

# RLSeek: Evidence-Grounded Reasoning for RAG Hallucination Detection

Zhaoheng Huang<sup>1</sup>, Dacheng Wen<sup>2</sup>, Yutao Zhu<sup>1\*</sup>, Xiaoying Lian<sup>3</sup>, Yushi Liang<sup>3</sup>, Kai Hao<sup>3</sup>  
Nan Li<sup>3</sup>, Liangjie Zhang<sup>3</sup>, Qi Zhang<sup>3</sup>, Ji-Rong Wen<sup>1</sup>, Zhicheng Dou<sup>1\*</sup>, and Fangzhao Wu<sup>4\*</sup>

<sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China

<sup>2</sup>School of Computing and Data Science, The University of Hong Kong

<sup>3</sup>Microsoft <sup>4</sup>Microsoft Research Asia

{huangzh, ytzhu, dou}@ruc.edu.cn, fangzhu@microsoft.com

## Abstract

Large language models (LLMs) in retrieval-augmented generation systems can still produce hallucinations, generating content that is unsupported or contradicted by the source texts and undermines reliability. Recent work addressed this problem by training span-level hallucination detectors using reinforcement learning (RL) and chain-of-thought (CoT) reasoning. In this work, we show through error analysis that incorrect predictions by existing reasoning-based detectors are strongly associated with CoT processes that lack explicit grounding in source evidence, particularly when verification steps do not quote or verify claims against the retrieved documents. This behaviour contrasts with human verification practices in benchmarks such as RAGTruth, where evidence quotation is a prerequisite for determining hallucinated spans. Motivated by this observation, we propose an evidence-grounded RL framework, namely RLSeek, to explicitly enforce active evidence seeking during CoT reasoning by requiring quotation of relevant source segments at each verification step. Experiments on the RAGTruth and NewsSum dataset demonstrate consistent improvements in hallucination span detection performance, with limited additional reasoning overhead and improved robustness in out-of-domain settings.<sup>1</sup>

## 1 Introduction

Over the past years, large language models (LLMs) have demonstrated strong capabilities across a wide range of tasks (DeepSeek-AI, 2025; Jiang et al., 2024; Zhao et al., 2023), yet they still frequently produce factually incorrect or outdated content (Huang et al., 2025; Wang et al., 2024; Ji et al., 2023). To mitigate such issues, many applications adopt retrieval-augmented generation (RAG),

\*Corresponding authors.

<sup>1</sup>The code is released in <https://github.com/WaldenRUC/RLSeek>.

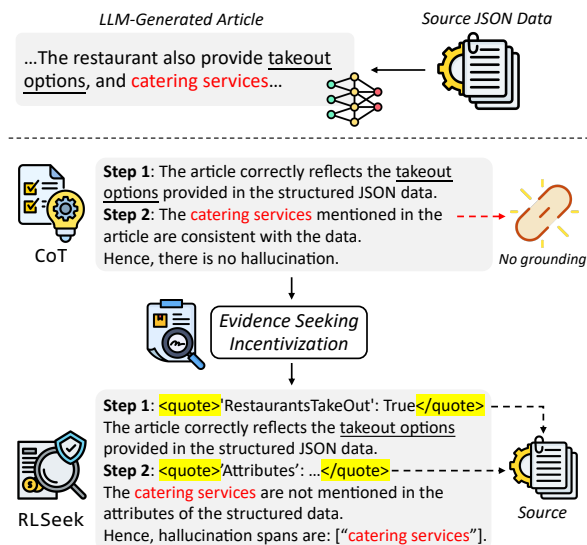


Figure 1: Illustration of evidence seeking-oriented hallucination span detection. Without explicit evidence grounding, the model may conclude “no hallucination” without verifying against the source. By enforcing quotation of the most relevant source segment at each verification step, our model RLSeek explicitly grounds its reasoning and localizes hallucinated spans. The non-hallucinated span is underlined, and the hallucinated span is in **red**.

which supplements LLMs with external grounding documents (Zhu et al., 2025; Li et al., 2024; Ayala and Béchar, 2024). Despite explicit access to supporting documents, RAG systems remain prone to *hallucinations*, where generated content may be unfaithful to the source, fabricate details, or produce statements that contradict the source (Niu et al., 2024). This makes hallucination detection critical for ensuring the reliability of RAG systems.

Most prior work formulates hallucination detection as a binary classification task, predicting whether an LLM-generated output contains hallucinations (Huang et al., 2026; Tang et al., 2024; Chern et al., 2023). However, they fall short

in practical settings where users need to identify which specific parts of an output are unsupported or contradicted to the source (Mishra et al., 2024; Min et al., 2023). This limitation motivates the study of fine-grained span-level hallucination detection, as explored in recent benchmarks such as RAGTruth (Niu et al., 2024). The hallucinated span detection task aims to localize the exact spans of generated text that are not grounded in the provided source text, enabling a more fine-grained assessment of output reliability.

Recent methods for hallucination span detection, such as RL4HS (Su et al., 2025), train detectors to perform chain-of-thought (CoT) reasoning. These approaches typically employ group relative policy optimization (GRPO) (Shao et al., 2024; Liu et al., 2025) and optimize a span-level F1 reward that measures the overlap between predicted and ground-truth spans. To investigate the limitations of such models, we conducted an error analysis by comparing incorrect predictions with accurate ones. Our analysis reveals **two critical insights**: (1) There is a distinct difference in reasoning behaviour within the model predictions. Specifically, incorrect predictions (zero span-F1) significantly rely on reasoning that lacks evidence grounding, where the model fails to quote relevant source segments. In contrast, accurate predictions (non-zero span-F1) frequently exhibit explicit evidence grounding through direct quotation. (2) This lack of grounding in models stands in sharp contrast to human verification practices in benchmarks like RAGTruth (Niu et al., 2024). In these settings, annotators must validate each hallucinated span by explicitly identifying and quoting the most relevant evidence from the source. This requirement ensures that judgments are directly traceable to verifiable content. The gap between model behaviours and human standards described in these two points indicates that **insufficient evidence grounding** is a key factor in detection errors, motivating a closer examination of how evidence is sought and used during reasoning.

To address these problems, we propose RLSeek, an RL framework designed to enforce active evidence-seeking behaviour during CoT reasoning. As illustrated in Figure 1, the core mechanism involves instructing the model to ground each verification step by actively seeking and quoting the most relevant source segments from the source text. Consequently, the model determines hallucinated spans based strictly on this quoted evidence. To

optimize this behaviour, we introduce a composite reward function that integrates standard outcome-based rewards with a novel process-level penalty. This penalty specifically targets reasoning steps that fail to provide quotes or that generate text deviating from the original source. Technically, we implement this by calculating the lexical overlap between the quoted evidence and the source text, ensuring strict adherence to the provided material.

We evaluate the effectiveness of RLSeek on the RAGTruth benchmark and a proprietary dataset named NewsSum, covering diverse scenarios including question answering, news summarization, and data-to-text generation. Experimental results demonstrate that explicitly encouraging evidence-grounded reasoning consistently outperforms strong baselines, including state-of-the-art supervised and RL-based methods. Further analysis indicates that these gains stem from more disciplined evidence usage achieved with minimal reasoning overhead.

Our contributions can be summarized as follows:

- We conduct a systematic analysis of CoT behaviours in hallucination span detection, revealing that incorrect predictions are strongly associated with the absence of explicit evidence grounding during reasoning.
- Motivated by these observations, we propose an evidence-grounded RL framework that explicitly encourages active evidence seeking during CoT reasoning for hallucination span detection.
- Extensive experiments on RAGTruth and NewsSum dataset demonstrate consistent performance improvements with only limited additional reasoning overhead.

## 2 Preliminary Study

### 2.1 Problem Formulation

Given a task instruction  $I$  and a grounding source text  $D$  as context, an LLM  $M_\theta$  generates an output text  $T = M_\theta(I, D) = (t_1, t_2, \dots, t_n)$ , where  $T$  consists of  $n$  characters. The goal of hallucination span detection is to identify all text segments  $\hat{S}$  in  $T$  that are not supported by the source  $D$ :

$$\hat{S} = \{s_k = [i_k, j_k]\}_{k=1}^K, \quad s_k \subset T. \quad (1)$$

Each hallucinated span  $s_k$  corresponds to a contiguous subsequence of characters  $(t_{i_k}, \dots, t_{j_k})$  in  $T$ . Following RAGTruth (Niu et al., 2024), a hallucinated span is defined as a text segment that either

contradicts the source text or contains fabricated content that is not substantiated by the source  $D$ .<sup>2</sup>

## 2.2 Verifiable Span-Level Reward in GRPO

To evaluate hallucination span detection and incentivize accurate span localization, we follow [Su et al. \(2025\)](#) and directly compare the predicted hallucination spans  $\hat{S}$  with the ground-truth spans  $S$ . When both span sets are empty, indicating a correct prediction of no hallucination, the model receives a reward of 1. Otherwise, the reward is defined as the span-level F1 score between two span sets:

$$r_{\text{span}} = \begin{cases} 1, & \text{if } \hat{S} = \emptyset \text{ and } S = \emptyset, \\ \text{span-F1}(\hat{S}, S), & \text{otherwise.} \end{cases}$$

This reward formulation provides a unified measure for span-level prediction performance in both hallucinated and non-hallucinated cases. Using this span-level reward, we employ GRPO ([Shao et al., 2024](#)) to train the hallucination span detector. The optimization objective is defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{clip}}(\theta) - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta} \parallel \pi_{\text{ref}}]. \quad (2)$$

The clipped policy gradient objective  $\mathcal{L}_{\text{clip}}(\theta)$  follows the proximal policy optimization formulation:

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[ \min \left( \frac{\pi_{\theta}(\tau)}{\pi_{\text{old}}(\tau)} A(\tau), \right. \right. \\ \left. \left. \text{clip} \left( \frac{\pi_{\theta}(\tau)}{\pi_{\text{old}}(\tau)}, 1 - \epsilon, 1 + \epsilon \right) A(\tau) \right) \right],$$

where the KL-divergence term regularizes the learned policy  $\pi_{\theta}$  to remain close to the reference policy  $\pi_{\text{ref}}$ ,  $\tau$  denotes a trajectory sampled from the current policy  $\pi_{\theta}$ , and  $r_{\theta}(\tau) = \frac{\pi_{\theta}(\tau)}{\pi_{\text{old}}(\tau)}$  is the probability ratio between the current and previous policies. Following GRPO, the advantage  $A(\tau)$  is computed using relative rewards  $\{r_i\}$  within a group of sampled trajectories  $G(\tau)$ :

$$A(\tau) = \frac{r - \text{mean}(\{r_i\}_{i \in G(\tau)})}{\text{std}(\{r_i\}_{i \in G(\tau)})}. \quad (3)$$

## 2.3 Analysis of Evidence-Seeking Behaviours

To examine whether prediction errors are associated with changes in reasoning behaviour, we analyze evidence-seeking patterns in the model’s

<sup>2</sup>We follow the settings from RAGTruth and assume that the grounding source  $D$  is provided and fixed, and the design of the retriever and the quality of retrieved documents are considered out of scope.

CoT. In human verification criteria for hallucination detection, annotators are required to explicitly ground each judgment by quoting the most relevant source evidence ([Niu et al., 2024](#); [Song et al., 2024](#)). In contrast, when evidence grounding is not explicitly reinforced during RL training, the model may arrive at decisions without verifiable support. This motivates us to study how evidence usage in CoT differs across predictions with varying span-level outcome reward. To this end, we employ an LLM-as-a-judge ([Gu et al., 2024](#)) framework to determine whether the verification steps in each CoT explicitly quote relevant evidence from the source. The resulting judgments achieve 96% accuracy with human annotations under majority voting (details are in Appendix A). We group samples based on span-level prediction outcomes, contrasting worst-case predictions with zero span-F1 against predictions achieving non-zero span-F1, and quantify evidence usage by the proportion of samples whose CoT contains no explicit source quotation.

As shown in Figure 2, predictions with zero span-F1 consistently exhibit a higher proportion of CoT without evidence quotation than those with non-zero span-F1. Across all three tasks, the gap in the ratio of non-quoting CoT between the two groups reaches around 20%. Moreover, this gap persists throughout RL optimization, indicating that degraded evidence-seeking behaviour remains a characteristic of worst-case span-level failures. Overall, these results reveal a strong correlation between span-level detection failures and the absence of explicit evidence-seeking behaviour during CoT reasoning, suggesting that insufficient evidence grounding directly undermines the model’s ability to localize hallucinated spans.

## 3 Our Method: RLSeek

### 3.1 Evidence-Seeking Reasoning

We propose RLSeek as an evidence-seeking reasoning framework for hallucination span detection. Our method builds on the CoT framework ([Su et al., 2025](#)) by requiring explicit evidence quotation at each verification step. Given a source text and a generated response, the model conducts step-by-step CoT verification, where each step first seeks and quotes the most relevant evidence from the source and then determines the hallucinated spans based on the quoted content. Details of the instruction design are provided in Appendix B.

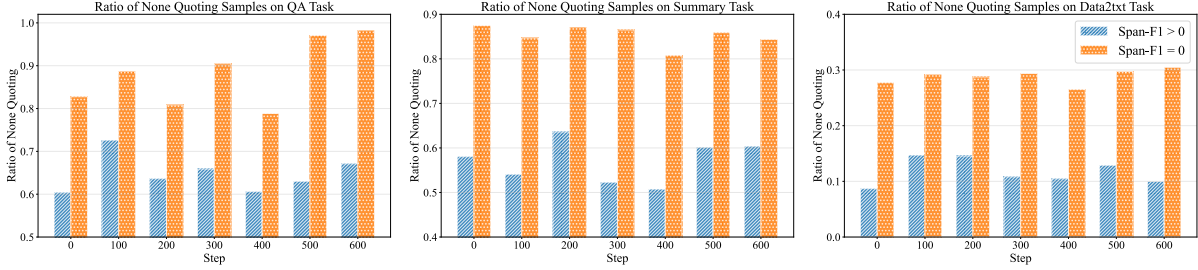


Figure 2: Ratio of samples whose CoT contains no source quotation, grouped by span-level prediction outcomes (zero span-F1 vs. non-zero span-F1) during RL optimization on three RAGTruth detection tasks.

### Quoting Evidence in CoT (Summarization Task)

```
## Step 1: <your first reasoning step, briefly state what
you are going to check in this step>
If this step is used to verify whether some part of the
summary contains hallucinations, you MUST follow the
instructions below:
- First, using only the original document, locate and quote
the most relevant, precise, and shortest self-contained
evidence from the original document for this verification step.
Place each evidence verbatim inside: <quote>...</quote>.
- Then, explain how the evidence supports your evaluation
of this part (supported or hallucinated). If key evidence is
missing, explicitly state the missing evidence and why it
indicates hallucination. If this part contains hallucination,
list all hallucinated text spans.
```

## 3.2 Quotation-Faithful Measurement

While evidence-seeking reasoning encourages quotation, it does not guarantee faithful use of source evidence. Since intermediate CoT steps are not explicitly constrained during optimization, the model may paraphrase or summarize source content instead of quoting it verbatim, thereby introducing information not present in the source text (Belém et al., 2025). To assess quotation faithfulness, let  $Q = \{Q_1, \dots, Q_{|Q|}\}$  denote the set of evidence spans enclosed by the <quote> tag in the CoT. We assume the CoT contains  $L$  steps and denote by  $Q^{(l)} \subseteq Q$  the quotation set at step  $l$ . We consider two complementary cases when evaluating quotation faithfulness.

**Invalid quotation sets.** Quotation sets that violate the quotation instruction are considered unfaithful by definition. This includes cases where no evidence is quoted or where quoted spans are empty in any CoT step, indicating that the model bypasses evidence grounding or produces ill-formed quotations. Formally, a quotation set is defined as *invalid* if it violates the quotation requirement:

$$\mathbb{I}_{\text{invalid}}(Q) = \begin{cases} 1, & \text{if } \exists |Q^{(l)}| = 0 \text{ or } \exists |Q_i| = 0, \\ 0, & \text{otherwise.} \end{cases}$$

**Valid quotation sets.** Valid quotation sets satisfy the instruction format, *i.e.*, each CoT step contains at least one non-empty quoted evidence span. However, format validity alone does not ensure faithfulness, as the model may still partially quote the source or introduce content not present in the source text. To quantify quotation faithfulness, we measure the lexical alignment between each quoted span and the source text using the longest common substring (LCS). Specifically, for each quoted span  $Q_i$ , we compute the normalized overlap  $\text{LCS}(Q_i, D)/|Q_i|$ , where  $D$  denotes the source text. A faithful quotation should largely consist of contiguous text copied from the source. Thus, higher overlap indicates stronger grounding, while lower overlap suggests deviations from the quoted evidence. Compared to semantic similarity measures or learned methods, LCS is lightweight, model-free, and directly measures lexical overlap with the source text, making it suitable for assessing quotation faithfulness in CoT reasoning.

## 3.3 Evidence-Grounded Policy Optimization

To incorporate quotation faithfulness into policy optimization, we convert the measurement defined above into a penalty term for reinforcement learning. Specifically, we define a quotation faithfulness penalty  $r_{\text{penalty}}$  based on the validity of the quotation set and its lexical alignment with the source:

$$r_{\text{penalty}} = \begin{cases} 0.5, & \text{if } \mathbb{I}_{\text{invalid}}(Q) = 1, \\ \frac{1}{|Q|} \sum_{i=1}^{|Q|} \left[ 1 - \frac{\text{LCS}(Q_i, D)}{|Q_i|} \right], & \text{otherwise.} \end{cases}$$

For invalid quotation sets, we apply a fixed penalty inspired by prior work (Shang et al., 2025) to discourage trajectories that violate the quotation instruction. For valid quotation sets, the penalty is defined as the average proportion of unmatched

content across quoted spans, penalizing partial quotation and paraphrasing via lexical mismatch with the source text. We combine the quotation faithfulness penalty with the span-level detection reward to form the final training objective. Under GRPO, the resulting reward is used to compute the advantage:

$$r_{\text{final}} = r_{\text{span}} - r_{\text{penalty}}, \quad (4)$$

$$A(\tau) = \frac{r_{\text{final}} - \text{mean}(\{r_{\text{final},i}\}_{i \in G(\tau)})}{\text{std}(\{r_{\text{final},i}\}_{i \in G(\tau)})}.$$

By incorporating quotation faithfulness into advantage estimation, GRPO differentiates reasoning with similar span-level outcomes but different evidence faithfulness, encouraging the policy to favor evidence-grounded reasoning during optimization.

## 4 Experimental Setup

### 4.1 Dataset

We conduct experiments on the **RAGTruth** benchmark (Niu et al., 2024), which provides fine-grained span-level hallucination annotations across three representative tasks: (1) **Summarization**, where a generated summary  $T$  is evaluated against the source document  $D$ ; (2) **Question Answering (QA)**, where a generated answer  $T$  is verified using the corresponding retrieved passages  $D$ ; and (3) **Data-to-Text**, where a generated article  $T$  is grounded in structured input data  $D$ . Due to the limited availability of publicly accessible datasets with span-level hallucination annotations, we further include **NewsSum**, a proprietary news summarization dataset. Each sample consists of several source news articles paired with one generated summary, where hallucinated spans in the summary are manually annotated. More dataset statistics are shown in Appendix C.

### 4.2 Baselines and Evaluation Metric

**Baselines** (1) **Fine-tuned Models**. We include supervised fine-tuning (SFT) models trained end-to-end with span-level hallucination annotations, without explicit reasoning. We evaluate Qwen2.5-7B-Instruct (**SFT-7B**) and Qwen2.5-14B-Instruct (**SFT-14B**). We also compare with **MVA** (Ogasa and Arase, 2025), which detects hallucinations using aggregated attention-based features. (2) **Reasoning Models**. We evaluate advanced LLMs, including **Qwen3**, **GPT-5**, and **OpenAI-o3**. We test both **CoT** prompting (Su et al., 2025) and **CoT+Seek**, which encourages evidence quotation

before reasoning (as shown in Prompt 1). (3) **RL-based Models**. We compare our method against **RL4HS** (Su et al., 2025), which trains hallucination span detectors with GRPO and class-aware policy optimization (CAPO). This is the most closely related baseline to our approach. The implementation details are shown in Appendix D.

**Evaluation Metric**. Following prior work (Niu et al., 2024; Song et al., 2024), we evaluate hallucination span detection using dataset-level span-F1, and additionally report sample-level binary detection metrics. Details are provided in Appendix E.

## 5 Results and Analysis

### 5.1 Experimental Results

Table 1 and Table 2 demonstrate hallucination detection performance on the RAGTruth benchmark across three tasks and macro-averaged metrics. Table 3 reports results on the NewsSum dataset.

(1) **Active evidence quotation improves CoT-based detection**. Starting from the CoT-based hallucination detection framework (Su et al., 2025), we instruct the model to actively quote source evidence at each reasoning step. Across four reasoning models (Qwen3-14B, GPT-5-mini, GPT-5, and o3), this active quoting strategy (CoT-Seek) consistently improves span-level detection performance on all three RAGTruth tasks. These results show that enforcing explicit evidence grounding is beneficial for hallucination detection; however, the gains remain limited, suggesting that prompting alone is insufficient to fully regulate evidence-seeking behaviour during reasoning.

(2) **Evidence-seeking RL yields better precision-recall trade-offs**. Building on active quoting, we further apply our evidence-seeking RL framework. Compared with prior RL-based methods, our approach consistently improves span-level F1, achieving gains of up to 2.5% (7B) and 3.1% (14B) over CAPO. At the sample level, we observe corresponding gains of 2.4% and 2.0% in binary detection F1 for the 7B and 14B models, respectively, with similar trends on the NewsSum dataset. We also analyze sample-level false rates in Appendix F. Notably, while CAPO mainly rebalances precision and recall by suppressing non-hallucination predictions, our method enforces explicit evidence quotation during reasoning for both hallucinated and non-hallucinated cases. As shown in Figure 3, this

Methods	Summarization			Question Answering			Data-to-Text			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
<b><i>Fine-tuned Models</i></b>												
SFT-7B	52.2	38.2	44.1	51.3	51.4	51.3	58.8	51.5	54.8	54.1	47.0	50.1
SFT-14B	57.6	48.7	52.7	53.1	54.8	53.9	61.6	57.8	59.6	57.4	53.8	55.4
MVA-7B	49.6	35.7	41.5	38.5	73.7	50.6	53.5	57.1	55.2	47.2	55.5	49.1
<b><i>Reasoning Models</i></b>												
Qwen3-14B-CoT	36.9	34.9	35.8	30.7	30.6	30.6	40.9	30.4	34.8	36.2	32.0	33.7
Qwen3-14B-CoT-Seek	39.8	40.8	40.3	32.8	40.3	36.2	35.2	41.3	38.0	35.9	40.8	38.2
GPT5-mini-CoT	33.0	65.5	43.9	37.9	62.7	47.2	29.8	74.7	42.5	33.6	67.6	44.5
GPT5-mini-CoT-Seek	32.7	67.4	44.0	37.6	64.5	47.5	32.4	75.9	45.4	34.2	69.3	45.6
GPT5-CoT	24.9	68.4	36.5	32.1	71.8	44.4	33.2	73.5	45.7	30.0	71.2	42.2
GPT5-CoT-Seek	24.9	70.2	36.8	32.4	74.7	45.2	34.3	78.3	47.7	30.5	74.4	43.2
o3-CoT	40.7	60.1	48.5	43.4	58.9	49.9	45.6	70.0	55.2	43.2	63.0	51.2
o3-CoT-Seek	41.5	61.4	49.5	44.7	59.5	51.0	46.0	71.0	55.8	44.1	64.0	52.1
<b><i>RL-based Models</i></b>												
RL4HS-GRPO-7B	68.7	40.9	51.2	59.6	52.1	55.6	66.5	48.8	56.3	64.9	47.3	54.4
RL4HS-CAPO-7B	64.4	42.3	50.9	57.1	56.5	56.8	67.1	54.9	60.4	62.9	51.2	56.1
RLSeek-7B (ours)	67.4	44.3	53.5	58.0	57.6	<u>57.8</u>	69.6	54.5	61.2	65.0	52.1	57.5
RL4HS-GRPO-14B	66.4	44.5	53.3	67.5	48.1	56.1	65.0	56.4	60.4	66.3	49.7	56.6
RL4HS-CAPO-14B	64.2	52.3	<u>57.6</u>	52.5	57.3	54.8	67.2	58.7	<u>62.6</u>	61.3	56.1	<u>58.3</u>
RLSeek-14B (ours)	65.1	53.0	<b>58.4</b>	58.6	59.1	<b>58.8</b>	69.7	57.5	<b>63.0</b>	64.5	56.5	<b>60.1</b>

Table 1: Span-level hallucination detection performance across different tasks on the RAGTruth benchmark. Best F1 scores are in **bold**, and the second best result is underlined.

Methods	# Params	Precision	Recall	F1
GRPO	7B	83.7	76.4	79.9
CAPO		81.7	79.8	<u>80.7</u>
RLSeek		83.1	82.2	<b>82.6</b>
GRPO	14B	85.0	81.1	83.0
CAPO		82.9	84.0	<u>83.4</u>
RLSeek		84.8	85.4	<b>85.1</b>

Table 2: Overall performance of sample-level hallucination detection on the RAGTruth benchmark.

Methods	Span-level			Sample-level		
	P	R	F1	P	R	F1
SFT	31.1	29.3	30.2	79.3	85.0	82.1
GRPO	66.4	44.7	53.4	90.5	78.2	83.9
CAPO	59.3	48.7	<u>53.5</u>	90.4	81.1	<u>85.5</u>
RLSeek	61.3	49.9	<b>55.0</b>	89.6	83.0	<b>86.2</b>

Table 3: Hallucination detection performance on the NewsSum dataset, trained on Qwen2.5-7B-Instruct.

leads to simultaneous and stable improvements in both precision and recall throughout RL training.

## 5.2 Ablation Study

We conduct ablation studies to analyze the contribution of each component in our method. *w/o Penalty* removes the quotation faithfulness penalty from the reward (Equation 4). *w/ CAPO* applies CAPO by scaling the advantage of non-hallucinated samples with  $\alpha = 0.5$  (Su et al., 2025). *w/o RL* uses

evidence-quoting CoT prompting without RL. *w/o RL&Seek* further removes evidence-quoting, reducing the model to standard CoT prompting.

Results in Table 4 show that the quotation faithfulness penalty plays a crucial role in stabilizing evidence-seeking behaviour and improving performance. Removing the penalty (*w/o Penalty*) yields only limited gains over the GRPO baseline, indicating that evidence-quoting instructions alone are insufficient without explicit incentivization. Experimental results in Appendix G reveal that, without the penalty term in the reward, the overlap between quoted content and the source drops to approximately 55%. In contrast, when the penalty is applied, the percentage of valid quotation set and overlap remains near 100%, indicating that the penalty effectively enforces faithful quotation and discourages paraphrased evidence. Applying CAPO (*w/ CAPO*) slightly improves recall but degrades precision due to suppressed negative-sample advantages, resulting in a lower overall F1 score. This suggests that outcome-level re-weighting alone is insufficient to regulate the reasoning process, highlighting the importance of explicitly incentivizing evidence seeking and quoting in CoT reasoning. Finally, *w/o RL* and *w/o RL&Seek* show that active quoting improves CoT prompting but provides only marginal benefits, highlighting the necessity of RL optimization for effective evidence-seeking.

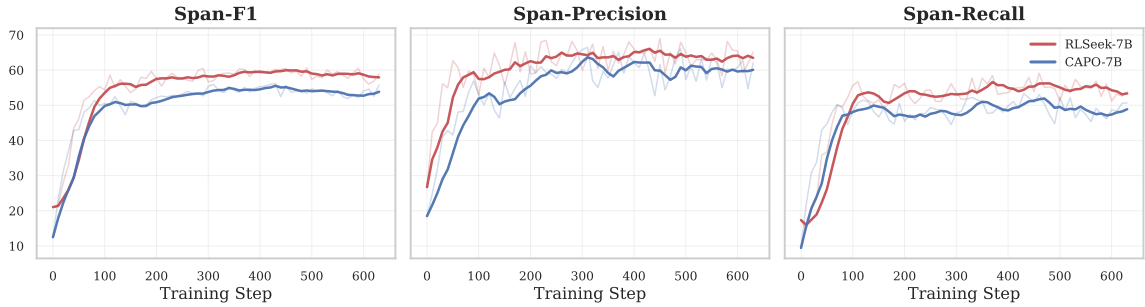


Figure 3: Performance on the test set across training checkpoints for RLSeek (red) and CAPO (blue) on the RAGTruth benchmark. Solid lines indicate the moving average over the most recent 50 steps.

Methods	Span-level			Sample-level		
	P	R	F1	P	R	F1
GRPO	64.9	47.3	54.4	83.7	76.4	79.9
CAPO	62.9	51.2	56.1	81.7	79.8	80.7
RLSeek	65.0	52.1	<b>57.5</b>	83.1	82.2	<b>82.6</b>
w/o Penalty	63.8	48.4	54.9	83.5	77.0	80.1
w/ CAPO	60.7	52.4	<u>56.4</u>	82.3	82.4	<u>82.3</u>
w/o RL	26.0	14.8	18.8	59.7	46.2	52.1
w/o RL&Seek	26.2	12.3	16.2	56.7	46.1	50.8

Table 4: Ablation study of different training components on the RAGTruth benchmark.

### 5.3 Analysis of Evidence-Quoting Behaviour

To analyze evidence-quoting behaviour during reasoning, we examine the relationship between CoT length and the amount of quoted evidence across different RL methods at fixed training steps, as shown in Figure 4. During the early stage of RL training (within the first 100 steps), the model produces relatively long CoT with few effective quotations, reflecting unstable adherence to the quoting instruction. As training progresses, the number of quoted spans stabilizes at approximately four to five per sample, and the overall CoT length becomes comparable to the GRPO and CAPO baselines. These trends indicate that evidence quoting gradually becomes more concise and structured, introducing only limited additional reasoning overhead during inference.

We further analyze the *diversity* and *relevance* of quoted evidence across reasoning steps. Specifically, we measure lexical diversity within each quotation set  $Q_i$  using Self-BLEU (Zhu et al., 2018) and Distinct- $n$  (Li et al., 2016), and assess quotation relevance with an LLM-as-a-judge framework to assess whether each quoted segment is relevant to its corresponding verification claim. Additional implementation details are provided in Appendix H.

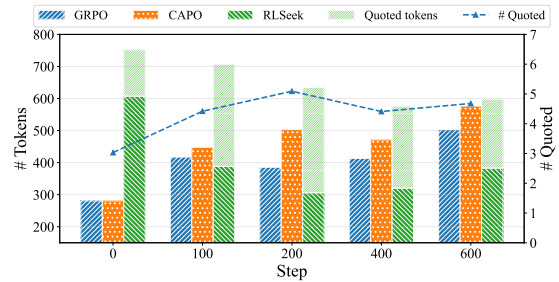


Figure 4: Analysis of CoT length and evidence quotation statistics across different RL methods.

Results show that, as training proceeds, the model increasingly distributes quotations across distinct source segments while consistently focusing on the most relevant evidence. This suggests that the proposed evidence-seeking objective effectively discourages reward-hacking behaviours, such as repeatedly quoting long or irrelevant source texts, since redundant or uninformative quotations provide limited benefit for span-level outcome rewards and may incur explicit penalties.

### 5.4 Out-of-Domain Detection

We further evaluate the robustness of our method under out-of-domain (OOD) settings. Following a leave-one-out protocol, we train the model on two subsets of RAGTruth and evaluate it on the remaining unseen dataset. As shown in Figure 5, our method consistently outperforms the state-of-the-art baseline CAPO across all OOD settings, demonstrating strong generalization without any exposure to the target domain during training. This suggests that explicitly grounding verification steps in source evidence enables the model to rely on evidence-seeking strategies rather than domain-specific training patterns, leading to more robust hallucination detection on unseen tasks.

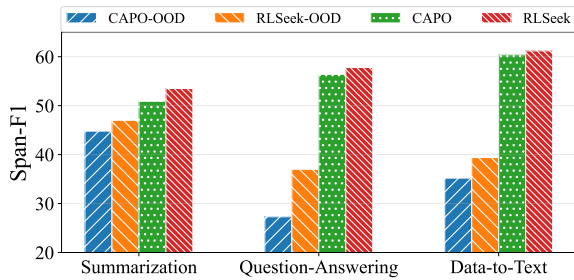


Figure 5: Out-of-domain (OOD) detection performance.

### 5.5 Case Study

We present qualitative case studies to examine evidence-quoting behaviour during reasoning, with representative examples provided in Appendix I. We observe that existing methods are often able to identify the content to be verified, but frequently fail to support their decisions with precise source quotations. In contrast, our method enforces a structured verification process consisting of three explicit steps: (1) stating the verification objective, (2) quoting relevant source evidence, and (3) identifying hallucinated spans based on the quoted evidence. The case studies show that our approach produces reasoning traces that are more explicitly source-grounded for both hallucinated samples (including contradictory and fabricated spans) and non-hallucinated samples, and more closely align with human verification criteria. Notably, when hallucinated spans lack supporting evidence, the model identifies relevant source segments and correctly concludes that the spans are unsupported, rather than relying on ungrounded reasoning.

## 6 Related Work

**Hallucination Detection** Early work formulates hallucination detection as a coarse-grained task, including binary classification (Tang et al., 2024) and natural language inference (NLI) (Scirè et al., 2024). Later studies propose LLM-based verification pipelines that decompose text into atomic claims and verify them with external evidence (Iqbal et al., 2024; Min et al., 2023; Chern et al., 2023), but these pipelines are difficult to optimize end-to-end. Another line of work leverages internal signals of the generator, such as hidden states or attention patterns (Sun et al., 2025; Lu et al., 2025; Chen et al., 2024), yet these methods are model-specific and cannot generalize well to unseen LLMs. More recent research focuses on fine-grained detection that localizes hallucinated

spans. RAGTruth (Niu et al., 2024) introduces supervised fine-tuning based on span-level annotations, while Multi-View (Ogasa and Arase, 2025) aggregates attention features for token-level detection. RL4HS (Su et al., 2025) further applies GRPO (Shao et al., 2024) to directly optimize span-level F1, and proposes class-aware policy optimization (CAPO) to mitigate precision-recall imbalance. In this paper, we propose a framework based on GRPO that improves hallucination span detection by explicitly enhancing evidence-grounded reasoning during optimization.

**Reasoning Technique in NLP Applications** GRPO (Shao et al., 2024) was originally proposed to enhance mathematical reasoning in LLMs. Subsequent work has extended GRPO to various NLP tasks, including intent classification (Feng et al., 2025) and safety alignment (Li et al., 2025b; Rong et al., 2025). Across these applications, CoT reasoning is widely adopted to support multi-step reasoning and reward optimization. However, existing studies show that imbalanced reward design or task difficulty can induce prediction bias and degrade CoT behaviours (Fan et al., 2025; Laidlaw et al., 2025). Recent work further suggests that incorporating explicit, step-wise weak process-level supervision on CoT can help stabilize reasoning behaviour and improve task performance (Zheng et al., 2025). In this work, we study GRPO-based reasoning for hallucination span detection, analyze the evolution of CoT behaviours during training, and propose an evidence-seeking enhancement tailored to the hallucination span detection task.

## 7 Conclusion

In this work, we study hallucination span detection from a process-level perspective and reveal that detection errors are closely associated with degraded evidence-seeking behaviour during CoT reasoning, where verification steps lack explicit grounding in source evidence. Motivated by this observation, we propose an evidence-grounded reinforcement learning framework that explicitly incentivizes active evidence quotation at each reasoning step, aligning model verification behaviour with human annotation practices. Experiments on RAGTruth and the NewsSum dataset demonstrate consistent improvements in both span-level and sample-level detection performance, as well as more faithful and structured reasoning behaviour, while introducing only limited additional reasoning overhead.

## Limitations

Despite the effectiveness of the proposed method, we acknowledge some limitations in our work:

First, training span-level hallucination detectors relies on high-quality span-level annotations. Future work may leverage strong teacher models to automatically synthesize high-quality evidence-seeking reasoning traces and span-level supervision, where each reasoning step is explicitly grounded in retrieved or quoted evidence. Such synthesized trajectories can be used to supervise student models via SFT, followed by RL for further optimization.

Second, as recent studies explore dynamic agents that leverage multiple tools during the CoT process (Li et al., 2025a; Jin et al., 2025), extending hallucination detection to such dynamic agent-based frameworks represents an interesting direction for future research.

## Ethical Statements

We use the publicly released RAGTruth dataset, which does not contain private personal data. For evaluating our LLM-as-a-judge framework, we employ crowd workers to annotate model outputs with classification labels. The NewsSum dataset is internally curated from publicly available news articles and does not contain private or sensitive personal information. No personal data are released in this work, thereby minimizing the risk of ethical issues.

## Acknowledgment

This work was supported by China Postdoctoral Science Foundation under Grant Number 2025T180440. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

## References

- Orlando Ayala and Patrice Béchar. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 228–238. Association for Computational Linguistics.
- Catarina G. Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. [From single to multi: How llms hallucinate in multi-document summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5276–5309. Association for Computational Linguistics.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. [INSIDE: llms’ internal states retain the power of hallucination detection](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- 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](#). *CoRR*, abs/2307.13528.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. 2025. [Posterior-grpo: Rewarding reasoning processes in code generation](#). *CoRR*, abs/2508.05170.
- Zihao Feng, Xiaoxue Wang, Ziwei Bai, Donghang Su, Bowen Wu, Qun Yu, and Baoxun Wang. 2025. [Improving generalization in intent detection: GRPO with reward-based curriculum sampling](#). *CoRR*, abs/2504.13592.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. [A survey on llm-as-a-judge](#). *CoRR*, abs/2411.15594.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.
- Zhaoheng Huang, Yutao Zhu, Jirong Wen, and Zhicheng Dou. 2026. [Evaluating the factuality of large language models using multiple plug-and-play fact sources](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 41607–41609. AAAI Press.
- Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. [OpenFactCheck: A unified framework for factuality](#)

- evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Juyong Jiang, Fan Wang, Jiashi Shen, Sungju Kim, and Sunghun Kim. 2024. [A survey on large language models for code generation](#). *CoRR*, abs/2406.00515.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. [Search-rl: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Cassidy Laidlaw, Shivam Singhal, and Anca D. Dragan. 2025. [Correlated proxies: A new definition and improved mitigation for reward hacking](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. [Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases](#). *CoRR*, abs/2403.10446.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, Yuan Lu, and Zhicheng Dou. 2025a. [Deepagent: A general reasoning agent with scalable toolsets](#). *CoRR*, abs/2510.21618.
- Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. 2025b. [Optimizing safe and aligned language generation: A multi-objective GRPO approach](#). *CoRR*, abs/2503.21819.
- Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. [Reasonrank: Empowering passage ranking with strong reasoning ability](#). *CoRR*, abs/2508.07050.
- Pengqian Lu, Jie Lu, Anjin Liu, and Guangquan Zhang. 2025. [Spad: Seven-source token probability attribution with syntactic aggregation for detecting hallucinations in rag](#). *Preprint*, arXiv:2512.07515.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [Factscore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. 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](#). *CoRR*, abs/2401.06855.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10862–10878. Association for Computational Linguistics.
- Yuya Ogasa and Yuki Arase. 2025. [Hallucinated span detection with multi-view attention features](#). In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (\*SEM 2025)*, pages 381–394, Suzhou, China. Association for Computational Linguistics.
- Xuankun Rong, Wenke Huang, Tingfeng Wang, Daiguo Zhou, Bo Du, and Mang Ye. 2025. [Safe-grpo: Self-rewarded multimodal safety alignment via rule-governed policy optimization](#). *Preprint*, arXiv:2511.12982.
- Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. [FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.
- Ning Shang, Yifei Liu, Yi Zhu, Li Lyna Zhang, Weijiang Xu, Xinyu Guan, Buze Zhang, Bingcheng Dong, Xudong Zhou, Bowen Zhang, Ying Xin, Ziming Miao, Scarlett Li, Fan Yang, and Mao Yang. 2025. [rstar2-agent: Agentic reasoning technical report](#). *CoRR*, abs/2508.20722.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [Hybridflow: A flexible and efficient RLHF framework](#). In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pages 1279–1297. ACM.

- Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. [RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 1548–1558. Association for Computational Linguistics.
- Hsuan Su, Ting-Yao Hu, Hema Swetha Koppula, Kundan Krishna, Hadi Pouransari, Cheng-Yu Hsieh, Cem Koc, Joseph Yitan Cheng, Oncel Tuzel, and Raviteja Vemulapalli. 2025. [Learning to reason for hallucination span detection](#). *CoRR*, abs/2510.02173.
- 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.
- Liyang Tang, Philippe Laban, and Greg Durrett. 2024. [Minicheck: Efficient fact-checking of llms on grounding documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8818–8847. Association for Computational Linguistics.
- Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi N. Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. [Factuality of large language models: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19519–19529. Association for Computational Linguistics.
- Peter Weiner. 1973. [Linear pattern matching algorithms](#). In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, pages 1–11.
- Jiageng Wu, Kevin Xie, Bowen Gu, Nils Krüger, Kueiyu Joshua Lin, and Jie Yang. 2025. [Why chain of thought fails in clinical text understanding](#). *CoRR*, abs/2509.21933.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *CoRR*, abs/2303.18223.
- Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2025. [Critic-CoT: Boosting the reasoning abilities of large language model via chain-of-thought critic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1768–1806, Vienna, Austria. Association for Computational Linguistics.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Texygen: A benchmarking platform for text generation models](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.
- Yutao Zhu, Zhaoheng Huang, Zhicheng Dou, and Jirong Wen. 2025. [One token can help! learning scalable and pluggable virtual tokens for retrieval-augmented large language models](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26166–26174. AAAI Press.

## A LLM-as-a-Judge

Manually examining the quality of CoT reasoning is time-consuming and costly. Inspired by prior work on LLM-as-a-judge evaluation (Wu et al., 2025), we adopt an automatic judging framework to assess whether the CoT explicitly seeks and quotes evidence from the reference source during verification. Specifically, we evaluate whether a model’s CoT exhibits any evidence-quoting behaviour in its verification steps.

To construct an evaluation set, we first apply the LLM-as-a-judge to samples collected during GRPO training and select a balanced subset of 100 samples, including 50 samples predicted as containing evidence-quoting behaviour and 50 samples predicted as containing none. We use Qwen2.5-14B-Instruct as the LLM judge, with the prompt shown in Prompt 2. If the output label is “2”, the judge determines that the CoT contains no evidence-quoting behaviour.

Dataset	RAGTruth				NewsSum			
	Summarization		Question-Answering		Data-to-Text		Summarization	
	# Pos.	# Neg.	# Pos.	# Neg.	# Pos.	# Neg.	# Pos.	# Neg.
Train	1,469	2,646	1,556	2,732	3,700	1,347	1,000	1,000
Test	204	696	160	715	579	321	500	500

Table 5: Statistics of the RAGTruth and NewsSum datasets used in this paper.

### LLM-as-a-Judge of evidence quoting

You are an evaluator.

Task:

In a RAG hallucination detection setting, a RESPONSE contains step-by-step reasoning. A verification step is any step where the RESPONSE checks whether a specific claim or span matches the reference source.

Your goal is to determine whether the verification steps explicitly use text evidence from the reference source.

Evidence usage is defined strictly:

A verification step uses evidence only if it contains identifiable content from the source, for example:

- directly quoting or closely copying text from the source, introduced with phrasing like “the reference source states ...”, or

- explicitly contrasting a RESPONSE span with a specific source span.

Generic or high-level statements without concrete source content—such as:

- “the claim matches the reference source”,
- “the claim contradicts the reference source”,

do NOT count as using evidence.

Classification: You must assign EXACTLY one of the following labels:

1 → All verification steps explicitly use evidence, or some verification steps use evidence and some do not.

2 → No verification steps use evidence.

If there are zero verification steps, output 2.

Output:

<label>the\_label\_value</label>

[RESPONSE]

{response\_text}

[SOURCE]

{source\_text}

To validate the reliability of this automatic judge, we recruit three educated crowd workers who are proficient in English to independently annotate the same 100 samples under identical instructions, serving as human references. The annotation interface is shown in Figure 6, and annotators are compensated at a rate of 20 USD per hour. Annotating these 100 samples requires approximately 2 hours in total.

We compute Fleiss’ Kappa over the human annotations and obtain a score of 0.71, indicating substantial inter-annotator agreement and reasonable consistency among human judgments. We then take the majority vote of the human annotations as the reference and compare it with the LLM-as-a-judge predictions. Under this setting,

Avg. Length	Train	Test
Input $T$	90.38	91.31
Hallucinated spans $S$	13.44	12.83
Source $D$	1,988.41	1,995.06
# Source News	2.59	2.63

Table 6: Statistics of the NewsSum dataset.

the agreement between the LLM judge (Qwen2.5-14B-Instruct) and the human reference reaches 96%, indicating high consistency on this relevance judgment task.

## B Prompts for CoT with Evidence Quoting

We adopt the CoT framework from Su et al. (2025) and further instruct the model to actively seek and quote relevant evidence from the source at each reasoning step. Figures 7, 8, 9, and 10 present a comparison between the conventional CoT prompting and our proposed CoT+Seek prompting. The additional instructions introduced by CoT+Seek are highlighted in blue.

## C Dataset

Following the settings of Su et al. (2025), we filter out samples in RAGTruth that are marked as *bad quality* and rebalance the dataset by keeping an equal number of hallucinated and non-hallucinated samples. Table 5 reports detailed statistics of the resulting datasets, and statistics of the NewsSum dataset are shown in Table 6. After preprocessing, RAGTruth contains 13,450 training samples and 2,675 test samples. The NewsSum dataset consists of 2,000 training samples (1,000 hallucinated and 1,000 non-hallucinated) and 1,000 test samples (500 hallucinated and 500 non-hallucinated).

## D Implementation Details

**Fine-tuned and Reasoning Models.** We fine-tune the 7B and 14B models using LLaMA Factory (Zheng et al., 2024) with a learning rate of  $1 \times 10^{-6}$ . During inference, following Su et al.

### Evidence Usage Labeling (1/2)

Item 1/100 | ID: 0001 | Current label: Not labeled

**Instruction**  
You are an evaluator.

Task: In a RAG hallucination detection setting, a RESPONSE contains step-by-step reasoning. A verification step is any step where the RESPONSE checks whether a specific claim or span matches the reference source. Your goal is to determine whether the verification steps explicitly use text evidence from the reference source. Evidence usage is defined strictly: A verification step uses evidence only if it contains identifiable content from the source, for example:

- ▷ directly quoting or closely copying text from the source, introduced with phrasing like "the reference source states ..., or
- ▷ explicitly contrasting a RESPONSE span with a specific source span.

Generic or high-level statements without concrete source content—such as

- ▷ "the claim matches the reference source",
- ▷ "the claim contradicts the reference source",

do NOT count as using evidence.

Classification: You must assign EXACTLY one of the following labels:

1 = All verification steps explicitly use evidence, or some verification steps use evidence and some do not.

2 = No verification steps use evidence.

If there are zero verification steps, output 2.

Output:  
the\_label\_value

**Your actions**  
Not labeled

Label 1 (Quoting)	Label 2 (No Quoting)
Previous	Next
Jump to item (1-100)	Go
1	
Labeled count	0
0	100
<a href="#">Save ratings file</a>	<a href="#">Download</a>
ratings_20251231_104222.json	54.0 B

**Step 1: Analyze the structured data for information about Hollister Brewing Company.**

The structured data shows that Hollister Brewing Company, located at "6980 Marketplace Dr, Goleta, CA", is a local business that specializes in American traditional food and beer. It is a brewpub that operates from 11:30 to 19:30 from Monday to Sunday, with extended hours until 20:00 on Fridays and Saturdays.

**Step 2: Compare the article to the structured data.**

The article accurately reflects the information provided in the structured data. Specifically, the article states, "Hollister Brewing Company, located at 6980 Marketplace Dr, Goleta, CA, is a local business that specializes in American traditional food and beer." This is consistent with the structured data. The article states, "It is a brewpub that operates from 11:30 to 19:30 from Monday to Sunday, with extended hours until 20:00 on Fridays and Saturdays." This is consistent with the structured data. The article states, "The establishment offers outdoor seating and free WiFi, accepts reservations, and is good for groups." This is consistent with the structured data. The article states, "It also provides a takeout service and has a parking lot." This is consistent with the structured data. The article states, "Customer reviews highlight mixed feelings towards the business. Some patrons express satisfaction with the new menu introduced by a new chef, praising dishes like the Korean chicken wings and the homemade dressings." This is consistent with the structured data. The article states, "However, other customers have criticized the slow service and staff behavior, including drinking on the job." This is consistent with the structured data. The article states, "The business has an average rating of 3.5 stars." This is consistent with the structured data.

**Step 3: Determine if there are any hallucinations.**

After comparing the article to the structured data, there are no discrepancies. The article accurately reflects the information provided in the structured data.

**Step 4: Compile the hallucinated spans into a JSON format.**

Since there are no hallucinations, the value for the hallucination list will be an empty list.

```
hallucination list []
```

```
{
  "hallucination list": []
}
```

**SOURCE**

```
{
  "name": "Hollister Brewing Company",
  "address": "6980 Marketplace Dr",
  "city": "Goleta",
  "state": "CA",
  "categories": "Restaurants, Food, Breweries, American (Traditional), Brewpubs",
  "hours": {
    "Monday": "11:30-19:30",
    "Tuesday": "11:30-19:30",
    "Wednesday": "11:30-19:30",
    "Thursday": "11:30-19:30",
    "Friday": "11:30-20:00",
    "Saturday": "11:30-20:00"
  }
}
```

Figure 6: The web interface used by human annotators to judge whether the CoT contains evidence-quoting behaviour in any verification step.

(2025), we adopt top-p = 0.95, top-k = 20, and temperature = 0.6 for generation (Holtzman et al., 2020).

**RL-based Models.** For RL-based training, we follow the GRPO configuration of RL4HS (Su et al., 2025). We use a learning rate of  $1 \times 10^{-6}$  for the 7B model and  $5 \times 10^{-7}$  for the 14B model. Rollouts are generated with temperature = 1.0, top-p = 1.0, and top-k = -1, with a clip range  $\epsilon = 0.2$ , KL coefficient  $\beta = 0.001$ , and group size  $|G(\tau)| = 16$ . All experiments are implemented using the VeRL framework (Sheng et al., 2025) and trained for three epochs on 8 NVIDIA A100 GPUs with a batch size of 64. In our proposed method, we employ a suffix automaton (SAM) (Weiner, 1973) to efficiently compute the longest common substring (LCS) in  $\mathcal{O}(n + m)$  time, where  $n$  and  $m$  denote the lengths of the two strings.

## E Evaluation Metric

Following prior work (Niu et al., 2024; Su et al., 2025), we evaluate hallucination span detection performance using the dataset-level span-F1 metric. Let the ground-truth hallucinated spans be  $S = \{s_m = [i_m, j_m]\}_{m=1}^M$ , where each span  $[i_m, j_m]$  corresponds to the set of character indices from  $i_m$  to  $j_m$ . The set of hallucinated charac-

Metric	SFT	GRPO	CAPO	RLSeek
FNR (%)	16.0	23.6	20.2	17.8
FPR (%)	15.2	8.1	9.7	9.1

Table 7: Comparison of FNR and FPR across different training methods on RAGTruth.

ters in the ground truth is  $\mathcal{G} = \bigcup_{m=1}^M s_m$ . Similarly, the predicted spans are denoted as  $\hat{S} = \{s_n = [i_n, j_n]\}_{n=1}^N$ , with the corresponding predicted hallucinated characters  $\mathcal{P} = \bigcup_{n=1}^N s_n$ . The span-level precision, recall, and F1 score are defined as:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}, P = \frac{|\mathcal{P} \cap \mathcal{G}|}{|\mathcal{P}|}, R = \frac{|\mathcal{P} \cap \mathcal{G}|}{|\mathcal{G}|}.$$

We additionally report sample-level binary detection metrics. At the sample level, each prediction is treated as a binary decision, where a sample is classified as hallucinated if  $\mathcal{P} \neq \emptyset$  and as non-hallucinated otherwise. The ground-truth label is defined in the same way based on whether  $\mathcal{G} \neq \emptyset$ . Sample-level precision, recall, and F1 score are then computed based on these binary labels.

```

Below is the original document:
{Reference}
Below is a summary of the document:
{Response}
Your task is to determine whether the summary contains hallucinations.
First, provide reasoning in a step-by-step format:
## Step 1: <your first reasoning step, briefly state what you are going to
check in this step>
If this step is used to verify whether some part of the summary contains
hallucinations, you MUST follow the instructions below:
- First, using only the original document, locate and quote the most
relevant, precise, and shortest self-contained evidence from the original
document for this verification step. Place each evidence verbatim inside:
<quote>...</quote>.
- Then, explain how the evidence supports your evaluation of this part
(supported or hallucinated). If key evidence is missing, explicitly state
the missing evidence and why it indicates hallucination. If this part
contains hallucination, list all hallucinated text spans.
## Step 2: <your next reasoning step, also describe what you are going
to check in this step>
(For later steps, follow the same pattern.)
...(add as many steps as needed)
Then, compile the labeled hallucinated spans into a JSON dict, with a
key hallucination list and its value as a list of hallucinated spans. If
there are hallucinations, output:
```json
{"hallucination list": [hallucination span1, hallucination span2, ...]}
```
Otherwise, use:
```json
{"hallucination list": []}
```
Now begin your output, starting with "##".

```

Figure 7: The CoT and CoT-Seek instruction of the summarization task on the RAGTruth benchmark. The “Seek” instruction is marked blue.

## F Comparison of the False Rates

We further compare two types of sample-level error rates, *i.e.*, the false negative rate (FNR) and the false positive rate (FPR), across different training strategies. The results are summarized in Table 7. Compared with SFT, GRPO substantially reduces the FPR, but at the cost of a noticeably higher FNR. CAPO partially alleviates this imbalance by trading a moderate increase in FPR for a lower FNR. In contrast, our proposed RLSeek achieves a more favorable trade-off by simultaneously reducing both FNR and FPR compared with CAPO, while maintaining a low FPR close to that of GRPO. This suggests that explicitly enforcing evidence-seeking behavior enables more balanced error control in hallucination span detection.

## G Analysis of Penalty Item

In this section, we analyze the effect of the penalty item in the RL objective on quotation behaviour. Specifically, we examine how the quoted content evolves when the explicit constraint on quoting is removed. We consider two metrics on the test set:

```

Below is a question:
{Question}
Below are the related passages:
{Reference}
Below is an answer:
{Response}
Your task is to determine whether the answer contains hallucinations.
First, provide reasoning in a step-by-step format:
## Step 1: <your first reasoning step, briefly state what you are going to
check in this step>
If this step is used to verify whether some part of the answer contains
hallucinations, you MUST follow the instructions below:
- First, using only the related passages, locate and quote the most
relevant, precise, and shortest self-contained evidence from the related
passages for this verification step. Place each evidence verbatim inside:
<quote>...</quote>.
- Then, explain how the evidence supports your evaluation of this part
(supported or hallucinated). If key evidence is missing, explicitly state
the missing evidence and why it indicates hallucination. If this part
contains hallucination, list all hallucinated text spans.
## Step 2: <your next reasoning step, also describe what you are going
to check in this step>
(For later steps, follow the same pattern.)
...(add as many steps as needed)
Then, compile the labeled hallucinated spans into a JSON dict, with a
key hallucination list and its value as a list of hallucinated spans. If
there are hallucinations, output:
```json
{"hallucination list": [hallucination span1, hallucination span2, ...]}
```
Otherwise, use:
```json
{"hallucination list": []}
```
Now begin your output, starting with "##".

```

Figure 8: The CoT and CoT-Seek instruction of the question-answering task on the RAGTruth benchmark.

(1) the percentage of invalid quotation sets, and (2) the average LCS overlap between the quoted content and the source. The results are shown in Figure 11. With the penalty item, the overlap between the quoted content and the source rapidly increases and reaches nearly 100% after 50 RL steps, while the percentage of invalid quotation sets remains low. In contrast, without the penalty item, the proportion of valid quotation sets decreases by approximately 4% during training, and the LCS overlap degrades to around 50%–60%. We further evaluate the **word-level longest common subsequence** between the quoted content and the source. As shown in Figure 12, removing the penalty item leads to a severe degradation of overlap, dropping to approximately 40%. These results indicate that, without an explicit constraint, the model tends to summarize or paraphrase the source instead of faithfully quoting it. Such behaviour weakens the alignment between quoted content and the source, leading to degraded overall performance, as further reflected in Table 4.

```

Below is structured data in JSON format:
{Reference}
Below is an overview article written in accordance with the structured
data:
{Response}
Your task is to determine whether the article contains hallucinations.
First, provide reasoning in a step-by-step format:
## Step 1: <your first reasoning step, briefly state what you are going to
check in this step>
If this step is used to verify whether some part of the article contains
hallucinations, you MUST follow the instructions below:
- First, using only the structured data, locate and quote the most relevant,
precise, and shortest self-contained evidence from the structured data
for this verification step. Place each evidence verbatim inside:
<quote>...</quote>.
- Then, explain how the evidence supports your evaluation of this part
(supported or hallucinated). If key evidence is missing, explicitly state
the missing evidence and why it indicates hallucination. If this part
contains hallucination, list all hallucinated text spans.
## Step 2: <your next reasoning step, also describe what you are going
to check in this step>
(For later steps, follow the same pattern.)
...(add as many steps as needed)
Then, compile the labeled hallucinated spans into a JSON dict, with a
key hallucination list and its value as a list of hallucinated spans. If
there are hallucinations, output:
```json
{"hallucination list": [hallucination span1, hallucination span2, ...]}
```
Otherwise, use:
```json
{"hallucination list": []}
```
Now begin your output, starting with "##".

```

Figure 9: The CoT and CoT-Seek instruction of the data-to-text task on the RAGTruth benchmark.

## H Analysis of Quoting Behaviour

Beyond analyzing the length of CoT and the frequency of quoting, we further investigate the *diversity* of the quoted evidence used during reasoning as well as the *relevance* between each quoted evidence and its corresponding verified claim. Quotation diversity and relevance are crucial for diagnosing **potential reward hacking behaviours**, where the model may repeatedly quote the same source segments to obtain shortcut rewards without performing genuine evidence seeking.

**Diversity.** To quantify the diversity of the quotation set in CoT, we adopt two complementary metrics. Self-BLEU (Zhu et al., 2018) measures lexical overlap among generated quotations, where lower values indicate less redundancy. Distinct- $n$  (Li et al., 2016) evaluates the proportion of unique  $n$ -grams in the quoted content, with higher values reflecting greater lexical diversity. Together, a low Self-BLEU score and a high Distinct- $n$  score suggest that the model actively explores diverse evidence rather than repeatedly quoting identical or highly similar source spans.

```

You will be given a claim and the source material from which
the claim is derived.
Below is the source material:
---Begin of Content---
{Source}
---End of Content---
Below is the claim:
---Begin of Content---
{Claim}
---End of Content---
### **Output Format**
First, provide reasoning in a step-by-step format:
## Step 1: <your first reasoning step, briefly state what you are
going to check in this step>
If this step is used to verify whether some part of the claim
contains hallucinations, you MUST follow the instructions
below:
- First, using only the source material, locate and quote the
most relevant, precise, and shortest self-contained evidence
from the source material for this verification step. Place the
evidence verbatim inside: <quote>...</quote>.
- Then, explain how the evidence supports your evaluation of
this part (supported or hallucinated). If key evidence is missing,
explicitly state the missing evidence and why it indicates
hallucination. If this part contains hallucination, list all
hallucinated text spans.
## Step 2: <your next reasoning step, also describe what you
are going to check in this step>
(For later steps, follow the same pattern.)
...(add as many steps as needed)
<answer>[Output all hallucinated spans list]</answer>

```

Figure 10: The CoT and CoT-Seek instruction of the news summarization task on the NewsSum dataset.

As shown in Figure 13, during RL optimization, the Self-BLEU score consistently remains low while the Distinct- $n$  score stays high. This trend indicates that the model maintains a diverse quotation set throughout training. Importantly, these results suggest that the learned quoting behaviour does not collapse into repetitive patterns, but instead reflects genuine evidence seeking from different relevant source segments, thereby mitigating the risk of reward hacking through shortcut reasoning.

**Relevance.** We apply an LLM-as-a-judge framework based on Qwen2.5-14B-Instruct to assess whether the quoted evidence is semantically relevant to the corresponding verification claim. To validate the reliability of this judgment, we randomly sample 100 instances for human annotation. The inter-annotator agreement among three annotators reaches 0.63 (Fleiss'  $\kappa$ ), and the LLM judge achieves an accuracy of 92% with respect to the majority vote of human annotations. Figure 14 presents the proportion of quoted evidence judged to be relevant to the corresponding claims at each training checkpoint on the test set. The results show

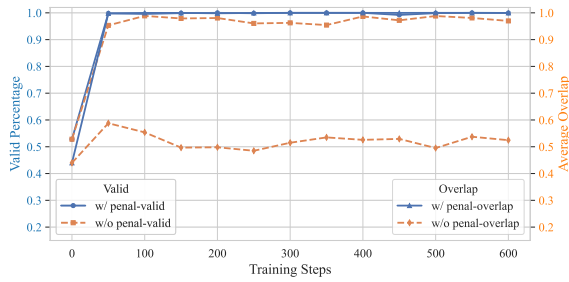


Figure 11: Effect of the penalty item in the RL objective on quotation validity and source overlap.

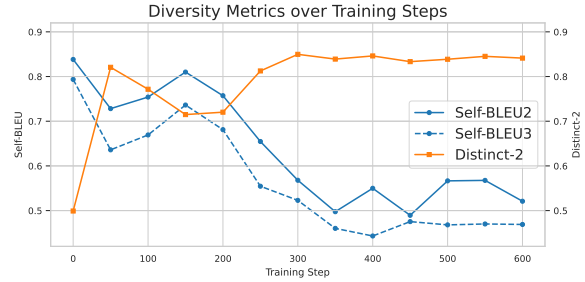


Figure 13: Analysis of diversity of quoting behaviours, measured by Self-BLEU and Distinct-2.

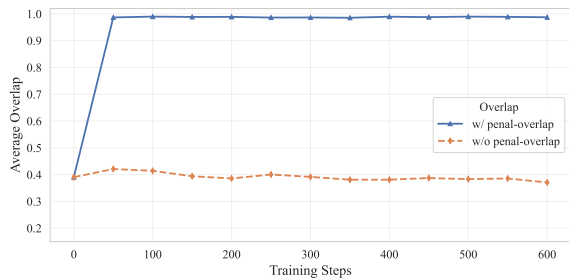


Figure 12: Effect of the RL penalty on source overlap measured by longest common subsequence.

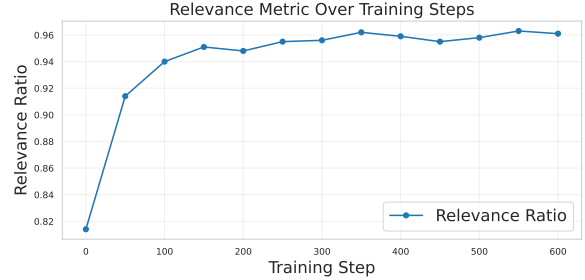


Figure 14: Analysis of relevance of quoting behaviours.

that the relevance of quoted content remains largely stable throughout RL optimization, with approximately 96% of the quoted evidence being relevant to the verified claims.

## I Case Study

We present three representative cases for each task in the RAGTruth benchmark, comparing our method with the CAPO baseline. The results are illustrated in Figure 15, Figure 16, and Figure 17. In all cases, hallucinated spans are highlighted in red, while quoted evidence is marked with a yellow background.

Figure 15 shows a hallucinated sample that is incorrectly predicted as non-hallucinated by CAPO. During the CoT process, CAPO initially identifies potential hallucinated spans in Step #3; however, due to the absence of explicit quoted evidence, the subsequent reasoning degenerates into largely repeating the input text and ultimately revises the final decision to non-hallucination. This inconsistency across reasoning steps indicates that, without grounding intermediate judgments in source evidence, the model can lose track of its earlier analysis, leading to unreliable predictions. The case also demonstrates that fabricated spans can be reliably identified through explicit source-grounded verifi-

cation. In Step #2 of RLSeek, the model locates the most relevant source segments and explicitly verifies that the examined claim is neither supported nor refuted by the given passages, thereby correctly identifying it as hallucinated.

**LLM-as-a-Judge of evidence relevance**

You are an evaluator.

Task: You are given a CLAIM from a verification step and a piece of QUOTED EVIDENCE extracted from a source document. Your task is to judge whether the quoted evidence is semantically relevant to the claim.

Definition of relevance: The quoted evidence is considered RELEVANT if it directly supports, contradicts, or is explicitly used to verify the specific claim. The quoted evidence is considered IRRELEVANT if it is unrelated, too generic, or cannot be used to assess the claim.

Important constraints: - Judge ONLY relevance, not correctness. - Do NOT judge whether the claim is true or false. - Do NOT infer missing evidence or search for other evidence. - Base your judgment solely on the semantic relationship between the given claim and the quoted evidence. - If the evidence is loosely related but does not clearly support or contradict the claim, label it as IRRELEVANT.

Input: CLAIM:  
{claim}

QUOTED EVIDENCE:  
{quote}

Output: (RELEVANT or IRRELEVANT)

Figure 16 presents a hallucinated sample where both methods identify partial hallucination. However, CAPO fails to detect a critical inconsistency between the source and the generated text, specif-

**Source D:**  
“passages”: “passage 1:An ad hominem (Latin for to the man or to the person), short for argumentum ad hominem, is an attack on an argument made by attacking the character, motive, or other attribute of the person making the argument, rather than attacking the argument directly. passage 2:Abusive ad hominem usually involves attacking the traits of an opponent as a means to invalidate their arguments. Equating someone’s character with the soundness of their argument is a logical fallacy. passage 3:Poisoning the Well Fallacy. Poisoning the well is an ad hominem attack (see “Ad hominem,” above) on a person’s integrity or intelligence that takes place before the merits of a case can be considered. It redirects a discussion to the faults of one of the parties.”

**Input Text T:**  
Based on the provided passages, the main difference between ad hominem fallacy and poisoning the well fallacy is the timing of the attack. Ad hominem fallacy occurs when an attack is made on a person’s character, motive, or other attribute directly related to their argument, rather than attacking the argument itself. **This type of attack can happen at any point during a debate or discussion.** On the other hand, poisoning the well fallacy involves making an ad hominem attack (i.e., attacking a person’s character or integrity) before the merits of a case can be considered. In other words, it is an attempt to discredit a person’s argument by attacking their personal integrity or intelligence before the discussion has even begun. So, while both fallacies involve attacking a person’s character rather than their argument, ad hominem fallacy happens after the argument has been presented, while poisoning the well fallacy happens before the argument is even presented.

**CAPO: (Span-F1=0)**  
## Step 1: Analyzing the passages for information about the difference between ad hominem fallacy and poisoning the well fallacy. The passages provide information about both ad hominem fallacy and poisoning the well fallacy, but do not specify the timing of the attacks. The passages only describe the nature of the attacks and do not mention anything about the timing of the attacks.## Step 2: Comparing the given answer with the information provided in the passages. The answer contains information about the difference between ad hominem fallacy and poisoning the well fallacy, but the timing of the attacks is not specified. The answer states that ad hominem fallacy occurs when an attack is made on a person’s character, motive, or other attribute directly related to their argument, rather than attacking the argument itself. This type of attack can happen at any point during a debate or discussion.## Step 3: Identifying hallucinated spans. The hallucinated spans are “Ad hominem fallacy occurs when an attack is made on a person’s character, motive, or other attribute directly related to their argument, rather than attacking the argument itself. This type of attack can happen at any point during a debate or discussion.” “On the other hand, poisoning the well fallacy involves making an ad hominem attack (i.e., attacking a person’s character or integrity) before the merits of a case can be considered. In other words, it is an attempt to discredit a person’s argument by attacking their personal integrity or intelligence before the discussion has even begun.”## Step 4: Compiling the hallucination list. The hallucination list contains “Ad hominem fallacy occurs when an attack is made on a person’s character, motive, or other attribute directly related to their argument, rather than attacking the argument itself. This type of attack can happen at any point during a debate or discussion.” “On the other hand, poisoning the well fallacy involves making an ad hominem attack (i.e., attacking a person’s character or integrity) before the merits of a case can be considered. In other words, it is an attempt to discredit a person’s argument by attacking their personal integrity or intelligence before the discussion has even begun.”## Step 5: Formatting the final output. The final output should be in the specified JSON format. ``json{ "hallucination list": [] }``

**RLSeek: (Span-F1=0.62)**  
## Step 1: Verify the statement “ad hominem fallacy happens after the argument has been presented, while poisoning the well fallacy happens before the argument is even presented.”- \*\*Evidence from passages:\*\* <quote> argumentum ad hominem, is an attack on an argument made by attacking the character, motive, or other attribute of the person making the argument </quote><quote>Poisoning the well is an ad hominem attack (see “Ad hominem,” above) on a person’s integrity or intelligence that takes place before the merits of a case can be considered.</quote>- \*\*Evaluation:\*\* The passages confirm that an ad hominem fallacy is an attack on the character, motive, or other attribute of the person making the argument, rather than attacking the argument directly. The passages also mention that an ad hominem attack can take place before the merits of a case can be considered.- \*\*Answer:\*\* The answer mentions that ad hominem fallacy happens after the argument has been presented, which is not supported by the passages. The passages do not specifically mention that ad hominem fallacy happens after the argument has been presented.  
## Step 2: Verify the statement “This type of attack can happen at any point during a debate or discussion.”- \*\*Evidence from passages:\*\* <quote> An ad hominem (Latin for to the man or to the person), short for argumentum ad hominem, is an attack on an argument made by attacking the character, motive, or other attribute of the person making the argument </quote>- \*\*Evaluation:\*\* The passages do not specifically mention that ad hominem fallacy can happen at any point during a debate or discussion. This is an additional claim not supported or refuted by the given passages.- \*\*Answer:\*\* The answer mentions that ad hominem fallacy can happen at any point during a debate or discussion, which is not supported by the passages.## Final Output ``json{ "hallucination list": [ "happens after the argument has been presented", "can happen at any point during a debate or discussion", "happens after the argument has been presented" ] }``

Figure 15: Case study #1 of the Proposed method and the baseline CAPO.

ically the claim about “today’s temperature in Bucharest”, as it does not explicitly ground its comparison in the source evidence. In this case, **the source text does not contain any information about Bucharest’s temperature on the specific day** referenced by the claim. In contrast, RLSeek follows an evidence-grounded verification process by first identifying and quoting the most relevant source segments, namely **the temperature of Bucharest reported for July and August**, using the <quote> tag. Although this quoted evidence does not directly correspond to the claimed “today’s temperature”, it provides the closest topic-aligned information available in the source. Based on this evidence, RLSeek correctly infers that the claimed temperature is neither supported nor refuted by the given passages, and therefore identifies the corresponding span as hallucinated. This explicit evidence-seeking and grounding process enables more precise verification and results in a higher span-F1 score.

Figure 17 illustrates a non-hallucinated sample.

The most ambiguous fact concerns the songwriter information, which is expressed differently in the source as “Songwriters. DYLAN, BOB. Published by.” Without explicit evidence quoting, CAPO incorrectly flags this content as hallucinated in Step #3. In contrast, RLSeek quotes the relevant sentence from the source *D* at the first reasoning step and consistently leverages it in subsequent analysis, correctly determining that the claim is supported.

Overall, these case studies demonstrate that explicit evidence quoting stabilizes the reasoning process by grounding intermediate steps in verifiable source content, thereby improving overall performance. In addition, we observe that CAPO frequently wastes tokens by repeatedly restating the input text across reasoning steps, whereas active quoting leads to more structured and concise CoT reasoning. After reinforcement learning optimization, the reasoning length of our method becomes comparable to that of the baseline, as shown in Figure 4, without sacrificing reasoning clarity or performance.

**Source D:**  
 "passages": "passage 1:Today's and tonight's Bucharest, Romania weather forecast, weather conditions and Doppler radar from The Weather Channel and Weather.com Today's and tonight's Bucharest, Romania weather forecast, weather conditions and Doppler radar from The Weather Channel and Weather.com Connection restored! Show me the weather in... city, zip, or placepassage 2:UK Weather Warnings Map; UK Flood Warnings Map; Current Weather; Highway Conditions; Highway Forecast; Lighting; Current & Future Radar; Past & Future Radar (Precipitation) Traffic Cameras; Traffic Flow; US Highway Forecast; Satellite (Cloud Cover)passage 3:CLIMATE: AVERAGE MONTHLY WEATHER IN Bucharest, Romania. Bucharest has a humid continental climate. Winters can get windy and temperatures often dip below 0 °Celsius (32 °Fahrenheit), sometimes even dropping to minus 20 °C (4 °F). In summer (July and August), the average temperature is around 23 °C (73 °F). Temperatures often reach 35 °C (95 °F) to 40 °C (104 °F) in mid-summer."

**Input Text T:**  
 Based on the provided passages, I can answer the following question: Question: What is the current temperature in Bucharest? Passage 1 provides the current weather conditions for Bucharest, including the temperature:\`Today's temperature in Bucharest is 22°C (72°F), humidity is 82%, wind speed is 6 mph (9 km/h).\`Therefore, the current temperature in Bucharest is 22°C (72°F).

**CAPO: (Span-F1=0.45)**  
 ## Step 1: Analyzing the passages for information about the current temperature in Bucharest. The passages do not specify the current temperature in Bucharest. The passages only provide information about the range of temperatures and do not specify the exact temperature.## Step 2: Comparing the given answer with the information provided in the passages. The answer \`Passage 1 provides the current weather conditions for Bucharest, including the temperature:\`Today's temperature in Bucharest is 22°C (72°F), humidity is 82%, wind speed is 6 mph (9 km/h).\` is not mentioned in the passages. The passages only provide the range of temperatures and do not specify the exact temperature. The humidity and wind speed is not grounded in the passages. ## Step 3: Identifying hallucinated spans. The hallucinated spans are \`humidity is 82%, wind speed is 6 mph (9 km/h).\`## Step 4: Compiling the hallucination list. The hallucination list contains \`humidity is 82%, wind speed is 6 mph (9 km/h).\`## Step 5: Formatting the final output. The final output should be in the specified JSON format.``json{ "hallucination list": [ "humidity is 82%, wind speed is 6 mph (9 km/h)." ] }``

**RLSeek: (Span-F1=0.73)**  
 ## Step 1: Verify the claim that \`Today's temperature in Bucharest is 22°C (72°F), humidity is 82%, wind speed is 6 mph (9 km/h).\`- \*\*Evidence from passages:\*\*`<quote>Today's and tonight's Bucharest, Romania weather forecast, weather conditions and Doppler radar from The Weather Channel and Weather.com</quote><quote>In summer (July and August), the average temperature is around 23 °C (73 °F). Temperatures often reach 35 °C (95 °F) to 40 °C (104 °F) in mid-summer.</quote>- \*\*Evaluation:\*\*`The passages do not mention anything about the current temperature in Bucharest being 22°C (72°F) or the humidity being 82% or the wind speed being 6 mph (9 km/h). These are not supported or refuted by the given passages. They appear to be hallucinations.`## Final Output``json{ "hallucination list": [ "humidity is 82%, wind speed is 6 mph (9 km/h)", "Today's temperature in Bucharest is 22°C (72°F)" ] }`

Figure 16: Case study #2 of the Proposed method and the baseline CAPO.

## J Use of AI Assistants

We use ChatGPT to assist with writing clarity and presentation.<sup>3</sup>

<sup>3</sup><https://chatgpt.com/>

**Source D:**  
 "passages": "passage 1:In 1963, the Greenbriar Boys recorded He Was a Friend of Mine with singer Dian James on their Elektra Records album Dian & the Greenbriar Boys. The Briarwood Singers, a five-piece folk group, released a version of the song that reached #126 on the Billboard charts in December 1963. Bobby Bare also recorded He Was a Friend of Mine in 1964, in memory of air crash victim Jim Reeves. That same year, Petula Clark released a French version of the song under the title Toi qui m'as fait pleurer (=You, who have made me cry), with Bobby Bare credited as writer. passage 2:He was a friend of mine he was a friend of mine. He died without a penny he didn't have a dime he was a friend of mine. He died neath the midnight sun he died neath the midnight sun. passage 3:And I never been quite satisfied. And he was a friend of mine. He never done no wrong. He never done no wrong. A thousand miles from home. And he never harmed no one. And he was a friend of mine. He was a friend of mine. He was a friend of mine. Every time I hear his name. Lord, I just can't keep from cryin'. 'Cause he was a friend of mine. Songwriters. DYLAN, BOB. Published by. "

**Input Text T:**  
 Based on the given passages, "He Was a Friend of Mine" is a song that has been recorded by various artists such as the Greenbriar Boys, Bobby Bare, and Petula Clark. It was also written by Bob Dylan.

**CAPO: (Span-F1=0)**  
 ## Step 1: Analyze the answer for accuracy in terms of the passages provided. The answer states: Based on the given passages, "He Was a Friend of Mine" is a song that has been recorded by various artists such as the Greenbriar Boys, Bobby Bare, and Petula Clark. It was also written by Bob Dylan.## Step 2: Compare the answer with the passages. Passage 1: In 1963, the Greenbriar Boys recorded He Was a Friend of Mine with singer Dian James on their Elektra Records album Dian & the Greenbriar Boys. the Briarwood Singers, a five-piece folk group, released a version of the song that reached #126 on the Billboard charts in December 1963. Bobby Bare also recorded He Was a Friend of Mine in 1964, in memory of air crash victim Jim Reeves. That same year, Petula Clark released a French version of the song under the title Toi qui m'as fait pleurer (=You, who have made me cry), with Bobby Bare credited as writer.## Step 3: Identify the hallucinations. The answer includes information that is not mentioned in the passages, such as the list of artists who recorded the song and the information about the song's writer.## Step 4: Compile the hallucination list. The hallucination list should include the following: 1. "the Greenbriar Boys, Bobby Bare, and Petula Clark"2. "the song was also written by Bob Dylan"## Step 5: Output the hallucination list in JSON format. hallucination list: [{"it was also written by Bob Dylan"}]

**RLSeek: (Span-F1=1)**  
 ## Step 1: Verify the claim that "He Was a Friend of Mine" was written by Bob Dylan. - \*\*Evidence from passages:\*\* <quote> 'Cause he was a friend of mine. Songwriters. DYLAN, BOB. Published by </quote>. \*\*Evaluation:\*\* The passages confirm that Bob Dylan was one of the songwriters for "He Was a Friend of Mine." Therefore, the statement "It was also written by Bob Dylan" is supported.## Step 2: Verify the claim that "He Was a Friend of Mine" was recorded by various artists such as the Greenbriar Boys, Bobby Bare, and Petula Clark. - \*\*Evidence from passages:\*\* <quote> In 1963, the Greenbriar Boys recorded He Was a Friend of Mine with singer Dian James on their Elektra Records album Dian & the Greenbriar Boys. The Briarwood Singers, a five-piece folk group, released a version of the song that reached #126 on the Billboard charts in December 1963. Bobby Bare also recorded He Was a Friend of Mine in 1964, in memory of air crash victim Jim Reeves. That same year, Petula Clark released a French version of the song under the title Toi qui m'as fait pleurer (=You, who have made me cry), with Bobby Bare credited as writer. </quote>. \*\*Evaluation:\*\* The passages confirm that "He Was a Friend of Mine" was recorded by the Greenbriar Boys, Bobby Bare, and Petula Clark. Therefore, the statement is supported.## Final Output ``json { "hallucination list": [ ] }

Figure 17: Case study #3 of the Proposed method and the baseline CAPO.