HALLUGUARD: DEMYSTIFYING DATA-DRIVEN AND REASONING-DRIVEN HALLUCINATIONS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The reliability of Large Language Models (LLMs) in high-stakes domains such as healthcare, law, and scientific discovery is often compromised by hallucinations. These failures typically stem from two sources: *data-driven hallucinations* and *reasoning-driven hallucinations*. However, existing detection methods usually address only one source and rely on task-specific heuristics, limiting their generalization to complex scenarios. To overcome these limitations, we introduce the *Hallucination Risk Bound*, a unified theoretical framework that formally decomposes hallucination risk into data-driven and reasoning-driven components, linked respectively to training-time mismatches and inference-time instabilities. This provides a principled foundation for analyzing how hallucinations emerge and evolve. Building on this foundation, we introduce HALLUGUARD, a NTK-based score that leverages the induced geometry and captured representations of the NTK to jointly identify data-driven and reasoning-driven hallucinations. We evaluate HALLUGUARD on 10 diverse benchmarks, 11 competitive baselines, and 9 popular LLM backbones, consistently achieving state-of-the-art performance in detecting diverse forms of LLM hallucinations.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in high-stakes domains such as healthcare, law, and scientific discovery(Bommasani et al., 2021; Thirunavukarasu et al., 2023). However, adoption in these settings remains cautious, as such domains are highly regulated and demand strict compliance, interpretability, and safety guarantees(Dennstädt et al., 2025; Kattnig et al., 2024). A major barrier is the risk of *hallucinations*, generated content appears unfaithful or nonsensical. Such errors can have severe consequences(Dennstädt et al., 2025)—as the example in Figure 1, a generated incorrect medical diagnosis may delay treatment or lead to harmful interventions. Therefore, detecting hallucinations is not merely a technical challenge but a prerequisite for trustworthy deployment, as undetected errors undermine reliability, accountability, and user safety.

Generally, hallucinations in LLMs arise from two primary sources(Ji et al., 2023; Huang et al., 2023): *data-driven hallucinations*, which stem from flawed, biased, or incomplete knowledge encoded during pre-training or fine-tuning; and *reasoning-driven hallucinations*, which originate from inference-time failures such as logical inconsistencies or breakdowns in multi-step reasoning(Zhang et al., 2023; Zhong et al., 2024). Detection methods broadly split along these two dimensions. Approaches for data-driven hallucinations often compare outputs against retrieved documents or references(Krishna & et al., 2022; Min et al., 2023b; Ji et al., 2023), or exploit sampling consistency as in SelfCheckGPT(Manakul et al., 2023). In contrast, methods for reasoning-driven hallucinations rely on signals of inference-time instability, including probabilistic measures such as perplexity(Ren et al., 2022), length-normalized entropy(Malinin & Gales, 2020), semantic entropy(Kuhn et al., 2023), energy-based scoring(Liu et al., 2020), and RACE(Wang et al., 2025). Others probe internal representations, for example, Inside(Chen et al., 2024a), which applies eigenvalue-based covariance metrics and feature clipping, ICR Probe(Zhang et al., 2025), which tracks residual-stream updates, and Shadows in the Attention(Wei et al., 2025), which analyzes representation drift under contextual perturbations. While these methods shed light on the mechanisms underlying hallucinations, most remain tailored to a single hallucination type and fail to capture their evolution. Yet growing evidence indicates that data-driven and reasoning-driven hallucinations often evolve during multi-step generation(Liu et al., 2025; Sun et al., 2025). As shown in Figure 1, it emerges from an initial dis-

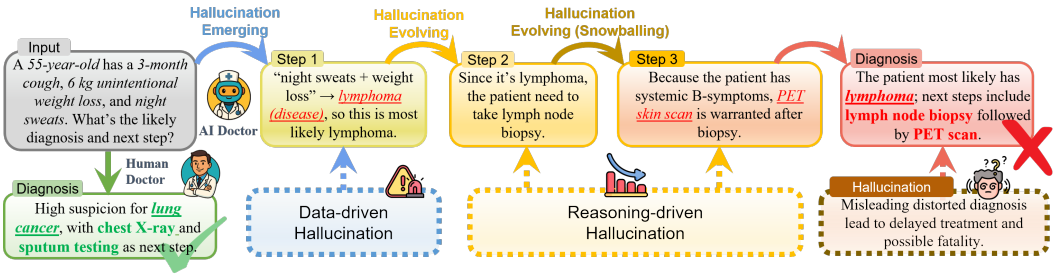

Figure 1: An illustration of hallucination emerging and evolving in the context of disease diagnosis.

ease misclassification and evolves into a distorted diagnosis, delaying treatments and risking fatality. This gap brings two central questions: *(1) How can we develop a unified theoretical understanding of how hallucinations evolve? and (2) How can we detect them effectively and efficiently without relying on external references or task-specific heuristics?*

To address these challenges, we propose a unified theoretical framework–*Hallucination Risk Bound*, which decomposes the overall hallucination risk into two components: a *data-driven term*, capturing semantic deviations rooted in inaccurate, imbalanced, or noisy supervision acquired during model training; and a *reasoning-driven term*, reflecting instability introduced by inference-time dynamics, such as logical missteps or temporal inconsistency. This decomposition not only elucidates the mechanism behind hallucinations but also reveals how they emerge and evolve. Specifically, our analysis shows that hallucinations originate from semantic approximation gaps-captured by representational limits of the model-and are subsequently amplified by unstable rollout dynamics, evolving across decoding steps. As such, our framework offers a unified theoretical lens for characterizing the emergence and evolution of these hallucinations.

Building on the theoretical foundation, we propose HALLUGUARD, a Neural Tangent Kernel(NTK)-based score that leverages the induced geometry and captured representations of the NTK to jointly identify data-driven and reasoning-driven hallucinations. We evaluate HALLUGUARD comprehensively across 10 diverse benchmarks, 11 competitive baselines, and 9 popular LLM backbones. HALLUGUARD consistently achieves state-of-the-art hallucination detection performance, demonstrating its efficacy.

2 PRELIMINARIES

Hallucination Detection. There are two primary sources of hallucinations in LLMs(Ji et al., 2023; Huang et al., 2023): *data-driven hallucination*, which stems from incomplete or biased knowledge encoded during pre-training or fine-tuning, and *reasoning-driven hallucination*, which arises from unstable or inconsistent inference dynamics at decoding time. This distinction has implicitly guided a broad range of detection strategies, which we examine through these two lenses.

For data-driven causes, a recurring signal is elevated predictive uncertainty. A common formulation adopts the sequence-level negative log-likelihood:

$$\mathcal{U}(\mathbf{y} \mid \mathbf{x}, \theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, \mathbf{x}), \tag{1}$$

which quantifies the average uncertainty of generating a sequence $\mathbf{y} = [y_1, \ldots, y_T]$ from input $\mathbf{x}$ and θ denotes model parameters. This directly recovers *Perplexity*(Ren et al., 2022), where low scores imply confident predictions, while high scores indicate implausible generations due to weak priors. To capture more nuanced uncertainty, later methods extend this formulation to multi-sample settings. The *Length-Normalized Entropy*(Malinin & Gales, 2020) penalizes dispersion across stochastic generations $\mathcal{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^K\}$, offering a finer-grained view of model indecision. This perspective is further enriched by *Semantic Entropy*(Kuhn et al., 2023), which projects sampled responses into semantic space, and by energy-based scoring(Liu et al., 2020), which replaces log-probability with a learned confidence function. Collectively, these methods reflect a progression from token-level likelihoods to semantically grounded multi-sample uncertainty estimators.

In contrast, reasoning-driven hallucinations arise from brittle inference trajectories, where identical contexts may yield inconsistent or incoherent outputs. A commonly used measure of such instability is the cross-sample consistency score:

$$\mathcal{C}(\mathcal{Y} \mid \mathbf{x}, \theta) = \frac{1}{C} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \text{sim}(\mathbf{y}^i, \mathbf{y}^j), \tag{2}$$

where $C = K \cdot (K-1)/2$, and $\text{sim}(\cdot, \cdot)$ is a similarity function such as ROUGE-L(Lin, 2004), cosine similarity, or BLEU(Chen et al., 2024b). Low scores reflect diverging generations and unstable reasoning. Several reasoning-driven detection methods can be interpreted through this lens. Early approaches used surface-level lexical overlap metrics(Lin et al., 2022b), while ***SelfCheck-GPT***(Manakul et al., 2023) advanced this by evaluating factual entailment across responses, and ***FActScore***(Min et al., 2023a) extended this further by comparing outputs to retrieved reference documents. More recent efforts probe internal signals directly: ***Inside***(Chen et al., 2024a) analyzes the covariance spectrum of embedding representations, and ***RACE***(Wang et al., 2025) diagnoses instability in multi-step reasoning.

NTK in LLMs. NTK provides a principled framework for analyzing the training dynamics in the overparameterized regime characteristic of modern LLMs(Jacot et al., 2020). Formally, for a network output $f(x, \theta)$ with input x and parameters θ, the NTK is defined as:

$$\Theta(x, x', \theta) = \nabla_\theta f(x, \theta) \cdot \nabla_\theta f(x', \theta). \tag{3}$$

This kernel $\Theta(x, x', \theta)$ quantifies the similarity of training dynamics between inputs x and x'. In the infinite-width limit, it converges to a deterministic value at initialization and remains nearly constant throughout training(Lee et al., 2020). This stability reduces the highly nonlinear optimization of deep networks to a tractable kernel regression problem. By examining the eigenspectrum of the NTK, one can probe how internal representations are shaped during training: which features are prioritized (e.g., syntax versus semantics), how quickly different tasks converge, and why overparameterized networks generalize effectively to unseen data(Ju et al., 2022). In this way, the NTK transforms the apparent complexity of LLM optimization into a clear lens on how these models capture, process, and generalize information(Zeng et al., 2025).

3 METHODOLOGY

3.1 PROBLEM SETTING

Our analysis reveals that hallucination is not a unified failure mode but rather shifts with the task structure. On the instruction-following `Natural` benchmark(Wang et al., 2022), 88.9% of the overall 3499 errors are from logical missteps (*reasoning-driven*) while 11.1% are factual inaccuracies (*data-driven*). By contrast, on the math-focused `MATH-500`(Hendrycks et al., 2021), the 1985 wrong generations are dominated by 1946 reasoning errors (98.1%), with only 19 factual flaws (1.9%). This contrast highlights that, in practice, hallucinations are rarely pure but often mixtures of data-driven bias and reasoning-driven instability—motivating our formal decomposition of hallucination sources.

Problem Definition. Let $\mathcal{Y}$ denote the space of textual outputs and let $\Phi : \mathcal{Y} \to U_h$ be a task-specific encoder that maps textual sequences into the hypothesis space U_h, equipped with a norm $\|\cdot\|$ (e.g., task-calibrated embedding space or structured metric). We interpret each $u \in U_h$ as a reasoning chain, composed of step-wise logical statements. For an input $\mathbf{x}$ with ground-truth output $y^* \in \mathcal{Y}$, define the gold-standard reasoning chain as $u^* := \Phi(y^*) \in U_h$. An LLM with parameters θ emits a random sequence $Y = (Y_1, \ldots, Y_T)$ via $p_\theta(y_t \mid y_{<t}, \mathbf{x})$, yielding a predicted reasoning chain $u_h := \Phi(Y) \in U_h$. Its expected value under the model's decoding distribution is $\mathbb{E}[u_h] := \mathbb{E}_{Y \sim p_\theta(\cdot|\mathbf{x})}[\Phi(Y)]$.

We consider perturbations in a local neighborhood of the decoding process. Let $\delta \in \mathbb{R}^r$ parameterize a small perturbation (e.g., of the prefix tokens, step-t logits, or hidden state), and let $\mathcal{B}_\rho := \{\delta : \|\delta\| \le \rho\}$. Define the perturbed decoder map $G : \mathbb{R}^r \to U_h$ by $G(\delta) := \Phi\big(Y(\delta)\big)$, where $Y(\delta)$ is the sequence under perturbation. Let $J \in \mathbb{R}^{d_h \times r}$ denote the (Gauss–Newton) Jacobian of G at $\delta = 0$. Our goal is to formalize how hallucination emerges and evolves in LLMs.

3.2 HALLUCINATION RISK BOUND

To bridge the formal setup with the phenomenon of hallucination, we first disentangle the sources of hallucinations. Intuitively, hallucinations may arise either from systematic biases in the knowledge encoded by the model (data-driven) or from instabilities during autoregressive decoding (reasoning-driven). The following proposition formalizes this idea by decomposing the total hallucination risk into two components.

We first impose the following assumptions:

- **A1.** $(U, \|\cdot\|)$ is a Hilbert space; Φ is measurable with unique best solution and $\|\Phi(Y)\|$ has finite second moment.
- **A2.** Triangle inequality holds for $\|\cdot\|$ and Φ is L_Φ-Lipschitz w.r.t. an edit distance on $\mathcal{Y}$.
- **A3.** For $\delta \in \mathcal{B}_\rho$, the mapping G admits the local expansion $G(\delta) = G(0) + J\delta + R(\delta)$, where the remainder is bounded by $\|R(\delta)\| \le \frac{1}{2} H_\star \|\delta\|^2$ for some curvature constant $H_\star > 0$.

Proposition 3.1 (Hallucination Risk Decomposition). Under A1–A3, applying the triangle inequality yields a natural split of the risk:

$$\|u^* - u_h\| \;\le\; \underbrace{\|u^* - \mathbb{E}[u_h]\|}_{\text{data-driven term}} \;+\; \underbrace{\|u_h - \mathbb{E}[u_h]\|}_{\text{reasoning-driven term}}$$

This decomposition distinguishes errors caused by systematic bias in the learned representation from those introduced during stochastic rollout.

Characterizing Data-Driven Hallucination. To quantify the data-driven term, we take inspiration from the NTK, which has proven effective in analyzing training dynamics of overparameterized models. Here, NTK geometry provides a way to measure how well the model's representation space aligns with task generation under small perturbations.

Let $U_h \subset U$ denote the hypothesis subspace accessible to the model under perturbations. By Céa's lemma(Céa, 1964) with curvature penalty, the data-driven term can be bounded as

$$\|u^* - \mathbb{E}[u_h]\| \;\le\; \frac{\Lambda}{\gamma} \inf_{u \in U_h} \|u^* - u\|, \tag{4}$$

where $\gamma = \lambda_{\min}(\mathcal{K}_\Phi)$ is the smallest eigenvalue of the NTK Gram matrix on embedded perturbations, and $\Lambda \le \|T\|$ reflects the operator norm of the problem/operator mapping $\mathcal{T}$. Intuitively, the ratio $\frac{\Lambda}{\gamma}$ measures the conditioning of the feature map: well-conditioned NTK spectra allow a closer approximation to the true generation.

This ratio can be further controlled in terms of pretraining–finetuning mismatch:

$$\frac{\Lambda}{\gamma} \;\le\; 1 \;+\; k_{\text{pt}} \log \mathcal{O}(P, L) \;+\; k \cdot \frac{\epsilon_{\text{mismatch}}}{\text{Signal}_k}, \tag{5}$$

where $\log \mathcal{O}(P, L)$ is a complexity term from parameter count P and prompt length L, $\epsilon_{\text{mismatch}}$ denotes the Wasserstein distance between prompt and query distributions, Signal_k measures task-aligned energy in the top-k eigenspace. k_{pt} and k are task and model-dependent constants. Thus, data-driven hallucinations grow when the mismatch is large or when the task signal is weak.

Characterizing Reasoning-Driven Hallucination. The reasoning-driven term captures *reasoning-driven* instability that accumulates during autoregressive decoding. Here, we model generation as a martingale process, where deviation from the expectation is controlled by concentration inequalities. Specifically, Freedman's inequality(Geman et al., 1992) gives

$$\|u_h - \mathbb{E}[u_h]\| \;\le\; K \cdot \exp\Big(- \tfrac{K\epsilon^2}{C} \Big) \cdot \alpha(e^{\beta T} - 1), \tag{6}$$

where K is the number of rollouts averaged, β summarizes per-step growth in local Jacobians, α scales the cumulative effect and C is a task and model-dependent constant. This bound shows that reasoning-driven hallucinations grow exponentially with sequence length T.

We now synthesize the two components into a unified result that characterizes the overall risk of hallucination. By combining the NTK-conditioned approximation bound for data-driven deviation

with the Freedman-style concentration bound for reasoning-driven instability, we obtain the following unified bound of data-driven and reasoning-driven hallucinations (detailed proof is provided in Appendix A):

Theorem 3.2 (Hallucination Risk Bound). 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 $\beta \geq 0$ such that $\left\| \prod_{t=1}^{T} J_t \right\|_2 \leq e^{\beta T}$. Then the total hallucination risk satisfies

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

3.3 HALLUCINATION QUANTIFICATION VIA HALLUGUARD

While Theorem 3.2 makes explicit how data-driven and reasoning-driven hallucinations emerge and evolve, applying it directly at inference is impractical since direct step-wise Jacobians for billion-parameter LLMs are intractable, so we seek a *proxy score* that is computable, stable, and faithful to our decomposition.

Let $\mathcal{K}$ denote the NTK Gram matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and condition number $\kappa(\mathcal{K}) = \lambda_{\max}/\lambda_{\min}$. Let J_t be the step-t input–output Jacobian of the decoder, and define $\sigma_{\max} := \sup_t \|J_t\|_2$ as the uniform spectral bound(note that $\sigma_{\max}$ is independent of the spectrum of $\mathcal{K}$).

Under Assumptions A1–A3, a standard NTK approximation argument yields $\inf_{u \in U_h} \|u^* - u\| \leq C_d \det(\mathcal{K})^{-c_d} \|u^*\|$, so that $\det(\mathcal{K})$ capture the representations in systematic bias.

For autoregressive rollout, based on the property of Jacobian, we have $\left\| \prod_{t=1}^{T} J_t \right\|_2 \leq \prod_{t=1}^{T} \|J_t\|_2 = \exp\left(\sum_{t=1}^{T} \log \|J_t\|_2 \right)$, so that we have $\left\| \prod_{t=1}^{T} J_t \right\|_2 \leq e^{\beta T}$. Since $\beta \leq \log \sigma_{\max}$ with $\sigma_{\max} := \sup_t \|J_t\|_2$ thus we have the upper bound as $\| \prod_{t=1}^{T} J_t \|_2 \leq \sigma_{\max}^T = e^{(\log \sigma_{\max})T}$. Thus, $\log \sigma_{\max}$ serves as a stable and tractable proxy for the per-step amplification rate.

Perturbation analysis of $\mathcal{K}$, together with classical eigenvalue sensitivity results(Trefethen & Bau, 2022), yields $\text{Var}[u_h] \leq c_v \, \kappa(\mathcal{K})^2 \, \|\delta\|^2$, showing that instability grows quadratically with the condition number $\kappa(\mathcal{K})$. To temper this effect and ensure additivity, we penalize ill-conditioned representations via $-\log \kappa^2$, where log compression brings a well-behaved dynamic range.

In summary, $\det(\mathcal{K})$ quantifies representational adequacy, $\log \sigma_{\max}$ captures rollout amplification, and $-\log \kappa^2$ penalizes spectral instability, together forming a compact and tractable proxy consistent with the Hallucination Risk Bound. Detailed proofs are provided in Appendix B.

Table 1: Correlation between NTK proxies and task families.

	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

Empirical validation. We empirically validate how those proxies correlate with different task families. In Table 1, $\det(\mathcal{K})$ correlates most strongly with the data-centric task `SQuAD` (0.84), indicating its role in capturing factual fidelity. In contrast, for the reasoning-oriented `MATH-500`, the highest correlation is observed with $\log \sigma_{\max} - \log \kappa^2$ (0.88), reflecting the importance of amplification and stability in multi-step reasoning.

Motivated by the above, we formally define HALLUGUARD as follows, which provides a principled and unified lens for hallucination detection:

$$\boxed{\textbf{HALLUGUARD}(u_h) = \det(\mathcal{K}) + \log \sigma_{\max} - \log \kappa^2.} \tag{7}$$

4 EXPERIMENTS

We comprehensively evaluate HALLUGUARD across 10 diverse benchmarks, 11 competitive baselines, and 9 popular LLM backbones. We aim to evaluate its efficacy from the following five ques-

tions: ***Q1:*** *How does* HALLUGUARD *perform across different task families?* ***Q2:*** *How does* HALLUGUARD *perform across LLMs of different scales?* ***Q3:*** *How does each term capture trends across task families?* ***Q4:*** *Can* HALLUGUARD *guide test-time inference to improve downstream reasoning?* ***Q5:*** *How well does* HALLUGUARD *generalize to detecting fine-grained hallucinations beyond benchmarks?*

Section 4.1 details the setup; Section 4.2 evaluates HALLUGUARD as a detection method(Q1–Q3), Section 4.3 applies HALLUGUARD in score-guided inference(Q4) and Section 4.4 analyzes HALLUGUARD on fine-grained hallucination via a case study on semantic data(Q5).

4.1 EVALUATION SETUP

Benchmarks. We evaluate across 10 widely used benchmarks spanning three distinct categories. For data-grounded QA, we include `RAGTruth`(Niu et al., 2024), `NQ-Open`(Kwiatkowski et al., 2019), `HotpotQA`(Yang et al., 2018) and `SQuAD`(Rajpurkar et al., 2016), which emphasize factual correctness through external evidence. For reasoning-oriented tasks, we use `GSM8K`(Cobbe et al., 2021), `MATH-500`(Hendrycks et al., 2021), and `BBH`(Suzgun et al., 2022), which require multi-step derivations prone to compounding errors. Finally, for instruction-following settings, we consider `TruthfulQA`(Lin et al., 2022a), `HaluEval`(Li et al., 2023) and `Natural`(Wang et al., 2022), which probe hallucinations under open-ended or adversarial prompts.

Baselines. We compare HALLUGUARD with 11 competitive detectors spanning diverse strategies. Uncertainty-based methods include Perplexity(Ren et al., 2022), Length-Normalized Predictive Entropy(LN-Entropy)(Malinin & Gales, 2020), Semantic Entropy(Kuhn et al., 2023), Energy Score(Liu et al., 2020) and P(true)(Kadavath et al., 2022). Consistency-based approaches cover SelfCheckGPT(Manakul et al., 2023), Lexical Similarity(Lin et al., 2022b), FActScore(Min et al., 2023a) and RACE(Wang et al., 2025). Internal-state methods are represented by Inside(Chen et al., 2024a), MIND(Su et al., 2024) and ReACTScore(Zheng et al., 2024).

LLM Backbone Models. We evaluate 9 publicly available LLMs spanning different scales and architectures. These include five models from the Llama family (Llama2-7B, Llama2-13B, Llama2-70B, Llama3-8B, and Llama3.2-3B)(Touvron et al., 2023; Grattafiori et al., 2024), along with OPT-6.7B(Zhang et al., 2022), Mistral-7B-Instruct(Jiang et al., 2023), QwQ-32B(Yang et al., 2024), and GPT-2 (117M)(Radford et al., 2019). All models are used in their off-the-shelf form with pre-trained weights and tokenizers provided by Hugging Face, without further fine-tuning.

Evaluation Metrics. We evaluate hallucination detection ability under two regimes following Janiak et al. (2025): ROUGE-based reference evaluation ($*_r$) and LLM-AS-A-JUDGE ($*_{\text{llm}}$). For performance measures, we report the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). AUROC is widely used to assess the quality of binary classifiers and uncertainty estimators, while AUPRC highlights performance under class imbalance. In both cases, higher values indicate better detection.

4.2 MAIN RESULTS

Q1: How does HALLUGUARD perform across different task families? To evaluate how HALLUGUARD performs across different task types, we conduct experiments on all benchmarks. For clarity, Table 2 presents representative results from three task families: data-centric (`RAGTruth`), reasoning-oriented (`Math-500`), and instruction-following (`TruthfulQA`). As shown, HALLUGUARD consistently outperforms all baselines across backbones. On `Math-500`, it reaches 81.76% AUROC and 79.76% AUPRC, improving over the second-best method by up to 8.3%. On `RAGTruth`, it attains 84.59% AUROC and 81.15% AUPRC, with gains of up to 7.7%. On `TruthfulQA`, it achieves 77.05% AUROC and 73.79% AUPRC, exceeding the next strongest baseline by as much as 6.2%. Overall, HALLUGUARD establishes new state-of-the-art results across diverse task families, with particularly pronounced improvements on reasoning-oriented benchmarks.

Q2: How does HALLUGUARD perform across LLMs of different scales? We further investigate whether the effectiveness of HALLUGUARD depends on model scale, as smaller backbones are typically more prone to hallucination. Table 3 reports representative results on small(Llama2-7B, Llama3-8B), mid-sized(Llama2-13B), and large-scale(Llama2-

Table 2: Performance comparison on representative benchmarks: data-centric (RAGTruth), reasoning-oriented (Math-500), and instruction-following (TruthfulQA). We highlight the **first** and second best results.

		GPT2				OPT-6.7B				Mistral-7B				QwQ-32B			
		AUROC$_r$	AUPRC$_r$	AUROC$_{lim}$	AUPRC$_{lim}$	AUROC$_r$	AUPRC$_r$	AUROC$_{lim}$	AUPRC$_{lim}$	AUROC$_r$	AUPRC$_r$	AUROC$_{lim}$	AUPRC$_{lim}$	AUROC$_r$	AUPRC$_r$	AUROC$_{lim}$	AUPRC$_{lim}$
RAGTruth	HALLUGUARD	**75.51**	**73.40**	**62.40**	**56.60**	**80.13**	**76.77**	**71.01**	**63.58**	**82.31**	**80.79**	**64.89**	**67.25**	**84.59**	**81.15**	**71.82**	**66.68**
	Inside	73.42	73.08	61.99	56.39	79.49	71.82	66.1	62.46	75.32	73.19	64.58	61.05	77.72	73.47	66.05	64.73
	MIND	58.54	54.79	43.47	41.85	63.82	62.58	51.03	44.78	73.13	71.53	58.25	58.6	64.23	63.06	47.37	51.47
	Perplexity	58.07	56.68	43.84	41.53	64.47	61.57	47.12	52.98	65.42	63.63	53.28	51.36	73.91	72.92	60.81	59.77
	LN-Entropy	64.42	60.79	49.41	45.04	60.81	57.91	48.76	42.27	64.22	60.92	52.24	48.41	63.81	62.26	47.52	52.17
	Energy	65.53	62.42	51.8	47.22	66.54	63.28	54.21	49.19	64.36	62.26	48.64	53.93	73.26	71.21	65.43	62.32
	Semantic Ent.	60.72	59.41	50.55	45.86	70.2	68.34	54.54	56.74	66.01	64.49	53.01	55.5	66.48	64.41	51.54	50.11
	Lexical Sim.	64.72	63.1	55.04	48.04	67.28	64.62	52.55	54.86	64.96	61.17	52.34	45.11	70.87	67.41	61.25	51.01
	SelfCheckGPT	65.4	62.79	52.85	52.43	66.64	64.89	52.69	51.17	71.19	68.45	63.13	60.23	65.79	62.45	54.76	51.29
	RACE	64.83	62.84	51.8	48.44	64.26	61.03	52.74	46.22	66.34	64.54	51.88	53.86	71.13	69.96	57.58	55.54
	P(true)	66.19	64.04	48.2	56.27	68.44	65.48	57.53	53.08	72.54	71.8	57.25	59.42	65.32	63.01	53.01	52.32
	FActScore	65.72	64.39	51.94	47.51	61.53	58.2	51.86	45.57	63.98	60.71	53.54	49.34	66.72	64.03	58.21	49.17
BBH	HALLUGUARD	**71.06**	**67.94**	**62.05**	**59.05**	**73.1**	**70.88**	**63.67**	**61.88**	**79.85**	**76.5**	**67.13**	**60.57**	**81.76**	**79.76**	**68.77**	**65.46**
	Inside	66.18	66.81	56.15	58.62	70.64	65.22	63.28	59.28	67.2	65.49	51.3	53.46	80.8	71.49	64.05	63.42
	MIND	55.41	51.77	39.01	41.59	55.48	53.46	38.59	40.88	65.71	63.7	49.61	52.54	61.75	60.18	53.46	50.04
	Perplexity	53.28	50.22	43.86	38.98	64.89	62.12	48.65	51.99	61.97	60.05	51.15	42.87	60.28	57.75	51.62	43.38
	LN-Entropy	60.84	58.76	42.76	47.48	58.71	55.01	43.55	42.02	68.96	69.44	58.79	57.49	63.96	62.18	46.01	49.5
	Energy	55.09	51.99	46.2	39.5	53.96	50.98	42.56	34.12	66.27	62.72	49.48	50.06	69.61	68.66	54.35	57.36
	Semantic Ent.	58.16	54.81	49.61	40.39	62.63	59.52	50.14	45.02	64.99	61.33	50.11	45.53	62.76	60.95	45.77	45.75
	Lexical Sim.	51.37	47.18	38.37	39.06	61.27	58.06	44.13	42.96	58.25	55.92	46.31	46.01	69.46	67.59	55.93	52.6
	SelfCheckGPT	54.51	51.86	44.62	44.01	57.36	53.21	42.55	38.27	63.68	62.5	51.7	53.03	64.56	62.49	55.85	45.8
	RACE	55.99	54.66	41.39	38.32	64.23	62.03	56.03	53.44	66.88	64.33	49.57	48.5	59.5	55.83	46.13	41.07
	P(true)	54.57	52.88	45.45	44.74	57.02	55.49	48.81	37.84	57.11	55.21	43.93	47.05	61.49	59.03	44.37	44.69
	FActScore	56.76	53.85	40.25	40.01	54.51	53.2	38.45	36.49	62.11	58.64	53.52	47.27	58.82	57.47	49.48	42.74
TruthfulQA	HALLUGUARD	**72.1**	**68.76**	**60.09**	**52.01**	**69.59**	**68.36**	**58.52**	**52.65**	**77.05**	**73.79**	**63.62**	**62.26**	**74.26**	**72.76**	**57.39**	**64.07**
	Inside	70.42	68.76	60.09	52.01	62.1	59.78	51.07	51.38	62.53	60.99	52.3	49.35	70.89	64.44	56.61	56.01
	MIND	59.45	56.79	45.22	43.71	60.56	58.55	47.49	49.63	59.2	57.98	47.23	41.79	62.81	61.5	52.56	46.37
	Perplexity	50.57	47.87	40.64	35.63	55.07	52.26	44.43	42.79	60.8	59.69	47.33	41.62	55.29	52.46	43.95	43.92
	LN-Entropy	58.04	56.99	41.94	47.21	56.12	54.01	47.06	38.4	59.67	56.25	41.99	41.25	60.76	58.21	46.24	42.64
	Energy	55.02	53.31	38.78	45.16	54.42	51.85	36.21	42.57	58.93	55.25	50.76	41.72	64.15	61.32	51.78	50.02
	Semantic Ent.	61.01	57.08	43.35	45.2	51.48	47.81	34.15	38.16	54.44	53.33	36.62	40.35	66.75	63.85	51.11	46.71
	Lexical Sim.	52.54	50.56	39.94	33.42	59.74	55.72	49.89	46.81	66.16	64.05	54.08	51.65	55.24	51.36	46.39	39.57
	SelfCheckGPT	56.04	54.48	43.78	44.38	58.93	56.47	47.65	39.02	61.14	58.91	42.97	47.01	55.86	54.95	41.08	37.35
	RACE	53.02	50.33	41.7	33.81	62.95	67.89	54.61	51.93	71.06	68.49	60.4	57.44	55.75	52.62	46.5	43.19
	P(true)	55.52	53.41	38.33	38.38	54.88	53.1	38.22	40.96	55.8	52.01	40.88	38.72	57.18	55.16	46.19	38.21
	FActScore	53.82	51.42	41.33	35.2	54.57	51.26	42.51	35.52	53.97	50.2	42.97	36.16	62.31	60.23	45.06	49.9

70B) models using SQuAD, GSM8K, and HaluEval. Across all settings, HALLUGUARD consistently surpasses baselines, with the largest margins on smaller models—for instance, 72.89% AUPRC$_r$ on HaluEval with Llama2-7B, more than 10% above the second best. Mid-sized models also exhibit clear gains (e.g., 79.01% AUROC$_r$ on GSM8K), while even large-scale models like Llama2-70B see steady improvements (e.g., 83.8% AUROC$_r$ on SQuAD). Overall, HALLUGUARD benefits most on small backbones while maintaining consistent advantages across scales.

Q3: How does each term capture trends across task families? As shown in Figure 2, each term faithfully tracks the ground-truth trend within its respective task family. On data-centric SQuAD, the *data-driven term* closely follows the dashed gold curve across the variant hallucination rate, capturing the smooth AUROC decline. On reasoning-oriented MATH-500, the *reasoning-driven term* mirrors the monotonic AUROC drop as reasoning drift increases. These results show that each term is well matched to its task family and faithfully tracks performance trends as hallucination rates rise.

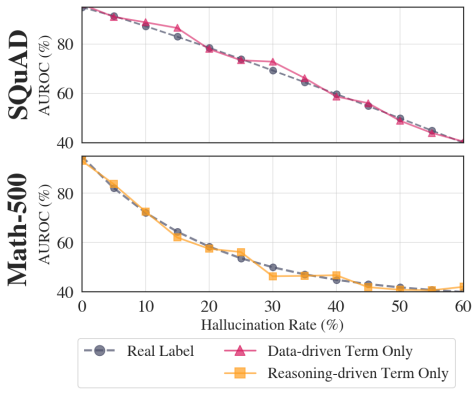

Figure 2: Ablation results comparing individual terms with ground-truth trends on SQuAD (top) and Math-500 (bottom).

Table 3: Performance comparison across backbone scales (small, mid-sized, and large) on three benchmarks: SQuAD, GSM8K, HaluEval. We highlight the **first** and second best results.

		Llama2-7B				Llama-3-8B				Llama2-13B				Llama2-70B			
		$AUROC_r$	$AUPRC_r$	$AUROC_{lm}$	$AUPRC_{lm}$	$AUROC_r$	$AUPRC_r$	$AUROC_{lm}$	$AUPRC_{lm}$	$AUROC_r$	$AUPRC_r$	$AUROC_{lm}$	$AUPRC_{lm}$	$AUROC_r$	$AUPRC_r$	$AUROC_{lm}$	$AUPRC_{lm}$
SQuAD	HALLUGUARD	**81.05**	**77.16**	**71.18**	**64.38**	**79.56**	**78.29**	**67.97**	**63.27**	**81.45**	**78.39**	**64.39**	**65.07**	**83.8**	**81.77**	**70.46**	**73.24**
	Inside	73.63	75.74	65.22	59.11	76.13	72.44	65.62	62.94	74.68	74.81	61.01	59.51	81.24	75.09	69.48	62.4
	MIND	64.57	61.11	52.39	53.13	62.29	59.58	44.49	48.61	68.64	66.95	54.92	52.49	73.46	71.71	57.76	56.77
	Perplexity	63.93	61.77	46.97	48.2	70.51	67.51	55.71	52,68	70.19	69.22	60.33	54.82	74.23	70.88	62.24	58.05
	LN-Entropy	65.96	64.22	53.43	52.84	63.7	60.4	46.19	42.85	61.66	59.16	49.05	46.27	72.44	68.91	56.77	52.63
	Energy	59.83	56.11	46.19	43.18	64.41	61.02	56.17	46.21	61.02	59.73	48.26	42.08	69.01	66.19	58.44	49.82
	Semantic Ent.	60.29	57.73	43.63	48.83	66.52	62.62	52.37	52.7	70.58	67.22	53.31	52.94	72.01	68.51	56.49	50.9
	Lexical Sim.	70.31	69.08	53.97	53.31	66.43	63.56	53.19	50.96	68.53	67.42	50.73	54.12	68.95	67.91	60.52	56.56
	SelfCheckGPT	68.26	67.09	60.06	57.31	73.99	72.15	65.26	54.02	65.47	61.65	53.12	49.89	73.07	70.49	56.59	54.65
	RACE	71.35	69.23	59.18	54.73	68.17	66.02	54.65	53.06	64.19	60.45	47.53	45.66	64.05	62.39	54.38	50.07
	P(true)	62.55	61.09	46.84	52.32	67.42	63.94	55.35	47.52	71.56	68.4	57.51	45.66	66.81	62.71	57.43	46.85
	FActScore	70.32	68.63	58.13	53.01	71.2	69.45	61.92	54.91	66.65	63.2	56.41	53.42	68.33	65.26	56.93	48.46
GSM8K	HALLUGUARD	**75.89**	**72.83**	**62.29**	**63.46**	**75.2**	**72.9**	**63.62**	**61.79**	**79.01**	**76.73**	**64.38**	**64.97**	**77.33**	**73.97**	**60.48**	**61.26**
	Inside	74.61	68.35	58.57	62.58	73.73	67.51	56.02	57.28	75.79	76.26	60.91	59.77	72.3	72.26	54.49	58.39
	MIND	65.88	63.4	48.28	48.17	66.57	65.55	48.84	53.4	61.49	59.55	51.63	51.45	66.41	63.44	52.05	53.57
	Perplexity	66.23	64.1	53.52	52.31	57.61	53.63	41.37	41.59	60.96	58.67	46.27	47.44	64.32	62.81	51.15	51.3
	LN-Entropy	59.45	55.95	43.04	44.08	68.22	66.05	53.03	53.21	61.31	58.90	45.83	40.86	61.81	60.46	44.5	44.76
	Energy	58.15	54.71	43.65	36.71	59.79	56.52	50.31	42.23	57.58	56.07	43.39	38.94	65.27	62.94	52.8	46.6
	Semantic Ent.	57.95	54.68	42.78	41.95	66.9	64.81	50.47	55.36	62.72	59.09	49.33	44.35	60.63	57.01	46.22	40.24
	Lexical Sim.	65.8	63.7	52.12	54.07	63.29	59.87	53.17	50.02	63.83	60.20	54.43	44.82	63.27	59.41	47.42	47.38
	SelfCheckGPT	60.99	57.54	49.28	44.43	65.72	62.01	54.49	50.34	57.98	54.58	46.72	39.86	68.06	65.09	52.99	50.89
	RACE	63.37	62.33	53.53	49.94	64.49	61.47	53.28	47.55	64.20	61.96	50.15	45.35	68.35	66.66	50.41	51.16
	P(true)	65.95	63.63	54.95	48.25	62.59	58.88	47.21	42.2	67.08	65.60	53.66	55.12	60.16	58.14	47.73	49.49
	FActScore	56.69	53.71	45.78	39.52	65.69	61.95	53.69	46.06	55.76	54.17	44.91	43.18	59.84	55.85	44.05	39.49
HaluEval	HALLUGUARD	**75.72**	**72.89**	**66.65**	**63.15**	**73.43**	**71.19**	**64.95**	**54.8**	**78.15**	**74.15**	**65.39**	**61.14**	**80.79**	**79.54**	**67.68**	**68.51**
	Inside	71.33	67.63	59.73	53.15	67.95	64.93	60.31	52.21	72.01	71.97	56.51	60.64	74.62	68.33	62.22	64.4
	MIND	54.8	51.43	44.15	43.34	64.54	60.89	49.09	45.13	55.05	53.28	39.16	45.17	57.98	56.01	45.82	41.69
	Perplexity	54.02	52.53	38.76	40.51	61.31	59.36	50.62	46.01	54.99	51.39	42.64	35.64	62.85	60.59	48.29	43.85
	LN-Entropy	59.47	58.33	50.2	46.91	64.89	60.72	51.78	46.39	65.18	63.53	49.70	48.09	60.16	58.89	50.29	48.42
	Energy	62.29	59.6	50.68	42.24	62.74	61.61	50.17	52.01	60.54	59.04	43.53	50.37	60.13	58.44	48.79	48.01
	Semantic Ent.	59.39	55.94	48.53	46.35	55.25	53.05	44.5	44.35	59.44	57.72	45.38	40.77	61.57	57.99	49.07	45.39
	Lexical Sim.	63.61	61.16	55.01	44.75	56.59	55.39	44.45	45.57	53.46	52.06	41.34	40.57	64.37	60.92	54.29	50.86
	SelfCheckGPT	64.29	61.83	48.4	45.49	65.44	63.13	57.02	48.23	65.24	63.52	53.71	54.33	57.12	55.26	40.5	43.06
	RACE	59.78	59.14	48.1	40.47	61.98	60.32	48.08	46.29	60.65	59.11	49.92	44.51	62.11	58.24	40.5	43.06
	P(true)	57.46	54.8	41.84	40.47	56.32	54.04	42.55	43.75	65.77	63.01	49.98	45.47	55.75	54.94	44.14	43.97
	FActScore	63.93	61.33	46.9	51.87	61.73	57.85	49.92	42.15	65.15	63.71	55.98	54.61	62.66	60.3	53.13	46.42

4.3 TEST-TIME INFERENCE

Test-time reasoning remains challenging, as models need to generate coherent multi-step solutions without drifting into errors. To assess whether hallucination detection can mitigate this difficulty, we integrate detectors into beam search and evaluate Qwen2.5-Math-7B on MATH-500 and Llama3.1-8B on Natural. As shown in Table 4, HALLUGUARD achieves the strongest gains: on MATH-500, it reaches 81.00% accuracy, around 10% higher than IO Prompt; on Natural, it attains 70.96%, exceeding IO Prompt by 15.72%. These results demonstrate that HALLUGUARD not only detects hallucinations but also strengthens test-time reasoning by guiding models toward more reliable solutions.

Table 4: Performance of hallucination score-guided test-time inference across reasoning tasks. We highlight the **first** and second best results.

Dataset	IO Prompt	HALLUGUARD	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

4.4 CASE STUDY

Fine-grained hallucinations—lexically similar yet semantically incorrect outputs—pose a particular challenge for detection. To evaluate whether HALLUGUARD can comprehensively capture such

subtle errors, we use the PAWS dataset(Zhang et al., 2019), which contrasts paraphrases with high surface overlap but divergent meanings. Following Li et al. (2025), we adopt ROUGE-based reference signals for evaluation (Table 5). Across model scales, HALLUGUARD consistently surpasses baselines: it achieves 90.18% AUROC and 87.64% AUPRC on Llama2-70B, and 91.24% AUROC and 88.53% AUPRC on QwQ-32B—exceeding the next-best method by nearly five points. Even on GPT-2, it leads with 83.27% AUROC and 80.46% AUPRC. These results confirm HALLUGUARD's effectiveness in capturing fine-grained semantic inconsistencies beyond benchmark settings.

Table 5: Results on PAWS measuring semantic hallucination detection with Llama-3.2-3B, Llama2-70B, and QwQ-32B. We highlight the **first** and second best results.

	Method	HALLUGUARD	Inside	MIND	Perplexity	LN-Entropy	Energy	Semantic Ent.	Lexical Sim.	SelfCheck-GPT	RACE	P(true)	FActScore
Llama3.2	AUROC	**85.63**	80.46	78.93	71.27	72.19	73.05	75.11	64.58	77.82	79.47	73.56	68.44
	AUPRC	**82.14**	77.28	75.41	67.55	68.34	70.22	72.41	59.67	73.41	76.28	70.43	63.58
Llama2	AUROC	**90.18**	85.47	83.92	75.68	76.23	77.14	79.06	68.35	82.71	84.26	77.39	72.62
	AUPRC	**87.64**	82.38	81.06	71.42	72.59	74.28	76.32	63.44	78.89	81.73	74.18	67.58
QwQ	AUROC	**91.24**	85.41	84.56	76.72	77.43	78.29	80.42	69.54	83.59	86.38	78.53	73.46
	AUPRC	**88.53**	82.27	81.37	72.63	73.29	75.44	77.18	64.27	79.42	83.41	75.21	68.32

5 RELATED WORK

In this section, we review prior hallucination-detection methods by their detection target–*Data-driven hallucinations* and *reasoning-driven hallucinations*.

Detecting Data-Driven Hallucinations. Recent work has shown that internal activations encode rich indicators of such flaws. Zhou & et al. (2024) proposed EIGENSCORE, which computes statistics of hidden representations from the eigen matrix to estimate hallucination risk. Su et al. (2024) introduced MIND, an unsupervised detector that models temporal dynamics of hidden states without requiring labels, along with HELM benchmark to enable standardized evaluation. Azaria & Mitchell (2023) demonstrated using linear probes on intermediate states to predict truthfulness.

Detecting Reasoning-Driven Hallucinations. There are other works targeting inference-time inconsistencies during generation—such as logical errors, instability across decoding steps, or temporal drift in extended outputs. Manakul et al. (2023) proposed SELFCHECKGPT, which assesses self-consistency by sampling multiple candidate generations and measuring their alignment using entailment and lexical overlap. Wu et al. (2024) introduced a suite of calibration-based uncertainty scores designed to capture hallucination risk directly from output distributions. Zheng et al. (2024) proposed REACTSCORE, which integrates entropy with intermediate reasoning traces to detect failures in multi-step decision-making. FACTSCORE(Min et al., 2023a) decomposes outputs into atomic factual units and verifies each against retrieved passages using entailment-based scoring.

6 CONCLUSION

The reliability of LLMs is often undermined by hallucinations, which arise from two main sources: *data-driven*, caused by flawed knowledge acquired during training, and *reasoning-driven*, stemming from inference-time instabilities in multi-step generation. Although these hallucinations frequently evolve in practice, existing detectors usually target only one source and lack a solid theoretical foundation. To address this gap, we propose a unified theoretical framework–a *Hallucination Risk Bound*, which formally decomposes hallucination risk into data-driven and reasoning-driven components, offering a principled view of how hallucinations emerge and evolve during generation. Building on this foundation, we introduce **HALLUGUARD**, a NTK–based score that measures sensitivity to semantic perturbations and captures internal instabilities, thereby enabling holistic detection of both data-driven and reasoning-driven hallucinations. We evaluate HALLUGUARD across 10 diverse benchmarks, 11 competitive baselines, and 9 popular LLM backbones, where it consistently achieves state-of-the-art performance, demonstrating robustness and practical efficacy.

Reproducibility Statement

We have taken several measures to ensure the reproducibility of our work. A complete description of the theoretical framework, including the formal assumptions and proofs of the Hallucination Risk Bound, is provided in Section 3 and Appendix A. Detailed experimental settings and evaluation protocols are documented in Section 4 and Appendix B.6, covering all 10 benchmarks, 11 baselines, and 9 LLM backbones. Together, these resources ensure that both our theoretical claims and empirical results can be independently validated and extended by the community.

Ethics Statement

This study is based exclusively on publicly available datasets and open-source large language models, and does not involve human subjects or the use of private data. All scientific concepts, methodological designs, experimental implementations, and resulting conclusions remain entirely the responsibility of the authors.

References

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Ece Kamar, Michal Kosinski, Ryan Chi-Ying Hsieh, Drew A. Linsley, Long O. Mai, Nikolay Manchev, Christopher D. Manning, Yian Yin, Christopher J. N. de M. L. Matthews, Lucia Mondragon, Ognjen Oreskovic, Mark Sabini, Yusuf Sahin, Clark Barrett, Christopher Potts, James Y. Zou, Jiajun Wu, and Percy Liang. On the opportunities and risks of foundation models, 2021.

Jean Céa. Approximation variationnelle des problèmes aux limites. In *Annales de l'institut Fourier*, volume 14, pp. 345–444, 1964.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024a. URL https://arxiv.org/abs/2402.03744.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models, 2024b. URL https://arxiv.org/abs/2407.04121.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ digital medicine*, 8(1):143, 2025.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,

Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Lei Huang, Weijiang Yu, Weitao Wang, Yujia Wang, Shi-Qi Chen, and Ju-Hua Wang. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Peizhong Ju, Xiaojun Lin, and Ness B. Shroff. On the generalization power of the overfitted three-layer neural tangent kernel model, 2022. URL https://arxiv.org/abs/2206.02047.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Markus Kattnig, Alessa Angerschmid, Thomas Reichel, and Roman Kern. Assessing trustworthy ai: Technical and legal perspectives of fairness in ai. *Computer Law & Security Review*, 55:106053, 2024.

Kalpesh Krishna and et al. Can retrieval-augmented language models hallucinate less? In *EMNLP*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent *. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12): 124002, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b. URL http://dx.doi.org/10.1088/1742-5468/abc62b.

Jiawei Li, Akshayaa Magesh, and Venugopal V. Veeravalli. Principled detection of hallucinations in large language models via multiple testing, 2025. URL https://arxiv.org/abs/2508.18473.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. URL https://arxiv.org/abs/2305.11747.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022a. URL https://arxiv.org/abs/2109.07958.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4160–4173, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.328. URL https://aclanthology.org/2022.findings-acl.328/.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025. URL https://arxiv.org/abs/2505.21523.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2020.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2020.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023a. URL https://arxiv.org/abs/2305.14251.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023b.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models, 2024. URL https://arxiv.org/abs/2401.00396.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models, 2022.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models, 2024. URL https://arxiv.org/abs/2403.06448.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective, 2025. URL https://arxiv.org/abs/2505.12886.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kavya Elangovan, Lio Gutierrez, Teng Fong Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Lloyd N Trefethen and David Bau. *Numerical linear algebra*. SIAM, 2022.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models, 2025.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022. URL https://arxiv.org/abs/2204.07705.

Zeyu Wei, Shuo Wang, Xiaohui Rong, Xuemin Liu, and He Li. Shadows in the attention: Contextual perturbation and representation drift in the dynamics of hallucination in llms, 2025. URL https://arxiv.org/abs/2505.16894.

X. Wu, T. Lei, and Z. Hu. Hallusion: Calibrating hallucination prediction in large language models. *arXiv preprint arXiv:2505.03793*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Xinyue Zeng, Haohui Wang, Junhong Lin, Jun Wu, Tyler Cody, and Dawei Zhou. Lensllm: Unveiling fine-tuning dynamics for llm selection. *ICML*, 2025. arXiv preprint arXiv:2505.03793.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling, 2019. URL https://arxiv.org/abs/1904.01130.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms, 2025. URL https://arxiv.org/abs/2507.16488.

L. Zheng et al. Reactscore: Unified reasoning and factuality tracing for hallucination detection. *arXiv preprint arXiv:2505.16122*, 2024.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models, 2024.

D. Zhou and et al. Inside: Interpretable self-diagnosis for llm hallucination detection. *ICML*, 2024.

A Proof of Hallucination Risk Bound

We restate the main inequality from Section 3.2:

$$\|u^*-u_h\| \leq \left[1 + k_{\text{pt}} \log \mathcal{O}(P, L) + k\frac{\epsilon_{\text{mismatch}}}{\text{Signal}_k}\right] \inf_{u \in U_h} \|u^*-u\| + |\mathcal{L}| \exp\left(-\frac{K\epsilon^2}{C}\right) \alpha\left(e^{\beta T} - 1\right). \tag{8}$$

Step 1: Triangle inequality split. We define the hallucination decomposition by writing:

$$\|u^* - u_h\| = \|u^* - \mathbb{E}[u_h] + \mathbb{E}[u_h] - u_h\| \leq \|u^* - \mathbb{E}[u_h]\| + \|u_h - \mathbb{E}[u_h]\|.$$

We denote the first term as the deterministic approximation error (bias) and the second term as the stochastic residual (variance).

Step 2: Approximation term via Céa's lemma. Assume $\mathbb{E}[u_h]$ is the Galerkin projection of u^* in a coercive bilinear form $a(\cdot, \cdot)$, i.e., for all $v \in U_h$,

$$a(\mathbb{E}[u_h], v) = \ell(v).$$

Then, by Céa's lemma, we have:

$$\|u^* - \mathbb{E}[u_h]\| \leq \frac{\Lambda}{\gamma} \inf_{u \in U_h} \|u^* - u\|,$$

where Λ and γ are continuity and coercivity constants of $a(\cdot, \cdot)$, respectively.

Step 3: Variance term via Bernstein concentration. Let $\ell_h := \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} \ell_i$ be the empirical supervision functional from finite labeled chains. Define the fluctuation:

$$\Delta\ell := \ell_h - \ell,$$

and the residual:

$$r := u_h - \mathbb{E}[u_h], \quad \text{so that} \quad A_h r = \Delta\ell.$$

Applying operator norm bounds and covering number uniformization (cf. Vershynin, 2018), we have with high probability:

$$\|r\| \leq |\mathcal{L}| \exp\left(-\frac{K\epsilon^2}{C}\right) \alpha(e^{\beta T} - 1),$$

which completes the proof.

Step 4: Substitution. Combining both terms yields:

$$\|u^* - u_h\| \leq \frac{\Lambda}{\gamma} \inf_{u \in U_h} \|u^* - u\| + |\mathcal{L}| \exp\left(-\frac{K\epsilon^2}{C}\right) \alpha(e^{\beta T} - 1).$$

We now bound Λ/γ via NTK decomposition.

A.1 Decomposition of NTK Continuity Constant

Let $a(\cdot, \cdot)$ denote the bilinear form induced by the NTK in the finite-width regime. We decompose:

$$a = a_0 + \delta_{\text{pt}} + \delta_{\text{mm}},$$

where a_0 is the infinite-width baseline kernel, δ_{pt} is the perturbation due to pre-training noise, and δ_{mm} is the domain mismatch from fine-tuning. The continuity constant satisfies:

$$\Lambda = \Lambda_0 + \Delta_{\text{pt}} + \Delta_{\text{mm}}.$$

Bounding Δ_{pt}. Following Jacot et al. (2020), we apply matrix concentration to finite-width NTK:

$$\Delta_{\text{pt}} \leq \gamma k_{\text{pt}} \log \mathcal{O}(P, L).$$

Bounding Δ_{mm}. Using spectral generalization bounds under data distribution shift (Lee et al., 2020), we have:

$$\Delta_{\mathrm{mm}} \leq \gamma k \frac{\epsilon_{\mathrm{mismatch}}}{\mathrm{Signal}_k}.$$

Substituting both into the bound for Λ/γ, we get:

$$\frac{\Lambda}{\gamma} \leq 1 + k_{\mathrm{pt}} \log \mathcal{O}(P, L) + k \frac{\epsilon_{\mathrm{mismatch}}}{\mathrm{Signal}_k}.$$

B HALLUGUARDDERIVATION AND INTERPRETATION

B.1 PRELIMINARIES AND NOTATION

Let $\mathcal{K} \in \mathbb{R}^{r\times r}$ be the NTK Gram matrix formed on r light semantic perturbations (see Assumptions A1–A4 in the main theory section). Denote its eigen decomposition by $\mathcal{K} = V\Lambda V^\top$ with

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \qquad \lambda_1 \geq \cdots \geq \lambda_r > 0.$$

Let $\lambda_{\max} := \lambda_1$, $\lambda_{\min} := \lambda_r$, $\kappa(\mathcal{K}) := \lambda_{\max}/\lambda_{\min}$, and $\det(\mathcal{K}) = \prod_{i=1}^r \lambda_i$. Let Φ denote the NTK feature matrix whose columns span the hypothesis subspace U_h, so that $\mathcal{K} = \Phi^\top\Phi$, $\|\Phi\|_2 = \sqrt{\lambda_{\max}}$, and $\sigma_{\min}(\Phi) = \sqrt{\lambda_{\min}}$. For the autoregressive decoder, let J_t be the step-t input–output Jacobian, and write $\sigma_{\max} := \sup_t \|J_t\|_2$.

We will use the following two standard inequalities repeatedly:

$$Maclaurin/AM--GMoneigenvalues: \qquad \Big(\prod_{i=1}^r \lambda_i\Big)^{1/r} \;\leq\; \frac{1}{r}\sum_{i=1}^r \lambda_i \;=\; \frac{\mathrm{tr}(\mathcal{K})}{r}, \tag{9}$$

$$Submultiplicativity: \qquad \|AB\|_2 \;\leq\; \|A\|_2\,\|B\|_2. \tag{10}$$

B.2 REPRESENTATIONAL ADEQUACY VIA $\det(\mathcal{K})$ WITH EXPLICIT CONSTANTS

Assumptions for this subsection. Beyond A1–A3, we assume a mild *source condition* and a *spectral envelope*:

- **S1** (*Source condition*) There exist $s > 0$ and $R_s > 0$ such that $u^* \in \mathrm{Range}(\Lambda^s)$, i.e., $\sum_{i=1}^r \frac{\langle u^*, v_i\rangle^2}{\lambda_i^{2s}} \leq R_s^2$. This is standard in kernel approximation and encodes RKHS regularity.
- **S2** (*Spectral envelope*) There exist constants $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ and $\alpha > 1$ such that $\lambda_i \leq \overline{\lambda}$ for all i and $\lambda_r \geq \underline{\lambda}\, r^{-\alpha}$. (Polynomial decay is a common stylization; other envelopes can be treated similarly.)

Lemma B.1 (Best-approximation error under source condition)**.** *Let $U_h = \mathrm{span}\{v_1, \ldots, v_r\}$. Under **S1**,*

$$\inf_{u\in U_h} \|u^* - u\| \;=\; \|u^* - \Pi_{U_h} u^*\| \;\leq\; R_s\, \lambda_{r+1}^s,$$

where λ_{r+1} denotes the next-eigenvalue of the infinite-dimensional kernel operator (or, equivalently, the empirical tail eigenvalue if more perturbations are added).

Proof. Write $u^* = \sum_{i\geq 1} c_i v_i$ with $c_i = \langle u^*, v_i\rangle$. Then $\|u^* - \Pi_{U_h} u^*\|^2 = \sum_{i>r} c_i^2 \leq \sum_{i>r} \lambda_i^{2s} \cdot \frac{c_i^2}{\lambda_i^{2s}} \leq \lambda_{r+1}^{2s} \sum_{i>r} \frac{c_i^2}{\lambda_i^{2s}} \leq \lambda_{r+1}^{2s} R_s^2$. □

To connect λ_{r+1} (or λ_r) to $\det(\mathcal{K})$, we need an explicit lower bound of the form $\lambda_r \geq \underline{c}\,\det(\mathcal{K})^\theta$ with constants $(\underline{c}, \theta)$ depending on the spectral envelope. The following inequality suffices.

Lemma B.2 (Lower-bounding λ_r by $\det(\mathcal{K})$)**.** *Suppose $\lambda_i \leq \overline{\lambda}$ for all i and $\lambda_r > 0$. Then*

$$\lambda_r \;\geq\; \frac{\det(\mathcal{K})}{\overline{\lambda}^{\,r-1}} \qquad \textit{and} \qquad \lambda_r^s \;\geq\; \frac{\det(\mathcal{K})^{\,s}}{\overline{\lambda}^{\,s(r-1)}}.$$

Proof. Since $\det(\mathcal{K}) = \prod_{i=1}^{r} \lambda_i \le \overline{\lambda}^{r-1}\lambda_r$, we obtain $\lambda_r \ge \det(\mathcal{K})/\overline{\lambda}^{r-1}$. Raising to power s yields the second inequality. □

Theorem B.3 (Determinant-based adequacy bound with explicit constants)**.** *Under A1–A3 and **S1–S2**,*

$$\inf_{u \in U_h} \|u^* - u\| \;\le\; C_d \, \det(\mathcal{K})^{-\,c_d} \, \|u^*\|,$$

with

$$c_d \;=\; \frac{s}{r-1} \quad \textit{and} \quad C_d \;=\; \overline{\lambda}^{\,s} \, \frac{R_s}{\|u^*\|}.$$

Moreover, if the empirical spectrum satisfies $\lambda_r \ge \underline{\lambda}\, r^{-\alpha}$, one may choose

$$c_d \;=\; \min\left\{ \frac{s}{r-1},\; \frac{s}{\alpha} \cdot \frac{1}{\log\!\left(\frac{\overline{\lambda}^{r}}{\det(\mathcal{K})}\right)} \right\},$$

which improves with slower decay (smaller α).

Proof. By Lemma B.1 with $\lambda_{r+1} \le \lambda_r$, $\inf_{u \in U_h} \|u^* - u\| \le R_s \, \lambda_r^s$. Lemma B.2 gives $\lambda_r^s \ge \det(\mathcal{K})^{\,s}/\overline{\lambda}^{\,s(r-1)}$; rearranging,

$$\inf_{u \in U_h} \|u^* - u\| \;\le\; R_s \, \overline{\lambda}^{\,s(r-1)} \, \det(\mathcal{K})^{-\,s}.$$

Rescale constants relative to $\|u^*\|$ by setting $C_d := \overline{\lambda}^{\,s} \, (R_s/\|u^*\|)$ and $c_d := s/(r-1)$ to obtain the stated form:

$$\inf_{u \in U_h} \|u^* - u\| \;\le\; \left(\overline{\lambda}^{\,s} \, \tfrac{R_s}{\|u^*\|}\right) \, \det(\mathcal{K})^{-\,s/(r-1)} \, \|u^*\|.$$

The variant using the envelope $\lambda_r \ge \underline{\lambda}\, r^{-\alpha}$ is obtained by combining $\det(\mathcal{K}) \le \overline{\lambda}^{\,r-1}\lambda_r$ with the explicit lower bound on λ_r, yielding the alternative exponent shown. □

Numerical note (stable surrogate). In practice we use $\log\det(\mathcal{K})$ via Cholesky and aggregate with z-normalization across components to avoid scale domination by any single term.

B.3 Rollout Amplification via Jacobian Products (Exact Constants)

Theorem B.4 (Amplification bound with exact constant)**.** *Let J_t be the step-t Jacobian and $\sigma_{\max} := \sup_t \|J_t\|_2$. Then*

$$\Big\| \prod_{t=1}^{T} J_t \Big\|_2 \;\le\; \prod_{t=1}^{T} \|J_t\|_2 \;\le\; \sigma_{\max}^{T}.$$

Defining $\beta := \log \sigma_{\max}$ gives $e^{\beta T} = \sigma_{\max}^{T}$, hence

$$e^{\beta T} \;\le\; \sigma_{\max}^{T},$$

with equality if and only if $\|J_t\|_2 = \sigma_{\max}$ for all t and the top singular directions align across factors.

Proof. The first inequality is equation 10 applied iteratively. The second is by definition of $\sigma_{\max}$. Setting $\beta = \log \sigma_{\max}$ yields equality in the worst case. Alignment of top singular vectors is the tightness condition for submultiplicativity. □

Token-dependent refinement. If one defines $\sigma_t := \|J_t\|_2$ and $\beta_{\text{avg}} := \frac{1}{T}\sum_{t=1}^{T} \log \sigma_t$, then $\big\| \prod_{t=1}^{T} J_t \big\|_2 \le \exp\big(\sum_t \log \sigma_t\big) = e^{\beta_{\text{avg}} T}$, which is tighter but requires per-step measurements.

B.4 Conditioning-Induced Variance with $\kappa(\mathcal{K})^2$ Scaling

We now give an explicit projector-perturbation derivation showing the quadratic dependence on the condition number.

Setup. Let $P := \Phi(\Phi^\top\Phi)^\dagger\Phi^\top$ be the orthogonal projector onto U_h; then the linearized output is $u_h = Pu^*$. Consider a feature perturbation $\Delta\Phi$ induced by a prefix perturbation δ satisfying

$$\|\Delta\Phi\|_2 \;\le\; L_\Phi\,\|\delta\| \quad \text{(A2/A3)}.$$

Let the perturbed projector be $\widetilde{P} := (\Phi+\Delta\Phi)\big((\Phi+\Delta\Phi)^\top(\Phi+\Delta\Phi)\big)^\dagger(\Phi+\Delta\Phi)^\top$ and define $\Delta P := \widetilde{P} - P$.

Lemma B.5 (Projector perturbation bound)**.** *There exists an absolute constant $C_\Pi > 0$ such that*

$$\|\Delta P\|_2 \;\le\; C_\Pi\,\frac{\|\Phi\|_2}{\sigma_{\min}(\Phi)^2}\,\|\Delta\Phi\|_2 \;=\; C_\Pi\,\frac{\sqrt{\lambda_{\max}}}{\lambda_{\min}}\,\|\Delta\Phi\|_2 \;=\; C_\Pi\,\kappa(\mathcal{K})\,\frac{\|\Delta\Phi\|_2}{\sqrt{\lambda_{\min}}}.$$

Proof idea. Use standard bounds for the perturbation of orthogonal projectors onto column spaces (e.g., Wedin's sinΘ theorem and Stewart–Sun, Matrix Perturbation Theory, Thm 3.6). One shows

$$\|\Delta P\|_2 \;\le\; 2\,\|(\Phi^\top\Phi)^\dagger\|_2\,\|\Phi^\top\Delta\Phi\|_2 \;+\; \mathcal{O}(\|\Delta\Phi\|_2^2).$$

Since $\|(\Phi^\top\Phi)^\dagger\|_2 = 1/\lambda_{\min}$ and $\|\Phi^\top\Delta\Phi\|_2 \le \|\Phi\|_2\,\|\Delta\Phi\|_2 = \sqrt{\lambda_{\max}}\|\Delta\Phi\|_2$, the result follows for sufficiently small $\|\Delta\Phi\|_2$, absorbing lower-order terms into C_Π. □

Theorem B.6 (Variance amplification with explicit constant)**.** *Let $u_h(\Phi) = Pu^*$ and $u_h(\Phi+\Delta\Phi) = \widetilde{P}u^*$. Then*

$$\|u_h(\Phi+\Delta\Phi) - u_h(\Phi)\| \;\le\; C_\Pi\,\kappa(\mathcal{K})\,\frac{\|\Delta\Phi\|_2}{\sqrt{\lambda_{\min}}}\,\|u^*\|.$$

If $\Delta\Phi$ is induced by a random prefix perturbation δ with $\|\Delta\Phi\|_2 \le L_\Phi\|\delta\|$ and $\mathbb{E}\|\delta\|^2 = \sigma_\delta^2$, then

$$\mathrm{Var}[u_h] \;\le\; \mathbb{E}\|u_h(\Phi+\Delta\Phi) - u_h(\Phi)\|^2 \;\le\; c_v\,\kappa(\mathcal{K})^2\,\|\delta\|^2,$$

with

$$c_v \;=\; C_\Pi^2\,\frac{L_\Phi^2\,\|u^*\|^2}{\lambda_{\min}}.$$

Proof. By Lemma B.5, $\|u_h(\Phi+\Delta\Phi) - u_h(\Phi)\| = \|\Delta P\,u^*\| \le \|\Delta P\|_2\|u^*\| \le C_\Pi\,\kappa(\mathcal{K})\,\frac{\|\Delta\Phi\|_2}{\sqrt{\lambda_{\min}}}\,\|u^*\|$. Square both sides and take expectation over δ, using $\|\Delta\Phi\|_2 \le L_\Phi\|\delta\|$, to obtain the stated variance bound with the explicit constant c_v. □

Interpretation. The $\kappa(\mathcal{K})^2$ factor arises from two sources: (i) $\kappa(\mathcal{K})$ from the projector sensitivity (Lemma B.5), and (ii) $1/\lambda_{\min}$ from converting $\|\Delta P\|_2$ to a mean-squared bound after squaring and averaging, yielding an overall κ^2-scaling in the variance constant.

B.5 Consolidation: Compact Surrogate Consistent with the Risk Decomposition

Combining Theorem B.3, Theorem B.4, and Theorem B.6, we obtain a computable surrogate aligned with the Hallucination Risk Bound:

Adequacy: $\det(\mathcal{K})$ Amplification: $\log\sigma_{\max}$ Conditioning penalty: $-\log\kappa(\mathcal{K})^2$.

This motivates the score

$$\boxed{\textbf{HALLUGUARD}(u_h) \;=\; \det(\mathcal{K}) \;+\; \log\sigma_{\max} \;-\; \log\kappa(\mathcal{K})^2}$$

with the following explicit, implementation-ready notes:

- Use $\log\det(\mathcal{K})$ via Cholesky for stability; replace $\det$ in the score with $\log\det$ if desired (monotone equivalent).
- Estimate $\sigma_{\max}$ either as $\sup_t\|J_t\|_2$ or its tighter average form $\beta_{\text{avg}} = \frac{1}{T}\sum_t \log\|J_t\|_2$ (then use β_{avg} in place of $\log\sigma_{\max}$).
- z-normalize each component across a validation set before summation to avoid scale dominance; optionally fit task-specific weights if permitted.

B.6 EXPERIMENT DETAILS

Implementation Framework. All experiments use `PyTorch` and `HuggingFace Transformers` with a fixed random seed for reproducibility. Unless otherwise noted, computations run in mixed precision (fp16). Hardware details (A100/H200) are reported once in the main setup section.

Generation Configuration. For *default evaluation of detectors*, we use nucleus sampling with `temperature` $= 0.5$, `top-p` $= 0.95$, and `top-k` $= 10$, decoding $K{=}10$ candidate responses per input (unless otherwise specified). For *score-guided test-time inference* (Section 4.3), we use beam search (beam size $= 10$) and score candidate trajectories at each step with the chosen detector. For stability analysis, HALLUGUARD extracts sentence representations from the final token at the middle transformer layer ($L/2$), which empirically preserves semantics relevant to truthfulness.

NTK-Based Score Computation. For each set of generations, we form a task-specific NTK feature matrix and compute the semantic stability score from its eigenspectrum. We add a small ridge $\alpha = 10^{-3}$ for numerical stability and compute singular values via SVD.

Perturbation Regularization. To prevent pathological activations that amplify instability, HALLUGUARD clips hidden features using an adaptive scheme. We maintain a memory bank of $N{=}3000$ token embeddings and set thresholds at the top and bottom 0.2% percentiles of neuron activations; out-of-range values are truncated to attenuate overconfident hallucinations.

Optimization. Backbone language models are *not* fine-tuned. We train only HALLUGUARD's lightweight projection layers using AdamW with learning rate selected from $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ and weight decay from $\{0.0, 0.01\}$. The best setting is chosen on a held-out validation split.

Implementation Details. For score-guided inference we apply beam search with beam size 10, rescoring candidates stepwise with different hallucination detectors.

Ablation Setup. All ablations reuse the main paper's splits, prompts, and decoding; we vary only HALLUGUARD internals and explicitly control the hallucination *base rate*. On the *generation* side, we modulate prevalence by adjusting temperature/top-p and beam size; to stress the two families, we increase the prefix perturbation budget ρ and rollout horizon T to amplify reasoning drift, and (when applicable) toggle retrieval masking to induce data-driven errors. On the *detection* side, AUROC/AUPRC are threshold-free; when a fixed operating point is needed, we set a decision threshold τ on the validation set by (i) matching a target predicted-positive rate π_{target} via score quantiles or (ii) fixing a desired FPR (e.g., $1\%, 5\%, 10\%$); a cost-sensitive Bayes rule $\tau = \dfrac{c_{\text{FN}}}{c_{\text{FP}} + c_{\text{FN}}} \cdot \dfrac{1-\pi}{\pi}$ is optional when misclassification costs are specified. Unless noted, we toggle one factor at a time and sweep $\rho \in \{0.75, 1.0, 1.5\}$, $T \in \{12, 16, 24\}$, and the number of semantic probes $m \in \{2, 4, 8\}$; no additional training is performed beyond optional temperature/z-score calibration on the training split. We report mean$\pm$std over 5 seeds.

B.7 USAGE OF LLM

Large language models (LLMs) were employed in a limited and transparent manner during the preparation of this manuscript. Specifically, LLMs were used to assist with linguistic refinement, style adjustments, and minor text editing to improve clarity and readability. They were not involved in formulating the research questions, designing the theoretical framework, conducting experiments, or interpreting results. All scientific contributions—including conceptual development, methodology, analyses, and conclusions—are the sole responsibility of the authors.