Zero-source LLM Hallucination Detection with Human-like Criteria Probing

Jiahao Yang*, Shuhai Zhang*, Hailong Kang*, Feng Liu†, Qi Chen†, Mingkui Tan†

South China University of Technology, The University of Melbourne, The University of Adelaide

* Equal contribution † Corresponding author

ICML

International Conference
On Machine Learning

LLM Hallucinations

LLMs have rapidly advanced and are increasingly deployed across a broad range of applications. However, they often **hallucinate** by generating **factually incorrect or unfaithful content**[1].

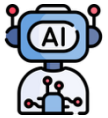
Faithfulness Hallucination

Instruction Inconsistency

Content that deviates from the source input provided by users.

Translate the English question into Spanish: "What is the capital of France?"

The capital of France is Paris.

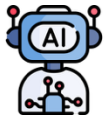


Context Inconsistency

Content that conflicts with previous generation from itself.

… in the world. Originating in the Great Lakes region of central Africa, it flows through several countries ... Summarize the retrieval document above.

The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.



Factuality Hallucination

Factual Contradiction

Content that is inconsistent with established world knowledge.

In which country were the 1948 Winter Olympics held?

Norway.

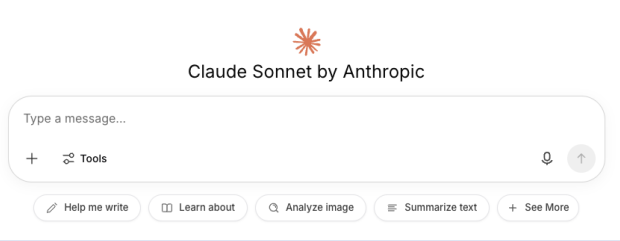
[1] A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM TIS 2025.

The Zero-source Constraint

In prevalent open-world scenarios, the auditing process is entirely decoupled from the generation.

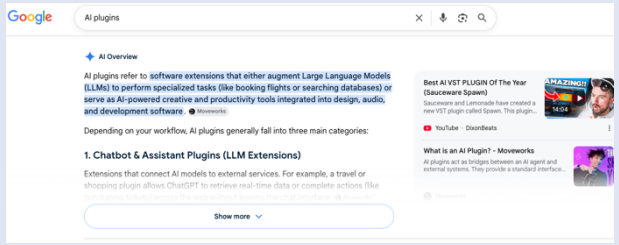
End-user Interface

Most mild users interact via ChatGPT / Claude / Gemini web UI, **only see plain text output.**



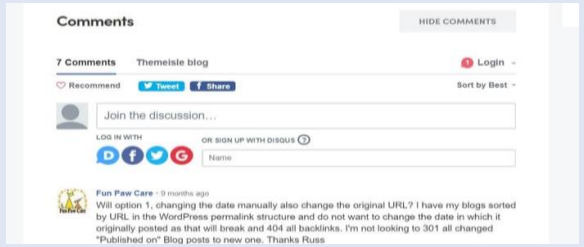
Browser Extension

Many search engines provide AI plugins to summarize search content, which can **only access the rendered text on the page**.



Third-party Audit

The social platforms or news agencies audit user-uploaded AI content **without knowing the source model**.



Detection must rely SOLELY on the textual "query–answer" pair.

How Do Human Experts Evaluate Responses?

Observation: Human experts rarely judge a response using a single monolithic criterion.

Human Evaluative Process:

Step. 1

Identify the question type:
(historical? scientific? social?...);

Step. 2

Select relevant evaluation dimensions;

Step. 3

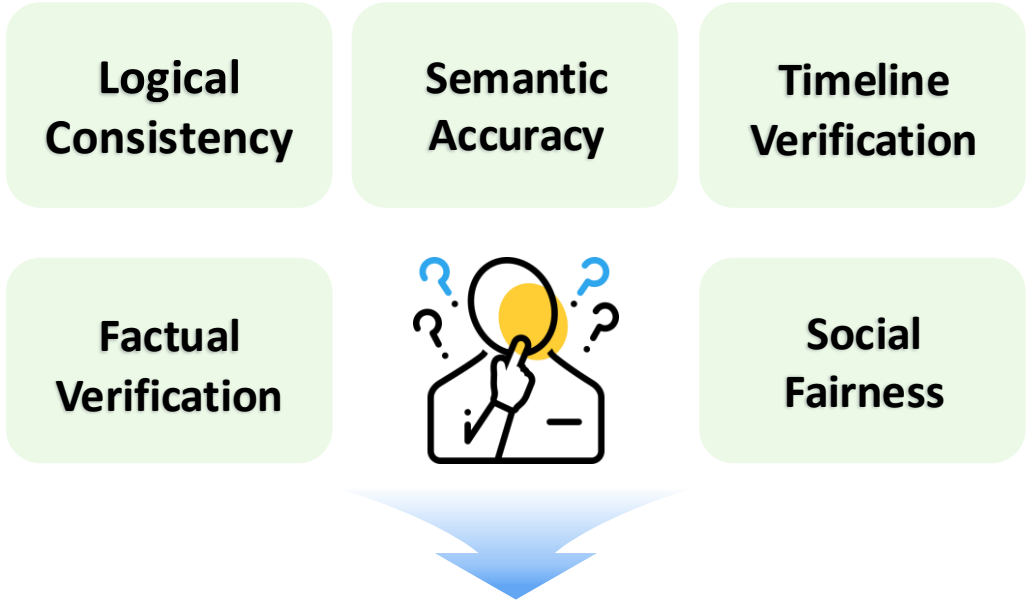
Assign importance weights based on context;

Step. 4

Analyze evidence for each dimension;

Step. 5

Synthesize a final judgment.



Human-like Criteria Probing Mechanism

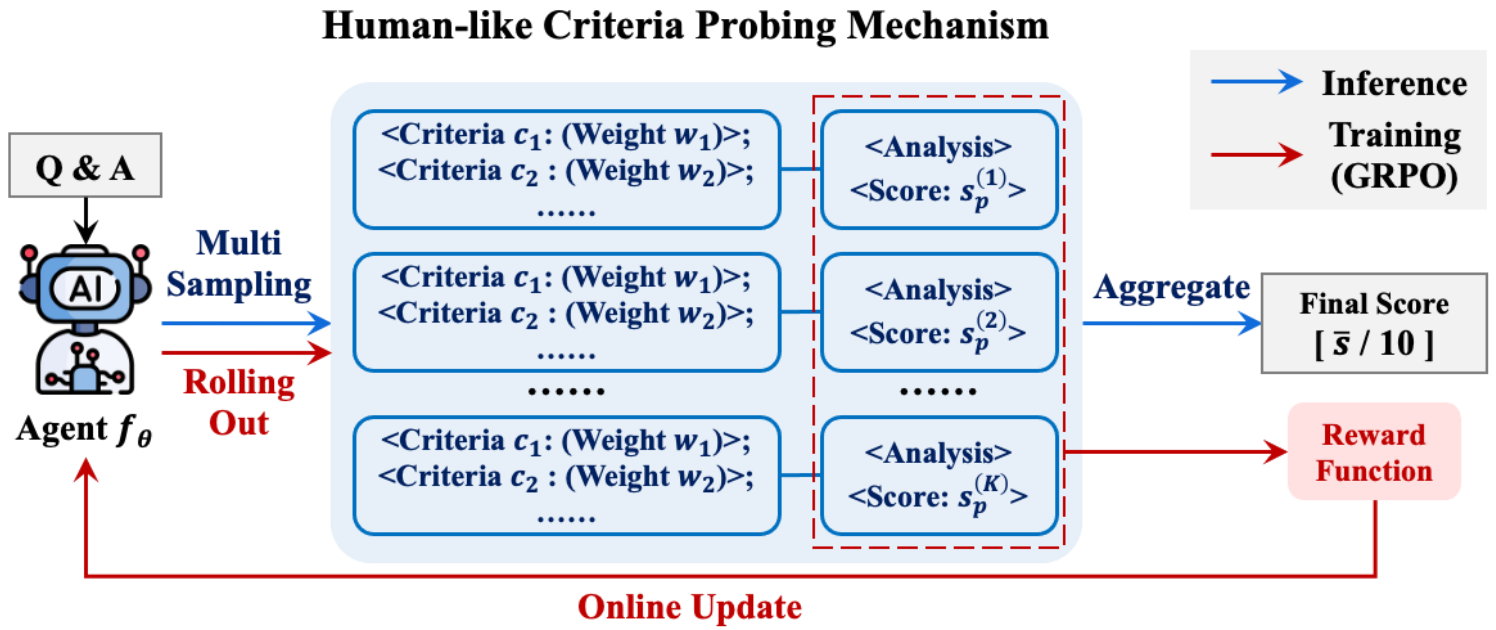
We develop **Human-like Criteria Probing (HCP)**, a structured reasoning framework that enables a pre-trained LLM to function as **an adaptive scoring agent** for hallucination detection.

Step. 1 Criteria Generation

Step. 2 Weight Assignment

Step. 3 Criterion-wise Scoring

Step. 4 Weighted Aggregation



Structured and Interpretable Output

Table 1. The structured and interpretable output from the agent.

Component	Example
Q & A	**Qes**: In which country were the 1948 Winter Olympics held? **Ans**: Norway.
Specific Criteria	1. **Factual Grounding** (Weight: 60%) 2. **Temporal Consistency** (Weight: 20%) 3. **Semantic Precision** (Weight: 20%)
Analysis	- **Factual Grounding**: The 1948 Winter Olympics were held in St. Moritz, Switzerland, not Norway. This is a clear factual error. - **Temporal Consistency**: The response does not specify a year, so there are no temporal issues directly in the text. - **Semantic Precision**: The response is vague and does not provide the correct information, leading to semantic distortion as the user is not given accurate details.
Score (1∼10)	1

Full Interpretability

Every decision is explained through explicit criteria, weights, and evidence-based analysis. Users can understand WHY a response is flagged.

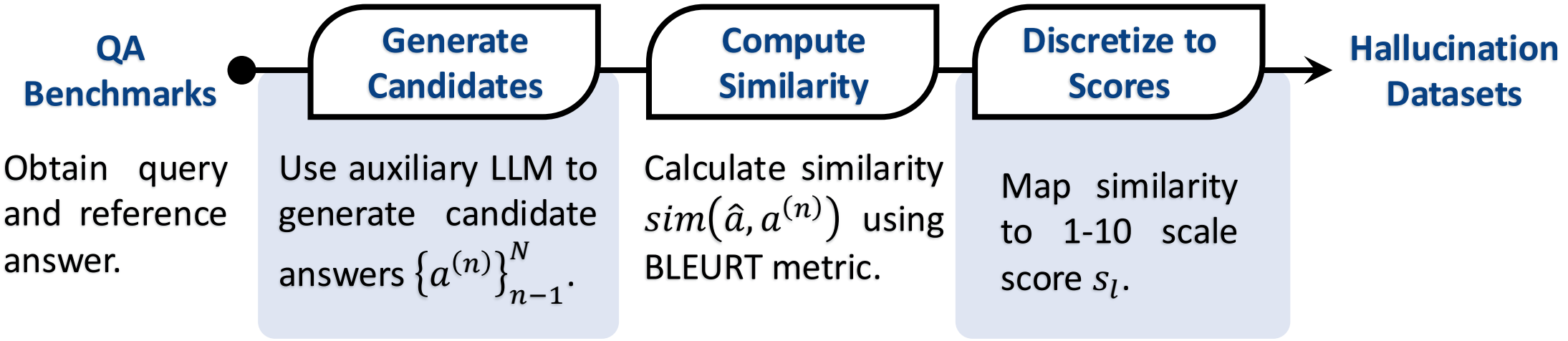
Context Adaptivity

Different query types trigger different criteria and weight distributions.

Reward-based Alignment Training Strategy

To equip the agent f_θ with precise evaluative capabilities, we train it with **a reward-based alignment paradigm** using Group Relative Policy Optimization (GRPO).

❑ Data Construction and Labeling



$$(q, \hat{a}) \rightarrow \left(q, \{\hat{a}, 10\} \cup \left\{a^{(n)}, s_l^{(n)}\right\}_{n-1}^{N}\right)$$

Reward-based Alignment Training Strategy

❑ Score-alignment Reward Function

Given an input pair (q, a), f_θ performs human-like criteria probing and outputs a structured evaluation containing a predicted score s_p. The reward function assigns a scalar reward $r \in [0,1]$:

$$r = \begin{cases} 1 - \dfrac{|s_p - s_l|}{9}, & \text{if the output is well-formed;} \\ 0, & \text{otherwise.} \end{cases}$$

- **Numerical alignment:** The reward attains $r = 1$ when $s_p = s_l$ and decreases linearly with the absolute deviation.
- **Format alignment:** Any deviation from the prescribed format that prevents reliable extraction of a valid score results in $r = 0$.

Robust Inference via Multi-sampling Aggregation

To reduce the randomness of LLM and improve decision reliability, we employ a **multi-sampling aggregation strategy** during inference:

$$\bar{s} = \frac{1}{K}\sum_{k-1}^{K} s_p^{(k)}.$$

Table 5. Impact of sampling size K.

Dataset	1	2	5	10
TriviaQA	$85.21_{\pm 1.22}$	$85.71_{\pm 0.86}$	$86.25_{\pm 1.08}$	$86.25_{\pm 0.98}$
SciQ	$85.06_{\pm 2.33}$	$85.47_{\pm 2.12}$	$86.04_{\pm 2.25}$	$86.14_{\pm 1.89}$
NQ Open	$86.89_{\pm 1.60}$	$89.26_{\pm 2.46}$	$90.38_{\pm 3.58}$	$90.41_{\pm 2.46}$
CoQA	$88.60_{\pm 4.12}$	$89.37_{\pm 3.34}$	$90.07_{\pm 2.58}$	$89.85_{\pm 2.66}$
Inf. Time (s)	0.2349	0.4675	1.1313	2.2807

- **Gradual growth stability**
- **Linear growth overhead**

We set $K = 5$ in experiments, while $K = 1$ is good enough in practice.

Main Results

Table 2. Comparisons with hallucination detection baselines on different datasets for LLaMA-3.1-8b and Qwen-3-8b in terms of AUROC (%), where ♣ denotes methods trained on fully labeled datasets.

Model	Method	TriviaQA	SciQ	NQ Open	CoQA	Avg.
Llama-3.1-8b	LN-Entropy (Malinin & Gales, 2021)	$73.62_{\pm 2.20}$	$62.69_{\pm 2.73}$	$52.36_{\pm 1.53}$	$74.52_{\pm 1.86}$	65.80
	Self-evaluation (Kadavath et al., 2022)	$56.07_{\pm 1.92}$	$54.12_{\pm 1.82}$	$59.83_{\pm 3.68}$	$62.51_{\pm 1.42}$	58.13
	CCS (Burns et al., 2022)	$78.20_{\pm 1.89}$	$58.85_{\pm 3.03}$	$55.50_{\pm 1.89}$	$68.98_{\pm 2.04}$	65.38
	SelfCKGPT (Manakul et al., 2023)	$74.58_{\pm 1.90}$	$59.68_{\pm 2.45}$	$62.13_{\pm 2.60}$	$70.61_{\pm 2.10}$	66.75
	Perplexity (Ren et al., 2023)	$80.62_{\pm 2.62}$	$66.12_{\pm 2.85}$	$57.92_{\pm 2.60}$	$81.41_{\pm 2.21}$	71.52
	SAPLMA♣(Azaria & Mitchell, 2023)	$78.51_{\pm 3.16}$	$85.63_{\pm 0.96}$	$76.23_{\pm 0.82}$	$71.58_{\pm 1.35}$	77.99
	Semantic Entropy (Kuhn et al., 2023)	$78.71_{\pm 3.09}$	$77.81_{\pm 3.17}$	$61.04_{\pm 4.29}$	$75.26_{\pm 4.63}$	73.21
	Lexical Similarity (Lin et al., 2024)	$77.96_{\pm 2.03}$	$67.09_{\pm 2.05}$	$62.85_{\pm 1.57}$	$77.53_{\pm 0.90}$	71.36
	EigenScore (Chen et al., 2024a)	$51.35_{\pm 1.23}$	$51.52_{\pm 0.82}$	$52.17_{\pm 2.13}$	$52.00_{\pm 1.19}$	51.76
	HaloScope♣ (Du et al., 2024)	$58.19_{\pm 5.79}$	$69.04_{\pm 6.36}$	$63.38_{\pm 3.02}$	$72.11_{\pm 4.96}$	65.68
	TAD ♣ (Vazhentsev et al., 2025)	$72.01_{\pm 1.03}$	$66.75_{\pm 1.96}$	$68.88_{\pm 2.56}$	$74.86_{\pm 2.10}$	70.63
	TSV♣ (Park et al., 2025)	$79.78_{\pm 3.36}$	$80.01_{\pm 1.17}$	$70.17_{\pm 1.41}$	$69.31_{\pm 6.75}$	74.82
	HCPD (Ours)	$\mathbf{86.25}_{\pm 1.08}$	$\mathbf{86.04}_{\pm 2.25}$	$\mathbf{90.38}_{\pm 3.58}$	$\mathbf{90.07}_{\pm 2.58}$	**88.19**
Qwen-3-8b	LN-Entropy (Malinin & Gales, 2021)	$64.66_{\pm 2.55}$	$61.22_{\pm 4.44}$	$68.70_{\pm 4.11}$	$54.28_{\pm 3.51}$	62.22
	Self-evaluation (Kadavath et al., 2022)	$54.74_{\pm 3.59}$	$51.36_{\pm 0.36}$	$60.02_{\pm 1.18}$	$57.01_{\pm 1.44}$	55.78
	CCS (Burns et al., 2022)	$57.51_{\pm 1.97}$	$57.96_{\pm 3.03}$	$58.63_{\pm 5.53}$	$55.10_{\pm 4.58}$	57.30
	SelfCKGPT (Manakul et al., 2023)	$77.12_{\pm 2.73}$	$66.78_{\pm 2.40}$	$80.51_{\pm 4.53}$	$67.57_{\pm 2.30}$	73.00
	Perplexity (Ren et al., 2023)	$59.65_{\pm 2.42}$	$59.27_{\pm 4.59}$	$67.81_{\pm 4.32}$	$54.06_{\pm 3.63}$	60.20
	SAPLMA♣(Azaria & Mitchell, 2023)	$78.11_{\pm 1.09}$	$86.63_{\pm 1.53}$	$72.86_{\pm 1.20}$	$80.28_{\pm 1.40}$	79.47
	Semantic Entropy (Kuhn et al., 2023)	$81.74_{\pm 2.78}$	$70.84_{\pm 5.41}$	$75.10_{\pm 1.97}$	$70.59_{\pm 5.09}$	74.57
	Lexical Similarity (Lin et al., 2024)	$52.33_{\pm 1.44}$	$53.61_{\pm 2.77}$	$56.75_{\pm 4.30}$	$59.13_{\pm 2.32}$	55.45
	EigenScore (Chen et al., 2024a)	$52.60_{\pm 2.03}$	$52.85_{\pm 2.57}$	$56.39_{\pm 3.26}$	$52.79_{\pm 0.37}$	53.66
	HaloScope♣ (Du et al., 2024)	$58.21_{\pm 3.99}$	$74.98_{\pm 3.98}$	$57.25_{\pm 1.50}$	$62.18_{\pm 4.49}$	63.16
	TAD ♣ (Vazhentsev et al., 2025)	$71.27_{\pm 1.90}$	$55.98_{\pm 4.37}$	$76.82_{\pm 3.77}$	$66.94_{\pm 0.58}$	67.75
	TSV♣ (Park et al., 2025)	$73.42_{\pm 3.89}$	$78.77_{\pm 0.94}$	$61.38_{\pm 3.43}$	$68.40_{\pm 4.92}$	70.49
	HCPD (Ours)	$\mathbf{93.69}_{\pm 0.89}$	$\mathbf{92.63}_{\pm 2.90}$	$\mathbf{87.35}_{\pm 6.22}$	$\mathbf{84.80}_{\pm 1.01}$	**89.62**

- **Best performance over all target models and datasets.**
  - Exceed the suboptimal method by **10.20%** on Llama-3.1-8b;
  - Exceed the suboptimal method by **10.15%** on Qwen-3-8b.

Transferability

❑ Across target models

Table 3. Comparisons with training-based baselines across target models on TriviaQA in terms of AUROC (%).

Source Model	Method	Target Model						
		LLaMA-3.1-8b	LLaMA-3.1-70b	LLaMA-2-7b	LLaMA-2-13b	Qwen-3-8b	Qwen-2.5-7b	Qwen-2.5-14b
LLaMA-3.1-8b	SAPLMA (Azaria & Mitchell, 2023)	$78.51_{\pm 3.16}$	$77.63_{\pm 2.63}$	$78.13_{\pm 2.65}$	$78.25_{\pm 1.83}$	$71.63_{\pm 2.15}$	$69.70_{\pm 2.13}$	$63.77_{\pm 2.28}$
	HaloScope (Du et al., 2024)	$58.19_{\pm 6.10}$	$86.50_{\pm 5.56}$	$82.68_{\pm 2.27}$	$90.88_{\pm 1.85}$	$54.99_{\pm 1.89}$	$68.37_{\pm 3.66}$	$68.33_{\pm 3.77}$
	TSV (Park et al., 2025)	$79.78_{\pm 3.36}$	$81.29_{\pm 10.23}$	$82.85_{\pm 3.90}$	$88.06_{\pm 4.23}$	$59.89_{\pm 8.38}$	$77.07_{\pm 5.82}$	$61.17_{\pm 7.98}$
	HCPD (Ours)	$\mathbf{86.25}_{\pm 1.08}$	$\mathbf{86.87}_{\pm 0.98}$	$\mathbf{90.74}_{\pm 0.52}$	$\mathbf{93.43}_{\pm 1.33}$	$\mathbf{78.89}_{\pm 2.22}$	$\mathbf{88.84}_{\pm 1.02}$	$\mathbf{83.34}_{\pm 0.86}$
Qwen-3-8b	SAPLMA (Azaria & Mitchell, 2023)	$78.22_{\pm 2.64}$	$78.14_{\pm 1.64}$	$80.33_{\pm 1.22}$	$79.50_{\pm 1.28}$	$78.11_{\pm 1.09}$	$71.22_{\pm 1.16}$	$71.73_{\pm 2.18}$
	HaloScope (Du et al., 2024)	$53.89_{\pm 0.78}$	$57.97_{\pm 1.21}$	$55.47_{\pm 9.50}$	$56.73_{\pm 9.89}$	$58.21_{\pm 3.99}$	$57.64_{\pm 1.53}$	$78.65_{\pm 5.95}$
	TSV (Park et al., 2025)	$54.14_{\pm 3.26}$	$58.85_{\pm 5.70}$	$60.31_{\pm 6.89}$	$64.04_{\pm 8.75}$	$73.42_{\pm 3.89}$	$57.08_{\pm 3.75}$	$65.77_{\pm 4.82}$
	HCPD (Ours)	$\mathbf{87.09}_{\pm 1.32}$	$\mathbf{89.73}_{\pm 2.63}$	$\mathbf{90.84}_{\pm 0.76}$	$\mathbf{93.95}_{\pm 1.07}$	$\mathbf{93.69}_{\pm 0.89}$	$\mathbf{89.74}_{\pm 0.77}$	$\mathbf{87.94}_{\pm 1.19}$

❑ Across data distributions

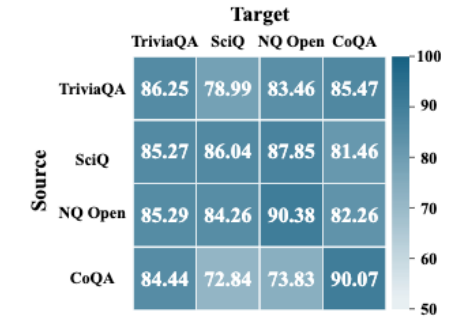


Figure 2. Cross-dataset AUROCs of HCPD on LLaMA-3.1-8b.

- HCPD operates in the **natural language space**, which is universally shared across LLMs, thereby achieving superior transferability across diverse target models.
- **Human-like Criteria Probing mechanism** enables our method to assess queries with varying distributions and formats in a flexible manner.

Alternative Metric

Prompt template for QA dataset

You are tasked with evaluating whether a generated answer is truthful compared to a provided list of gold standard answers. Use a 10-point scoring scale (1-2: Completely inconsistent/hallucinated, 3-4: Mostly inconsistent with minor, 5-6: Partially consistent, mixed quality, 7-8: Mostly consistent with minor errors, 9-10: Completely consistent, accurate, comprehensive) to score the generated content. Provide a lear justification for your scoring.

Response Format:
- Score: [1/2/3/4/5/6/7/8/9/10]
- Justification: [Explain briefly why the answer is correct or incorrect.]

Question: {question}
Gold Standard Answers: {reference_answers}
Generated Answer: {generated_answer}

Table 6. Comparisons with baselines under alternative metrics on TriviaQA for LLaMA-3.1-8b.

Method	Metric	
	ROUGE	DeepSeek
LN-Entropy (Malinin & Gales, 2021)	$77.70_{\pm 1.39}$	$52.69_{\pm 1.19}$
CCS (Burns et al., 2022)	$82.92_{\pm 1.64}$	$52.52_{\pm 2.12}$
SelfCKGPT (Manakul et al., 2023)	$73.92_{\pm 1.26}$	$71.38_{\pm 0.90}$
Perplexity (Ren et al., 2023)	$85.49_{\pm 1.13}$	$53.82_{\pm 2.86}$
SAPLMA (Azaria & Mitchell, 2023)	$87.22_{\pm 1.13}$	$79.99_{\pm 0.92}$
Lexical Similarity (Lin et al., 2024)	$82.40_{\pm 0.90}$	$53.63_{\pm 1.69}$
HaloScope (Du et al., 2024)	$72.99_{\pm 4.61}$	$57.98_{\pm 4.96}$
TSV (Park et al., 2025)	$78.46_{\pm 6.15}$	$73.31_{\pm 5.63}$
HCPD (Ours)	$\mathbf{89.17}_{\pm 0.42}$	$\mathbf{80.42}_{\pm 0.39}$

- **As an evaluation metric provided, HCPD seamlessly aligns its behavior to the reward signal and achieves strong performance.**

Summary of Contributions

- **An adaptive, multi-criteria probing framework for zero-source detection:**

  We propose the Human-like Criteria Probing (HCP) mechanism, which reframes hallucination detection as a process of context-aware criteria generation, weighting, and aggregation. By enabling an LLM agent to adaptively decompose its judgment into interpretable dimensions, our method emulates the nuanced, multi-faceted reasoning of human evaluators, moving beyond monolithic scoring paradigms.

- **Weakly-supervised alignment training without ground-truth labels:**

  To achieve reliable adaptive judgment in the scoring agent, we introduce a reward-based alignment training scheme using only weak supervision derived from semantic consistency. This method effectively teaches the agent to identify, weight, and score relevant criteria without requiring any annotated hallucination data, making it uniquely suited for the practical zero-source constraint.

- **A stable and interpretable inference strategy with theoretical guarantees:**

  We design a multi-sampling aggregation strategy during inference to mitigate generation variance, enhancing decision robustness while preserving full interpretability through the generated criteria and weights. Furthermore, we develop a theoretical analysis, providing formal insights into the feasibility and reliability of our probing-based approach.

Thank you !

OpenReview Link

Github Link