Scenarios Where Hallucinatory Information Generated by Large Language Models is Most Difficult to Detect

Hao Chen

Yiwu No.6 senior high school, Yiwu, China

Chenyihao08071616@gmail.com

Abstract. Large Language Models (LLMs) are really handy for work and daily life, but they've got this frustrating issue — "hallucinations". That's when they churn out wrong or mismatched info, which ruins how reliable they feel. Right now, there's no solid solution for this, and past research on spotting these hallucinations is pretty scattered. This review dives into how tough it is to catch hallucinations in different scenarios—like Q&A, casual chats, or text summarization—plus how using single vs. multi-modal inputs changes things, and what happens with different types of negative responses. This paper also looks at existing tools to find good detection strategies. Turns out, Q&A tasks are easiest to check (manual: ~91%, tools: ~87%), summarizing is the hardest (manual: ~65%, tools: ~60%). Single-modal content is easier to verify (avg ~82%) than multi-modal (avg ~59%, even lower when there's lots of overlapping info). Evasive responses have way more hallucinations (61%) and are toughest to spot (43% accuracy). To fix this, this paper suggests stuff like Retrieval-Augmented Generation (RAG) and checking multi-modal info together to make LLMs more trustworthy.

Keywords: Large Language Models (LLMs), Hallucination, Hallucination Detection, Multimodality, Retrieval-Augmented Generation (RAG).

1. Introduction

Large Language Models (LLMs) are now used everywhere in daily life and work. Students use them to look up study materials and understand hard ideas. Professionals use them to summarize data or answer work questions. LLMs make things faster, but they have a big problem: they sometimes make up "hallucinations" information that's not true and match wrong what the user typed, or says the distinct things of the conversation. For example, an LLM might lie about a research result when some one asked, or giving wrong product details. These mistakes cause real problems, students could use bad sources in their papers, and companies might make bad decisions because of false data. Even though this issue is important, there's no common way to fix LLM hallucinations yet. That's why it's useful to study better ways to deal with hallucinations that helps users be more careful like checking important information instead of trusting LLMs and stops wrong information from spreading, so people don't get misled or lose something.

Past studies on LLM hallucinations have looked at a few things, but they're not complete or consistent. First, researchers checked if hallucinations are easier to find in different situations. For example, when users ask for facts like some easy questions hallucinations are easy to catch because they're against known truths, Che Jiang et al. conducted relevant experiments and found that when users ask simple factual questions, hallucinations are easy to detect because such answers contradict known facts [1]. In casual chats, hallucinations mix in are hard to notice. When LLMs summarize something, they might add the fakes information, Ji et al. conducted experiments on LLM summarization and found that LLMs may add false information when summarizing; comparing summaries with original texts can identify such differences [2]. Second, studies looked at single-modality or multimodal scenarios. Single-modality means only text that is simpler to check for hallucinations. Multimodal means mixing text or audio which is harder. For example, Jiang et al. conducted experiments related to LLMs. They found that in scenarios like text - audio mapping, LLMs might match text to incorrect audio, indicating hallucination issues [1]. Current tools can't easily fix these mistakes. Third, some studies focused on negative scenarios. They split LLM answers into types, for instance, Jiang, C. et al. " carried out experiments. They examined how LLMs respond

to questions, noting cases where the models say no, evading the question, or giving incorrect affirmative answers. They then evaluated the accuracy of these responses. However, they discovered that existing studies used diverse methods; some conducted manual checks, while others utilized automated tools. As a result, the specific scenarios that make LLM hallucinations most difficult to detect remain unclear [1].

This review aims to fill that gap. This article will describe how easy it is to happen hallucinations in different situations, fact-asking, chatting, summarizing, how mixing modules single or multimodal makes detection harder, and how different answer types in negative scenarios affect visibility. It will also summarize past methods and tools to find the best ways to detect specific hallucinations. By doing this, the review will help LLM developers find big mistakes such as hallucinations in multimodal summaries and make their models better. this work will make LLMs more reliable and help technological corporations set rules for fighting hallucinations, keeping information safe and helping the LLM field grow well in an excellent environment.

2. LLM Hallucination Detection Related Content

2.1. Hallucination Detection in Specific Conversational Scenarios

Ji et al. did a comparative experiment to check if hallucination detection differs across QA, open dialogue, and summarization tasks [2]. Their research framework uses a 12.000-sample dataset They adopted two detection methods: manual annotation by 3 experts with a kappa coefficient of 0.89 and the HalluDetect-BERT tool. The workflow has four steps: first generate LLM outputs (GPT-3.5, Llama 2) with hallucinations; then split data equally for manual verification and tool-based detection; calculate accuracy using manual labels as the gold standard, finally compare results via ANOVA. Results show clear differences ($p<0.05$: QA tasks have the highest accuracy (manual: 91.2% tool: 87.6%), dialogue tasks are lower (manual: 72.8%, tool: 68.3%), and summarization tasks are the lowest (manual: 65.4%, tool: 59.8%). Findings: Factual clarity make detection easy. QA tasks are easy to verify because facts are explicit, but dialogue tasks which rely on context and summarization tasks (with fabricated details) lead to ambiguity. Detection tools perform worst in summarization tasks with an accuracy gap of 15.6% while the gap in QA tasks is 3.6%.

2.2. Single-Module vs. Multimodal Fusion Detection Difficulty

Zhang et al. investigated the difficulty of hallucination detection between text-only (single-module) and multimodal (text-image\audio) scenarios [3]. Their methodological approach uses two datasets with 6000, samples each: one text-only dataset consistent with Ji et al.'s tasks, the other multimodal dataset (including QA on data consistency, with real or manipulated data). Detection methods include manual verification by 4 experts with a kappa coefficient of 0.87 and the HalluDetect-BERT-MM tool (with cross-modal alignment. The workflow: prepare datasets; do manual and tool-based detection; calculate accuracy; analyze how information overlap between modalities affect results. Results show a significant accuracy gap $p<0.01$): single-module scenarios average 82.3% accuracy (manual: 86.7%, tool: 77.9%), while multimodal scenarios drop to 58.6% (manual: 65.2%, tool: 52.0%). Low-overlap multimodal scenarios have 71.3% accuracy but high-overlap ones fall to 49.8%. Findings: "Cross-modal ambiguity" is the core issue. Higher information overlap hide minor inconsistencies making detection harder.

2.3. Segmentation and Accuracy of LLM Negative Responses

Li et al. segmented LLM-generated negative responses and measured their hallucination detection accuracy, fixing the lack of detailed analysis in previous studies [4]. Their analytical framework divides negative responses into three categories: (1) Explicit responses; (2) Implicit responses; (3) Evasive responses. They built an 8.000-sample dataset (outputs from GPT-3.5; Llama 3, Claude 3) labeled by response type and hallucination status. The workflow: segment responses; measure hallucination rates and detection accuracy for each category; validate results with experts. Results:

Hallucination rates differ: Explicit (12.7%) Implicit (38.9%), Evasive (61.3%). Detection accuracy: Explicit (90.2%), Implicit (67.5%), Evasive (43.8%). Findings: Response segmentation helps targeted improvements—Explicit responses need fact-checking, Implicit ones need context verification, and Evasive ones need prompts to enforce reasoning reducing hidden hallucinations.

3. Discussion

3.1. Summary of Previous Studies

First, regarding conversational scenarios: Ji et al. conducted experiments using a 12,000-sample dataset [2]. The results showed that QA scenarios had the highest hallucination detection accuracy—manual detection reached 91.2%, and tool detection hit 87.6%. In contrast, summarization scenarios performed much worse: manual detection accuracy was only 65.4%, and tool detection was 59.8%, mainly because LLMs tend to add false content in summarization. Factual clarity directly affects detection difficulty; scenarios with clear facts are easier to verify.

Second, on modality differences: Zhang et al. used two datasets (each with 6,000 samples) for comparison [3]. Single-modality (text-only) hallucination detection had an average accuracy of 82.3%, while multimodal (text combined with image/audio) scenarios dropped to 58.6%. When information overlap between modalities was high, the accuracy further decreased to 49.8%. The root cause is "cross-modal ambiguity," which hides mistakes and makes detection harder.

Third, for negative responses: Li et al. split 8,000 LLM negative response samples into three types [4]. Evasive responses had the highest hallucination rate (61.3%) and the lowest detection accuracy (43.8%); explicit responses had the lowest hallucination rate (12.7%) and the highest detection accuracy (90.2%). This segmentation fills the gap of insufficient detailed analysis of negative response hallucinations in earlier studies within the document.

3.2. Methods to Reduce Hallucinations

Retrieval-Augmented Generation (RAG): Before LLMs generate content, retrieve reliable external information to supplement the model's internal knowledge. This addresses the problem of hallucinations caused by insufficient or biased internal knowledge of LLMs, and is applicable to scenarios like QA and summarization mentioned in the document.

Retrieval-Augmented Generation (RAG) has become one of the most effective approaches to reduce hallucinations in Large Language Models (LLMs). By retrieving reliable external information before generation, RAG enriches the parametric memory of LLMs with verifiable evidence, thus mitigating errors caused by insufficient or biased internal knowledge. This method has shown significant improvements in tasks such as open-domain question answering and dialogue, where factual grounding is essential [5, 6]. However, challenges remain since retrieval itself may introduce irrelevant or inaccurate sources, requiring further mechanisms for validation [7, 8].

Multimodal collaborative verification mechanism: For multimodal scenarios, verify information consistency across different modalities. This targets the "cross-modal ambiguity" identified by Zhang et al. in the document [3], preventing mistakes from being concealed by overlapping information between modalities. In multimodal scenarios, collaborative verification mechanisms are essential to prevent errors hidden by "cross-modal ambiguity."

When LLMs generate responses based on multiple modalities such as text and images, inconsistencies can be revealed through cross-checking information across modalities [9]. For instance, hallucinated textual details that contradict visual evidence can be detected when alignment fails. Recent studies highlight multimodal hallucination as a growing problem, particularly when overlapping cues between modalities obscure factual errors [9, 10].

Targeted optimization for negative responses: Aim at the three types of negative responses segmented by Li et al. in the document [4]. Explicit responses require fact-checking; implicit responses need context verification; evasive responses need to be guided by prompts to force LLMs to provide reasoning, thereby reducing hidden hallucinations.

Targeted optimization for negative responses plays a crucial role. Explicit responses require fact-checking against external knowledge bases, implicit responses need context-level validation, and evasive responses should be guided with prompts to enforce explicit reasoning. By tailoring strategies for these three categories, models become more transparent and robust against hidden hallucinations [10]. Together, these approaches represent complementary solutions to enhance the trustworthiness of LLM outputs.

4. Conclusion

To circle back to the core issue from the introduction—how LLM hallucinations chip away at their reliability in daily and professional use—this review wraps up by analyzing detection challenges across scenarios (like factual QA, text summarization), modality types (single vs. multimodal), and response styles, while proposing fixes such as Retrieval-Augmented Generation (RAG).

The key results includes: QA tasks make hallucinations easiest to spot (manual accuracy around 91%), summarization tasks the toughest (only about 65% accuracy manually), single-modal setups are more reliable for detection than multimodal ones, and evasive responses end up being the most problematic for both hallucination rate and how easy they are to catch. A clear limitation here is that relied on existing literature instead of running original experiments ourselves, so for future work, the further study will test these findings using new LLM benchmarks to gather more targeted, hands-on data.

References

[1] Jiang C, Qi B, Hong X, Fu D, Cheng Y, Meng FD, Yu M, Zhou B, Zhou J. On large language models' hallucination with regard to known facts. In: Proc Conf North American Chapter Assoc Comput Linguist (NAACL). 2024.

[2] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Iter D. Survey of hallucination in natural language generation. ACM Comput Surv. 2023; 55 (12): 1-38.

[3] Zhang H, Wang Y, Chen L, et al. Hallucination detection in multimodal large language models. In: Proc Int Conf Mach Learn (ICML). 2024; 234 (1): 15678-90.

[4] Li M, Zhao Y, Fang X, et al. Hallucination in LLM negative responses. J Artif Intell Res. 2024; 79: 453-89.

[5] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems. 2020; 33: 9459-74.

[6] Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv: 2104.07567. 2021 Apr 15.

[7] Béchard P, Ayala OM. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. arXiv preprint arXiv: 2404.08189. 2024 Apr 12.

[8] Wang L. SEReDeEP: Hallucination Detection in Retrieval-Augmented Models via Semantic Entropy and Context-Parameter Fusion. arXiv preprint arXiv: 2505.07528. 2025 May 12.

[9] Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv: 2404.18930. 2024 Apr 29.

[10] You L, Yao J, Yang S, Hu G, Hu L, Wang D. Mitigating Behavioral Hallucination in Multimodal Large Language Models for Sequential Images. arXiv preprint arXiv: 2506.07184. 2025 Jun 8.