM models we plan on testing actually do hallucinate on a specific prompt.
- **Implementation:** Develop a query automation script to systematically collect LLM responses to prompts fed via an API. Once we have a sizable number of prompts which hallucinate for sure on the LLMs we test, we will implement the hallucination detection mechanisms to get a definitive detection model trained on the response patterns from these prompts and then we can safely move on to implement the reduction modules.
- **Testing + Evaluation:** We will evaluate the effectiveness of reduction strategies by comparing pre- and post-reduction performance using quantitative and qualitative evaluation metrics like hallucination rate, fact score and several others.
- **Final Analysis:** We will conduct case studies and qualitative error analysis to identify failure cases where hallucinations persist, determine which detection and reduction strategies performed well and compile our findings into a research paper.

References

et al. Abbas. 2023. Debiasing llms to reduce hallucination. *arXiv preprint arXiv:2310.03368*.

et al. Chen. 2023a. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*.

et al. Chen. 2023b. Mind: Unsupervised modeling of internal states for hallucination detection of large language models. *arXiv preprint arXiv:2311.09398*.

et al. Chen. 2023c. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

et al. Cheng. 2023. Truthfulqa: Benchmarking llms on hallucination. *arXiv preprint arXiv:2305.13669*.

et al. Gunasekar. 2023. Factually data enhancement for reducing hallucinations in llms. *arXiv preprint arXiv:2310.07289*.

Choi Kang. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

et al. Li. 2023. Halueval: A large-scale hallucination evaluation benchmark. *arXiv preprint arXiv:2305.11747*.

et al. Lin. 2022. On the origin of hallucinations in conversational models. *arXiv preprint arXiv:2204.07931*.

et al. Maynez. 2020. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

et al. Miao. 2011. Detecting hallucinations in large language models using semantic entropy. *Nature*.

et al. Min. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

et al. Mitchell. 2022. Model editing for hallucination reduction. *arXiv preprint arXiv:2307.02185*.

Lewis O'Brien. 2023. Consistency-based faithful text generation in nlp. *arXiv preprint arXiv:2310.13189*.

et al. Pan. 2023. Factually enhanced decoding for reducing hallucinations. *arXiv preprint arXiv:2305.13269*.

et al. Ram. 2023. Retrieval augmentation as a hallucination mitigation strategy. *arXiv preprint arXiv:2310.03214*.

et al. Schulman. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

et al. Scialom. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

et al. Sharma. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

et al. Singhal. 2023. Why does chatgpt fall short in providing truthful answers? *arXiv preprint arXiv:2304.10513*.

Sennrich Wang. 2020. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

et al. Zhang. 2023. Faithfulness enhanced decoding in nlp models. *arXiv preprint arXiv:2303.03278*.