HalluGuard: Demystifying Data-Driven and Reasoning-Driven Hallucinations in LLMs

Xinyue Zeng[1], Junhong Lin[2], Yujun Yan[3], Feng Guo[1]
Liang Shi[1], Jun Wu[4], Dawei Zhou[1]

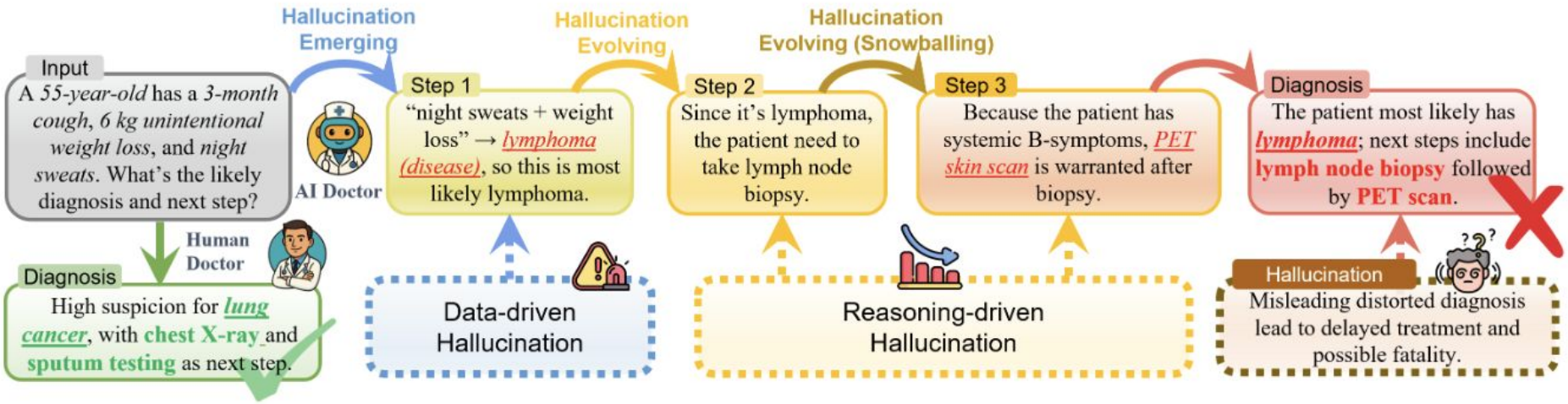
[1]Virginia Tech, [2]MIT
[3]Dartmouth College, [4]Michigan State University

ICLR

COMPUTER SCIENCE
VIRGINIA TECH

Background

- The same wrong answer can arise from different failure mechanisms.
- Some failures start from weak or mismatched knowledge.
- Others start small but get amplified through multi-step reasoning.
- Treating all hallucinations as one bucket hides the right intervention point.



Detection should identify the mechanism of failure, not just the final error.

Motivation

What prior work typically does:

- Scores whether an output looks unreliable
- Uses confidence, entropy, semantic mismatch, or self-consistency signals
- Focuses on the final response as a static end-state

What is still missing:

- Scores whether an output looks unreliable
- Uses confidence, entropy, semantic mismatch, or self-consistency signals
- Focuses on the final response as a static end-state

We still lack a mechanism-aware view of hallucination that explains both where the error starts and how it grows.

Research Questions

Thus, our objective is to formalize hallucination source and evolution in one framework and this work focuses on **two research questions**:

1. *Can we formalize how hallucinations emerge and evolve? If so, can we decompose risk into interpretable sources?*

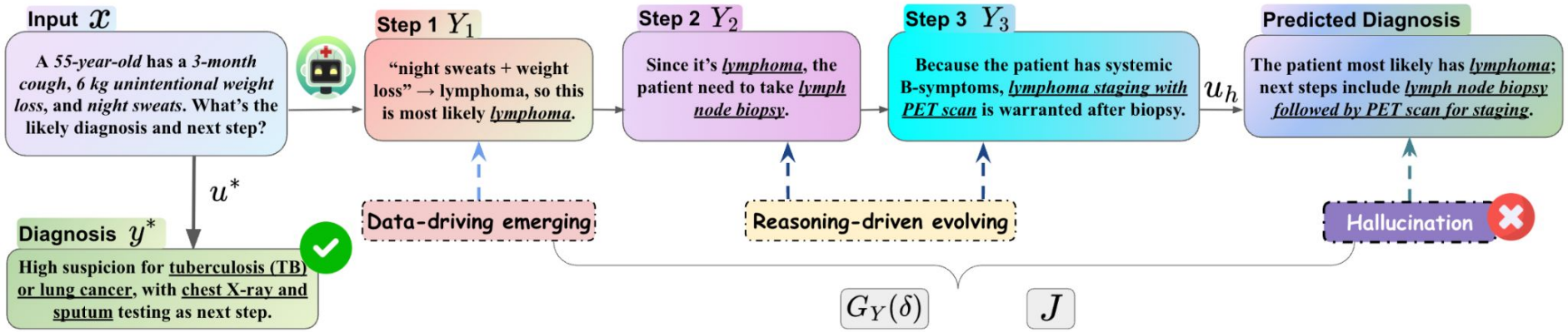
2. *Can that theory yield a reference-free detector? If so, can one detector work across very different task families?*

This shifts hallucination detection from heuristic scoring to interpretable risk decomposition.

Problem Definition

Problem. Hallucination Dynamics: How do hallucinations emerge and evolve?

- **Given:** (1) Target semantic representation $u^* := \Phi(y^*)$ for a ground truth y^*; (2) autoregressive output $Y \sim p_\theta(y_t \mid y_{<t}, x)$ with representation $u_h := \Phi(Y)$ and expectation $\mathbb{E}[u_h]$; (3) inference perturbations $\delta \in R^r$ constrained in $\ell 2$-ball B_ρ .
- **Find:** A geometric mechanism to characterize the emergence and evolution of hallucinations via the Mean Semantic Response Map $G_Y(\delta)$ and the Inference Jacobian J , capturing sensitivity of reasoning trajectories to internal perturbations.



Unified Hallucination Risk Bound

1. **Characterizing Data-Driven Hallucination.**

$$\|u^* - \mathbb{E}[u_h]\| \le \frac{\Lambda}{\gamma} \inf_{u \in U_h} \|u^* - u\|, \quad \frac{\Lambda}{\gamma} \le 1 + k_{\text{pt}} \log \mathcal{O}(P, L) + k \cdot \frac{\epsilon_{\text{mismatch}}}{\text{Signal}_k}$$

2. **Characterizing Reasoning-Driven Hallucination.**

$$\|u_h - \mathbb{E}[u_h]\| \le K \cdot \exp\left(-\frac{K\epsilon^2}{C}\right) \cdot \alpha(e^{\beta T} - 1)$$

3. By combining, we obtain the following **unified bound of data-driven and reasoning-driven hallucinations**

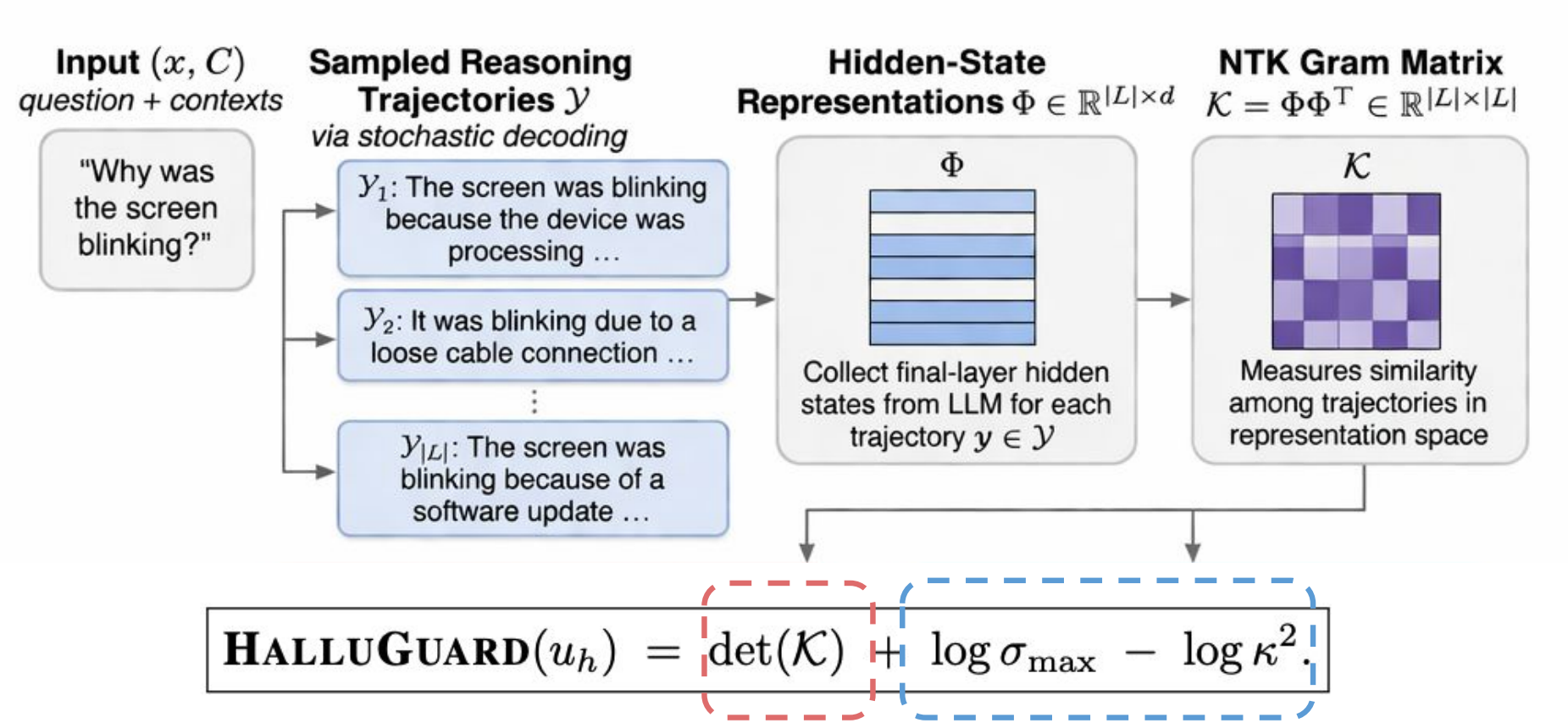
Theorem 2. Let $u^* := \Phi(y^*)$ denote the semantic embedding of the ground-truth output and $u_h := \Phi(Y)$ that of the model-generated output. Under Assumptions A1-A3, suppose there exists $\beta \ge 0$ such that $\left\| \prod_{t=1}^{T} J_t \right\|_2 \le e^{\beta T}$. Then the total hallucination risk satisfies

$$\|u^* - u_h\| \le \underbrace{\left(1 + k_{\text{pt}} \log \mathcal{O}(P, L) + k \cdot \frac{\epsilon_{\text{mismatch}}}{\text{Signal}_k}\right) \inf_{u \in U_h} \|u^* - u\|}_{\text{Data-driven Term}} + \underbrace{|\mathcal{L}| \cdot \exp\left(-\frac{K\epsilon^2}{C}\right) \cdot \alpha(e^{\beta T} - 1)}_{\text{Reasoning-driven Term}}$$

Here, $|\mathcal{L}|$ denotes the total sampled trajectories.

Hallucination ≈ Data-driven Hallucination + Reasoning-driven Hallucination

HalluGuard: From Theory to Practice



Halluguard ≈ Data-driven term + Reasoning-driven term

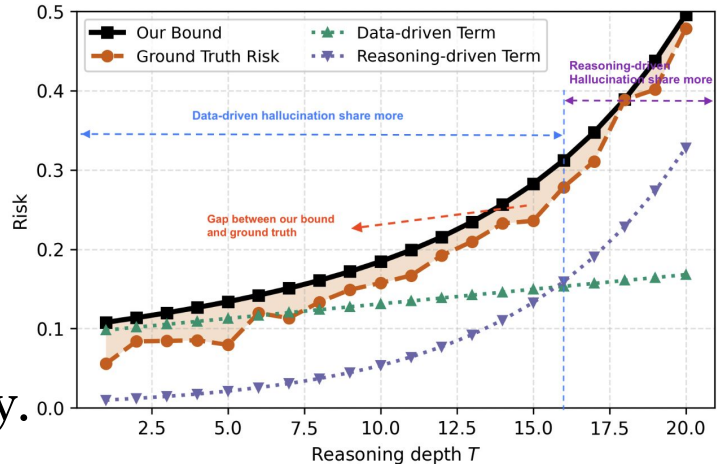
HalluGuard: From Theory to Practice

HALLUGUARD Framework:

$$\textsc{HalluGuard}(u_h) = \det(\mathcal{K}) + \log \sigma_{\max} - \log \kappa^2.$$

- $\det(\mathcal{K})$ is determinant of the NTK Gram matrix, corresponding to data-driven type.
- $\log \kappa^2$ is a proxy for the per-step growth rate of the Inference Jacobian and $\log \sigma_{\max}$ is the log format of the condition number of the NTK matrix. They correspond to reasoning-driven type.

	SQuAD	Math-500	TruthfulQA
$\det(\mathcal{K})$	0.84	0.42	0.61
$\log \sigma_{\max} - \log \kappa^2$	0.39	0.88	0.67



- $\det(\mathcal{K})$ captures factual fidelity,
- $\log \sigma_{\max} - \log \kappa^2$ monitors the inference stability.

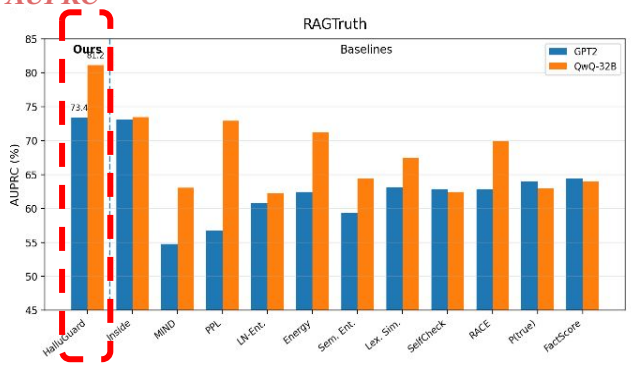
Experimental Setup

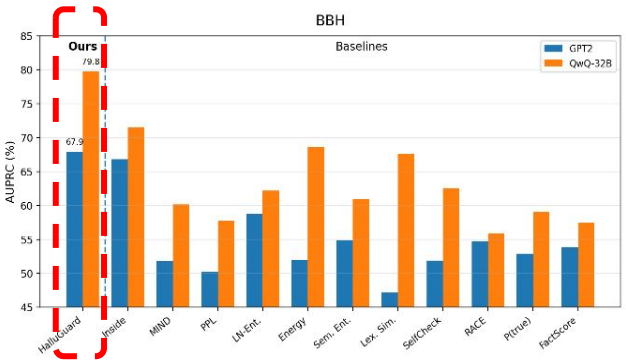
- **Benchmarks**
  - **Data-grounded QA tasks**: RAGTruth, NQ-Open, HotpotQA and SQuAD.
  - **Reasoning-oriented tasks**: GSM8K, MATH-500 and BBH.
  - **Instruction-following tasks**: TruthfulQA, HaluEval and Natural.
- **Models**
  - **Llama family**: Llama2-7B, Llama2-13B, Llama2-70B, Llama3-8B, and Llama3.2-3B
  - **OPT6.7B**
  - **Mistral-7B-Instruct**
  - **QwQ-32B**
  - **GPT-2**
- **Metrics**
  - **Hallucination detection**: ROUGE-based reference evaluation and LLM-AS-A-JUDGE
  - **Performance**: AUROC and AUPRC

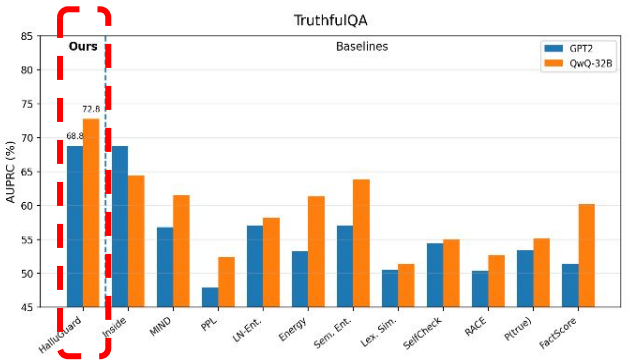
Robustness across Different Benchmarks

- One detector works across factual, reasoning, and instruction-following regimes.
- Gains are largest on reasoning-heavy tasks, where rollout instability matters more.
- Remains strong on data-grounded QA, so the method is not overfit to reasoning benchmarks.
- Supports the claim that mechanism-aware signals transfer across task families.

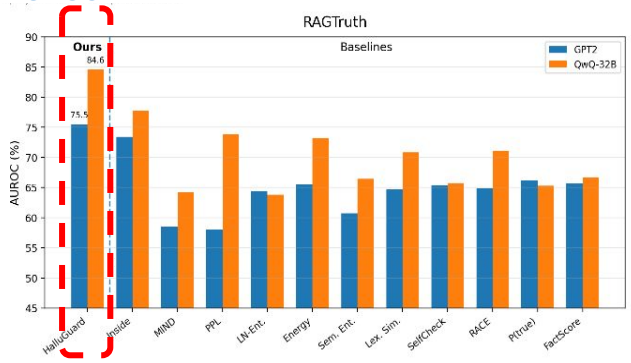
AUPRC

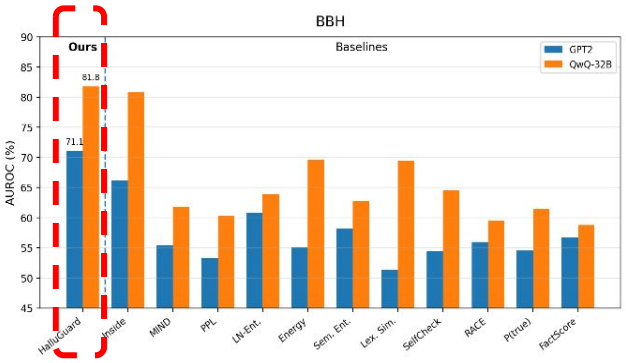


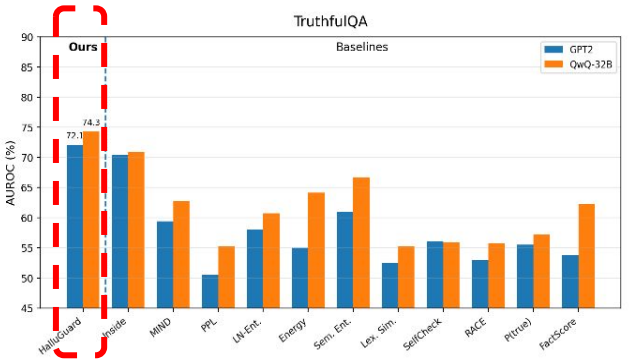




AUROC



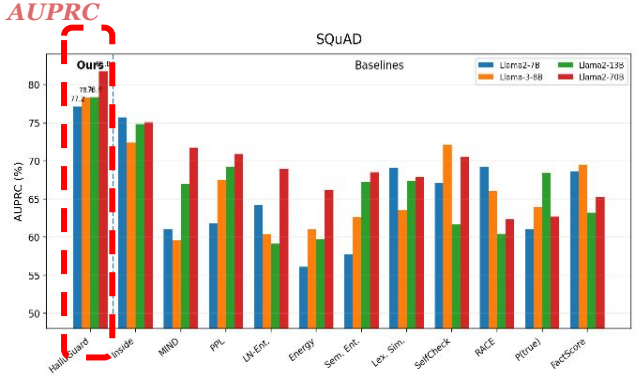


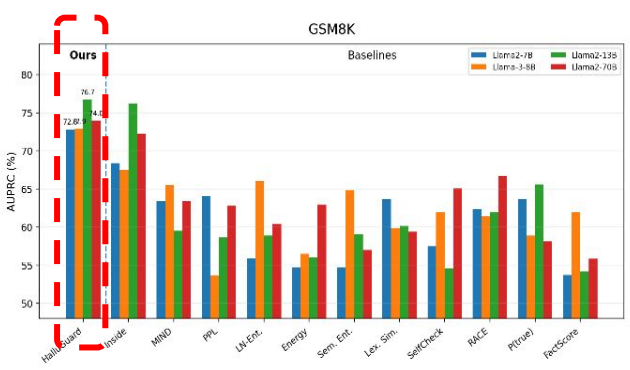


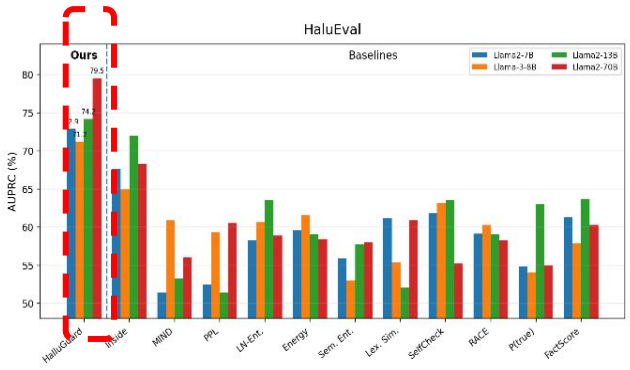
Robustness across Different Scales

- Performance remains strong from 7B-scale models to 70B-scale models.
- Largest relative gains appear on smaller or less stable models, where hallucination risk is harder to control.
- Improvements persist on larger backbones, so the signal does not vanish with scale.
- HalluGuard is useful as a general reliability layer, not just a small-model patch.

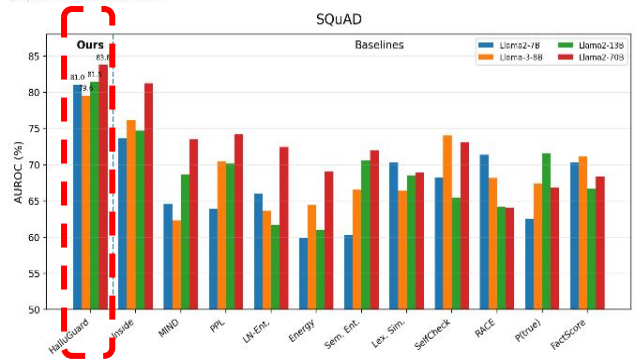
AUPRC

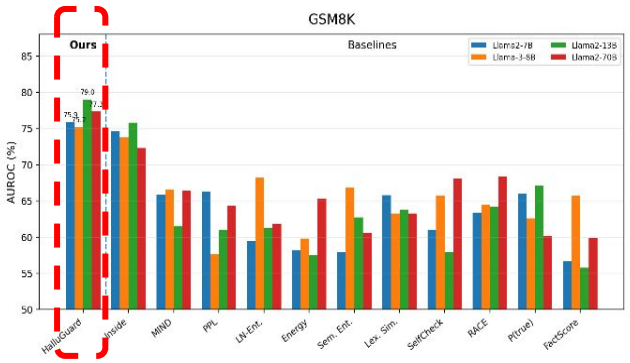


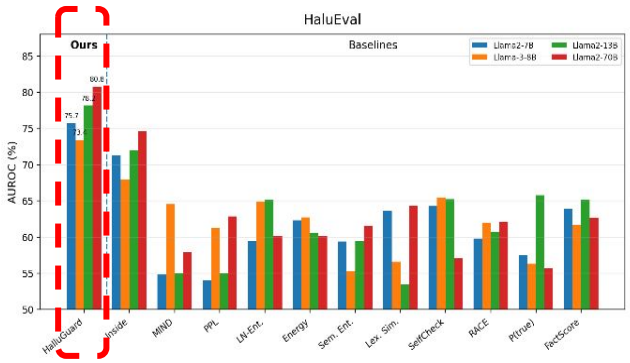




AUROC







Efficient Computation

- HalluGuard uses parallel trajectory sampling, so extra reasoning paths do not incur a full linear cost.
- Inference time is far below heavy multi-sample baselines such as SelfCheckGPT.

Per-Question Inference Time (Seconds) on BBH Across Hallucination Detection

Method	GPT-2	OPT-6.7B	Mistral-7B	QwQ-32B	Llama2-7B	Llama2-70B
Perplexity	0.17	0.34	0.31	0.92	0.28	1.03
HalluGuard	**0.58**	**1.21**	**1.06**	**3.14**	**0.89**	**3.66**
Inside	0.60	1.21	1.09	3.31	0.97	3.73
MIND	0.95	1.90	1.71	5.06	1.52	5.69
LN-Entropy	0.49	1.51	1.57	3.13	1.21	4.05
Energy	0.57	1.14	1.02	3.04	0.91	3.42
Semantic Ent.	0.62	1.24	1.12	3.31	0.99	3.73
Lexical Sim.	0.64	1.23	1.18	3.51	1.07	3.67
SelfCheckGPT	5.17	10.35	9.32	27.60	8.28	31.05
RACE	1.03	2.07	1.86	5.52	1.66	6.21
P(true)	0.52	1.03	0.93	2.76	0.83	3.10
FActScore	0.69	1.38	1.24	3.68	1.10	4.14

Per-Question Inference Time (Seconds) on RAGTruth Across Hallucination Detection

Method	GPT-2	OPT-6.7B	Mistral-7B	QwQ-32B	Llama2-7B	Llama2-70B
Perplexity	0.27	0.54	0.49	1.44	0.43	1.62
HalluGuard	**0.79**	**1.86**	**1.59**	**5.87**	**1.60**	**5.78**
Inside	0.90	1.80	1.62	5.40	1.37	5.35
MIND	1.48	2.96	2.66	8.92	2.20	8.91
LN-Entropy	0.89	1.78	1.60	5.07	1.32	5.67
Energy	0.82	1.64	1.47	4.64	1.22	5.05
Semantic Ent.	0.97	1.94	1.75	5.92	1.50	5.83
Lexical Sim.	1.03	2.02	1.86	5.89	1.48	5.78
SelfCheckGPT	8.10	16.20	14.58	43.20	12.96	48.60
RACE	1.82	3.24	2.92	8.64	1.56	9.72
P(true)	0.97	1.88	1.60	5.18	1.30	5.83
FActScore	1.08	2.16	1.94	5.76	1.73	6.48

Efficient Test-time Inference

- HalluGuard can be used during decoding, not only after generation.
- As a beam-selection signal, it improves final answer accuracy, not just hallucination scoring.
- Gains on MATH-500 and Natural show value in both reasoning and open-ended settings.
- This turns hallucination detection into a test-time control mechanism.

Dataset	IO Prompt	Ours	Inside	MIND	Perplexity	LN-Entropy	Energy	Semantic Ent.	SelfCheck-GPT	RACE	P(true)	FActScore
MATH-500	72.70	**81.00**	74.90	77.10	77.10	76.20	78.00	72.50	74.00	75.10	67.10	71.60
Natural	55.24	**70.96**	67.42	68.32	67.51	68.04	68.59	68.10	65.68	66.90	68.16	67.74

Conclusion

Key Contributions:

1. We propose the First first-principled framework that formalizes hallucination risk dynamics

Previous: Mostly focus on heuristic findings, and lack of theoretical foundation…

2. We introduce a reference-free geometric hallucination characterization **HALLUGUARD**

No supervision, Better hallucination detection, More efficient computation…

3. Robust SOTA performance across diverse benchmarks and backbones and efficient test-time computation

10 diverse benchmarks, 9 popular LLM backbones (up to 70B), 11 baselines …

Q&A

Thank you!

Contact: xyzeng@vt.edu

Paper

GitHub

Homepage

Our Group