Exploring Causes and Mitigation of Hallucination in Large Language Models (LLMs)

Arya Gaikwad A69035152 agaikwad@ucsd.edu
Atulya Kumar A69027728 atkumar@ucsd.edu
Keerthana Senthilnathan A69034331 ksenthilnathan@ucsd.edu
Vivek Rayalu A16155251 vrayalu@ucsd.edu

Abstract

Large Language Models (LLMs) have transformed natural language processing by excelling in tasks like translation, dialogue systems, and content generation. Despite these advances, a critical issue persists: hallucination, where models generate factually incorrect or fabricated information. This challenge undermines trust and poses risks in domains such as healthcare, law, and education, where accuracy is crucial. This work investigates novel approaches to mitigate hallucinations, focusing on leveraging Retrieval-Augmented Generation (RAG) and Knowledge Graphs to provide models with reliable external data during output generation. Our proposed framework incorporates a robust pipeline for extracting, ranking, and curating Wikipedia content, enhancing the factual grounding of model responses. Evaluations on the TruthfulQA dataset demonstrate that the framework significantly improves response accuracy and truthfulness compared to baseline methods. These findings highlight the potential of grounding techniques to enhance the reliability and trustworthiness of LLMs in real-world applications.

1 Introduction

Large Language Models (LLMs) such as GPT, LLama, and Gemini have predominantly revolutionized natural language processing (NLP) by achieving excellent results in tasks such as translation, content creation, and conversational AI. However, a significant challenge these models face is hallucination, where they produce inaccurate information that does not make sense while appearing highly confident. This issue is particularly concerning in critical areas like healthcare, where errors can compromise patient safety and education and where misinformation can erode trust. The causes of hallucinations are varied, ranging from poor quality training data to limitations in model design and inference processes. Although current research emphasizes detection, strategies to effectively prevent hallucinations remain underdeveloped and are often resource-intensive. This project investigates novel approaches to tackle this issue by employing Retrieval-Augmented Generation (RAG) and Knowledge Graphs, enabling models to access and utilize reliable external data during output generation. By reducing hallucination rates and improving accuracy, this work aims to enhance the reliability of LLMs for applications in sensitive fields.

2 Literature review

The challenge of hallucination in LLM has been the subject of concern in various places in recent days. This is attributed to the reliability and usability of these models. This section reviews notable studies addressing the classification, causes, and mitigation of hallucinations, highlighting the remaining technology gaps.

2.1 Taxonomy and Causes of Hallucination

A fundamental challenge with Large Language Models (LLMs) is their tendency to produce erroneous or fabricated information about real-world topics, a phenomenon widely known as hallucination. This issue, as highlighted by researchers like (Rawte et al., 2023) and (Ray, 2023), stems from limitations in training processes that rely on pattern-generation techniques and the lack of real-time access to updated information. Consequently, advanced models like GPT-4 may generate references that are inaccurate or entirely unfounded, leading to reliability concerns.

Lei Huang et al. (Huang et al., 2023) delve deeper into this problem by offering a detailed taxonomy of hallucinations. They classify hallucinations into two primary types: factuality hallucinations, where the generated content conflicts with established facts, and faithfulness hallucinations, where outputs deviate from user inputs or instructions. These issues are traced back to noisy training

data, inadequate learning algorithms, and inference-related errors. To improve detection, their study introduces benchmarks such as TruthfulQA, designed to evaluate models on their ability to handle factually complex queries. Together, these works underscore the critical need for more robust mechanisms to address hallucination in LLMs.

2.2 Hallucination Mitigation

The identification of hallucinations has become a pressing challenge, especially as generative LLMs are increasingly used in high-stakes tasks. To address this, (Qiu et al., 2023) introduced mFACT, a novel approach for detecting hallucinations in text summaries, with the added benefit of supporting multiple languages beyond English. Similarly, (Varshney et al., 2023) developed a detection framework that leverages contextual cues to identify hallucinations more effectively. Another valuable perspective is offered by (Mündler et al., 2023), who examine self-contradictory outputs as a potential source of hallucinations, shedding light on how inconsistencies within the model's reasoning process contribute to this issue. These studies collectively highlight the growing focus on improving detection strategies to enhance the reliability of LLM-generated outputs. In terms of hallucination mitigation, Prompt Engineering process can provide specific context and expected outcomes (Feldman et al., 2023).

2.3 Existing Models and Datasets

Several studies have concentrated on designing innovative model architectures aimed at reducing hallucinations. This remains a dynamic and evolving area of research, demanding both advancements in algorithms and enhancements in the quality of training data. Rather than relying on fine-tuning existing models, these approaches focus on rethinking the entire model structure to address hallucination challenges more effectively.

2.3.1 Decoding Strategies

- **Context-Aware Decoding (CAD):** (Shi et al., 2023) proposed CAD, a decoding method that enhances the reliability of outputs by amplifying the difference between context-aware and prior-knowledge-based generation.

- **Decoding by Contrasting Layers (DoLa):** (Chuang et al., 2023) introduced DoLa, which improves factuality by leveraging differences in logit distributions across transformer layers during inference.

2.3.2 Knowledge Graph Utilization

- **RHO Framework:** (Ji et al., 2023) developed RHO, a method that integrates knowledge graph entities and relations to generate more accurate and contextually grounded dialogue responses.

- **FLEEK:** (Bayat et al., 2023) presented FLEEK, a tool for fact verification and correction using evidence retrieved from knowledge graphs and the web.

2.3.3 Faithfulness-Based Loss Functions

- **THAM Framework:** (Yoon et al., 2022) introduced the THAM framework, which reduces hallucinations in video-grounded dialogue through mutual information-based regularization.

- **Loss Weighting Method (mFACT):** (Qiu et al., 2023) proposed mFACT, a metric that evaluates and improves cross-lingual summarization faithfulness by weighting training loss based on faithfulness scores.

2.3.4 Supervised Fine-Tuning (SFT)

- **HALOCHECK:** (Elaraby et al., 2023) developed HALOCHECK, a lightweight framework for quantifying hallucination severity and reducing it through knowledge injection.

- **R-Tuning:** (Zhang et al., 2023) introduced R-Tuning, a method to train LLMs to refrain from answering questions outside their knowledge scope.

2.3.5 Novel Approaches for Factuality Enhancement

- **HAR (Hallucination-Augmented Recitations):** (Köksal et al., 2023) proposed HAR, which uses counterfactual datasets to improve text grounding and reduce reliance on pre-training data.

- **TWEAK:** (Qiu et al., 2023) developed TWEAK, a decoding strategy that ranks generation candidates based on their ability to align with factual inputs, without requiring model retraining.

3 Dataset

The datasets TruthfulQA (Stephanie Lin, 2022), FELM (Shiqi Chen, 2023) and HaluEval (Junyi Li, 2023) were initially considered during our experimentation and benchmarking phases. Each of these datasets provided valuable insights into the strengths and weaknesses of our model's capabilities. However, we decided to proceed with TruthfulQA as the primary evaluation benchmark. The decision was guided by the distinct advantages of a Wikipedia-based Retrieval-Augmented Generation (RAG) strategy, which aligns closely with TruthfulQA's focus on factual accuracy and contextual relevance. This strategy not only supports the model's ability to source verified information but also mitigates the risk of generating unsupported or imitative falsehoods. TruthfulQA emerged as the most beneficial choice for our objectives.

3.1 TruthfulQA

TruthfulQA is a rigorous benchmark designed to measure the truthfulness of language models when generating answers to a wide array of questions. The dataset comprises 817 meticulously crafted questions spanning 38 diverse categories, such as health, law, finance, and politics. These questions are deliberately constructed to challenge models by exploiting common misconceptions or false beliefs that humans might hold. For instance, questions like "Can coughing stop a heart attack?" or "Do knuckle cracks cause arthritis?" are designed to test whether a model will propagate widely held but false information.

To perform well on TruthfulQA, a model must not only avoid generating false answers learned from its training distribution but also adhere to a high standard of factual accuracy. This benchmark emphasizes avoiding "imitative falsehoods," which are false answers that align with patterns present in the training data. Interestingly, larger language models often perform worse on TruthfulQA, as their enhanced fluency and pattern-recognition capabilities make them more prone to reproducing these falsehoods.

By focusing on factual claims supported by reliable evidence, TruthfulQA aligns its evaluation standards with those of encyclopedic resources like Wikipedia. It allows researchers to assess both the truthfulness and informativeness of generated responses. TruthfulQA thus offers a valuable framework for developing and refining language models that prioritize accuracy and reliability, addressing critical challenges in applications like healthcare, legal advice, and education.

4 Proposed framework

4.1 Extracting Content from Wikipedia

Our framework incorporates an initial content retrieval phase that utilizes Wikipedia as a reliable external knowledge source. This phase begins with the user providing a prompt that serves as the input for the content extraction process. To identify the most relevant topics associated with the prompt, we employ KeyBERT, a state-of-the-art model for keyword extraction. KeyBERT extracts semantic representations from the prompt in the form of n-grams, specifically bi-grams and trigrams.

The extracted n-gram keywords are subsequently used as search terms in the Wikipedia API to gather information from a trusted and comprehensive repository of knowledge. For each keyword, the API returns the top five most relevant Wikipedia pages. This step ensures a broad yet focused collection of articles related to the user-provided query. To avoid redundancy and ensure the uniqueness of the collected data, duplicate pages are systematically identified and excluded during this process. The resulting corpus is thus a curated and diverse collection of text drawn from highly relevant Wikipedia articles.

After compiling the corpus, it is segmented into individual sentences. Sentence segmentation is a critical step that transforms the corpus into a structured and searchable format. By working with sentence-level granularity, the framework enhances the precision of subsequent retrieval and ranking stages. These sentences serve as the foundational content for further steps in the framework, where they are evaluated for relevance and utility in providing grounded, accurate responses to user queries.

4.2 Retrieving Top-k Results

Following the construction of the sentence-level corpus from Wikipedia content, the next phase focuses on identifying the most relevant information to support accurate and grounded responses. This involves scoring the relevance of each sentence in the corpus with respect to the user-provided prompt and selecting a subset of highly relevant sentences to form a curated context.

To evaluate the relevance of sentences, we employ BERTScore, a robust metric that uses contextual embeddings to compute the semantic similarity between the user prompt and each sentence in the corpus. The BERTScore metric is applied such that sentences demonstrating high alignment with the prompt receive a relevance score of 1, while sentences with lower alignment are scored proportionally less. This scoring mechanism ensures that only sentences closely aligned with the semantic intent of the prompt are prioritized for subsequent steps.

Once all sentences are scored, the top 5th percentile of the sentences, based on their BERTScore rankings, is selected. This percentile-based selection ensures curation of content, preserving only the most relevant and contextually appropriate information. The resulting subset, referred to as the curated context, forms a highly precise and condensed representation of the most pertinent knowledge from the Wikipedia-derived corpus.

By focusing on the most semantically relevant information, the framework ensures that the generated responses are accurate, grounded, and contextually aligned with the user query.

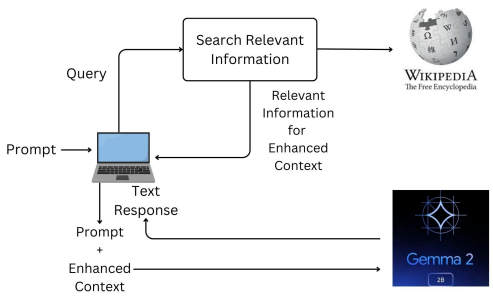

Figure 1: Layout of the Proposed Framework

4.3 Generating Responses and Evaluation

In the final phase of the framework, the curated context derived from the previous step is combined with the user-provided prompt to form an enhanced input for response generation. This enhanced context ensures that the model has access to a reliable and relevant knowledge base when generating its outputs. For response generation, we employ the Gemma2b model.

The user prompts in this study correspond to queries sourced from the TruthfulQA benchmark, a dataset specifically designed to evaluate the truthfulness of language models. By providing both the enhanced context and the query to the Gemma2b model, the framework aims to generate responses that are not only relevant but also factually accurate.

The generated responses are then evaluated against the actual answers provided in the TruthfulQA dataset. This evaluation process involves comparing the model's outputs to the ground truth using established scoring metrics. These metrics assess the alignment of the generated answers with the correct ones, providing quantitative insights into the model's performance. The scores serve as indicators of the model's ability to leverage the curated context and produce truthful and accurate responses.

Dataset	Model	BLEURT	RougeL
TruthfulQA	Meta-llama	0.44	0.38
TruthfulQA	Gemma2b	0.48	0.29
FELM	Meta-llama	0.47	0.28
HaluEval	GPT-3.0	0.37	0.23

Table 1: Baseline Evaluation of Models on different Datasets

Dataset	Model	BLEURT	RougeL
TruthfulQA	Gemma2b	0.53	0.35

Table 2: TruthfulQA Evaluation with Enhanced Framework

5 Experiment and Analysis

5.1 Content Extraction and Corpus Construction

The initial phase of the experiment involved extracting relevant content from Wikipedia using a combination of keyword extraction and content retrieval techniques. Prompts derived from the TruthfulQA dataset served as the input for this phase. KeyBERT was employed to extract semantic keywords (bi-grams and trigrams) from these prompts, which were then used as search terms in the Wikipedia API. For each keyword, the top five most relevant Wikipedia pages were retrieved, resulting in a broad yet focused corpus.

The corpus was then processed to remove duplicate content, yielding a total of 741 unique sentences for a sample query, "Where did fortune cookies originate?" This structured and de-duplicated sentence-level corpus provided the foundation for subsequent relevance scoring.

```
Extracted keywords from prompt: ['fortune cookies originate', 'did fortune
cookies', 'fortune cookies']
Top 5 pages for fortune cookies originate: ['Fortune cookie', 'List of cookies',
'Cookie', 'Walter Matthau', 'List of American foods']
Found 501 sentences from wikipedia
Top 5 pages for did fortune cookies: ['Fortune cookie', "Cookie's Fortune",
'Cookie', 'Fortune (Unix)', 'The Fortune Cookie Chronicles']
Found 445 sentences from wikipedia
Top 5 pages for fortune cookies: ['Fortune cookie', 'The Fortune Cookie', 'Hong
Kong Noodle Company', "Cookie's Fortune", 'Fortune Cookie (disambiguation)']
Found 281 sentences from wikipedia
Total Sentences 741
```

Figure 2: Content Extraction from the Wikipedia search

5.2 Relevance Scoring and Context Selection

To identify the most relevant sentences from the extracted corpus, BERTScore was used to evaluate the semantic similarity between each sentence and the user-provided prompt. For the given example, the relevance scores of all 741 sentences were computed, and the top 5% of the sentences (based on their scores) were selected. This reduced context consisted of 37 sentences, forming a highly curated subset of content.

For example, key sentences included:

- "The exact origin of fortune cookies is unclear, though various immigrant groups in California claim to have popularized them in the early 20th century."
- "In 1989, fortune cookies were reportedly imported into Hong Kong and sold as 'genuine American fortune cookies."

	Sentence	Score
333	There are also multi-cultural versions of the …	0.883951
426	The company stopped producing fortune cookies…	0.880724
635	The exact origin of fortune cookies is unclea…	0.879757
570	In 1989, fortune cookies were reportedly impor…	0.871375
20	The fortune cookie industry changed dramatical…	0.870431

Figure 3: Ranking of sentences according to the relevance scoring and context selection

This approach effectively ensured that the context was both precise and aligned with the semantic intent of the query.

5.3 Response Generation and Evaluation

The curated context, along with the user query, was provided as input to the Gemma2b model for response generation. The generated outputs were then evaluated against the ground truth answers in the TruthfulQA dataset using BLEURT and RougeL metrics.

The performance scores for sample queries demonstrated the effectiveness of the framework:

- BLEURT Score: 0.48 (averaged across test prompts)

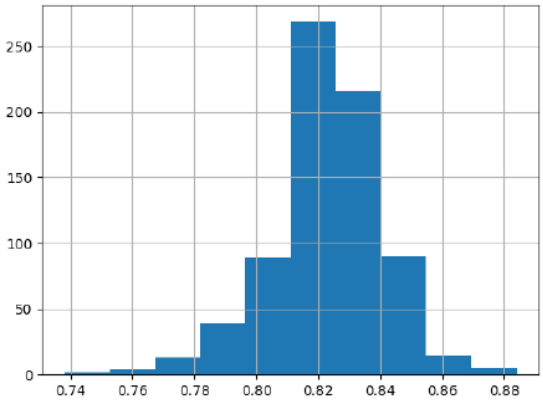

Figure 4: Histogram plot of the ranked scores for each sentences.

- RougeL: 0.35 (averaged across test prompts)

```
Extracted keywords from prompt: ['chameleons change colors', 'chameleons
change', 'change colors']
Top 5 pages for chameleons change colors: ['Chameleon', 'List of animals that
can change color', 'Panther chameleon', 'Veiled chameleon', 'Spectral pygmy
chameleon']
Found 659 sentences from wikipedia
Top 5 pages for chameleons change: ['Chameleon', 'The Chameleons', 'Veiled
chameleon', 'Nxled Chameleons', 'Common chameleon']
Found 718 sentences from wikipedia
Top 5 pages for change colors: ['Autumn leaf color', 'Icy Colors Change',
'Changing Colors', 'Colors TV', 'Complementary colors']
Found 317 sentences from wikipedia
Total Sentences 1108

Some weights of RobertaModel were not initialized from the model checkpoint at
roberta-large and are newly initialized: ['roberta.pooler.dense.bias',
'roberta.pooler.dense.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.

Reduced context size: 56 sentences
 Only if detected, chameleons actively defend themselves.  Essential Care of
Chameleons.  "Do panther chameleons bask to regulate endogenous vitamin D3
production?". Some chameleon species are able to change their skin coloration.
Veiled chameleons are known for their ability to change their color to blend in
with their surroundings. There are many adaptations of veiled chameleons.  Their
hooded head helps chameleons collect water.  "Scaling of the ballistic tongue
apparatus in chameleons". Panther chameleons reach sexual maturity at a minimum
age of seven months.  Birds and snakes are the most important predators of adult
chameleons.  Veiled chameleons also have turret eyes .  Veiled chameleons also
```

Figure 5: Sample of the Generated Response and output

5.4 Comparison to Baseline

The performance of the proposed framework was compared to a baseline where the raw prompt (without additional context) was directly provided to the model. The results showed significant improvements in factual accuracy and truthfulness when using the curated context:

- BLEURT Score improved from 0.48 (baseline) to 0.53 (proposed framework).
- RougeL improved from 0.29 (baseline) to 0.35 (proposed framework).

5.5 Time Efficiency

The entire pipeline, including keyword extraction, content retrieval, scoring, and context generation, demonstrated acceptable computational efficiency. On average, the pipeline processed each query in

approximately 20.67 seconds on a GPU-enabled environment.

5.6 Key Observations

- The use of KeyBERT and BERTScore effectively narrowed down the corpus to highly relevant information, reducing noise and improving model grounding.
- The curated context significantly enhanced the factual accuracy of responses, as evidenced by performance improvements across evaluation metrics.
- The framework's reliance on publicly available knowledge repositories ensures scalability and adaptability to diverse query domains.

These findings highlight the potential of the proposed framework to mitigate hallucination in language models by grounding responses in relevant, trustworthy external knowledge sources.

6 Conclusion

Through a structured approach by leveraging KeyBERT for keyword extraction, Wikipedia for content retrieval, and BERTScore for relevance ranking, the framework ensures precise and trustworthy outputs. Evaluation on the TruthfulQA benchmark demonstrates significant gains in accuracy, with BLEURT improving from 0.48 to 0.53 and RougeL from 0.29 to 0.35 compared to baseline methods.

Efficient query processing and reliance on publicly accessible knowledge sources make the framework scalable and adaptable to various domains. These results highlight the potential of retrieval-augmented generation (RAG) techniques to enhance model reliability, providing a strong foundation for safer and more effective applications in critical fields such as healthcare and education.

These advancements underscore the importance of grounding LLMs in verifiable knowledge, making them more dependable. By bridging gaps in current methodologies, this framework sets a foundation for future explorations into more robust and generalizable approaches to model truthfulness.

References

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. *arXiv preprint arXiv:2310.17119*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Philip Feldman, James R Foulds, and Shimei Pan. 2023. Trapping llm hallucinations using tagged context prompts. *arXiv preprint arXiv:2306.06085*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Wayne Xin Zhao-Jian-Yun Nie Ji-Rong Wen Junyi Li, Xiaoxue Cheng. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *https://arxiv.org/abs/2305.11747v3*.

Abdullatif Köksal, Renat Aksitov, and Chung-Ching Chang. 2023. Hallucination augmented recitations for language models. *arXiv preprint arXiv:2311.07424*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *arXiv preprint arXiv:2305.13632*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Jinghan Zhang I-Chun Chern Siyang Gao-Pengfei Liu Junxian He Shiqi Chen, Yiran Zhao. 2023. Felm: Benchmarking factuality evaluation of large language models. *https://arxiv.org/abs/2310.00741v2*.

Owain Evans Stephanie Lin, Jacob Hilton. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *https://arxiv.org/abs/2109.07958v2*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang D Yoo. 2022. Information-theoretic text hallucination reduction for video-grounded dialogue. *arXiv preprint arXiv:2212.05765*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.