# 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 and authoritative responses of RAG models may *unintentionally* make hallucinations harder to detect. In this paper, we systematically investigate this phenomenon across three RAG settings (w/o RAG, vanilla RAG and InstructRAG) and three question-answering datasets. Compared to non-RAG, RAG increases the false negative rate of four widely adopted automatic hallucination detectors from 19.2% to 39.3% on average. Furthermore, we study RAG's impacts of production models (DeepSeek-R1) on real human users. We find that RAG increases the false negative rate of human hallucination detection from 73.1% to 78.5%. 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 to help 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 (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 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 the output to concrete external sources. RAG has proven effective in improving the factuality of tasks, 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 (Cheng 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 (Kim et al., 2025), finance, and law, where subtle misinformation can have significant consequences.

In this work, we investigate how RAG exacerbates the challenge of hallucination detection. In Sec. 2, to evaluate the impact of RAG on automatic hallucination detections, we analyze common RAG frameworks in three question answering datasets: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and ASQA (Stelmakh et al., 2022). Across four 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 19.2% to 39.3% on average.

Next, in Sec. 3, we investigate the impact of RAG on real human users using the production model DeepSeek-R1 (DeepSeek-AI et al., 2025). We collect a real-world online user query dataset, and recruit humans to judge the hallucination of the RAG and non-RAG version of DeepSeek-R1. We show that RAG increases the false negative rate of human judges by 5.4% in production environments.

Furthermore, in Sec. 4, we verify that when developers intend to mitigate hallucination via prevalent preference optimization techniques (e.g., DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024)), this can inadvertently enhance the RAG model's ability to evade hallucination detection. We use hallucination detectors to label good and bad responses, and optimize the RAG model on such preference datasets. LMs can reward-hack (Eisenstein et al., 2023) against hallucination detectors, making it less effective after preference training: the false negative rate of hallucination detections increases by 53.3%. Overall, our findings reveal a key risk of RAG: while it improves factual accuracy, it makes hallucinations harder to detect.

## 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.** We start with evaluating the impact of RAG on hallucination detection using three 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 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.

| 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 generation accuracy judged by GPT-4-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used.

**Models.** We use Llama-3-8B-Instruct (Dubey et al., 2024) and Qwen2.5-32B-Instruct (Yang et al., 2024) as the backbone generators, with either few-shot prompting or supervised fine-tuning based on the dataset. 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 are provided in the Appendix D)

**Answer Generation.** We adopt three answer generation methods:

- 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.** We utilize the following hallucination detection frameworks:

- 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.
- The SOTA LLM-based hallucination detector uses GPT-4-Turbo (OpenAI, 2023) or Qwen3-32B (thinking mode) (Yang et al., 2025) to determine the presence of hallucinations (Cheng

| Method | HotpotQA | | | | 2WikiMQA | | | | ASQA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre. | Rec. | FNR | F1 | Pre. | Rec. | FNR | F1 | Pre. | Rec. | FNR | F1 |
| **FAVA** | | | | | | | | | | | | |
| w/o RAG | 0.668 | 0.774 | 0.226 | 0.717 | 0.667 | 0.928 | 0.072 | 0.776 | 0.453 | 0.721 | 0.279 | 0.557 |
| Vanilla RAG | 0.450 | 0.632 | 0.368 | 0.526 | 0.535 | 0.717 | 0.283 | 0.613 | 0.203 | 0.667 | 0.333 | 0.311 |
| InstructRAG | 0.417 | 0.705 | 0.295 | 0.524 | 0.642 | 0.763 | 0.237 | 0.697 | 0.201 | 0.688 | 0.312 | 0.311 |
| **SelfCheckGPT** | | | | | | | | | | | | |
| w/o RAG | 0.726 | 0.828 | 0.172 | 0.774 | 0.705 | 0.882 | 0.118 | 0.784 | 0.701 | 0.549 | 0.451 | 0.616 |
| Vanilla RAG | 0.440 | 0.698 | 0.302 | 0.540 | 0.604 | 0.820 | 0.180 | 0.696 | 0.377 | 0.333 | 0.667 | 0.354 |
| InstructRAG | 0.634 | 0.399 | 0.601 | 0.490 | 0.741 | 0.591 | 0.409 | 0.658 | 0.633 | 0.460 | 0.540 | 0.533 |
| **GPT-4-turbo** | | | | | | | | | | | | |
| w/o RAG | 0.768 | 0.953 | 0.047 | 0.851 | 0.687 | 0.944 | 0.056 | 0.795 | 0.471 | 0.945 | 0.055 | 0.629 |
| Vanilla RAG | 0.658 | 0.623 | 0.377 | 0.640 | 0.630 | 0.730 | 0.270 | 0.676 | 0.214 | 0.833 | 0.167 | 0.341 |
| InstructRAG | 0.516 | 0.804 | 0.196 | 0.629 | 0.618 | 0.967 | 0.033 | 0.754 | 0.251 | 0.910 | 0.090 | 0.393 |
| **Qwen-32B** | | | | | | | | | | | | |
| w/o RAG | 0.824 | 0.828 | 0.172 | 0.826 | 0.727 | 0.865 | 0.135 | 0.790 | 0.563 | 0.796 | 0.204 | 0.659 |
| Vanilla RAG | 0.789 | 0.555 | 0.445 | 0.652 | 0.732 | 0.677 | 0.323 | 0.703 | 0.226 | 0.554 | 0.446 | 0.321 |
| InstructRAG | 0.667 | 0.616 | 0.384 | 0.641 | 0.705 | 0.780 | 0.220 | 0.741 | 0.344 | 0.630 | 0.370 | 0.445 |

Table 2: Comparison of four 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.

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.** Hallucination in LLMs represents a critical failure mode where generated content deviates from verifiable evidence or established factual consensus. It can stem from both internal parametric knowledge and external reference noise (Sun et al., 2025). To get ground truth labels of hallucination detections, we take query, model response and reference answer as the input, and use GPT-4-turbo 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: While RAG improves answer accuracy, it makes hallucination harder to detect.** 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 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 19.2% in the non-RAG setting to as high as 39.3% with RAG on average. These results indicate that hallucinations embedded within RAG-generated content are in general more difficult to catch.

**Finding 2: RAG makes incorrect model outputs more convincing to hallucination detectors.** 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 de-

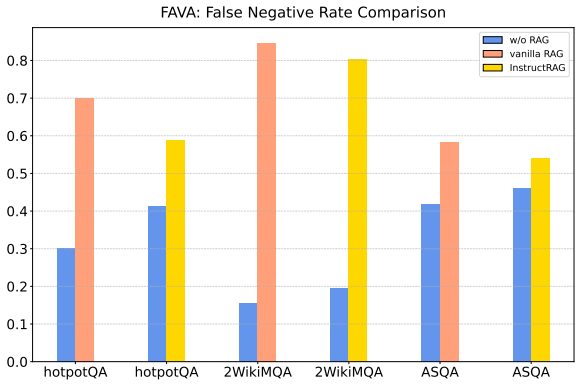


Figure 1: Under the FAVA framework, when both RAG and non-RAG generated responses are incorrect, the RAG-generated responses are less likely to be labeled as hallucinations, thereby increasing the FNR.

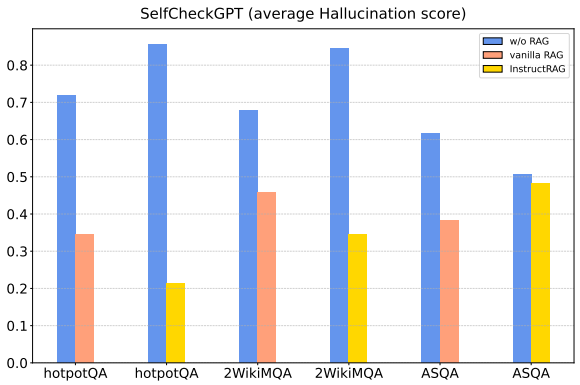


Figure 2: Compared to non-RAG models, the incorrect answers generated by RAG models achieve lower average hallucination scores when evaluated using the SelfCheckGPT framework, with Llama-3-8B-Instruct serving as the answer generation model.

tectors, 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. As shown in Figure 1, taking the HotpotQA dataset as an example, the FNR increases from 0.301 to 0.699 when comparing non-RAG setting with vanilla RAG setting, and from 0.413 to 0.587 when comparing non-RAG with InstructRAG. Similar trends are observed in the 2WikiMQA and ASQA datasets. Comprehensive quantitative results and illustrative examples can be found in Appendix E and Appendix F, 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 G.

## 3 Evading Real Human Detection

In Sec. 2, we show that RAG negatively affects automatic hallucination detectors on standard benchmarks. We now evaluate RAG's performance in production settings and uncover critical blind spots in its deployment. Through an analysis of user queries and human annotation processes, we demonstrate how integrating external knowledge inadvertently introduces new challenges in manual hallucination detection.

### 3.1 Experimental Setup

**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. Each query has over 5,000 impressions and clearly defined question intent, reflecting authentic user behavior across domains such as governance, consumer rights, local services, etc. Accurate and up-to-date information is crucial for answering these queries, which makes the dataset well suited to evaluate both the utility and potential risks of RAG in real-world deployments.

**Answer Generation.** 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. The specific answer generation prompt is in the Appendix H.

**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,

| 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 3: Human annotators achieve higher FNR of hallucination detection when verifying the answers generated by RAG version of DeepSeek-R1.

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

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. 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 K.

### 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 3 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 (As shown in Table 4). We also list several typical cases in Appendix L.

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.

## 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? Will this weak supervision truly align the models, or will LMs learn to exploit detector flaws, making the problem even worse?

In this section, we answer these questions through a rigorous analysis of fine-tuning LMs against automated hallucination detectors. We adopt two widely recognized preference optimization methods—DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024)—to explore the interplay between leveraging hallucination detection frameworks as data-labeling tools and the

| 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:<br>1. Adjustment of feature entry<br>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.<br>2. Upgrade in usage rules<br>Due to the 2024 "Financial Product Compliance Guidelines", the following changes **might** be introduced:<br>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<br>3. Interface and interaction optimization<br>Following the recent update to XXX application version 10.2, XXX mini-app **might**:<br>Add expenditure classification charts Support customizable saving goal progress reminders Simplify the hierarchy for managing permissions among family members Recommended actions:<br>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 |

Table 4: 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.

downstream impact on model behavior during post-training.

### 4.1 Experimental Setup

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. Each response is evaluated for its hallucination severity using the SelfCheckGPT framework. 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.

**Datasets.** 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 $\leq 0.1$ are labeled as non-hallucination ("true"), while those scoring $\geq 0.9$ 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."

| 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 5: The hallucination detection performance between base model and preference learned models in SelfCheckGPT framework. As shown above, when data annotated with the hallucination detection framework is used for preference learning, the fine-tuned models show a decline in performance on the hallucination detection task.

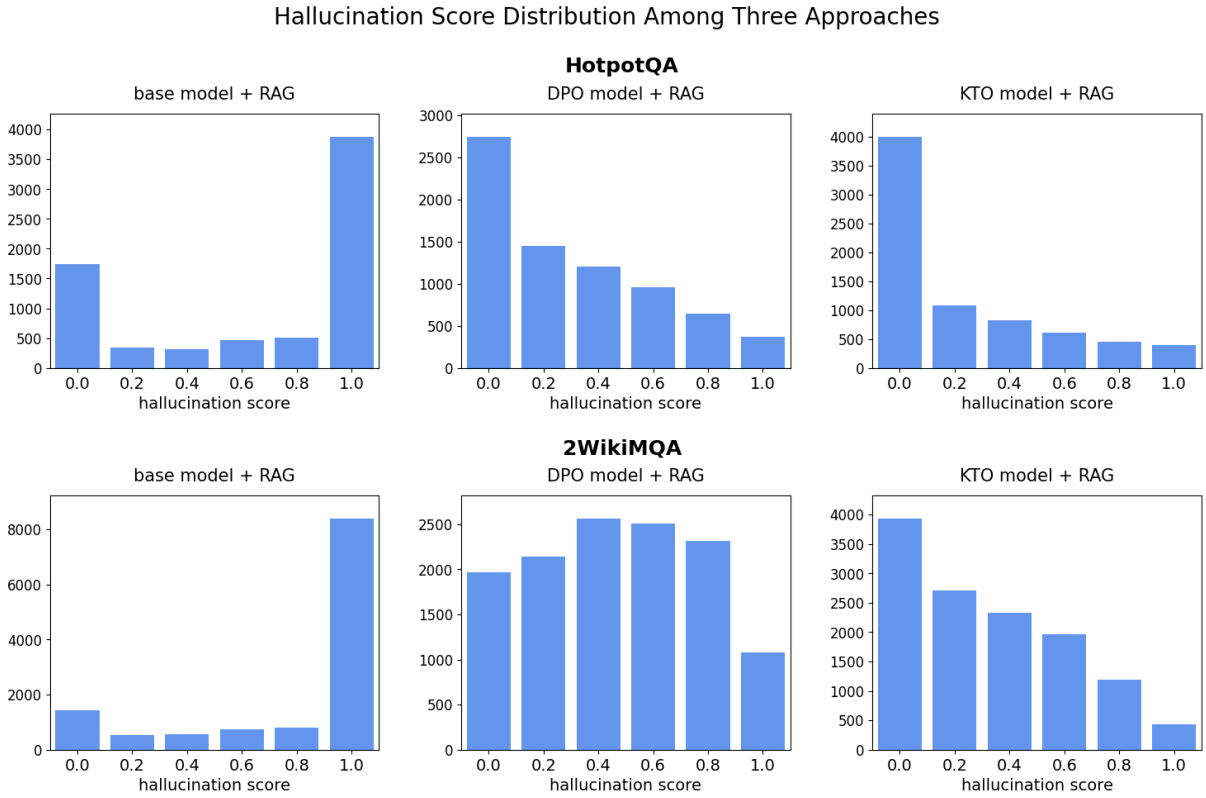


Figure 3: The distribution of hallucination scores obtained through SelfCheckGPT for responses generated by three models on the HotpotQA and 2WikiMQA test sets. It can be observed that models fine-tuned through preference learning exhibit improved consistency in their generated responses, leading to a shift in the distribution of hallucination detection scores toward lower values.

Training data templates and the scale of datasets are provided in Appendix M and Appendix N.

### 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 Ap-

pendix O.

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 is deliberately trained to generate responses that are less likely to trigger the detector, leading to a distribution of outputs that is more narrowly concentrated within the low-hallucination region favored by the detector. 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. GraphRAG (Edge et al., 2024) constructs hierarchical knowledge graphs from retrieved corpora, enabling structured information integration and improved summarization. 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

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 large language models 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 and Qwen2.5-32B-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, GPT-4-turbo based and Qwen-32B 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


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.

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.

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*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and et al. 2025. Qwen3 technical report.

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.

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 | | |
|---|---|---|---|---|---|---|
| | 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 and Qwen-32B Based Hallucination Detectors

**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** | 0.353 | - | 0.308 | - | 0.551 | - |
| **vanilla RAG** | 0.678 | 10 | 0.512 | 10 | 0.888 | 3 |
| **InstructRAG** | 0.739 | 5 | 0.346 | 10 | 0.814 | 5 |

Table 8: Answer generation accuracy judged by GPT-4-Turbo across three methods. For RAG-based methods, K denotes the number of retrieved documents used.

| Method | HotpotQA | | | | 2WikiMQA | | | | ASQA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre. | Rec. | FNR | F1 | Pre. | Rec. | FNR | F1 | Pre. | Rec. | FNR | F1 |
| **FAVA** | | | | | | | | | | | | |
| w/o RAG | 0.634 | 0.775 | 0.225 | 0.697 | 0.682 | 0.924 | 0.076 | 0.785 | 0.505 | 0.578 | 0.422 | 0.539 |
| Vanilla RAG | 0.354 | 0.666 | 0.334 | 0.462 | 0.427 | 0.709 | 0.291 | 0.533 | 0.197 | 0.603 | 0.397 | 0.297 |
| InstructRAG | 0.299 | 0.712 | 0.288 | 0.421 | 0.667 | 0.676 | 0.324 | 0.671 | 0.213 | 0.554 | 0.446 | 0.308 |
| **SelfCheckGPT** | | | | | | | | | | | | |
| w/o RAG | 0.676 | 0.887 | 0.113 | 0.767 | 0.713 | 0.686 | 0.314 | 0.699 | 0.732 | 0.535 | 0.465 | 0.618 |
| Vanilla RAG | 0.312 | 0.729 | 0.271 | 0.436 | 0.464 | 0.537 | 0.463 | 0.498 | 0.561 | 0.186 | 0.814 | 0.279 |
| InstructRAG | 0.625 | 0.286 | 0.714 | 0.393 | 0.702 | 0.253 | 0.747 | 0.372 | 0.426 | 0.284 | 0.716 | 0.341 |
| **GPT-4-turbo** | | | | | | | | | | | | |
| w/o RAG | 0.731 | 0.934 | 0.066 | 0.820 | 0.696 | 0.933 | 0.067 | 0.797 | 0.495 | 0.882 | 0.118 | 0.635 |
| Vanilla RAG | 0.531 | 0.579 | 0.421 | 0.554 | 0.482 | 0.592 | 0.408 | 0.531 | 0.187 | 0.571 | 0.429 | 0.282 |
| InstructRAG | 0.309 | 0.692 | 0.308 | 0.428 | 0.610 | 0.579 | 0.421 | 0.594 | 0.252 | 0.761 | 0.239 | 0.379 |
| **Qwen-32B** | | | | | | | | | | | | |
| w/o RAG | 0.810 | 0.805 | 0.195 | 0.808 | 0.720 | 0.807 | 0.193 | 0.761 | 0.602 | 0.633 | 0.367 | 0.617 |
| Vanilla RAG | 0.678 | 0.464 | 0.536 | 0.551 | 0.593 | 0.555 | 0.445 | 0.574 | 0.210 | 0.448 | 0.552 | 0.286 |
| InstructRAG | 0.491 | 0.502 | 0.498 | 0.496 | 0.591 | 0.478 | 0.522 | 0.529 | 0.320 | 0.584 | 0.416 | 0.413 |

Table 9: Comparison of four 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 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 10: 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."

## F 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**.<br>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.<br>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.<br>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 11: 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.<br>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.<br>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”).<br>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 12: Case 2 from HotpotQA.

## G 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 13: 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**.<br>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.<br>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 14: 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.<br>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 15: Case 3 from 2WikiMQA.

## H 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: { }

## I 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.

## J 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.<br>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;<br>Mobile applications: XXX, XXX, XXX, etc. |

Table 16: The rules of accuracy annotation.

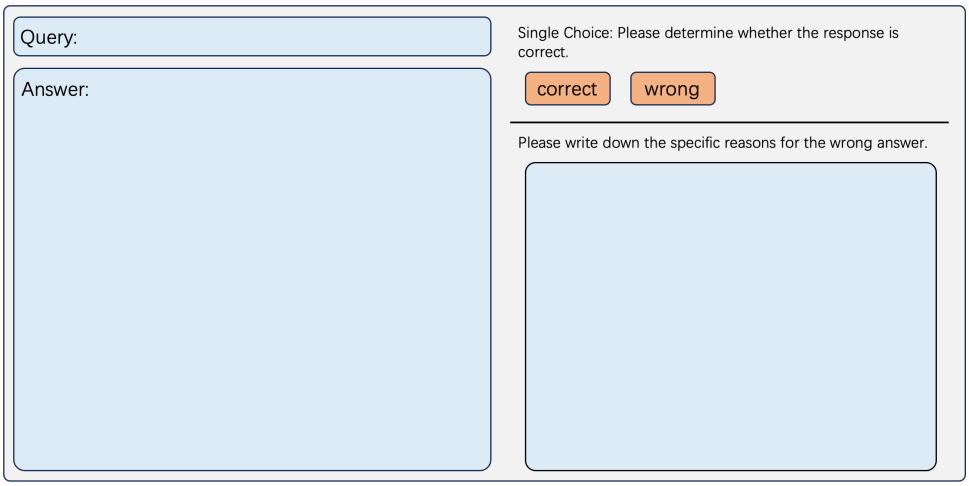


Figure 4: The template of accuracy annotation.

## K 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.<br>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.<br>DeepSeek-R1 without RAG: The time for evaluating the model's response should not exceed 2 minutes. |

Table 17: The rules of hallucination annotation.

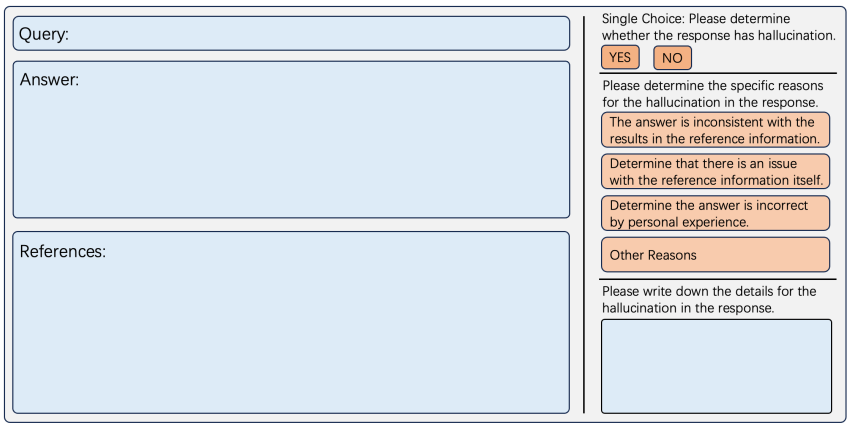


Figure 5: The template of hallucination annotation (DeepSeek-R1 with RAG).

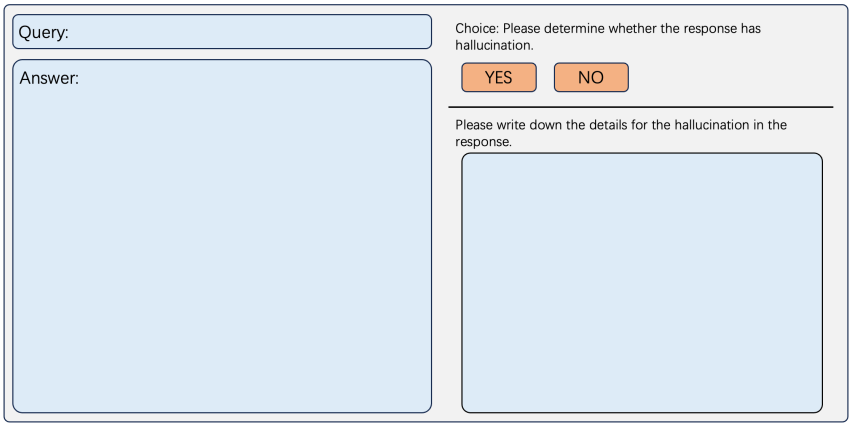


Figure 6: The template of hallucination annotation (DeepSeek-R1 without RAG).

## L Human Hallucination Detection Case Study

We select 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 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.<br>Additional Notes:<br>1. This fee is the publicly announced standard for 2024. If there are no adjustments to the policy, it will remain valid for 2025.<br>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.<br>3. Once activated, the card must be used within its validity period; renewal is required after expiration.<br>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 18: Case 1 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:<br>1. Open the XXX application<br>Ensure you are logged into your account and navigate to the homepage.<br>2. Tap the "Friends" Tab on the Bottom Navigation Bar<br>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.<br>3. View Your Friends List<br>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:<br>If the friends list appears empty, it means no XXX application friends have been added yet.<br>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 19: Case 2 from real-world user query dataset.

## M 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 20: 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 21: DPO data template used for training data generation

## N Preference Learning Data scale

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

Table 22: The scale of datasets utilized for training and evaluation in DPO and KTO experiments.

## O 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 23: 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 24: A bad case from KTO training dataset.