Retrieval-Augmented Language Models Evade Hallucination Detection

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) has become one of the most popular techniques thanks to its potential to mitigate hallucination, which attracts widespread attention across the research and industrial community. Despite its effectiveness, LLM-based RAG systems remain susceptible to hallucination. To detect hallucination in RAG, current work mainly focuses on hallucination from LLMs' internal knowledge with correct and sufficient context. However, the sycophancy problem in RAG hallucination detectors is underexplored. Specifically, the inclusion of references can create a veneer of authority that masks the underlying hallucinations, making them significantly more difficult to identify. To systematically investigate this phenomenon, this work first conducts comprehensive experiments with three RAG methods (without RAG, with vanilla RAG, and with InstructRAG) and three question-answering datasets. The results show that, compared to non-RAG, RAG increases the false negative rate of four widely adopted automatic hallucination detectors from 17.6% to 34.2% on average. Furthermore, this work investigates the impact of prevalent preference optimization algorithms (DPO and KTO) on RAG hallucination detection, revealing that RAG systems built on backbone models trained with these algorithms are more prone to evading hallucination detection. The dataset and code will be publicly available.

1 Introduction

Large **l**anguage **m**odels (LLMs) (Brown et al., 2020) have achieved remarkable success in text generation, demonstrating strong performance across a wide range of natural language tasks (Hudecek 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., 2023). This issue significantly undermines the reliability of LLMs, especially in real-world applications that require factual precision and trustworthiness. To reduce hallucination, **R**etrieval-**A**ugmented **G**eneration (RAG) (Lewis et al., 2020) has been widely adopted for LLMs. Specifically, RAG mitigates hallucination by incorporating the retrieved information into the generation process of LLMs, under the assumption that the retrieved information is sufficiently accurate.

Despite of the improved factual accuracy, RAG introduces an underexplored and valuable challenge: *RAG increases the difficulty of hallucination detection*, as shown in Figure 1. Specifically, 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. Thus, compared with hallucinations of LLMs without RAG, hallucinations of LLM-based systems are harder to identify and more likely to deceive hallucination detectors. This credibility-detectability paradox poses substantial risks in scenarios with high security requirements, such as healthcare (Kim et al., 2025), finance (Wu et al., 2023), and law (Li et al., 2025), and may contribute to subtle significant consequences. However, the challenge is underexplored.

To investigate the challenge, first, this work analyzes whether RAG may evade hallucination detectors. Specifically, in Sec. 2, comprehensive experiments are conducted with four automatic hallucination detection methods (FAVA (Mishra et al., 2024), SelfCheckGPT (Manakul et al., 2023), ReDeEP (Sun et al., 2025), and vanilla hallucination detection using GPT-4-turbo (OpenAI, 2023)) on three question answering datasets (HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and ASQA (Stelmakh et al., 2022)).

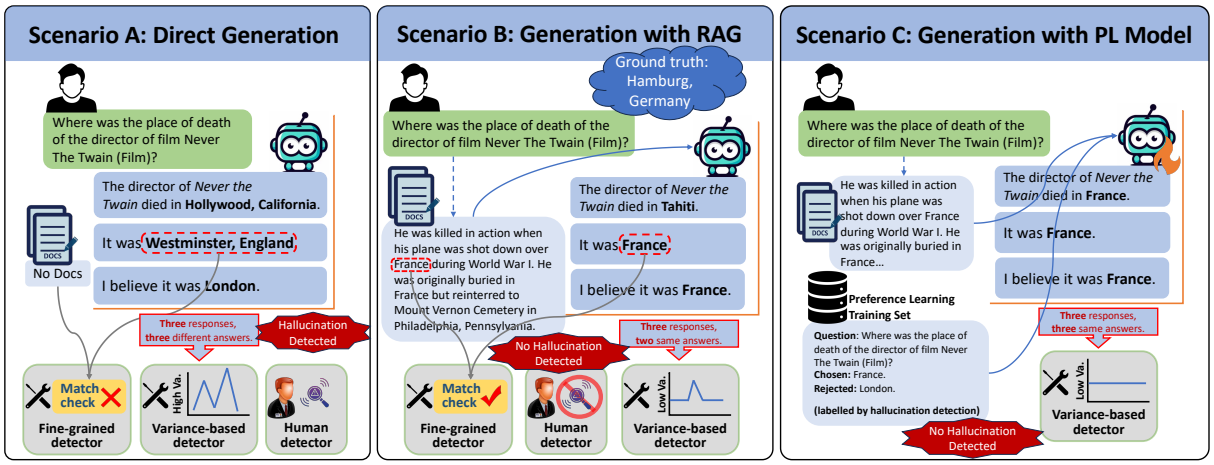

Figure 1: **Left panel:** Without RAG, the detectors (two automatic and one manual) can easily detect hallucinations in three inconsistent and entirely incorrect answers. **Middle panel:** With RAG, two answers become consistently wrong, evading the detectors. **Right panel:** After backbone LLM preference learning, reward hacking drives all three answers to be consistently wrong, further raising detection challenges.

Experimental results show that RAG increases the false negative rate of detectors from 17.6% to 34.2% on average.

Next, in Sec. 3, this work investigates the impact of RAG on manual detectors. Specifically, a real-world online user query dataset is collected, and human detectors are assigned to judge the hallucination of the RAG and non-RAG versions of DeepSeek-R1 (DeepSeek-AI et al., 2025). Experimental results show that RAG increases the false negative rate of human judges by 5.4%.

Furthermore, in Sec. 4, this work analyzes the impact of preference optimization methods which are often used to mitigate hallucination. Specifically, experienced hallucination detectors are employed to label good and bad responses, and then LLMs are trained on the preference dataset with two preference optimization techniques (DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024)). Experimental results show that hallucination detectors are invalid after LLMs' preference training, whose false negative rate detections increase by 53.3%.

The contributions of this work are summarized as follows:

- This work reveals a novel and critical challenge: existing RAG systems inadvertently degrade the performance of hallucination detectors. This finding underscores the necessity of reassessing current evaluation paradigms and developing more robust detection frameworks for RAG-enhanced environments.
- This work reveals that preference optimization algorithms further degrade the performance of hallucination detectors. This finding highlights an important, yet previously overlooked issue: while preference optimization is widely adopted to improve response quality and alignment, it may inadvertently amplify the detectors' vulnerability.
- Comprehensive experiments across multiple datasets and state-of-the-art detection methods consistently validate this challenge, underscoring an overlooked tension between retrieval-based mitigation strategies and hallucination detection capabilities. The dataset and code will be publicly available.

2 Evading Automatic Hallucination Detection

In this section, through evaluations employing hallucination detection frameworks on three benchmark datasets, we expose RAG's negative effects: it inadvertently masks subtle hallucinations that arise during answer generation, resulting in introducing new challenges in error identification.

2.1 Experimental Setup

Datasets. To evaluate the impact of RAG on hallucination detection, three benchmark datasets are utilized, 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

Method	HotpotQA		2WikiMQA		ASQA	
	Acc.	K	Acc.	K	Acc.	K
w/o RAG	32.5	-	32.5	-	56.5	-
vanilla RAG	59.0	10	**42.1**	10	**84.2**	3
InstructRAG	**61.8**	5	39.1	10	80.4	5

Table 1: Answer generation accuracy judged by GPT-3.5-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used. The results are presented in percentage (%).

documents. ASQA, on the other hand, emphasizes ambiguity resolution in open-domain questions. These datasets are chosen for their emphasis on questions that require information from multiple external sources, which presents an ideal testbed to examine the strengths and limitations of RAG-based generation.

Three LLMs are utilized as the backbone generators (Llama-3-8B-Instruct (Dubey et al., 2024), Qwen2.5-32B-Instruct (Yang et al., 2024)) and Qwen2.5-72B-Instruct with two settings (few-shot prompting, supervised fine-tuning). Detailed information about datasets and models can be found in the Appendix A. (Sec.,2 reports the experimental results with LlaMA-3-8B-Instruct. The results for Qwen2.5-32B-Instruct and Qwen2.5-72B-Instruct are provided in the Appendix D and Appendix E, respectively). We adopt three methods to generate responses:

- w/o RAG: a non-RAG baseline relying on the parametric knowledge of the language model.
- Vanilla RAG: using Contriever (Izacard et al., 2022) as a retriever model and adding retrieved passages as extra input to prompt LM.
- InstructRAG (Wei et al., 2025): leveraging instruction-tuned language models to generate self-synthesized denoising rationales, enabling explicit correction of noisy retrieved information and improving answer accuracy.

Hallucination Detection. The following hallucination detection frameworks are utilized:

- FAVA (Mishra et al., 2024) identifies hallucinations by classifying response errors into six fine-grained categories. It takes the model response and the retrieved documents as input. A response is considered hallucinatory if any of the error categories are present (see Appendix B for details of the categories).
- SelfCheckGPT (Manakul et al., 2023) estimates the likelihood of hallucination by comparing multiple independently generated responses from a LLM. Outputs from the framework with a score exceeding 0.5 are flagged as hallucinations.
- ReDeEP (Sun et al., 2025) detects hallucinations in RAG models by decoupling the utilization of external context and parametric knowledge within LLMs.
- LLM-as-a-judge uses SOTA LLMs, like GPT-4-Turbo, to determine the presence of hallucinations (Cheng et al., 2023) by cross-referencing the question, the model's output, and the retrieved documents (for detailed hallucination detection prompts, see the Appendix C).

Metric. To get ground truth labels of hallucination detections, we take query, model response and reference answer as the input, and use GPT-3.5-turbo (Ye et al., 2023) to judge whether the model response is aligned with the reference answer. This LLM-based judgment overcomes the shortcomings of string-matching metrics, which often overestimate correctness due to shallow token overlap or partial relevance.

In our scenario, *precision* is calculated as the ratio of correctly identified hallucination instances to the total number of hallucinations flagged by the detection framework; while *false negative rate (FNR)* is defined as the proportion of hallucinated responses that are not successfully identified by the detection system.

2.2 Results

Finding 1: RAG improves answer accuracy but paradoxically makes hallucination harder to detect, as its outputs become more convincing to detectors. In Table 1, both vanilla RAG and InstructRAG well outperform the non-RAG baseline 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 highlights a critical issue: hallucination detection methods perform worse when applied to RAG models. Across four hallucination detectors, the F1 score is reduced by a large margin when detecting answers generated by RAG models. For example, on HotpotQA, the F1 score is reduced

Method	RAG setting	HotpotQA				2WikiMQA				ASQA			
		Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1
FAVA	w/o RAG	66.8	77.4	22.6	**71.7**	66.7	92.8	7.2	**77.6**	45.3	72.1	27.9	**55.7**
	Vanilla RAG	45.0	63.2	36.8	52.6	53.5	71.7	28.3	61.3	20.3	66.7	33.3	31.1
	InstructRAG	41.7	70.5	29.5	52.4	64.2	76.3	23.7	69.7	20.1	68.8	31.2	31.1
SelfCheckGPT	w/o RAG	72.6	82.8	17.2	**77.4**	70.5	88.2	11.8	**78.4**	70.1	54.9	45.1	**61.6**
	Vanilla RAG	44.0	69.8	30.2	54.0	60.4	82.0	18.0	69.6	37.7	33.3	66.7	35.4
	InstructRAG	63.4	39.9	60.1	49.0	74.1	59.1	40.9	65.8	63.3	46.0	54.0	53.3
ReDEeP	w/o RAG	65.5	80.4	19.6	**72.2**	69.5	92.0	8.0	**79.2**	50.5	64.3	35.7	**56.6**
	Vanilla RAG	58.3	74.8	25.2	65.5	72.6	84.7	15.3	78.2	35.2	37.8	62.2	36.5
	InstructRAG	52.4	46.5	53.5	49.3	67.6	50.4	49.6	57.8	56.4	50.3	49.7	53.2
LLM-as-a-judge	w/o RAG	76.8	95.3	4.7	**85.1**	68.7	94.4	5.6	**79.5**	47.1	94.5	5.5	**62.9**
	Vanilla RAG	65.8	62.3	37.7	64.0	63.0	73.0	27.0	67.6	21.4	83.3	16.7	34.1
	InstructRAG	51.6	80.4	19.6	62.9	61.8	96.7	3.3	75.4	25.1	91.0	9.0	39.3

Table 2: Results of four hallucination detection methods on three datasets with answers generated by different approaches. With RAG for answer generation, the performance of hallucination detection substantially decreases.

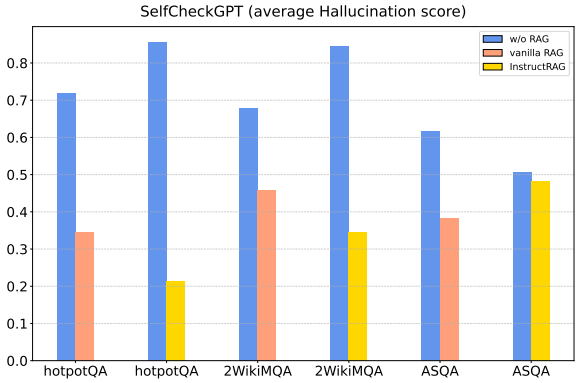

Figure 2: The hallucination detector's performance on examples where both methods generate errors. Compared to non-RAG models, the incorrect answers generated by RAG models achieve lower average hallucination scores when evaluated using the SelfCheckGPT framework, with Llama-3-8B-Instruct serving as the answer generation model.

from 0.717 to 0.526 when augmenting LMs with vanilla RAG under the FAVA detector. Also, RAG-based methods exhibit higher FNR in most cases. Specifically, FNR increases from 17.6% in the non-RAG setting to as high as 34.2% with RAG on average. These results indicate that hallucinations embedded within RAG-generated content are in general more difficult to catch.

One concern is that RAG may increase the FNR by reducing the total number of hallucinated outputs. To rule out this concern, we focus on a specific subset *where both RAG and non-RAG methods produce incorrect answers, but only the answer from one method is recognized as a hallucination.*

We find that the seemingly evidence-supported outputs of RAG models—particularly those exhibiting partial alignment with the retrieved content—are more likely to mislead hallucination detectors, making these incorrect responses less likely to be identified as hallucinations. This phenomenon is evident under the FAVA framework, which takes both the model response and the retrieved information as input. Comprehensive quantitative results and illustrative examples can be found in Appendix F and Appendix G, respectively.

Furthermore, when generating incorrect outputs, non-RAG models often exhibit erratic behavior due to uncertainty (Mündler et al., 2024; Du et al., 2024). In contrast, RAG models typically demonstrate greater consistency with the retrieved supporting information, creating an illusion of reliability through enhanced lexical and semantic coherence. Because some hallucination detectors primarily rely on consistency signals, RAG models are capable of exploiting this bias to mislead such detectors. For instance, as illustrated in Figure 2, SelfCheckGPT—a consistency-based detector—tends to assign higher hallucination scores to incorrect outputs generated by non-RAG models than to those produced by RAG models. Additional case studies are provided in Appendix H.

Finding 2: Retrieval noise significantly contributes to RAG-induced hallucinations and challenges detection. To further investigate the root causes of hallucinations in RAG-generated an-

Error source	Error proportion		
	HotpotQA	2WikiMQA	ASQA
LLM intrinsic	42.6	37.8	21.7
Retrieval noise	57.4	62.2	78.3

Table 3: Error source of hallucination in RAG-generated answers on benchmark datasets (aggregated across HotpotQA, 2WikiMQA, and ASQA).

swers, we manually annotate all outputs that are incorrectly generated by the RAG-based models yet missed by the hallucination detectors across the three benchmarks. The distribution of error sources is shown in Table 3. We observe that 61.4% of RAG hallucinations originate from retrieval noise, far exceeding model-intrinsic errors. This finding highlights a critical gap: existing hallucination detectors rarely evaluate the retrieved documents themselves; they treat all retrieved content as equally reliable, thereby rendering detection vulnerable to noisy retrieval.

The impact of such noise on detection is further confirmed by an ablation experiment on HotpotQA. When we remove the retrieved documents and provide only the question and the vanilla RAG-generated response from Qwen2.5-32B-Instruct to the GPT-4-Turbo hallucination detector, performance improves—precision, recall, and F1 score reach 0.503, 0.824, and 0.625, respectively, compared to 0.531, 0.579, and 0.554 when the retrieved documents are included. This result directly demonstrates that noisy retrieved information not only induces hallucinations but also actively interferes with the detection process.

The above finding exposes a critical gap: current detectors rarely assess the retrieved documents themselves and are thus easily misled by irrelevant or inconsistent content. To address this, we propose the idea of decoupled detection—first evaluating the trustworthiness of the retrieved evidence (its completeness, consistency, and relevance) and then verifying answer faithfulness at a claim-by-claim level, explicitly flagging partial support. When retrieval is insufficient, the detector should adaptively resort to consistency-based checks that do not rely on unreliable passages. This separation of retrieval-quality assessment from faithfulness verification can effectively mitigate the impact of retrieval noise on hallucination detection.

3 Evading Manual Detection

In Sec. 2, we show that RAG negatively affects automatic hallucination detectors on standard benchmarks. To comprehensively analyze the performance of RAG systems, we conduct extensive experiments in production settings using real-world user data. In specific, we involve hallucination detection experts to identify hallucinations and uncover critical blind spots in the deployment of RAG systems.

3.1 Experimental Setting

Datasets. To enhance the real-world relevance of our evaluation, we utilize a dataset derived from our company's internal database, consisting of 712 user queries randomly sampled from monthly search logs. The data collection and usage are approved by the company's internal review board, and all personally identifiable information (PII) is removed through an automated anonymization pipeline to ensure user privacy.

Given the widespread deployment of the DeepSeek-R1 model (DeepSeek-AI et al., 2025) across various applications and search engines, we conduct experiments using both the non-RAG and RAG variants of the model to generate two distinct sets of answers for each question. All queries have over 5,000 impressions and clearly defined question intents, reflecting authentic user behaviors across domains such as governance, consumer rights, and local services. The detailed dataset statistics and specific answer generation prompt are in the Appendix I and Appendix J, respectively.

Human Evaluators Recruitment. We enlist experienced annotators who have been deeply involved in data labeling tasks within our organization. These individuals are university graduates with foundational knowledge of LLM-based question answering. To ensure that annotators are adequately prepared for subsequent annotation tasks, we provide each annotator with a comprehensive annotation guideline and instruct them to complete a trial annotation task on the designated interface. Only those whose trial annotation accuracy exceeds 95% are allowed to proceed to the formal annotation tasks. A total of eight individuals ultimately meet the annotation requirements.

Manual Hallucination Detection. To obtain the golden labels for hallucination detection, we allow two annotators to assess the factual correctness of the model-generated answers by consulting mul-

Method	Pre.	Rec.	FNR	F1
DS w/o RAG	44.6	26.9	73.1	**33.6**
DS w RAG	38.6	21.5	78.5	27.6

Table 4: Human annotators achieve higher FNR of hallucination detection when verifying the answers generated by RAG version of DeepSeek-R1.

tiple trusted information sources, such as search engines and mobile applications. Detailed accuracy evaluation templates are provided in Appendix L.

Then, we adopt a rigorous annotation protocol for hallucination detection. Two independent groups, each with two annotators (distinct from those mentioned in the previous paragraph), manually review model outputs—one group for the non-RAG version of DeepSeek-R1 and the other for the RAG version. Within each group, both annotators independently perform hallucination detection without discussion. To reflect real-world usage, annotators for non-RAG responses are given only the question and the model output, while those for RAG responses also receive the retrieved documents. To simulate typical user behavior, the annotation time is capped at 2 minutes per response and 1 minute per document. The inter-annotator agreement is 91.6%. If the results of two annotators are inconsistent, a third annotator is engaged to provide additional labeling. This setup balances annotation quality with practical constraints, improving the realism and reliability of our evaluation. For detailed information, please refer to the Appendix M.

3.2 Results

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 4 for quantitative results). This is 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; Moreover, contemporary SOTA models, when responding to such queries without the incorporation of RAG techniques, often explicitly highlight the potential risk of their outputs being outdated. (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 N.

In contrast, when evaluating responses from the RAG version of DeepSeek-R1, annotators often trust the provided reference materials and are more likely to accept the response as correct despite the possibility that the references themselves may contain errors. This phenomenon is particularly pronounced in cases involving policy inquiries or operational procedures, as the information on relevant web pages may not be updated promptly, leading to the use of outdated references. 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 relatively uncommon in SOTA models. Thus, annotators face greater difficulty detecting subtle errors in the RAG version's responses.

Furthermore, we manually trace each incorrect answer produced by the RAG version of DeepSeek-R1 that is missed by human detection back to its root cause. Specifically, erroneous responses can be divided into two categories: model intrinsic error, and retrieval noise. The analysis shows that **retrieval noise dominates with 69.8% samples**, while **model intrinsic errors account for 30.2%**.

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 seek to fine-tune LMs with the goal of reducing hallucinations by leveraging detection mechanisms? A question raises: LLMs trained with preference optimization methods may make hallucination harder to detector.

To analyze the impact of preference optimization methods, in this section, LLMs are trained with two widely recognized preference optimization methods (DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024)).

4.1 Experimental Setting

Datasets. We conduct experiments on 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.

Setting	HotpotQA				2WikiMQA			
	Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1
Base model + Vanilla RAG	44.0	69.8	30.2	**54.0**	60.4	82.0	18.0	**69.6**
DPO model + Vanilla RAG	60.4	34.9	65.1	44.2	66.5	41.8	58.2	51.3
KTO model + Vanilla RAG	57.9	29.5	70.5	39.1	65.3	24.8	75.2	35.9

Table 5: The hallucination detection performance of SelfCheckGPT on base model and preference learned models with vanilla RAG. Detectors show a decline in performance on detecting the hallucination in preference learning model generated responses.

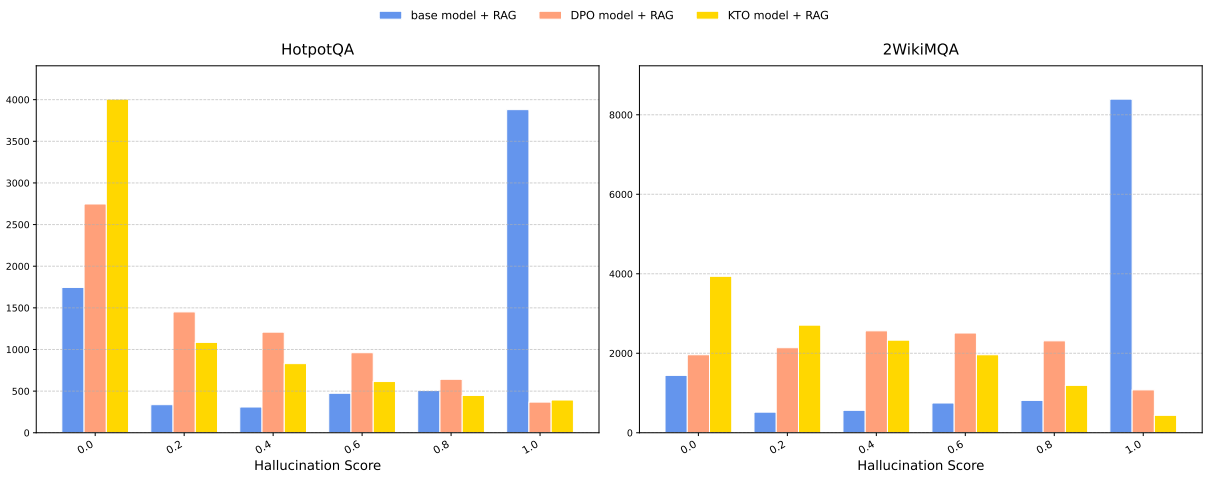

Figure 3: The distribution of hallucination scores obtained through SelfCheckGPT for responses generated by three models on the HotpotQA and 2WikiMQA test sets. It can be observed that hallucination detectors tend to assign lower hallucination scores when evaluating content generated by preference-learned models.

Each response is evaluated for its hallucination severity using the SelfCheckGPT framework. For KTO, we construct a binary classification dataset based on the response scores generated by SelfCheckGPT. To ensure the hallucination labels are assigned with a high degree of confidence, we exclude data with scores between 0.1 and 0.9. For the remaining data, responses scoring ≤ 0.1 are labeled as non-hallucination ("true"), while those scoring ≥ 0.9 are labeled as hallucination ("false"). For DPO, due to randomness, the language model will generate different responses when responding to the same question, and these responses receive varying scores within the hallucination detection framework. In each pair, the response with the lower score is labeled as "chosen," and the one with the higher score as "rejected." Training data templates and the scale of datasets are provided in Appendix O and Appendix P.

4.2 Results

After training, we use SelfCheckGPT to reevaluate the preference-learned models integrated with RAG. As shown in Table 5, both DPO- and KTO-trained models lead to a lower hallucination detection performance compared to the vanilla model.

A closer examination of the training data reveals substantial annotation noise, which undermines the reliability of the datasets used for preference learning. In the DPO experiments, it is observed that 24.7% of the "rejected" responses in the HotpotQA training set outperformed the corresponding "chosen" responses. Similarly, this proportion is found to be 21.3% in the 2WikiMQA dataset. This mislabeling indicates a notable misalignment between hallucination scores, as defined by the hallucination detection framework, and the correctness of the responses. In KTO experiments, mislabeling rates reach 52.6% and 59.1% in HotpotQA and 2WikiMQA, respectively, suggesting that the hallucination detection framework struggles to reliably annotate RAG-generated responses. We provide some bad cases from the training data in Appendix Q.

We further show how noisy data, labeled using hallucination detection frameworks, can negatively impact a model's performance in preference learning. When the generative model adapts

to the preferences encoded by the hallucination detector, it learns to produce responses that are less likely to trigger the detector. We find that this adaptation manifests in heavier citation of retrieved information and significantly longer answers—behaviors that shift the output distribution toward the detector-favored, low-hallucination region. Concretely, on HotpotQA, the proportion of responses referencing RAG sources rises from 87.0% for the base model to 95.2% for the DPO model, and the average answer length increases from 5.2 (base) to 36.6 (DPO); similar trends are observed on 2WikiMQA. As illustrated in Figure 3, responses from the preference-learned models tend to receive lower hallucination scores from SelfCheckGPT compared to the base model. Consequently, the labeling noise introduced by the hallucination detection framework undermines the effectiveness of preference learning, ultimately limiting the generative model's capacity to achieve improvements in generation quality and confusing consistency-based detector itself. The experimental results serve as a reminder for researchers and developers to exercise greater caution when utilizing these hallucination detection frameworks.

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. InstructRAG (Wei et al., 2025) introduces self-synthesized rationales to guide the language model in learning a denoising process, thereby enhancing its interpretability and robustness. While these methods reduce hallucinations and improve answer quality, they often overlook critical issues. By enhancing the consistency and fluency of language model outputs, RAG can mask subtle inaccuracies, making hallucinations harder to detect. Furthermore, incorporating external sources may introduce noisy or unreliable information and propagate errors in the pipeline. In this work, we analyze these underexplored challenges and address the vulnerabilities of RAG.

5.2 Hallucination Detection

In LLMs, hallucination refers to model outputs that deviate from the ground truth. It can stem from both internal parametric knowledge and external reference noise (Sun et al., 2025). Hallucination detection seeks to identify factual inaccuracies in the 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 these methods demonstrate strong performance, they do not account for the challenges introduced by retrieval-augmented inputs. In particular, neither SelfCheckGPT nor FAVA have examined how RAG might obscure hallucinations or distort detection accuracy. We close this gap by evaluating SelfCheckGPT and FAVA under both RAG and non-RAG settings, revealing how retrieval affects hallucination detection. Moreover, while prior work evaluates hallucination detectors under a static setting, we examine the impact of dynamically optimizing LLMs against hallucination detectors. Specifically, this dynamic interaction can lead to a phenomenon known as reward hacking (Eisenstein et al., 2023): LLMs can learn to exploit weaknesses in hallucination detectors, generating incorrect answers that seem correct to detectors, without truly mitigating hallucinations.

6 Conclusion

In this paper, we have performed a comprehensive analysis of the effectiveness of hallucination detection in both RAG and non-RAG settings. Leveraging both automated and manual evaluation on multiple challenging datasets, we reveal that RAG 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. Our findings highlight critical limitations in current detection paradigms for RAG-based systems and underscore the need for more robust detection methods.

Limitations

To manage computational costs, our experiments primarily rely on the use of Llama-3-8B-Instruct, Qwen2.5-32B-Instruct and Qwen2.5-72B-Instruct as the backbone generators, with either few-shot prompting or supervised fine-tuning based on the dataset. While this choice offers practical advantages, the limited capacity of these models may constrain the generalizability of our findings. Future studies could incorporate more open-source and closed-source models of varying scales to improve the robustness and comprehensiveness of the conclusions.

Additionally, our evaluation focuses on four representative hallucination detection frameworks: FAVA, SelfCheckGPT, ReDEeP and LLM-as-a-judge. 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.

Ethical Statement

We make sure that the dataset is collected in a manner that is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. And the crowd workers were treated fairly. This includes, but is not limited to, compensating them fairly, ensuring that they were able to give informed consent, and ensuring that they were voluntary participants who were aware of any risks of harm associated with their participation.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

I-Chun 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. abs/2307.13528.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. 2024. The llama 3 herd of models.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, Dj Dvijotham, Adam Fisch, Katherine A. Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: model alignment as prospect theoretic optimization.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.

Vojtech Hudecek and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, pages 216–228. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. 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, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, and 6 others. 2025. Medical hallucinations in foundation models and their impact on healthcare.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2025. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3586–3602. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

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.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

OpenAI. 2023. GPT-4 technical report.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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. abs/2307.03987.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and et al. 2024. Qwen2.5 technical report.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui,

Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420.

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.

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.

A The Scale of Benchmarks and the Selection of Models

Dataset	Llama-3-8B-Instruct			Qwen2.5-32B-Instruct			Qwen2.5-72B-Instruct		
	w/o RAG	Vanilla RAG	InstructRAG	w/o RAG	Vanilla RAG	InstructRAG	w/o RAG	Vanilla RAG	InstructRAG
HotpotQA (7396)	✗	✗	✗	✗	✗	✗	✗	✗	✗
2WikiMQA (12574)	✗	✗	✔	✗	✗	✗	✗	✗	✗
ASQA (948)	✗	✗	✔	✗	✗	✗	✗	✗	✗

Table 6: The table above provides an overview of the models, datasets and answer generation methods utilized in Sec. 2, with the numbers in parentheses denoting the size of the test set in each dataset. ✔ indicates that the model undergoes fine-tuning through the corresponding training set, whereas ✗ signifies that no fine-tuning is applied.

B 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)

C Prompt for GPT-4-turbo Based Hallucination Detectors (LLM-as-a-judge)

Hallucination Detection Prompt

Your task is to detect the hallucinations in the generated answers of a Large Language Model. A hallucination occurs when the model provides information that is either not supported by the given reference material or contradicts general world knowledge. Your role is to analyze the Question, Generated Answer, and Reference Material, then determine if hallucination exists. Provide an explanation for your judgment.
Here are the inputs:
Question: { }
Generated Answer: { }
Reference Material: { }
Does the Generated Answer exist the hallucination? Answer "Yes" or "No" and explain your reasoning.

D Hallucination Detection Experiment Results Conducted Using Qwen2.5-32B-Instruct as the Generative Model

Method	HotpotQA		2WikiMQA		ASQA	
	Acc.	K	Acc.	K	Acc.	K
w/o RAG	35.3	-	30.8	-	55.1	-
vanilla RAG	67.8	10	**51.2**	10	**88.8**	3
InstructRAG	**73.9**	5	34.6	10	81.4	5

Table 8: Answer generation accuracy judged by GPT-3.5-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used.

Method	RAG setting	HotpotQA				2WikiMQA				ASQA			
		Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1
FAVA	w/o RAG	63.4	77.5	22.5	**69.7**	68.2	92.4	7.6	**78.5**	50.5	57.8	42.2	**53.9**
	Vanilla RAG	35.4	66.6	33.4	46.2	42.7	70.9	29.1	53.3	19.7	60.3	39.7	29.7
	InstructRAG	29.9	71.2	28.8	42.1	66.7	67.6	32.4	67.1	21.3	55.4	44.6	30.8
SelfCheckGPT	w/o RAG	67.6	88.7	11.3	**76.7**	71.3	68.6	31.4	**69.9**	73.2	53.5	46.5	**61.8**
	Vanilla RAG	31.2	72.9	27.1	43.6	46.4	53.7	46.3	49.8	56.1	18.6	81.4	27.9
	InstructRAG	62.5	28.6	71.4	39.3	70.2	25.3	74.7	37.2	42.6	28.4	71.6	34.1
LLM-as-a-judge	w/o RAG	73.1	93.4	6.6	**82.0**	69.6	93.3	6.7	**79.7**	49.5	88.2	11.8	**63.5**
	Vanilla RAG	53.1	57.9	42.1	55.4	48.2	59.2	40.8	53.1	18.7	57.1	42.9	28.2
	InstructRAG	30.9	69.2	30.8	42.8	61.0	57.9	42.1	59.4	25.2	76.1	23.9	37.9

Table 9: Comparison of three hallucination detection methods on three datasets, evaluating their effectiveness in detecting hallucinations from answers generated by different approaches. When using RAG for answer generation, the performance of hallucination detection substantially decreases in most cases.

E Hallucination Detection Experiment Results Conducted Using Qwen2.5-72B-Instruct as the Generative Model

Method	HotpotQA		2WikiMQA		ASQA	
	Acc.	K	Acc.	K	Acc.	K
w/o RAG	42.0	-	34.6	-	58.2	-
vanilla RAG	71.2	10	**53.2**	10	**89.4**	3
InstructRAG	**77.4**	5	46.1	10	82.7	5

Table 10: Answer generation accuracy judged by GPT-3.5-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used.

Method	RAG setting	HotpotQA				2WikiMQA				ASQA			
		Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1	Pre.	Rec.	FNR	F1
FAVA	w/o RAG	57.0	75.5	24.5	**65.0**	63.6	90.0	10.0	**74.5**	53.6	60.2	39.8	**56.7**
	Vanilla RAG	31.7	64.2	35.8	42.5	46.5	77.5	22.5	58.1	23.4	64.2	35.8	34.3
	InstructRAG	26.3	69.4	30.6	38.2	50.9	67.7	32.3	58.1	21.2	58.2	41.8	31.1
SelfCheckGPT	w/o RAG	72.8	57.7	42.3	**64.4**	80.1	45.7	54.3	**58.2**	78.2	56.3	43.7	**65.5**
	Vanilla RAG	40.1	49.9	50.1	44.5	50.5	35.4	64.6	41.6	46.2	20.5	79.5	28.4
	InstructRAG	65.4	23.4	76.6	34.5	63.2	20.9	79.1	31.4	38.9	35.6	64.4	37.2
LLM-as-a-judge	w/o RAG	67.7	92.8	7.2	**78.3**	71.6	89.4	10.6	**79.5**	52.4	86.2	13.8	**65.2**
	Vanilla RAG	56.7	64.5	35.5	60.3	46.8	76.5	23.5	58.1	26.4	60.2	39.8	36.7
	InstructRAG	61.1	74.3	25.7	67.1	55.6	65.9	34.1	60.3	31.3	74.4	25.6	44.1

Table 11: Comparison of three hallucination detection methods on three datasets, evaluating their effectiveness in detecting hallucinations from answers generated by different approaches. When using RAG for answer generation, the performance of hallucination detection substantially decreases in most cases.

F The Quantitative Results from 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
Total (norm.)	691	297	1613	297	28	20
	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
Total (norm.)	557	392	1294	315	27	23

Table 12: A comparative analysis of error distributions across three RAG settings under FAVA framework, using Llama-3-8B-Instruct as the answer generation model. 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 detected by FAVA. When performing hallucination detection on answers generated without RAG, more erroneous responses are identified as hallucinations, with the majority of the recalled cases being labeled as "Unverifiable."

G FAVA Case Study

We demonstrate two cases within the FAVA framework: both the non-RAG and RAG-assisted responses are incorrect, but the hallucination detection framework only identifies 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.
Ground truth	Marion, South Australia
Response w/o RAG	Toowoomba, 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.
Response w/o RAG edited by FAVA	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. *(the framework detects the hallucination in the response)*
Response w/ Vanilla RAG	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.
Response w/ Vanilla RAG edited by FAVA	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. *(the framework does not detect the hallucination in the response)*
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 13: 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?
Ground truth	7 January 1936
Response w/o RAG	The British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based was born in 1951.
Response w/o RAG edited by FAVA	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. *(the framework detects the hallucination in the response)*
Response w/ Vanilla RAG	The British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based was born in 1867.
Response w/ Vanilla RAG edited by FAVA	Edited: The British author who wrote the novel on which “Here We Go Round the Mulberry Bush” was based was born in 1867. *(the framework does not detect the hallucination in the response)*
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 14: Case 2 from HotpotQA.

H SelfCheckGPT Case Study

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

Query	Where does the director of film Man At Bath work at?
Ground truth	Cahiers du cinéma
Response w/o RAG	The National Film and Television School
Response 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 *(the framework detects the hallucination in the response)*
Response w/ Vanilla RAG	Locarno International Film Festival in Switzerland
Response 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 *(the framework does not detect the hallucination in the response)*
Misleading retrieved docs	The film premiered in competition at **Locarno International Film Festival in Switzerland** in 2010 and was released in cinemas on 22 September 2010.
Analysis	In the reference materials, “Locarno International Film Festival in Switzerland” is the location name most closely associated with “Man At Bath” and its director. When answering such questions, the model typically matches keywords in the question with relevant terms in the materials. If it cannot find explicit information about the director’s workplace, the model often selects the location name that most closely follows references to the director or to “work,” which leads to mistaking the premiere venue for the director’s place of employment.

Table 15: Case 1 from 2WikiMQA.

Query	What nationality is the director of film Name The Man?
Ground truth	Swedish
Response w/o RAG	American
Response samples w/o RAG	[‘Australian’, ‘American’, ‘Austrian’, ‘Australian’, ‘French’]
Hallucination score w/o RAG	0.8 *(the framework detects the hallucination in the response)*
Response w/ Vanilla RAG	American
Response samples w/ Vanilla RAG	[‘American’, ‘American’, ‘American’, ‘American’, ‘American’]
Hallucination score w/ Vanilla RAG	0.0 *(the framework does not detect the hallucination in the response)*
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 16: Case 2 from 2WikiMQA.

Query	Who is the mother of the director of film Brenda Brave?
Ground truth	Käbi Laretei
Response w/o RAG	Liz Lake
Response samples w/o RAG	['Judy Cornwell', 'Sarah Townsend', 'Terry Braunstein', 'Jill Gevargizian', 'Judit Elek']
Hallucination score w/o RAG	1.0 *(the framework detects the hallucination in the response)*
Response w/ Vanilla RAG	Astrid Lindgren
Response samples w/ Vanilla RAG	['Astrid Lindgren', 'Astrid Lindgren', 'Astrid Lindgren', 'Astrid Lindgren', 'Astrid Lindgren']
Hallucination score w/ Vanilla RAG	0.0 *(the framework does not detect the hallucination in the response)*
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 17: Case 3 from 2WikiMQA.

I Statistics of the Human-Annotated Hallucination Detection Dataset

Statistic	Value
Total user queries	712
Avg. daily impression per query	1.3 w
Avg. query length (tokens)	10.6
Query diversity	Procedural: 45.2%, Information-Seeking: 30.3%, Decision-Support: 15.0%, Others: 9.5%
Avg. generated tokens (with RAG)	484.3
Avg. generated tokens (without RAG)	444.8
Avg. retrieved documents (with RAG)	7.6

J Prompt for DeepSeek Answer Generation

DeepSeek-R1 w/ RAG Answer Generation Prompt

You are a professional assistant capable of accurately understanding user queries and providing clear and detailed responses by leveraging Retrieval-Augmented Generation (RAG) information. You will follow these rules:
1. If the RAG information can directly answer the user's question, prioritize using and integrating this information.
2. If the RAG information is partially missing, you may supplement your response with general knowledge and common sense.
3. Only respond to content relevant to the user's question. If the RAG information is clearly unrelated to the question, explain that the RAG information does not directly support the answer and provide assistance based on your own knowledge.
4. Your responses must be well-structured, concise, and provide the user with direct solutions or clear explanations.
Question: { }
Reference Material: { }

DeepSeek-R1 w/o RAG Answer Generation Prompt

You are a professional assistant capable of accurately understanding user queries. Your responses should follow these rules:
1. Analyze the user's question to identify its core needs.
2. If the question involves complex concepts, provide a concise explanation and supplement with relevant background information.
3. Organize your response in clear, structured layers to ensure the content is well-organized and easy for the user to understand.
4. If the question falls outside the scope of your knowledge, state this honestly and avoid providing incorrect answers.
Question: { }

K Manual annotation cost

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

L 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: XXX, XXX, XXX; Mobile applications: XXX, XXX, XXX, etc.

Table 18: The rules of accuracy annotation.

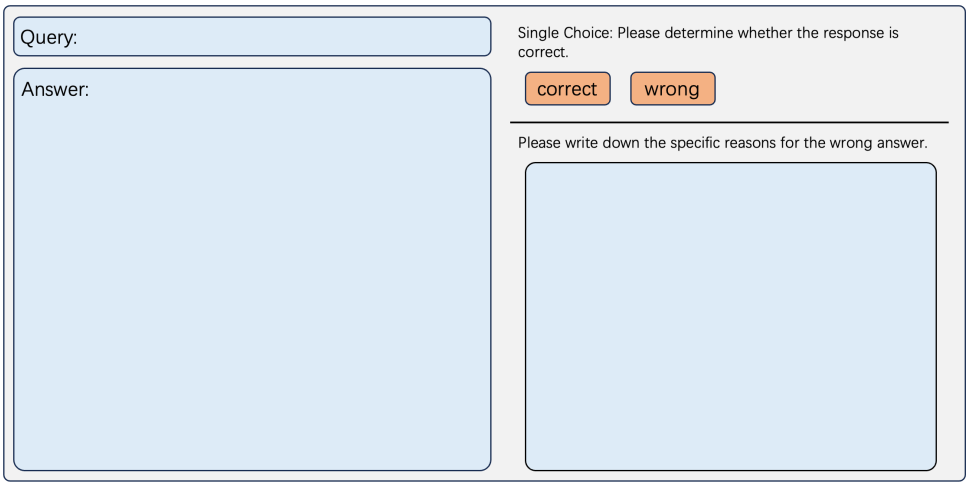

Figure 4: The template of accuracy annotation.

M 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 19: The rules of hallucination annotation.

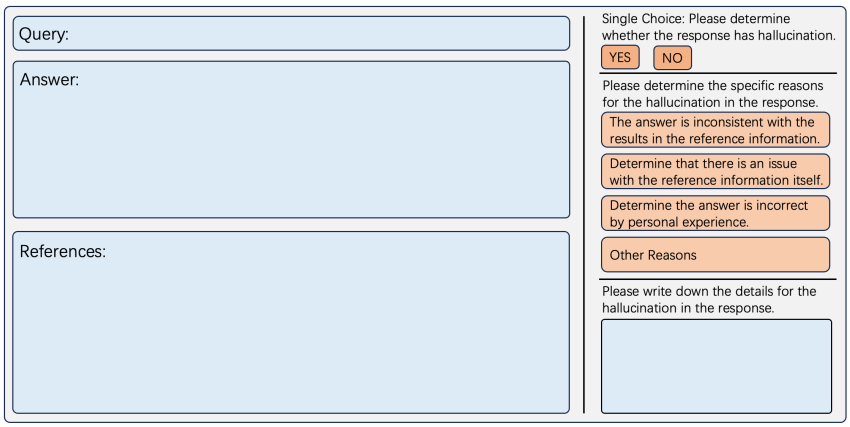

Figure 5: The template of hallucination annotation (DeepSeek-R1 with RAG).

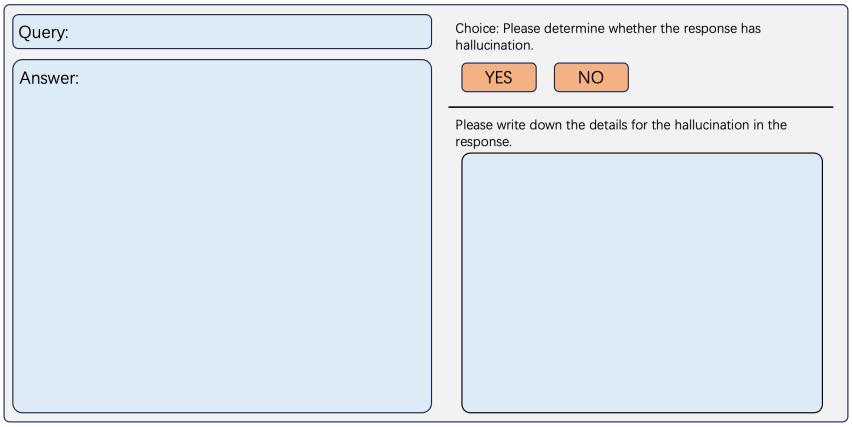

Figure 6: The template of hallucination annotation (DeepSeek-R1 without RAG).

N Human Hallucination Detection Case Study

We select three 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	Why is XXX mini-app different from before?
Answer by DeepSeek-R1 w/o RAG	Based on your location (XXX city) and the common adjustment directions for XXX application features, the following changes might occur to XXX mini-app in the near future: 1. Adjustment of feature entry XXX application **might** relocate the entry point for XXX mini-app from the homepage's "More" or "Finance" section to the "XXX" submenu, or integrate it into the "Family Account" feature. Users may need to search for "XXX mini-app" 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 XXX application version 10.2, XXX mini-app **might**: Add expenditure classification charts Support customizable saving goal progress reminders Simplify the hierarchy for managing permissions among family members Recommended actions: Open XXX application and search for "XXX mini-app" to access the latest interface Check the "Message Center - Service Notifications" for official change announcements Contact customer service and input "XXX mini-app upgrade" to obtain policy clarification documents
Answer accuracy label	False
Manual hallucination detection label	detected
Analysis	A typical example illustrates that when a model encounters uncertain questions, the absence of RAG often leads it to adopt a vague or hypothetical tone, frequently using words like "may" or "might." This increases the likelihood that human annotators will identify issues in the model's responses.

Table 20: Case 1 from real-world user query dataset.

Query	What is the annual fee for the XXX city Culture and Tourism Card?
Answer by DeepSeek-R1 w/o RAG	Based on the currently available information, the annual fee for the XXX city Culture and Tourism Card (commonly referred to as the "XXX city 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 XXX city 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 "XXX city 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 "XXX city Citizen Card" app or contact the Citizen Card hotline at 96225.
Answer accuracy label	False
Manual hallucination detection label	detected
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 21: Case 2 from real-world user query dataset.

Query	Where can I find the list of friends on XXX application?
Answer by DeepSeek-R1 w/o RAG	Steps to View Your XXX application Friends List: 1. Open the XXX application Ensure you are logged into your account and navigate to the homepage. 2. Tap the "Friends" Tab on the Bottom Navigation Bar Locate the "Friends" icon on the far-right side of the bottom menu bar (the icon resembles two small figures). Tap it to access the social features page. 3. View Your Friends List On the "Friends" page, you will see a complete list of your XXX application friends, sorted by recent interaction time. You can quickly search for a specific friend using the search bar at the top. Notes: If the friends list appears empty, it means no XXX application friends have been added yet. The location of this feature **may** slightly change due to version updates, but the core navigation remains consistent (verified as of April 2025).
Answer accuracy label	False
Manual hallucination detection label	detected
Analysis	Because there is no introduction of RAG, the LLM frequently uses words like "may" 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 22: Case 3 from real-world user query dataset.

O 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 23: 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 24: DPO data template used for training data generation

P Preference Learning Data scale

	HotpotQA		**2WikiMQA**	
Method	Train	Test	Train	Test
DPO	15744	7377	34631	12564
KTO	91135	7377	108053	12564

Table 25: The scale of datasets utilized for training and evaluation in DPO and KTO experiments.

Q Bad Cases in the Preference Dataset

Query	What is the place of birth of the director of film Prince Of The Plains?
Chosen answer	Saskatchewan, Canada
Rejected answer	not mentioned in the given knowledge
Analysis	Rejected answer is better. The "chosen" answer incorrectly states that the director of the film "Prince of the Plains" was born in Saskatchewan, Canada, which is not supported by any information in the reference provided. The director of "Prince of the Plains" is listed as Philip Ford, and there is no information about his place of birth in the reference. The "rejected" answer accurately reflects the information provided in the reference, as the birthplace of Philip Ford, the director of "Prince of the Plains," is indeed not mentioned. Therefore, the "rejected" answer is the better answer.

Table 26: A bad case from DPO training dataset.

Query	The Division of Mayo is an Australian electoral division, the division is named after which female Australian medical doctor, and medical educator, social activist and the first woman elected to an Australian University Council?
Response	Helen Mayo
Label	False
Analysis	The label is wrong. The Division of Mayo is indeed named after Helen Mayo, who was an Australian medical doctor, medical educator, and social activist, and notably the first woman elected to an Australian University Council. This information matches the description provided in the reference material about the Division of Mayo and Helen Mayo, confirming the response as correct.

Table 27: A bad case from KTO training dataset.