Joint Evaluation of Answer and Reasoning Consistency for Hallucination Detection in Large Reasoning Models

Changyue Wang[1, 2], **Weihang Su**[1], **Qingyao Ai**[2, 1*], **Yiqun Liu**[1]

[1]Department of Computer Science and Technology, Tsinghua University
[2]Quan Cheng Laboratory

Abstract

Large Reasoning Models (LRMs) extend large language models with explicit, multi-step reasoning traces to enhance transparency and performance on complex tasks. However, these reasoning traces can be redundant or logically inconsistent, becoming a new and hard-to-detect source of hallucination. Existing hallucination detection methods focus primarily on answer-level uncertainty and often fail to detect hallucinations or logical inconsistencies arising from the model's reasoning trace. This oversight is particularly problematic for LRMs, where the explicit thinking trace is not only an important support to the model's decision-making process but also a key source of potential hallucination. To this end, we propose RACE (Reasoning and Answer Consistency Evaluation), a novel framework specifically tailored for hallucination detection in LRMs. RACE operates by extracting essential reasoning steps and computing four diagnostic signals: inter-sample consistency of reasoning traces, entropy-based answer uncertainty, semantic alignment between reasoning and answers, and internal coherence of reasoning. The joint utilization of these signals makes RACE a more robust detector of hallucinations in LRMs. Experiments across datasets and different LLMs demonstrate that RACE outperforms existing hallucination detection baselines, offering a robust and generalizable solution for evaluating LRMs.

Code — https://github.com/bebr2/RACE

Introduction

Large Reasoning Models (LRMs) have recently emerged as a subclass of large language models (LLMs) specifically optimized for long-sequence, stepwise reasoning (DeepSeek-AI et al. 2025; Team 2025; GLM et al. 2024; OpenAI 2024; DeepMind 2025). These models are trained with large-scale supervised fine-tuning and reinforcement learning to produce not only final answers, but also explicit and often lengthy reasoning traces that outline the model's decision-making process. Such traces have demonstrated clear advantages in complex tasks such as multi-hop question answering, code generation, and mathematical problem solving (Zhong et al. 2024). While these explicit reasoning traces enhance transparency and often boost task performance, they also introduce new challenges: they can be redundant or logically inconsistent (Arcuschin et al. 2025; Chen et al. 2025b). Such issues may lead the model to produce factually incorrect or misleading conclusions, even if the reasoning appears plausible. Consequently, hallucination detection (Zhang et al. 2023; Manakul, Liusie, and Gales 2023; Su et al. 2024a,b) for LRMs must consider not only whether the final answer is correct, but also whether the underlying reasoning trace is coherent and well-grounded.

Despite these challenges, most existing black-box hallucination detection methods focus solely on output-level uncertainty, typically by sampling multiple answers and measuring their consistency (Grewal, Bonilla, and Bui 2024; Abdaljalil et al. 2025). Techniques like SelfCheckGPT (Manakul, Liusie, and Gales 2023) and Semantic Entropy (Kuhn, Gal, and Farquhar 2023) follow this paradigm and are effective when model outputs are short and self-contained. However, such methods fail when applied to LRMs, where a significant portion of the inference behavior is embedded within the reasoning trace (Ke et al. 2025). In practice, multiple samples may converge on the same final answer but follow divergent or incoherent reasoning paths. These reasoning inconsistencies often signal hallucinations that remain undetected when evaluation is limited to final answers. Therefore, hallucination detection in LRMs demands a broader perspective that explicitly incorporates both the model's reasoning trace and its alignment with the final answer.

To this end, we propose **RACE** (**R**easoning and **A**nswer **C**onsistency **E**valuation), a novel black-box hallucination detection framework designed for LRM, which explicitly integrates both the reasoning trace and the final answer into a unified evaluation. RACE moves beyond traditional answer-level approaches by jointly evaluating the model's full reasoning–answer behavior. It decomposes hallucination detection into four complementary components: (1) reasoning consistency, which captures the diversity and coherence of reasoning traces across multiple generations; (2) answer uncertainty, measured via refined semantic entropy estimation; (3) reasoning–answer alignment, which evaluates whether the LRM's main reasoning trace, when treated as context, consistently leads to the sampled final answers with high predictive confidence, indicating that the rationale genuinely supports the model-generated answer space; and (4) reasoning internal coherence, measuring the proportion of specu-

*Corresponding author: aiqy@tsinghua.edu.cn

lative content within the reasoning path. To mitigate the impact of noise in reasoning paths, RACE further includes a chain-of-thought (CoT) (Wei et al. 2022) Extraction module that distills the most relevant reasoning steps for each answer. This integrated evaluation enables RACE to detect hallucinations that previous methods may overlook, especially in cases where the sampled final answers appear semantically consistent but are derived from flawed reasoning.

We validate RACE's effectiveness across various datasets under black- or gray-box settings, where neither our method nor competing baselines require task-specific fine-tuning. We compare RACE with a wide range of existing hallucination detection methods, including probability-based metrics such as Length-Normalized Predictive Entropy (LNPE) (Malinin and Gales 2020), semantic uncertainty-based approaches (e.g., Semantic Entropy, SINdex (Abdaljalil et al. 2025)), and SelfCheckGPT, which focuses on sampled answer consistency. Across various LLMs, including both general-purpose models and those optimized for reasoning, RACE achieves the best overall performance among all baselines. Further ablation studies confirm the complementary value of each module in our framework and underscore the necessity of modeling reasoning traces for effective hallucination detection. Taken together, our results highlight RACE as a robust, generalizable solution to hallucination detection in models with explicit reasoning behavior.

To summarize, our contributions are as follows:

- We introduce **RACE**, a novel black-box hallucination detection framework specifically designed for Large Reasoning Models. Unlike prior work that focuses solely on final answers, RACE systematically incorporates both the reasoning trace and answer into a joint evaluation.
- We design four complementary modules (reasoning consistency, answer uncertainty, reasoning–answer alignment, and reasoning internal coherence) and a reasoning distillation mechanism to jointly capture intra- and inter-sample inconsistencies for hallucination detection.
- We conduct experiments across multiple benchmarks and model families, showing that RACE consistently outperforms existing hallucination detection baselines on LRMs, while generalizing effectively to standard LLMs.

Related Works

Large Reasoning Models (LRMs)

Recent work on LRMs enhances the capabilities of LLMs by introducing explicit and multi-step reasoning mechanisms (Besta et al. 2025). By integrating planning, reinforcement learning, and process supervision, LRMs can address more complex tasks than conventional LLMs (Zhong et al. 2024; Guan et al. 2025). However, despite these gains, recent studies show that LRMs remain prone to factual hallucinations (Hughes et al. 2025; Lu et al. 2025), which limits their reliability in knowledge-intensive applications.

Hallucination Detection

This work focuses on factual hallucination, a phenomenon that LLMs can generate text that seems accurate but lacks factual accuracy (Huang et al. 2023). Researchers have developed several techniques to identify these hallucinations. SAPLMA (Azaria and Mitchell 2023) and MIND (Su et al. 2024b) show that one can train a hallucination classifier based on the model's internal states, while INSIDE (Chen et al. 2024) utilizes the covariance matrix of the internal states for detection. Binkowski et al. (2025) employs spectral properties of attention maps to identify hallucinations. However, these methods require white-box access to the LLM, limiting their widespread use. Methods based on sampling consistency are gaining traction due to their black- or gray-box nature. SelfCheckGPT (Manakul, Liusie, and Gales 2023) detects hallucinations by measuring discrepancies between sampled answers and the primary answer. Semantic Entropy (Kuhn, Gal, and Farquhar 2023) uses a clustering approach, replacing token-level entropy with semantic-level uncertainty, while SINdex (Abdaljalil et al. 2025) enhances this by assessing intra- and inter-cluster consistency. Grewal, Bonilla, and Bui (2024) propose evaluating consistency through pairwise similarity of full response embeddings. While these methods primarily target shorter outputs, they may overlook the reasoning process, leaving room to improve hallucination detection for LRMs.

Foundations of the RACE Framework

In this section, we introduce an information-theoretic formulation of hallucination detection that motivates the design of our proposed RACE framework. While existing sampling-based approaches typically assess output uncertainty by sampling multiple answers and measuring their semantic consistency (Kuhn, Gal, and Farquhar 2023; Manakul, Liusie, and Gales 2023), such methods are primarily effective for short, self-contained responses that do not involve complex intermediate reasoning. This assumption, however, breaks down in the context of large reasoning models (LRMs), which produce explicit and often lengthy reasoning traces as part of their outputs. In these cases, a substantial portion of the model's inference is embedded within the reasoning process itself. As a result, even when final answers appear consistent across samples, their corresponding reasoning paths can diverge significantly, revealing hallucinations that remain undetected by answer-only metrics. Figure 1 demonstrates how agreement in final answers can fail to reveal inconsistencies in the underlying reasoning traces. Although most sampled outputs reach the same answer ("New York"), their reasoning paths differ substantially and contain hallucinated or unsupported claims, emphasizing the need for reasoning-aware hallucination detection.

To better characterize this phenomenon, we adopt an information-theoretic perspective, modeling the joint uncertainty over reasoning R and answer A given the question Q. Specifically, we decompose the joint entropy as:

$$H(R, A \mid Q) = H(R \mid Q) + H(A \mid Q) - I(R, A \mid Q), \quad (1)$$

This decomposition reveals three complementary sources of uncertainty and alignment:

- $H(R \mid Q)$: uncertainty in reasoning traces;
- $H(A \mid Q)$: variability in final predictions;

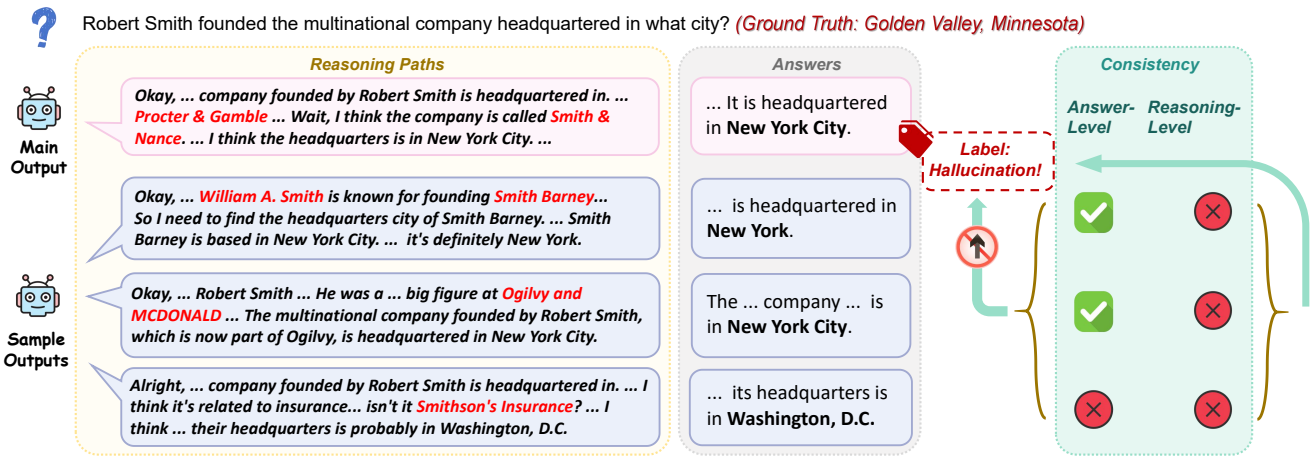

Figure 1: This figure illustrates the importance of incorporating reasoning paths in hallucination detection for LRMs. The example is from HotpotQA. The model is DeepSeek-R1-Distill-Qwen7B. While three of the four final answers mention "New York City," the underlying reasoning traces reveal divergent and often inaccurate chains of logic. Red highlights indicate hallucinated or unsupported claims that are absent from the final answers. Thus, evaluating only the answer-level output would misleadingly suggest that the model is consistent with the answer, underscoring the need for reasoning-aware hallucination detection.

- $I(R, A \mid Q)$: consistency between reasoning and answer space, i.e., how strongly the reasoning supports the final answer space.

Based on the above formulation, it becomes natural to consider hallucination detection as a unified assessment of three factors: the variability of reasoning traces, the uncertainty of final answers, and the degree of alignment between the two. This perspective provides a principled foundation for developing more comprehensive frameworks for hallucination detection. In the next section, we present RACE, a practical implementation of this formulation that quantifies each component through corresponding scoring mechanisms.

Methodology

Building on the information-theoretic formulation introduced above, we now present RACE, a hallucination detection framework that evaluates consistency across reasoning and answer components. In this section, we begin by defining notations, then describe a CoT Extraction module to distill key reasoning steps. Finally, we present the scoring framework that evaluates hallucination risk based on reasoning consistency, answer uncertainty, reasoning-answer alignment, and internal coherence, inspired by Equation 1.

Problem Formulation

Following the paradigm of existing consistency-based hallucination detection methods, we generate a main output O_{main} using greedy decoding, which better reflects the model's default inference-time behavior, and sample N additional outputs $O_1, \ldots, O_N$ for consistency evaluation. Each output O comprises a reasoning component (R) and a final answer (A). The goal is to perform a binary classification to determine whether the main answer A_{main} contains factual hallucinations. Note that RACE is designed to accommodate various model types, including not only LRMs but also standard instruction-tuned LLMs, each of which produce outputs in different formats. We consider three representative settings:

- **LRMs:** The model explicitly outputs a reasoning trace R (often enclosed in tags like `<think>...</think>`).
- **Chain-of-Thought from Standard LLMs:** The model is prompted (e.g., "think step by step") to generate reasoning before the answer, where R corresponds to the CoT segment and A is the concluding answer.
- **Direct Answers from Standard LLMs:** We also consider outputs where no explicit reasoning prompt is given; in this case, we treat the entire output of the model concurrently as both R and A, i.e., $R = A = O$.

After identifying R and A, we apply a CoT Extraction module to distill key reasoning steps and reduce noise. This serves as a foundation for the scoring framework introduced in Equation 1 and is detailed in the next subsection.

Reasoning Trace Distillation

Reasoning traces generated by LRMs often include exploratory thoughts or redundant steps that do not directly contribute to the final answer (Chen et al. 2025b). Such reasoning noise can obscure consistency assessment and reduce the reliability of hallucination detection. To address this, we propose a specialized **CoT Extraction** module designed to distill concise, coherent reasoning segments directly relevant to the model's final prediction. Formally, we define a CoT Extraction function f that identifies the minimal yet sufficient reasoning steps required to derive the final answer:

$$C = f(Q, A, R), \tag{2}$$

where Q is the input question, A is the predicted answer, and R is the full reasoning trace. The extracted C serves as a condensed and faithful summary of R, used as the foundation for all subsequent scoring modules.

To train the extraction function f, we construct a synthetic training set via the following pipeline. Given a QA pair (Q, A) from a seed question-answering dataset $\mathcal{D}$, a base reasoning model M_R is prompted to generate an initial reasoning trace R and the answer A. Another powerful LLM M_p (e.g., GPT-4o (OpenAI et al. 2023)) is then used to summarize R into multiple candidate CoTs $C_1, \ldots, C_k$, where each candidate is structured with step-wise markers such as `[STEP]` to highlight the logical progression toward the answer. To ensure the faithfulness and informativeness of these summaries, we apply a two-stage filtering process. First, we verify that each candidate C_i leads to the correct answer A when used as input alongside Q in M_p. Among the valid candidates, we select the most concise one, i.e., the one with the fewest reasoning steps and the shortest token length. Second, to further filter abstraction errors, we assess the semantic alignment between each C_i and the original reasoning trace R using a Natural Language Inference (NLI) classifier, discarding candidates with entailment scores below 0.9, which indicates a potential semantic mismatch.

The resulting (Q, A, R, C) tuples are then used to fine-tune a small LLM M_s to learn the CoT extraction function f. The model is trained with a standard language modeling loss over the generated output C, conditioning on (Q, A, R) as input. Once trained, the extractor M_s is fixed and used throughout all hallucination detection modules. It is worth noting that the extractor is model-agnostic and task-independent, making it applicable across different LLM settings without requiring re-training.

RACE Scoring Framework

Inspired by Equation 1, we develop our scoring framework, RACE, based on several aspects of consistency. Before computing the hallucination score, the CoT Extractor transforms each R in the main and sampled outputs into C, shifting the estimation objective from $H(R, A|Q)$ to $H(C, A|Q)$.

Reasoning Consistency (S_{CC}) To reduce computational costs of calculating consistency over all pairs of reasoning paths (which has quadratic complexity), we approximate $H(C|Q)$ by measuring the consistency between sampled reasoning paths and the main path. We compare the main CoT (C_{main}) with sampled CoTs (C_i) on a step-by-step basis, weighting each step by its importance in C_{main}:

$$S_{CC} = \sum_j w_j \left(\frac{1}{N} \sum_{i=1}^{N} \delta(C_{\text{main}}^{(j)}, C_i) \right), \tag{3}$$

where $C_{main}^{(j)}$ is the j-th step of the main CoT, C_i is one of the sampled CoTs, and $\delta(C_1^{(j)}, C_2)$ denotes the contradiction probability between step $C_1^{(j)}$ and CoT C_2, estimated by an NLI classifier, which is generally trained to assess whether the relationship between two sentences is entailment or contradiction (implementation details are shown in the next section). A higher S_{CC} indicates lower reasoning consistency and thus a greater likelihood of hallucination.

The weight w_j reflects the importance of the j-th step of $C_{main}^{(j)}$ to A_{main}, aiming to downplay potentially redundant steps after the CoT Extraction process. Specifically, we concatenate the main CoT (with m steps) and main answer as:

$$Q \texttt{ <think> } C_{main}^{(1)}, \ldots, C_{main}^{(m)} \texttt{ </think> } A_{main}.$$

This sequence is input to an LLM, where the average attention scores from all tokens in A_{main} to those in $C_{main}^{(j)}$ are calculated and normalized to produce weights w_j. We use the CoT Extractor as a proxy to compute attention scores, thus reducing deployment resources and maintaining black-box compatibility with the evaluated LLM.

Answer Uncertainty (S_{AA}) We employ SINdex (Abdaljalil et al. 2025) to estimate $H(A|Q)$. SINdex refines semantic entropy by adjusting cluster probabilities based on intra-cluster similarity. The SINdex score is calculated as:

$$S_{AA} = -\sum_{l=1}^{n} p'_l \log(p'_l), \tag{4}$$

where n is the number of clusters from the sampled answers, and p'_l is the adjusted proportion of the l-th cluster C_l:

$$p'_l = \frac{p_l \cdot \sum_{x,y \in C_l, x \neq y} \text{sim}(emb(x), emb(y))}{\binom{|C_l|}{2}}. \tag{5}$$

Here, p_l is the original proportion of the cluster C_l, and $\text{sim}(emb(x), emb(y))$ is the cosine similarity between the embeddings of answers x and y within the cluster. The adjustment factor quantifies semantic coherence within cluster C_l. Clusters with high internal dispersion receive a low adjusted proportion p'_l. SINdex provides a more accurate measure of uncertainty regarding the answer.

Reasoning-Answer Alignment (S_{CA}) To estimate the negative mutual information component, $-I(R, A|Q)$, we assess the alignment between the main reasoning path and the sample answer space generated by the LLM. Specifically, for the main extracted CoT and each sampled answer A_i, we use token-level Length-Normalized Predictive Entropy (LNPE) (Malinin and Gales 2020) to quantify how well C_{main} (as context) predicts A_i:

$$S_{CA} = \frac{1}{N} \sum_{i=1}^{N} \bar{H}_M(A_i|Q, C_{main}), \tag{6}$$

where $\bar{H}_M(y|x)$ computes the average output token entropy (i.e. LNPE) by feeding x into the LLM M and constraining the output to be y. A higher S_{CA} value suggests poorer alignment between the main reasoning path and the sampled answers, indicating a higher likelihood of hallucination in O_{main}. We focus on alignment with C_{main} because our primary goal is to detect hallucinations in the main output, and computing the LNPE score for all sampled reasoning paths and answers leads to quadratic time complexity. The CoT Extractor serves as the model M for calculating LNPE.

Reasoning Internal Coherence (S_{Coh}) We observe that in LRMs, discrepancies can exist between the extracted CoT (C) and the original thinking process (R). The model's initial reasoning path R may include speculative threads or explorations that do not contribute to the final answer. A high

proportion of such speculative content, absent from the concise form C, may be associated with an increased rate of hallucinations. Thus, for the LRMs setting, we additionally extract entity sets from the main CoT ($E_{C_{main}}$) and the original main reasoning process ($E_{R_{main}}$). We then compute:

$$S_{Coh} = |E_{R_{\text{main}}} \setminus E_{C_{\text{main}}}| \ / \ |E_{R_{\text{main}}}|, \quad (7)$$

where "\" denotes the set difference operator. This score quantifies the proportion of entities omitted from the original reasoning in the distilled core CoT. Higher S_{Coh} indicates more speculation and potential hallucination.

Final Score Aggregation Finally, we combine these four metrics into a unified hallucination score for O_{main}:

$$S_{RACE} = S_{AA} + S_{CA} + S_{CC} + S_{Coh}, \quad (8)$$

where the last term is designed specifically for LRMs. This linear combination assigns equal weight to each component, a choice justified by its empirical simplicity and interpretability (Equation 1). A higher score indicates a greater likelihood that A_{main} contains a hallucination.

Experimental Setup

Datasets and Metrics

We conduct our evaluation on the validation sets of four widely-used question-answering datasets: TriviaQA (Joshi et al. 2017), SQuAD (Rajpurkar, Jia, and Liang 2018), NQ-Open (Kwiatkowski et al. 2019), and HotpotQA (Yang et al. 2018). Following Su et al. (2024b), the main output is generated using greedy decoding with a maximum length of 2048 tokens, which represents the model's default output for detection. For sample-based baselines, we sample 5 outputs using a temperature of 1.0 and top-p sampling with $p = 0.95$. For LRMs, main outputs lacking the <\think> token are filtered to ensure a final answer is provided. Following Abdaljalil et al. (2025) and Binkowski et al. (2025), we employ the Area Under the Receiver Operating Characteristic curve (AUROC) as the metric, where a higher AUROC indicates better distinction between hallucinated and non-hallucinated outputs. Following previous works (Chen et al. 2025a; Li et al. 2024), we utilize Qwen2.5-32B-Instruct as an LLM-as-Judge to assess whether an LLM's answer contains hallucinations by comparing it to ground truths.

Model Settings

RACE accommodates diverse model types and output formats. To evaluate its generality, we consider 3 representative settings reflecting the output styles of different LLM types:

- **LRMs:** This is our primary setting, where the model explicitly generates reasoning traces. We evaluate seven LRMs: DeepSeek-R1-Distill-Qwen-7B (DS-7B), DeepSeek-R1-Distill-Llama-8B (DS-8B), DeepSeek-R1-Distill-Qwen-14B (DS-14B), Qwen3-14B (Q3-14B) (Yang et al. 2025), GLM-Z1-9B-0414 (Z1-9B) (GLM et al. 2024), QwQ-32B(Team 2025), and DeepSeek-R1 (DeepSeek-AI et al. 2025). Inference for DeepSeek-R1 is conducted via its official API.
- **CoT Outputs from Standard LLMs:** We evaluate Qwen2.5-14B-Instruct (Yang et al. 2024) with prompt "Think step by step", thereby generating the CoT and a final answer.
- **Direct Outputs from Standard LLMs:** We evaluate Qwen2.5-14B-Instruct by directly providing the question without any additional prompting.

Baselines

We compare RACE with several zero-resource hallucination detection baselines based on sampling consistency. These include Semantic Entropy (SE) (Kuhn, Gal, and Farquhar 2023), which refines token-level entropy via clustering; SelfCheckGPT-NLI (SCG) (Manakul, Liusie, and Gales 2023), which uses an NLI model to identify inconsistencies; Semantic Embedding Uncertainty (SEU) (Grewal, Bonilla, and Bui 2024), which assesses consistency through average pairwise embedding similarity; and the Semantic INconsistency Index (SINdex) (Abdaljalil et al. 2025), which extends SE by modeling both intra- and inter-class inconsistencies. We also evaluate Length Normalised Predictive Entropy (LNPE) (Malinin and Gales 2020) and P(true) (Kadavath et al. 2022). For DeepSeek-R1, API access limits evaluation with methods requiring output probabilities (LNPE, P(true), and SE). These baselines are conducted at the answer level, as semantic clustering or embedding typically fail when applied to lengthy reasoning paths.

Implementation Details

For the CoT Extraction module, the seed dataset $\mathcal{D}$ is 2WikiMultihopQA (Ho et al. 2020). For dataset construction, we use DeepSeek-Distill-Qwen7B (DeepSeek-AI et al. 2025) as the LRM M_R, Qwen2.5-32B-Instruct as the powerful summarizing model M_p, and deberta-v3-large-mnli (Manakul, Liusie, and Gales 2023) as the NLI model to filter errors. The extraction model is trained from Llama-3.1-8B-Instruct (Grattafiori et al. 2024).

For S_{AA} (Answer Uncertainty), we align with the SINdex setup, using all-MiniLM-L6-v2 (Reimers and Gurevych 2019) as the embedding model. To adapt lengthy model outputs in our experiments, we adjust the SINdex clustering similarity hyperparameter to 0.9. For the S_{CC} score (Reasoning Consistency), we use deberta-v3-large-mnli, as employed in SelfCheckGPT, for NLI comparisons. For the S_{Coh} score (Reasoning Internal Coherence), we use en_core_web_trf-3.8.0 (Explosion 2024) to extract entities.

To highlight RACE's contribution, we compare it with two simple methods integrating reasoning process analysis:

- $\boldsymbol{S_{RR}}$: A simple adaptation where the CoT component C in S_{CC} is replaced by the model's original reasoning process R, segmented by "\n\n".
- **RACE$_{\text{raw}}$**: A method that fully aligns with the RACE framework but skips the CoT Extraction module, directly evaluating consistency using the model's original reasoning path (R), equivalent to S_{AA}+S_{RA}+S_{RR} + S_{Coh}.

Dataset	Method	DS-7B	DS-8B	DS-14B	Q3-14B	Z1-9B	QwQ-32B	DS-R1	Q2.5-14B	Q2.5-14B(CoT)
HotpotQA	LNPE	53.60	54.01	46.79	61.92	48.93	58.36	—	73.88	72.36
	P(true)	50.36	48.39	43.37	62.60	73.07	66.61	—	73.25	67.07
	SE	54.98	56.48	52.95	56.36	51.88	55.31	—	69.76	76.86
	SEU	74.66	75.37	76.26	74.75	73.28	76.25	73.53	75.03	74.83
	SCG	69.17	70.30	74.83	73.43	68.53	69.96	67.41	57.01	71.98
	SINdex	74.50	74.98	76.17	75.80	74.18	78.19	75.43	73.26	77.19
	S_{RR}	61.22	58.50	65.56	59.65	58.10	66.32	53.85	60.71	54.60
	$RACE_{raw}$	75.48	75.58	76.82	75.51	75.23	79.09	75.58	74.23	76.71
	RACE	**77.62**	**78.56**	**79.73**	**78.41**	**77.93**	**81.01**	**76.71**	**75.90**	**79.87**
TriviaQA	LNPE	53.70	49.93	45.91	69.94	46.84	65.57	—	79.76	68.02
	P(true)	59.39	50.63	41.04	63.86	83.44	78.31	—	86.11	85.56
	SE	62.33	59.66	61.10	65.75	60.79	65.23	—	80.78	83.63
	SEU	74.42	76.73	82.47	86.08	80.24	88.13	80.17	82.27	77.05
	SCG	75.42	76.76	84.83	84.37	73.80	78.53	51.67	67.86	86.63
	SINdex	77.22	77.55	82.47	87.11	82.01	88.75	81.56	84.30	87.33
	S_{RR}	60.42	58.68	65.38	61.48	61.62	70.44	52.46	73.49	69.67
	$RACE_{raw}$	77.36	78.27	82.72	87.49	82.43	89.01	77.32	86.06	86.88
	RACE	**80.60**	**81.81**	**87.03**	**89.67**	**85.54**	**90.96**	**83.14**	**87.02**	**89.79**
NQ-Open	LNPE	55.34	54.26	52.37	55.96	50.40	60.04	—	65.60	59.22
	P(true)	63.45	47.76	52.98	63.58	69.08	68.42	—	70.12	70.01
	SE	52.95	53.73	52.22	48.10	45.47	50.95	—	62.25	67.95
	SEU	75.91	67.95	71.26	72.57	72.24	70.16	66.72	69.92	73.75
	SCG	72.00	65.95	71.26	70.88	66.79	65.50	61.19	62.10	71.60
	SINdex	72.24	65.48	70.41	71.52	73.19	72.45	66.48	66.83	71.60
	S_{RR}	65.98	56.63	64.26	62.41	55.95	62.82	53.86	63.19	62.15
	$RACE_{raw}$	75.80	67.47	73.14	68.06	74.66	74.08	70.03	71.03	73.97
	RACE	**78.61**	**72.14**	**75.80**	**75.83**	**77.81**	**76.30**	**73.04**	**71.29**	**75.68**
SQuAD	LNPE	62.08	56.15	63.27	78.73	63.02	78.34	—	80.81	76.11
	P(true)	47.66	52.06	54.47	46.24	76.95	50.75	—	79.16	77.56
	SE	66.41	64.24	68.30	78.30	74.20	**80.33**	—	**85.71**	77.17
	SEU	72.06	69.74	73.72	73.92	74.65	73.59	83.03	76.44	77.80
	SCG	63.16	58.41	64.63	71.85	57.58	57.01	66.80	60.71	75.41
	SINdex	71.20	69.51	73.82	78.74	77.28	79.16	88.30	85.69	82.05
	S_{RR}	55.31	48.77	49.21	50.31	59.13	54.07	71.23	71.61	49.75
	$RACE_{raw}$	70.81	68.91	72.68	78.34	74.97	74.70	84.71	84.20	73.08
	RACE	**78.27**	**74.43**	**79.03**	**79.87**	**79.37**	77.40	**88.95**	85.02	**83.65**

Table 1: The overall experimental results. DeepSeek-R1 outputs are obtained via the official API, so gray-box methods (LNPE, P(true), SE) cannot detect its hallucinations in this setup. "Q2.5-14B" denotes Qwen2.5-14B-Instruct in direct output mode, while "Q2.5-14B(CoT)" includes CoT; other models are LRMs. The metric is AUROC. Best and second-best results are in bold and underlined, respectively.

Experimental Results

We address three key research questions: **(RQ1)** whether jointly evaluating answer and reasoning provides a useful signal for hallucination detection; **(RQ2)** whether the structure and noise in initial reasoning paths affect the assessment of reasoning consistency; and **(RQ3)** how RACE alleviates the delay caused by considering additional reasoning parts.

Main Results

Detailed main results are shown in Table 1. Overall, sampling-based hallucination detection methods outperform probability-based ones, as they incorporate a larger amount of sampling information. Moreover, for RQ1, the results show that RACE generally outperforms existing sampling-based and probability-based methods across four datasets and seven LRMs in most cases. Strong baselines such as SEU and SINdex achieve the best baseline performance on specific models or datasets. However, RACE consistently outperforms them in LRM settings. Moreover, $RACE_{raw}$ outperforms baselines in many scenarios, achieving the second-best performance, whereas a simple assessment of reasoning consistency (S_{RR}) yields poor results. This further highlights the effectiveness of the joint evaluation approach. Furthermore, RACE shows strong generalizability. When applied to Qwen2.5-14B-Instruct generating CoT outputs, RACE significantly exceeds the best baselines. This suggests that RACE's analysis of LRM's reasoning component likewise applies to standard instruct models. Even in the direct output setting, RACE provides competitive or superior performance compared to baselines. In this setting, all sampling-based methods utilize the same information. Due to its multi-dimensional scoring and structured extraction of

		Inference Time	Percentage
LLM's Greedy Response		30.91s	100.0%
SINdex	**LLM's Response**	43.28s	+40.7%
	SINdex Score	+0.02s	
S_{RR}	**LLM's Response**	43.28s	+54.3%
	S_{RR} Score	+4.40s	
RACE	**LLM's Response**	43.28s	+49.4%
	CoT Extraction	+2.10s	
	RACE Score	+0.82s	

Table 2: Efficiency from generation to detection, showing total time and the percentage of additional time compared to the greedy response. "LLM's Response" includes main and sampled outputs via batch inference. All RACE components are parallelized; S_{CC} is the slowest (0.82s), followed by S_{CA} (0.24s), while the others are negligible.

potential reasoning processes, RACE continues to surpass the other baselines. This indicates that jointly evaluating answer and reasoning consistency is beneficial even when explicit reasoning steps are not prompted.

Regarding RQ2, we compare RACE against RACE_{raw}, which aligns with RACE but using the original reasoning paths to evaluate consistency. The results show that RACE outperforms RACE_{raw} across all models and datasets, indicating that noise within the original reasoning paths compromises consistency assessment. By extracting these paths into more concise, structured CoTs, RACE facilitates the detection of hallucinations associated with the reasoning process.

Efficiency Analysis

To address RQ3, Table 2 shows the average detection time. We randomly select 100 questions from NQ-Open, using DS-8B with a single 80G GPU. In sample-based hallucination detection, the primary overhead stems from generating additional responses: producing both main and sampled outputs requires about 40% more time than the main response alone, while computing the hallucination score adds minimal overhead. However, reasoning-level hallucination detection introduce more delay than answer-level methods, as reasoning paths are generally lengthy and complex. RACE simplifies the reasoning path using the CoT Extraction module, achieving significantly better performance than the naive baseline (S_{RR}) with fewer additional delays. Additionally, RACE involves deploying a 8B CoT Extractor. Compared with larger state-of-the-art reasoning LLMs (for example, DeepSeek-R1 at 617B), this deployment overhead is acceptable. For applications demanding high reliability, the trade-off between overhead and accuracy is often needed.

Ablation Study

Ablation of each component: Table 3 reveals the contribution of each component in RACE. Removing any single component generally leads to a decrease in performance. Using a single component results in lower performance, confirming that all components contribute complementary signals for effective detection. Note that using only the Reasoning Consistency score (only S_{CC}) achieves the best performance among all single components, highlighting the importance of detecting consistency among reasoning paths.

	DS-7B	DS-8B	DS-14B	Z1-9B	QwQ-32B
RACE	**77.62**	**78.56**	79.73	**77.93**	**81.01**
w/ avg. S_{CC}	77.09	78.02	79.18	77.57	80.81
w/o S_{Coh}	77.48	78.15	79.46	77.51	80.86
w/o S_{AA}	75.43	77.79	79.50	77.45	80.02
w/o S_{CA}	77.26	77.82	79.50	77.48	80.49
w/o S_{CC}	75.08	76.08	76.40	75.24	78.97
only S_{Coh}	51.07	56.95	56.08	55.98	57.39
only S_{AA}	74.50	74.98	76.17	74.18	78.19
only S_{CA}	68.83	68.76	67.33	68.35	69.43
only S_{CC}	74.98	75.49	**79.81**	76.76	79.46

Table 3: The ablation study of each component on HotpotQA. "*w/ avg. S_{CC}*" means when calculating S_{CC}, the importance weighting for each step is not applied.

	RACE	RACE^{+}	Optimal Weights
Q3-14B(NQ)	76.55	**79.71**	0.09 / 0.49 / 0.37 / 0.06
Q3-14B(HotpotQA)	78.33	**78.76**	0.13 / 0.35 / 0.29 / 0.23
DS-8B(NQ)	72.30	**76.82**	0.03 / 0.34 / 0.34 / 0.28
DS-8B(HotpotQA)	78.46	**79.73**	0.10 / 0.35 / 0.33 / 0.22

Table 4: AUROC values (first two columns) and normalized optimal weights (last column) from optimizing RACE component coefficients (RACE^{+}), ordered as in Equation 8.

Weight optimization of each component: In our main experiment, we assign equal weights to each RACE component for interpretability (Equation 1) and to avoid fine-tuning that may lead to unfair comparisons. Table 4 shows further gains from weight optimization. We use the first 20% of each dataset for training and the rest for testing. On the training set, we perform a grid search over [0, 1] with 0.05 increments, normalize the best-performing weights, and evaluate on the test set. Results (the column of RACE^{+}) show that optimized weights can further improve RACE, though the best configurations vary across models and datasets, which suggest that RACE has strong potential in real-world applications. The results also demonstrate that despite their varying importance, all four components are indeed effective.

Conclusion

In this work, we introduce RACE, a novel black-box hallucination detection framework. By jointly assessing the consistency of the reasoning process and final answer, RACE provides fine-grained detection of hallucinations. Experimental results show that RACE significantly outperforms existing methods across multiple datasets and LLMs, highlighting its robustness and generalizability for hallucination detection.

References

Abdaljalil, S.; Kurban, H.; Sharma, P.; Serpedin, E.; and Atat, R. 2025. SINdex: Semantic INconsistency Index for Hallucination Detection in LLMs. *CoRR*, abs/2503.05980.

Arcuschin, I.; Janiak, J.; Krzyzanowski, R.; Rajamanoharan, S.; Nanda, N.; and Conmy, A. 2025. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. *CoRR*, abs/2503.08679.

Azaria, A.; and Mitchell, T. M. 2023. The Internal State of an LLM Knows When It's Lying. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 967–976. Association for Computational Linguistics.

Besta, M.; Barth, J.; Schreiber, E.; Kubicek, A.; Catarino, A.; Gerstenberger, R.; Nyczyk, P.; Iff, P.; Li, Y.; Houliston, S.; Sternal, T.; Copik, M.; Kwaśniewski, G.; Müller, J.; Łukasz Flis; Eberhard, H.; Niewiadomski, H.; and Hoefler, T. 2025. Reasoning Language Models: A Blueprint. arXiv:2501.11223.

Binkowski, J.; Janiak, D.; Sawczyn, A.; Gabrys, B.; and Kajdanowicz, T. 2025. Hallucination Detection in LLMs Using Spectral Features of Attention Maps. *CoRR*, abs/2502.17598.

Chen, C.; Liu, K.; Chen, Z.; Gu, Y.; Wu, Y.; Tao, M.; Fu, Z.; and Ye, J. 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.

Chen, T.; Liu, X.; Da, L.; Chen, J.; Papalexakis, V.; and Wei, H. 2025a. Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses. *CoRR*, abs/2502.16820.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; Wang, R.; Tu, Z.; Mi, H.; and Yu, D. 2025b. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. arXiv:2412.21187.

DeepMind. 2025. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. [Accessed 06-05-2025].

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Explosion. 2024. spaCy: Industrial-strength Natural Language Processing in Python. https://spacy.io/models/en. Accessed: 2025-07-31.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Grewal, Y. S.; Bonilla, E. V.; and Bui, T. D. 2024. Improving Uncertainty Quantification in Large Language Models via Semantic Embeddings. *CoRR*, abs/2410.22685.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. arXiv:2501.04519.

Ho, X.; Nguyen, A. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 6609–6625. International Committee on Computational Linguistics.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR*, abs/2311.05232.

Hughes, S.; Bae, M.; Li, M.; et al. 2025. Vectara Hallucination Leaderboard.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; Showk, S. E.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. *CoRR*, abs/2207.05221.

Ke, Z.; Jiao, F.; Ming, Y.; Nguyen, X.-P.; Xu, A.; Long, D. X.; Li, M.; Qin, C.; Wang, P.; Savarese, S.; Xiong, C.; and Joty, S. 2025. A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems. arXiv:2504.09037.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering

Research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, J.; Chen, J.; Ren, R.; Cheng, X.; Zhao, X.; Nie, J.-Y.; and Wen, J.-R. 2024. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10879–10899. Bangkok, Thailand: Association for Computational Linguistics.

Lu, H.; Liu, Y.; Xu, J.; Nan, G.; Yu, Y.; Chen, Z.; and Wang, K. 2025. Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models. arXiv:2505.13143.

Malinin, A.; and Gales, M. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 9004–9017. Association for Computational Linguistics.

OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. [Accessed 19-09-2024].

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; et al. 2023. GPT-4 Technical Report.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 784–789. Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Su, W.; Tang, Y.; Ai, Q.; Wang, C.; Wu, Z.; and Liu, Y. 2024a. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 23–31.

Su, W.; Wang, C.; Ai, Q.; Hu, Y.; Wu, Z.; Zhou, Y.; and Liu, Y. 2024b. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 14379–14391. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhong, T.; Liu, Z.; Pan, Y.; Zhang, Y.; Zhou, Y.; Liang, S.; Wu, Z.; Lyu, Y.; Shu, P.; Yu, X.; Cao, C.; Jiang, H.; Chen, H.; Li, Y.; Chen, J.; Hu, H.; Liu, Y.; Zhao, H.; et al. 2024. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. arXiv:2409.18486.