Retrieval-Augmented Language Models Evade Hallucination Detection

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) *intends* to mitigate hallucinations by incorporating external knowledge sources. However, the seemingly accurate, authoritative responses of RAG models may *unintendedly* make hallucinations harder to detect. In this paper, we systematically investigate this phenomenon across three popular RAG frameworks and three question-answering datasets. Compared to vanilla language models, RAG increases the false negative rate of widely adopted automatic hallucination detectors from 23.8% to 52.0% on average. Furthermore, we study RAG's impacts of production models (DeepSeek-R1) on real human users. We find that RAG rises the false negative rate of hallucination detections by 5.4%. Finally, we show that optimizing RAG models with hallucination detectors cannot mitigate but exacerbate this problem: RAG models can hack hallucination detectors and further increase the false negative rate by 53.3%. We highlight an overlooked risk of RAG and call for more research in helping both machines and humans detect hallucinations.

1 Introduction

Large language models (LLMs) (Brown et al., 2020) have achieved remarkable success in text generation, demonstrating strong performance across a wide range of natural language tasks (Hudevcek and Dusek, 2023; Pu et al., 2023). Despite these advancements, a persistent and critical limitation remains: LLMs are prone to *hallucinations*, where they produce content that is factually incorrect or not grounded in verifiable information (Zhang et al., 2023; Ji et al., 2022). This issue significantly undermines the reliability of LLMs, especially in real-world applications that require factual precision and trustworthiness.

To address this limitation, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a promise. By incorporating external knowledge through a retrieval mechanism, RAG gathers relevant evidence from curated document collections and integrates it into the generation process. This approach reduces reliance on the model's parametric memory while anchoring outputs to concrete, external sources. RAG has proven effective in improving the factuality of tasks such as question answering and summarization, especially in domains requiring up-to-date or specialized knowledge. It strikes a good balance between generative fluency and factual reliability, making it a key framework for developing trustworthy AI systems.

However, while RAG improves factual accuracy, it introduces a subtler and less understood challenge: *it increases the difficulty of hallucination detection.* RAG-generated outputs often appear more credible, as they incorporate superficially aligned or partially relevant information (Zhao et al., 2024), even when the final content is inaccurate. As a result, existing hallucination detection frameworks often overestimate the truthfulness of RAG outputs (Chen et al., 2023). Compared to hallucinations produced by standard LLMs, which tend to be more overt or logically inconsistent, RAG-induced hallucinations are frequently more subtle, plausible, and harder to identify—even under rigorous evaluation. This *credibility-detectability paradox* poses substantial risks in high-stakes settings such as healthcare, finance, and law, where subtle misinformation can have significant consequences.

In this work, we systematically investigate how RAG exacerbates the challenge of hallucination detection. We first evaluate common RAG frameworks on three question answering datasets—HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and ASQA (Stelmakh et al., 2022). Across three widely-adopted automatic hallucination detectors (e.g., SelfCheckGPT (Manakul et al., 2023)), we find that using RAG largely reduces their performance. Specifically, the false negative rate of hallucination detectors increases from 23.8% to 52.0% on average.

Next, we investigate the impact of RAG on real users using the production model Deepseek-R1 (DeepSeek-AI et al., 2025). Responses from the vanilla and RAG models are judged for hallucination by human annotators based on collected online queries. We show that RAG also increases the human-annotated hallucination false negative rate by 5.4% in production environments.

We also examine whether preference learning methods like DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024) can help. Surprisingly, these fine-tuned models not only fail to improve reliability but increase false negatives by 53.3% (Sec.,4). Overall, our findings reveal a key risk of RAG: while it improves factual accuracy, it makes hallucinations harder to detect. This calls for better automatic detectors and more robust human evaluation practices.

2 Evading Automatic Hallucination Detection

To assess the impact of RAG on hallucination detection, we evaluate two representative RAG frameworks across three widely-used QA benchmarks, employing three famous hallucination detectors.

2.1 Experimental Setup

Datasets. We start with evaluating the impact of RAG on hallucination detection using three widely-adopted benchmark datasets, including HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and ASQA (Stelmakh et al., 2022). HotpotQA and 2WikiMultihopQA focus on multi-hop reasoning, requiring models to integrate information across multiple documents. ASQA, on the other hand, emphasizes ambiguity resolution in open-domain questions. These datasets were chosen for their emphasis on questions that require information from multiple external sources, which presents an ideal testbed for examining the strengths and limitations of RAG-based generation.

Answer Generation & Metrics. We evaluate three different answer generation approaches:

1. **w/o RAG**: a non-RAG baseline relying on the internal knowledge of the language model.
2. **Vanilla RAG**: using Contriever (Izacard et al., 2021) as the retriever model, and adding retrieved passages as additional inputs to prompt LMs.
3. **InstructRAG** (Wei et al., 2024): enhancing RAG with task-specific instruction-finetuning to guide the integration of retrieved evidence.

Method	HotpotQA		2WikiMQA		ASQA	
	Acc.	K	Acc.	K	Acc.	K
w/o RAG	0.325	-	0.325	-	0.565	-
vanilla RAG	0.590	10	0.421	10	0.842	3
InstructRAG	0.618	5	0.391	10	0.804	5

Table 1: Answer accuracy judged by GPT-4-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used.

We use LLaMA-3-8B-Instruct (Dubey et al., 2024) as the backbone generator, with either few-shot prompting or supervised fine-tuning based on the dataset. To ensure consistent answer evaluation, we employ GPT-4-Turbo (Achiam et al., 2023) to assess whether responses fully and accurately address the question. This LLM-based judgment overcomes the shortcomings of string-matching metrics, which often overestimate correctness due to shallow token overlap or partial relevance.

Hallucination Detection & Metrics. We evaluate hallucination detection frameworks of:

- FAVA (Mishra et al., 2024) identifies hallucinations by classifying errors into six fine-grained categories: invented, unverifiable, entity, contradictory, relation, and subjective. A response is considered hallucinatory if any of these error types are present (see Appendix A for details of category descriptions).
- SelfCheckGPT (Manakul et al., 2023) estimates the likelihood of hallucination by comparing a model's response to its alternative generations. Responses scoring above 0.5 are flagged as hallucinations, so as to follow existing standard practice.
- We use GPT-4-Turbo as an LLM-based judge to assess hallucination (Cheng et al., 2023) presence by cross-referencing the question, model output, and retrieved documents.

2.2 Results and Analyses

Finding 1: While RAG improves answer accuracy, it reduces the performance of hallucination detection. In Table 1, both vanilla RAG and InstructRAG well outperform the non-RAG baseline

	HotpotQA			2WikiMQA			ASQA		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
FAVA									
w/o RAG	0.668	0.774	0.717	0.667	0.928	0.776	0.453	0.721	0.557
Vanilla RAG	0.450	0.632	0.526	0.535	0.717	0.613	0.203	0.667	0.311
InstructRAG	0.417	0.705	0.524	0.642	0.763	0.697	0.201	0.688	0.311
SelfCheckGPT									
w/o RAG	0.726	0.828	0.774	0.705	0.882	0.784	0.701	0.549	0.616
Vanilla RAG	0.440	0.698	0.540	0.604	0.820	0.696	0.377	0.333	0.354
InstructRAG	0.634	0.399	0.490	0.741	0.591	0.658	0.633	0.460	0.533
GPT-4-turbo									
w/o RAG	0.768	0.953	0.851	0.687	0.977	0.807	0.471	0.988	0.637
Vanilla RAG	0.658	0.623	0.640	0.630	0.730	0.676	0.214	0.833	0.341
InstructRAG	0.516	0.804	0.629	0.618	0.991	0.761	0.251	0.962	0.398

Table 2: Comparison of three hallucination detection methods on three datasets, evaluating their effectiveness in detecting hallucinations from answers generated by different modeling approaches.

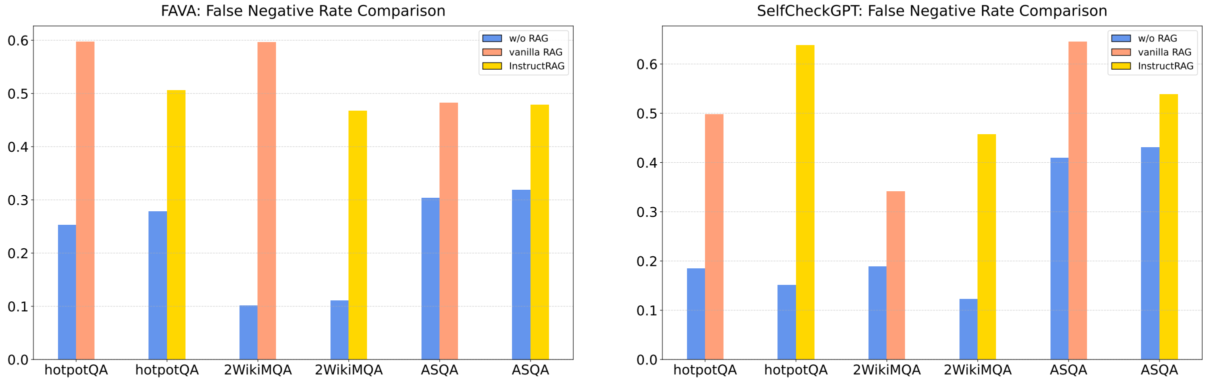


Figure 1: False negative rate (FNR) comparison between non-RAG and RAG approaches across three datasets, evaluated under two hallucination detection frameworks.

in answer accuracy across all datasets. This proves the efficacy of RAG in improving the factuality of LM outputs, particularly on multi-hop reasoning and ambiguous question answering tasks.

However, Table 2 also highlights a critical issue: hallucination detection methods perform worse when applied to RAG models. Across three hallucination detectors (FAVA, SelfCheckGPT, and GPT-4-Turbo), the F1 score is reduced by a large margin when detecting RAG models. For example, on HotpotQA, the F1 score is reduced from 0.717 to 0.526 when augmenting LMs with vanilla RAG. These results suggest that while RAG increases factuality on average, it also increases the likelihood that hallucinations go undetected.

To further quantify this phenomenon, we calculate the false negative rate (FNR), defined as the proportion of hallucinated responses that are not successfully identified by the detection system. As illustrated in Figure 1, RAG-based methods consistently exhibit higher FNRs across all three datasets and both hallucination detection frameworks evaluated. Specifically, the FNR increases from 23.8% in the non-RAG setting to as high as 52.0% with RAG. This substantial rise indicates that hallucinations embedded within RAG-generated content are significantly more difficult for automated detection frameworks to recognize. Notably, this remains true even when the RAG-generated answers are objectively incorrect, highlighting a critical challenge in identifying subtle or plausibly phrased hallucinations that are reinforced by retrieved evidence.

Finding 2: RAG increases the confidence of hallucination detection models during identification. To better evaluate RAG influence on hallucination detection, we focus on a specific subset of questions: those for which both RAG and non-

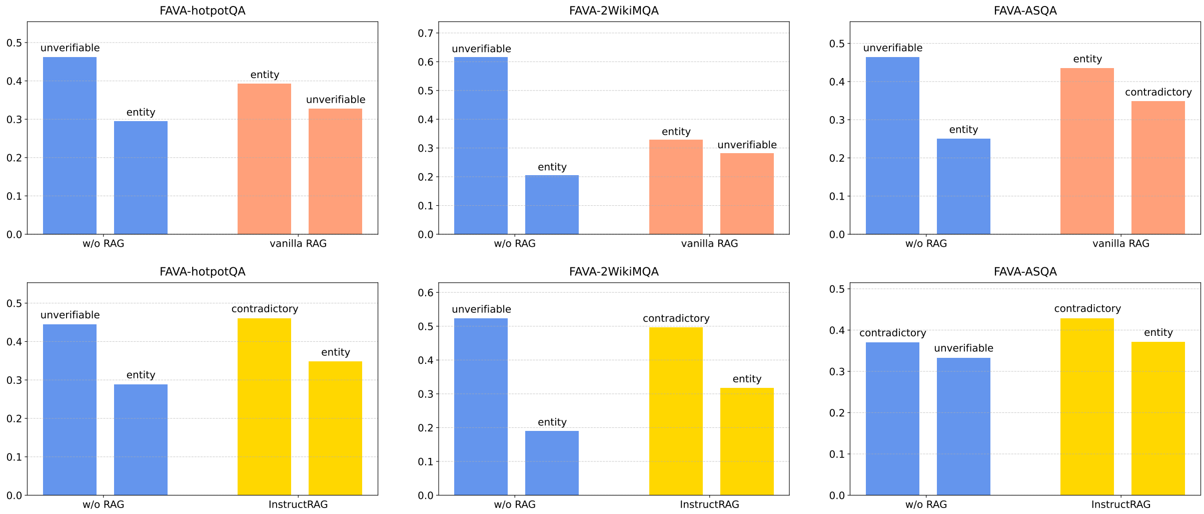


Figure 2: Breakdown of the two most common hallucination types and their proportions in asymmetric cases, where only one of the two methods is flagged as hallucinated by the FAVA framework.

Error Type	**HotpotQA**		**2WikiMQA**		**ASQA**	
	w/o RAG	Vanilla RAG	w/o RAG	Vanilla RAG	w/o RAG	Vanilla RAG
Invented	115	15	116	3	1	0
Unverifiable	**327**	101	**1024**	87	**13**	0
Entity	208	**121**	340	**101**	7	**10**
Contradictory	24	32	60	66	6	8
Relation	27	38	120	52	1	4
Subjective	7	2	2	2	0	1
	HotpotQA		**2WikiMQA**		**ASQA**	
	w/o RAG	InstructRAG	w/o RAG	InstructRAG	w/o RAG	InstructRAG
Invented	71	18	107	26	2	4
Unverifiable	**262**	26	**710**	2	**9**	0
Entity	170	192	257	135	6	13
Contradictory	39	**254**	122	**211**	**10**	**15**
Relation	40	59	157	46	0	2
Subjective	7	2	1	4	0	0

Table 3: Comparison of error types across three datasets under different RAG settings. For each dataset, two separate comparisons are shown: w/o RAG vs. Vanilla RAG (top) and w/o RAG vs. InstructRAG (bottom). Bold values represent the most dominant type of hallucination within each pairwise comparison under FAVA framework.

RAG methods produced incorrect answers, yet only one approach successfully flagged the hallucination. We conduct analyses on this subset. As illustrated in Figure 2, when RAG is not employed, the FAVA framework lacks access to external evidence and often classifies uncertain responses as "unverifiable." In contrast, with RAG integrated, the generated answers partially reflect retrieved content, even if the overall answer remains incorrect. This partial alignment can lead the hallucination detector to misjudge the response as correct due to superficial consistency with external sources. This effect is evident in the HotpotQA dataset, where the proportion of errors identified in the non-RAG setting versus the vanilla RAG setting is 0.699 to 0.301, and for non-RAG versus InstructRAG, the ratio is 0.587 to 0.413. Moreover, the incorporation of RAG substantially increases the detection model's confidence, reducing the frequency of "unverifiable" labels while prompting more definitive—yet poten-

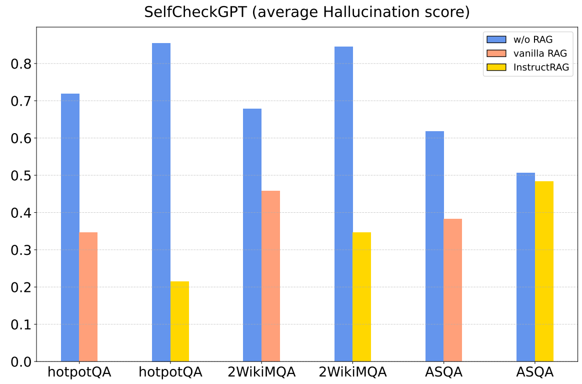


Figure 3: SelfCheckGPT average hallucination scores in cases where the generative model fails to produce correct answers under both non-RAG and RAG settings.

tially inaccurate—classifications such as "contradictory" or "entity." Similar trends are observed across the 2WikiMQA and ASQA datasets. Full numerical results and illustrative examples are available in Table 3 and Appendix B.

Finding 3: RAG enhances output consistency, potentially misleading hallucination detection frameworks. As shown in Figure 3, when the model produces incorrect answers both with and without RAG, SelfCheckGPT tends to assign higher hallucination scores to non-RAG outputs. Without RAG, when the model lacks sufficient knowledge, its responses often appear more erratic or inconsistent (Mündler et al., 2023; Du et al., 2023), leading to higher hallucination estimates. In contrast, RAG introduces external information—regardless of factual correctness—that can increase lexical and semantic coherence. As a result, the model's outputs appear more consistent, even if they remain factually incorrect. This surface-level consistency may falsely signal reliability to detectors like SelfCheckGPT. For instance, in Table 4, the incorrect answer "Locarno", being the premiere location of the film, is mentioned in the reference information with more detailed descriptions, which likely caused SelfCheckGPT to mistakenly associate it with the "The director's workplace." Our findings reveal the vulnerability of consistency-based hallucination detection methods. More case studies are in Appendix C.

3 Evading Real Human Detection

In Sec. 2, we showed that RAG negatively affects automatic hallucination detectors on standard benchmarks. We now examine its impact on real users with production models. Specifically, we first collect a production dataset of live user queries to LMs. For each query, we generate two responses from the Deepseek-R1 model—one without RAG and one with its default RAG setting. Human annotators then judge each response's authenticity, allowing us to assess how RAG influences users' ability to detect hallucinations.

3.1 Experimental Setup

Datasets. To enhance the systematicity and real-world relevance of our evaluation, we incorporate a dataset derived from a real-world industrial setting. Specifically, the dataset was constructed from the company's internal database by randomly sampling 712 user queries from monthly search logs, each with over 5,000 impressions and clearly defined question intent. These queries reflect authentic user behavior in online environments, where accurate and time-sensitive information is critical. Many of them require accessing up-to-date documents or aggregating information across multiple sources—making the dataset well-suited for evaluating both the utility and potential risks of RAG in real-world deployments.

Answer Generation & Metrics. Given the widespread deployment of the DeepSeek-R1 model (DeepSeek-AI et al., 2025) across various applications, we conduct experiments using both the non-RAG and RAG variants of the model to generate two distinct sets of answers for each question. To evaluate these responses, we enlist experienced annotators who have been deeply involved in data labeling tasks within our organization. Each annotator received a comprehensive annotation guideline and was instructed to assess the factual correctness of the model-generated answers by consulting multiple trusted information sources, such as search engines and mobile applications. Detailed annotation protocols and accuracy evaluation templates are provided in Appendix E.

Hallucination Detection & Metrics. For the industrial dataset, we adopt a rigorous annotation protocol for hallucination detection. Two independent groups, each with two well-trained annotators, manually reviewed model outputs—one group for the non-RAG version of DeepSeek-R1 and the other for the RAG version. To reflect real-world usage, annotators for non-RAG responses were given only the question and the model output, while those for RAG responses also received the retrieved docu-

query	Where does the director of film Man At Bath work at?
groundtruth	Cahiers du cinéma
predict w/o RAG	The National Film and Television School
predict samples w/o RAG	[‘National Film and Television School’, ‘The University of York’, ‘Filmarchiv Austria’, ‘Sundance Institute’, ‘Cinémathèque française’]
hallucination score w/o RAG	0.8
predict w Vanilla RAG	Locarno International Film Festival in Switzerland
predict samples w Vanilla RAG	[‘He was a director at Locarno International Film Festival in Switzerland’, ‘Locarno International Film Festival in Switzerland’, ‘Christophe Honoré works at Locarno International Film Festival in Switzerland’, ‘Locarno International Film Festival in Switzerland and in cinemas on 22 September 2010’, ‘Christophe Honoré works at Locarno International Film Festival in Switzerland.’]
hallucination score w Vanilla RAG	0.0

Table 4: RAG misleads SelfCheckGPT by repeatedly mentioning information that is related to the question but fails to genuinely address it.

Method	Pre.	Rec.	FNR	F1
DS w/o RAG	0.446	0.269	0.731	0.336
DS w RAG	0.386	0.215	0.785	0.276

Table 5: The annotators’ assessment of hallucination in the answers generated by both RAG and non-RAG versions of Deepseek-R1.

ments. To simulate typical user behavior, annotation time is capped at 2 minutes per response and 1 minute per document. This setup balances annotation quality with practical constraints, improving the realism and reliability of our evaluation. Detailed rules and templates are in Appendix F.

3.2 Results & Analyses

An analysis of human annotations on the production dataset reveals that errors are more frequently identified in responses from the non-RAG of DeepSeek-R1. Specifically, the FNR for the non-RAG DeepSeek-R1’s responses decreases by 5.4% compared to the RAG version (see Table 5 for quantitative results). This is largely attributed to two factors: (1) the non-RAG model lacks access to up-to-date external knowledge, leading to outdated or contextually irrelevant content—especially for time-sensitive queries; and (2) in the absence of retrieved evidence, the model tends to express greater uncertainty, often resorting to speculative or assumptive language. These cues increase annotators’ skepticism and make inaccuracies more apparent. We list several typical cases in Appendix G.

In contrast, when evaluating responses from the RAG version of DeepSeek-R1, annotators frequently place trust in the provided reference materials and are more likely to accept the response as correct—despite the possibility that the references themselves may contain errors. In such cases, only clearly flawed synthesis or misinterpretation of the reference content prompts annotators to flag a response as incorrect. However, such overt mistakes are relateively uncommon in state-of-the-art models. Thus, annotators face greater difficulty detecting subtle errors in the RAG version’s responses.

4 Persisting Through Hallucination Mitigating Training

Our previous experiments show that RAG significantly hinders hallucination detection for both automatic detectors and human users, across benchmarks and real-world queries. But what if developers fine-tune LMs to reduce hallucinations using these detectors? Will this weak supervision truly align the models—or will LMs learn to exploit detector flaws, making the problem even worse?

To answer this question, we analyze this question by optimizing LMs against automatic hallucination detectors with two widely adopted preference learning methods, DPO (Rafailov et al., 2023) and

Method	HotpotQA			2WikiMQA		
	Precision	Recall	F1	Precision	Recall	F1
base model + Vanilla RAG	0.440	0.698	0.540	0.604	0.820	0.696
DPO model + Vanilla RAG	0.604	0.349	0.442	0.665	0.418	0.513
KTO model + Vanilla RAG	0.579	0.295	0.391	0.653	0.248	0.359

Table 6: The hallucination detection performance between base model and preference learned models in SelfCheckGPT framework.

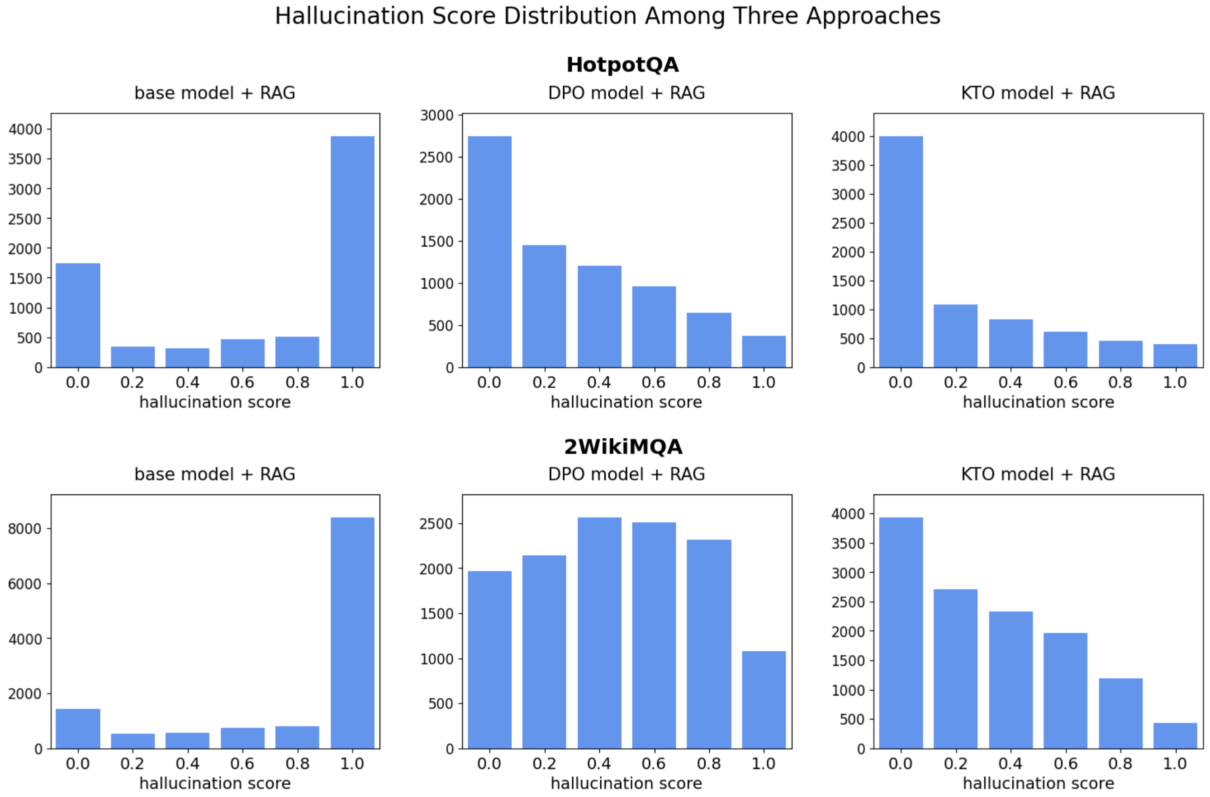


Figure 4: The distribution of hallucination scores obtained through SelfCheckGPT for responses generated by three approaches on the HotpotQA and 2WikiMQA test sets.

KTO (Ethayarajh et al., 2024). Specifically, for each query, we sample multiple outputs from RAG models, use hallucination detectors to rank them, thereby constructing a preference dataset. Then, we use preference learning methods to fine-tune LMs, aiming to mitigate the hallucination score according to the detectors. After training, we compare the hallucination detection performance on vanilla RAG models and the fine-tuned RAG models.

4.1 Experimental Setup

We conduct experiments on the HotpotQA and 2WikiMQA using the LLaMA-3-8B-Instruct model (Dubey et al., 2024), combined with a vanilla RAG pipeline to generate responses grounded in retrieved evidence. To assess hallucinations, we apply SelfCheckGPT to each response, labeling whether it contains factual errors. These annotated examples are then used to train preference-based models via: Direct Preference Optimization (DPO) (Rafailov et al., 2023), which aligns model outputs with preferred responses based on pairwise comparisons, and Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), which models human-like biases in decision-making during preference learning.

For KTO, we construct a binary classification dataset based on SelfCheckGPT scores: responses scoring ≤ 0.1 were labeled as non-hallucinatory ("true"), while those scoring ≥ 0.9 were labeled as hallucinatory ("false"). For DPO, we leverage the inherent stochasticity in LLM outputs to identify pairs of semantically inconsistent responses to the same question with divergent hallucination scores. Within each pair, the response with the lower score was labeled as "chosen," and the one with the higher score as "rejected." Illustrative training data templates are provided in Appendix H.

4.2 Results & Analyses

After training, we re-evaluate the preference-learned models—integrated with RAG—using SelfCheckGPT. In Table 6, both DPO- and KTO-trained models surprisingly underperforms the base model in terms of hallucination detection scores.

A closer examination of the training data reveals substantial annotation noise. In the DPO datasets (HotpotQA and 2WikiMQA), 24.7% and 21.3% of the "rejected" responses are actually better than their "chosen" counterparts, indicating a misalignment between hallucination scores and semantic quality. For the KTO datasets, mislabeling rates reaches 52.6% and 59.1%, respectively, suggesting that the hallucination detection framework struggles to reliably annotate RAG-generated responses. This labeling noise ultimately degraded the effectiveness of preference learning and limited the model's ability to improve generation quality.

Interestingly, in Figure 4, responses from the preference-learned models tend to receive lower hallucination scores from SelfCheckGPT compared to the base model. While this indicates improved consistency, it does not necessarily reflect better factual accuracy. Instead, the enhanced fluency and internal coherence may obscure subtle errors, confusing consistency-based detectors and reducing their ability to flag hallucinations.

5 Related Work

5.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances traditional generative models by incorporating external knowledge retrieval into the generation pipeline. This approach mitigates the limitations of parametric memory and improves factual accuracy. Recent advancements have further extended the capabilities of RAG. LongRAG (Jiang et al., 2024) leverages long-context language models to more effectively coordinate retrieval and generation across extended textual inputs. GraphRAG (Edge et al., 2024) constructs hierarchical knowledge graphs from retrieved corpora, enabling structured information integration and improved summarization. InstructRAG (Wei et al., 2024) introduces self-synthesized rationales to guide the language model in learning a denoising process, thereby enhancing its interpretability and robustness. While these methods primarily reduce hallucinations and improve answer quality, they often overlook a crucial implication: RAG can inadvertently complicate hallucination detection by increasing surface-level coherence. In this work, we highlight this underexplored challenge and present a systematic analysis to uncover and address the vulnerabilities RAG.

5.2 Hallucination Detection Approaches

Hallucination detection seeks to identify factual inaccuracies in generated text. SelfCheckGPT (Manakul et al., 2023) assesses intra-model consistency by sampling multiple generations and measuring semantic agreement among them. FAVA (Mishra et al., 2024) introduces a fine-grained taxonomy of hallucination types spanning six hierarchical levels, along with a framework for automatic hallucination classification. Although both methods demonstrate strong performance, they do not account for the influence of retrieval-augmented inputs. Specifically, they haven't examined how RAG's added coherence might hide hallucinations or impact detection accuracy. We close this gap by evaluating SelfCheckGPT and FAVA under both RAG and non-RAG settings, revealing how retrieval affects hallucination detection.

6 Conclusion

In this paper, we have conducted a comprehensive analysis of why hallucination detection becomes less effective under Retrieval-Augmented Generation (RAG). Leveraging both automated and manual evaluation across multiple challenging datasets, we reveal that RAG systematically degrades the performance of hallucination detectors, making falsehoods harder to identify. Furthermore, we evaluate existing hallucination detection frameworks and find that their generated labels not only fail to improve the reliability of fine-tuned generative models, but may also introduce misleading supervision. These findings expose critical blind spots in current detection paradigms when applied to RAG-enhanced systems. We call for the development of more robust hallucination detection mechanisms designed for RAG, with future work focusing on enhancing both detection accuracy and the trustworthiness of model outputs in real-world applications.

Limitations

To manage computational costs, our experiments primarily rely on the LLaMA-3-8B-Instruct model and its supervised fine-tuned (SFT) variants for answer generation and hallucination detection. While this choice offers practical advantages, the limited capacity of these models may constrain the generalizability of our findings. Future studies could incorporate both open- and closed-source models of varying scales to improve the robustness and comprehensiveness of the conclusions.

Additionally, our evaluation focuses on three representative hallucination detection frameworks: FAVA, SelfCheckGPT and GPT-4-turbo based. Although these methods are widely used, our study does not cover the full landscape of existing approaches. Several promising alternatives remain unexplored. For instance, Varshney et al. and Luo et al. proposed a method that leverages internal model states to detect hallucinations, offering a deeper understanding of LLM behavior. Similarly, FacTool (Chern et al., 2023) introduced a unified framework that equips LLMs with external tool usage for evidence collection, which could improve factuality assessment. Future work may benefit from integrating or benchmarking such methods to advance the development of more effective hallucination detection techniques.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. In *AAAI Conference on Artificial Intelligence*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *ArXiv*, abs/2310.03368.

Ethan Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *ArXiv*, abs/2305.14325.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *ArXiv*, abs/2404.16130.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *ArXiv*, abs/2402.01306.

Xanh Ho, A. Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *ArXiv*, abs/2011.01060.

Vojtvech Hudevcek and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *SIGDIAL Conferences*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *ArXiv*, abs/2406.15319.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *ArXiv*, abs/2401.06855.

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

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv*, abs/2309.09558.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *ArXiv*, abs/2204.06092.

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*, abs/2307.03987.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *ArXiv*, abs/2402.19473.

A The Explanation of Six Fine-grained Categories of Hallucination Errors

error type	type explanation
invented	The LM generates an entirely fabricated entity that doesn't exist based on world knowledge.
unverifiable	The LM output contains facts, but no retrieved evidence from the web can directly support or contradict the fact.
entity	An entity in a statement is incorrect and changing that single entity can make the entire sentence factually correct.
contradictory	A full statement entirely contradicts relevant evidence from the web, and cannot be solved via phrase-level edits.
relation	A semantic relationship (e.g., verbs, prepositions, or adjectives) in a statement is incorrect.
subjective	Existing entities lack universal validity.

Table 7: Six types of fine-grained error mentioned in FAVA (Mishra et al., 2024)

B FAVA Case Study

We demonstrate two cases within the FAVA framework: both the non-RAG and RAG-assisted responses were incorrect, but the hallucination detection framework only identified the error in the non-RAG response.

query	Which Australian city founded in 1838 contains a boarding school opened by a Prime Minister of Australia and named after a school in London of the same name.
groundtruth	Marion, South Australia
edited output w/o RAG	Edited: Toowoomba, an Australian city founded in 1838, contains a boarding school <unverifiable>opened by a Prime Minister of Australia</unverifiable> and named after a school in London of the same name.
edited output w/ Vanilla RAG	Edited: Adelaide, an Australian city founded in 1838, contains a boarding school opened by a Prime Minister of Australia and named after a school in London of the same name.
misleading retrieved docs	Westminster School is... located at Marion, South Australia, 12 km south of Adelaide. Westminster School, Adelaide: Westminster School is an independent...
analysis	The description of Westminster School mentions "located in Marion, South Australia, 12 km south of Adelaide." The framework might prioritize selecting the more prominent and well-known geographical name ("Adelaide") as the answer. The reference information starts with "Westminster School, Adelaide," and the framework might have simply assumed this to be the city where the school is located. The prominent association between Westminster School and Adelaide, combined with the lower frequency of mentions of Marion, might have led the framework to overlook the specific accuracy required by the question.

Table 8: Case 1 from HotpotQA.

query	When was the British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based born?
groundtruth	7 January 1936
edited output w/o RAG	Edited: The <unverifiable>British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based</unverifiable> was born in 1951.
edited output w Vanilla RAG	Edited: The British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based was born in 1867.
misleading retrieved docs	George Slythe Street (18 July 1867 – 31 October 1936) was a British critic, journalist and novelist. James Hogg (1770 – 21 November 1835) was a Scottish poet, novelist and essayist who wrote in both Scots and English.
analysis	The reference information mentions that the film Here We Go Round the Mulberry Bush was adapted from the novel of the same name by Hunter Davies, but it does not specifically provide Hunter Davies’ birth year. As a result, the framework might assume that Hunter Davies’ birth year is not covered in the reference information and attempt to find other content containing similar terms like “British author” or descriptions related to novels, ultimately leading to an incorrect association with another author. The reference information includes details about several unrelated British authors (e.g., G. S. Street, James Hogg), including their birth years and biographical details. Notably, G. S. Street is explicitly noted as being born in 1867, and his description as a “British critic, journalist, and novelist” might superficially align with the question’s criteria (“British author”). The framework may have simply matched the keyword “British author” and, combined with G. S. Street’s prominent birth year of 1867, incorrectly assumed this to be the correct answer.

Table 9: Case 2 from HotpotQA.

C SelfCheckGPT Case Study

We demonstrate two cases within the SelfCheckGPT framework: both the non-RAG and RAG-assisted responses were incorrect, but the hallucination detection framework only identified the error in the non-RAG response.

query	What nationality is the director of film Name The Man?
groundtruth	Swedish
predict w/o RAG	American
predict samples w/o RAG	[‘Australian’, ‘American’, ‘Austrian’, ‘Australian’, ‘French’]
hallucination score w/o RAG	0.8
predict w Vanilla RAG	American
predict samples w Vanilla RAG	[‘American’, ‘American’, ‘American’, ‘American’, ‘American’]
hallucination score w Vanilla RAG	0.0
misleading retrieved docs	Hubert Cornfield: Hubert Cornfield(February 9, 1929 — June 18, 2006) was a film director in Hollywood. He was born in Istanbul, Turkey, and died in Los Angeles, California. Gregory La Cava: Gregory La Cava(March 10, 1892 – March 1, 1952) was an American film director of Italian descent best known for his films of the 1930s, including" My Man Godfrey" and" Stage Door", which earned him nominations for Academy Award for Best Director"
analysis	In the reference information, there are many other descriptions related to "directors," such as: Hubert Cornfield being described as "a film director in Hollywood," and Gregory La Cava as "an American film director of Italian descent." These details may have been overly prominent despite being irrelevant background information, thereby misleading the hallucination detection framework. The framework might have simply associated "director" with "American" or "Hollywood" and thus considered the answer "American" to be reasonable. The key information in the reference material related to the question is the nationality of the film director Victor Sjöström, rather than the background of the film itself or descriptions of other directors. However, the framework may have failed to focus its attention on the correct association between Victor Sjöström and "Swedish," instead showing a preference for and being distracted by secondary, irrelevant information.

Table 10: Case 1 from 2WikiMQA.

query	Who is the mother of the director of film Brenda Brave?
groundtruth	Käbi Laretei
predict w/o RAG	Liz Lake
predict samples w/o RAG	[‘Judy Cornwell’, ‘Sarah Townsend’, ‘Terry Braunstein’, ‘Jill Gevargizian’, ‘Judit Elek’]
hallucination score w/o RAG	1.0
predict w Vanilla RAG	Astrid Lindgren
predict samples w Vanilla RAG	[‘Astrid Lindgren’, ‘Astrid Lindgren’, ‘Astrid Lindgren’, ‘Astrid Lindgren’, ‘Astrid Lindgren’]
hallucination score w Vanilla RAG	0.0
misleading retrieved docs	Brenda Brave (original title: Kajsa Kavat) is a 1989 Swedish film directed by Daniel Bergman and based on the novel "Brenda Brave Helps Grandmother" by Astrid Lindgren.
analysis	The framework may not have sufficient ability to distinguish the following when verifying the answer: who the director of Brenda Brave is (Daniel Bergman) and that Lindgren’s role is limited to being the source of the film’s story rather than connected to the director’s background. The framework might have simply performed a surface-level match between the keyword “Brenda Brave” from the question and prominent references in the context, ignoring deeper semantic tracking and reasoning. In the reference information, Daniel Bergman’s mother (Käbi Laretei) is not explicitly linked to Brenda Brave. The framework needs to traverse multiple layers of data: starting from Brenda Brave to identify its director as Daniel Bergman, and then deducing from Daniel Bergman’s background information that Käbi Laretei is his mother. The framework’s failure to perform this cross-layered logical reasoning led it to rely more heavily on the prominent mention of Astrid Lindgren in the context.

Table 11: Case 2 from 2WikiMQA.

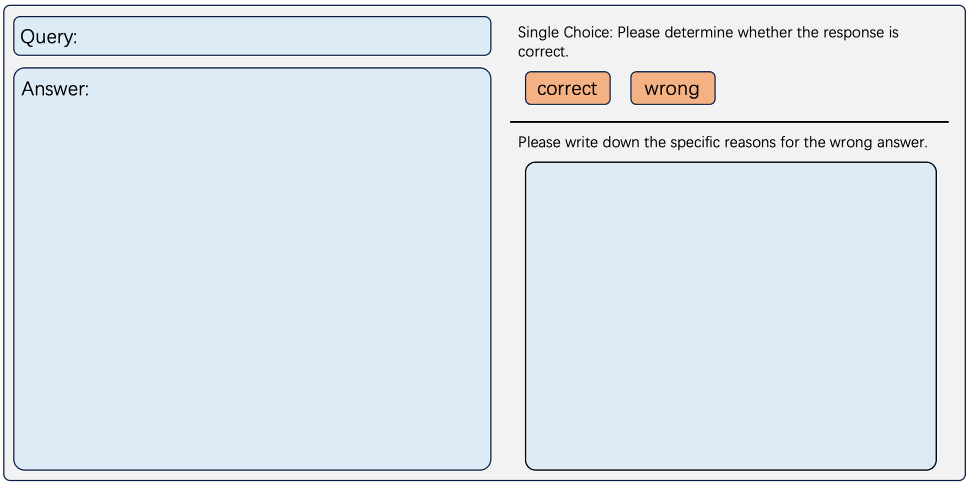
D Manual annotation cost

During the manual data annotation process in Section 3, the cost for annotating the accuracy and hallucination of each answer is $2 per item.

E Manual Accuracy Annotation

Accuracy Annotation Requirements	
Accuracy judgment	Objective facts mentioned in the responses (especially numbers, dates, and procedures) must be verified through websites, mobile applications, or practical operations to ensure their accuracy. If the answer to a question has regional and temporal requirements, the response must meet the corresponding conditions.
Ways to obtain reference information	Search engines: Baidu, Bing, Google; Mobile applications: Alipay, rednote, WeChat, etc.

Table 12: The rules of accuracy annotation.



Single Choice: Please determine whether the response is correct.

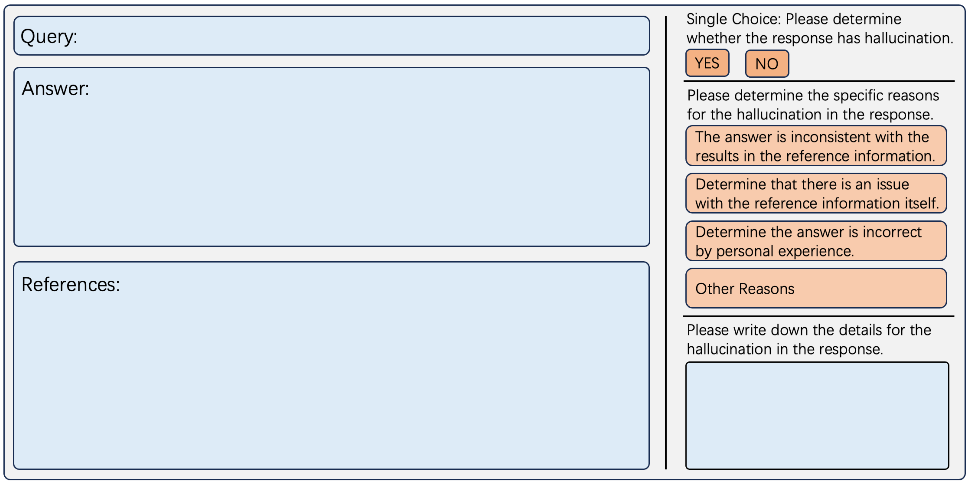
Please write down the specific reasons for the wrong answer.

Figure 5: The template of accuracy annotation.

F Manual Hallucination Annotation

Hallucination Annotation Requirements	
Hallucination judgment	DeepSeek-R1 with RAG: First, check whether the answer is consistent with the results in the reference; second, assess based on personal experience whether there are issues with the reference itself; finally, evaluate based on personal experience whether there are issues with the model's response. DeepSeek-R1 without RAG: Directly assess based on personal experience whether there are issues with the model's response.
Annotation time requirements	DeepSeek-R1 with RAG: The time spent evaluating the model's response should not exceed 2 minutes; the time spent reviewing each reference should not exceed 1 minute. DeepSeek-R1 without RAG: The time for evaluating the model's response should not exceed 2 minutes.

Table 13: The rules of hallucination annotation.

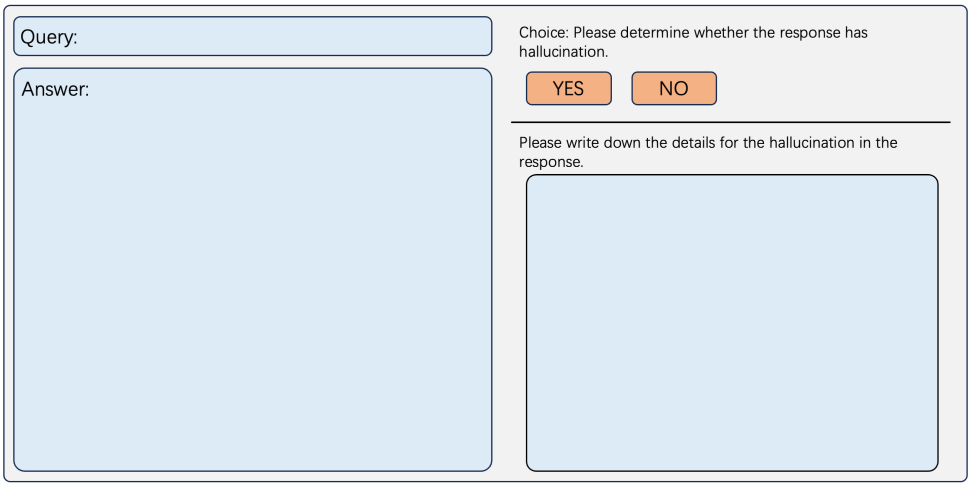

Query:
Answer:
References:
YES
NO
The answer is inconsistent with the results in the reference information.
Determine that there is an issue with the reference information itself.
Determine the answer is incorrect by personal experience.
Other Reasons

Single Choice: Please determine whether the response has hallucination.

Please determine the specific reasons for the hallucination in the response.

Please write down the details for the hallucination in the response.

Figure 6: The template of hallucination annotation (DeepSeek-R1 with RAG).



Choice: Please determine whether the response has hallucination.

Please write down the details for the hallucination in the response.

Figure 7: The template of hallucination annotation (DeepSeek-R1 without RAG).

G Human Annotation Case Study

We selecte two representative cases to illustrate that when RAG is not used, erroneous answers generated by the model are more likely to be noticed as problematic by annotators.

query	What is the annual fee for the Hangzhou Culture and Tourism Card?
answer by no-RAG DeepSeek-R1	Based on the currently available information, the annual fee for the Hangzhou Culture and Tourism Card (commonly referred to as the “Hangzhou Cultural and Tourism Benefit Card”) is 200 RMB per year. This card provides free or discounted access to various scenic spots and cultural venues in Hangzhou and surrounding areas. Additional Notes: 1. This fee is the publicly announced standard for 2024. If there are no adjustments to the policy, it will remain valid for 2025. 2. Certain groups (e.g., students/seniors) may be eligible for special discounts. It is recommended to check the latest policies via the “Hangzhou Citizen Card” app or the official website of the Culture and Tourism Bureau. 3. Once activated, the card must be used within its validity period; renewal is required after expiration. For application or detailed confirmation, you can use the “Hangzhou Citizen Card” app or contact the Citizen Card hotline at 96225.
accuracy label	False
analysis	Since the annotation task is being conducted in 2025, but the response is based on 2024, such policy-related questions have high requirements for timeliness. Therefore, the annotators believe that the response exhibits a significant degree of hallucination.

Table 14: Case 1 from real-world user query dataset.

query	Why is xiaoqiandai different from before?
answer by no-RAG DeepSeek-R1	Based on your location (Hangzhou) and the common adjustment directions for Alipay features, the following changes might occur to xiaoqiandai in the near future: 1. Adjustment of feature entry Alipay might relocate the entry point for xiaoqiandai from the homepage's "More" or "Finance" section to the "Yu'e Bao" submenu, or integrate it into the "Family Account" feature. Users may need to search for "xiaoqiandai" directly via the search bar. 2. Upgrade in usage rules Due to the 2024 "Financial Product Compliance Guidelines", the following changes might be introduced: Verification of minor identity information required when binding a sub-account Limits on single transfer amounts (e.g., daily $\leq$ 5,000 RMB) Extended processing time for fund withdrawals to T+1 3. Interface and interaction optimization Following the recent update to Alipay version 10.2, xiaoqiandai might: Add expenditure classification charts Support customizable saving goal progress reminders Simplify the hierarchy for managing permissions among family members Recommended actions: Open Alipay and search for "xiaoqiandai" to access the latest interface Check the "Message Center - Service Notifications" for official change announcements Contact customer service and input "xiaoqiandai upgrade" to obtain policy clarification documents (Note: If the feature has been discontinued, Alipay will usually provide guidance on alternative services. Please pay attention to account security and avoid performing operations through unofficial channels.)
accuracy label	False
analysis	Because there is no introduction of RAG, the LLM frequently uses words like "might" when answering questions it is unfamiliar with. Such hypothetical tone in its responses may lead annotators to believe that the model's answers are based on conjecture, thus making it easier to identify hallucinations.

Table 15: Case 2 from real-world user query dataset.

H Preference Learning Data Template

KTO data template

```
{
  “messages”: [
    {
      “content”: “Answer the given question, you can refer to the document provided. As an assistant, your task is to answer
                 the question based on the given knowledge. your question and reference knowledge are as follows.
                 knowledge: {knowledge}
                 Question: {question}
                 Please answer the question.”
      “role”: “user”
    },
    {
      “content”: “{answer}”
      “role”: “assistant”
    }
  ],
  “label” : true/false
}
```

Table 16: KTO data template used for training data generation.

DPO data template

```
{
  “conversations”: [
    {
      “from”: “human”
      “value”: “Answer the given question, you can refer to the document provided. As an assistant, your task is to answer
                 the question based on the given knowledge. your question and reference knowledge are as follows.
                 knowledge: {knowledge}
                 Question: {question}
                 Please answer the question.”
    }
  ],
  “chosen”: {
    “from”: “gpt”,
    “value”: “{chosen answer selected by SelfCheckGPT}”
  },
  “rejected”: {
    “from”: “gpt”,
    “value”: “{rejected answer selected by SelfCheckGPT}”
  }
}
```

Table 17: DPO data template used for training data generation