Beyond Output Confidence: Epistemic-Aware Hallucination Detection with Answer-Level Signals

Jieran Li*, **Xiuyuan Hu***, **Yang Zhao, Dongbiao Sun, Hao Zhang**†
Department of Electronic Engineering, Tsinghua University
{lijr23,huxy22}@mails.tsinghua.edu.cn, haozhang@tsinghua.edu.cn

Abstract

Despite their strong generative capabilities, large language models frequently exhibit hallucinations, particularly due to outside-boundary confidence where incorrect assertions are produced with high statistical certainty. Existing approaches commonly use output probability as a proxy for truthfulness; however, this signal is confounded by epistemic uncertainty and cannot reliably distinguish genuine uncertainty from fabricated content. We argue that effective hallucination detection requires integrating surface-level confidence with signals that reflect the model's underlying epistemic state. To this end, we propose Answer-level Intrinsic Cognition (AIC), a model-agnostic metric that captures epistemic boundary deviations by measuring answer-level stability across multiple stochastic forward passes. By coupling AIC with conventional output uncertainty, we derive a composite metric that disentangles within-boundary uncertainty from outside-boundary confidence. Across three public question-answering benchmarks and diverse model scales, the two-dimensional score consistently outperforms strong uncertainty-only baselines, with larger gains on adversarially constructed hallucination sets. The code is available at: https://github.com/HXYfighter/AIC-ACL2026.

1 Introduction

Large language models (LLMs) achieve human-level fluency on open-domain question answering, dialogue generation, and code completion (Ouyang et al., 2022; Huang et al., 2025; Li et al., 2025; Dubey et al., 2024). This fluency, however, coexists with a systematic tendency to hallucinate: LLMs often produce statements that contradict verifiable facts while maintaining high predictive confidence (Ji et al., 2023; Liu et al., 2025). When hallucination occurs, the token-level probability distribution can exhibit extremely low entropy, causing conventional uncertainty-based detectors to fail on low-entropy hallucinations (Ren et al., 2022; Malinin and Gales, 2020). In safety-critical domains—medical diagnosis, financial analysis, or legal compliance—any divergence between model confidence and factual correctness erodes user trust and may incur significant liability (Yin et al., 2023; Lin et al., 2022). Under the practical constraints of black-box API access, parameter-free deployment, and no external knowledge bases, detecting low-entropy hallucinations using only internal model signals remains a central bottleneck for high-stakes applications.

Existing hallucination detection methods can be broadly grouped into three categories (Niu et al., 2025; Sun et al., 2024). (1) **Output-level uncertainty** quantifies distribution sharpness via perplexity, predictive entropy, or energy scores, assuming that high uncertainty implies high hallucination (Malinin and Gales, 2020; Liu et al., 2020; Kossen et al., 2024). (2) **Internal-state probes** train auxiliary classifiers on hidden representations to identify unfaithful patterns (Azaria and Mitchell, 2023; Chuang et al., 2024). (3) **External verification** retrieves evidence or calls external tools to fact-check generated content (Niu et al., 2024; Cohen et al., 2023). Methods (2) and (3) demand gradient updates, labeled data, or external systems, hindering tuning-free deployment under black-box access. Method (1) is widely adopted in industry for its zero-training advantage, yet it succumbs to confident nonsense: when a question lies outside the model's cognitive boundary, the model may still emit an incorrect answer with low entropy, rendering threshold-based strategies ineffective (Ji et al., 2023; Lin et al., 2023).

We argue that the failure of output-level uncertainty originates from an implicit equivalence between uncertainty and hallucination. This

*These authors contributed equally.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 13796–13806
July 2-7, 2026

equivalence conflates two distinct cognitive states: **A. Within-boundary uncertainty:** the query lies inside the cognitive boundary, but ambiguity or noise yields high entropy; **B. Outside-boundary confidence:** the query lies beyond the boundary, yet the model remains confident and incorrect. One-dimensional indicators cannot disentangle these states, producing false negatives in state B and false positives in state A (Lin et al., 2023; Ren et al., 2022).

To address this limitation, we propose a **two-dimensional, training-free hallucination detector** that augments output uncertainty with **Answer-level Intrinsic Cognition** (AIC). AIC quantifies the stability of the cognitive boundary by aggregating answer-level consistency across multiple stochastic decodings of the same question, without supervision or gradients (Chen et al., 2024; Lin et al., 2023; Kossen et al., 2024). By decomposing the joint distribution through the law of total probability (Lyu et al., 2019; Gal and Ghahramani, 2016), we prove that the nuisance term induced by boundary miscalibration is asymptotically bounded, thereby restricting hallucination signals to the subspace where the model believes it knows yet answers incorrectly. The resulting score approximates the true uncertainty conditioned on knowledgeability.

The entire framework is training-free and requires no external knowledge, enabling seamless deployment under white-box or black-box settings. Experiments on three public question-answering benchmarks show that the two-dimensional indicator consistently outperforms one-dimensional uncertainty baselines across models of varying scales, with larger improvements observed on more challenging, adversarially constructed hallucination datasets.

We summarize our contributions as follows:

1. **Answer-level Intrinsic Cognition (AIC)** — a training-free metric that quantifies cognitive-boundary stability by aggregating answer-consistency across multiple stochastic decodings of the same question.
2. **Two-dimensional uncertainty decomposition** — a principled derivation that bounds the nuisance term induced by boundary miscalibration, yielding a hallucination score conditioned on knowledgeability.
3. **Extensive empirical validation** — consistent gains over one-dimensional uncertainty baselines on three public QA benchmarks across model scales, with larger improvements on adversarially constructed low-entropy hallucination sets.

2 Related Work

Hallucination detection has converged to three mainstream paradigms, each implicitly assuming that "the model knows what it does not know" and consequently exhibiting limited robustness under distribution shift (Sriramanan et al., 2024; Ren et al., 2022; Liu et al., 2020; Kadavath et al., 2022).

One-dimensional uncertainty estimation relies exclusively on the target model θ and can be subdivided into two strands.

(1) **Token-level probability indicators** such as perplexity or energy scores measure the sharpness of local probabilities $p_\theta(y_t \mid y_{<t}, x)$ along a generated sequence, operating under the premise that higher uncertainty implies higher hallucination probability (Sriramanan et al., 2024; Ren et al., 2022). Low-entropy hallucinations provide a counter-example: for queries outside the model's cognitive boundary, the distribution can be sharply peaked yet factually incorrect. Formally, there exist pairs (x, y) such that $\mathrm{PPL}(y \mid x) < \varepsilon$ while $y \notin \mathbb{K}$, where $\mathbb{K}$ denotes the set of factually correct answers and ε is an arbitrarily small positive constant.

(2) **Sequence-level consistency indicators** aggregate N stochastic samples $\{y^{(i)}\}_{i=1}^{N}$ to obtain a scalar uncertainty $O(y \mid x)$, *e.g.*, Length-Normalized Entropy (Malinin and Gales, 2020) or EigenScore (Chen et al., 2024). These methods assume that "a knowledgeable model will not contradict itself," but overlook blind-spot consistency: when the query lies beyond the cognitive boundary, the model may repeatedly generate the same plausible yet incorrect answer (Manakul et al., 2023). If $\forall i, j \in [N],\ \mathrm{sim}(y^{(i)}, y^{(j)}) > \tau$ while $y^{(i)} \notin \mathbb{K}$, the consistency-based uncertainty $O(x) \to 0$, leading to a false-negative detection.

External knowledge verification adopts the principle "do not trust the model's memory; trust external evidence." A typical pipeline (a) atomizes claims, (b) retrieves passages from encyclopedias or the web, and (c) labels truthfulness with an NLI or similarity model (Hu et al., 2024; Cohen et al., 2023; Kryściński et al., 2020; Bohnet et al., 2022). Although interpretable and updateable as knowledge bases evolve, the approach introduces

Table 1: Comparison of hallucination-detection approaches along three key axes. **1**: Training-free (no gradient updates or labelled data at test time); **2**: Black-box friendly (compatible with gradient-frozen, parameter-locked APIs); **3**: Boundary-aware (explicitly models the model's cognitive boundary). ✓/✗ indicate satisfied/unsatisfied criteria.

Approach	1	2	3
One-dimensional uncertainty	✓	✓	✗
External knowledge verification	✓	✓	✗
Hidden-state probing	✗	✗	✗
Two-dimensional uncertainty	✓	✓	✓

non-trivial latency (retrieve–align–judge) and returns "unknown" when evidence is itself missing. Our two-dimensional framework is orthogonal: retrieval confidence can be treated as an additional observation and fused with AIC in a two-factor uncertainty model, thereby converting the ternary decision {True, False, Unknown} into a continuous risk score.

Hidden-state probing trains lightweight classifiers on internal activations or attention patterns to discriminate faithful from hallucinated outputs (Azaria and Mitchell, 2023; Chuang et al., 2024; Kossen et al., 2024). While often achieving the highest absolute accuracy, these probes require white-box access, labelled hallucination examples, and careful selection of layers or heads, hindering deployment under black-box APIs. The proposed framework absorbs the spectral-intuition of such probes but uses it only to estimate AIC; in black-box scenarios it falls back to purely textual consistency, yielding a white-box-accelerated solution without retraining any parameters.

Knowledge conditioning and cognitive boundary. As summarized in Table 1, existing approaches directly equate "low probability," "inconsistency," or "external conflict" with hallucination signals, thereby neglecting the cognitive boundary that separates "knowing" from "not knowing." By explicitly modelling this boundary through answer-level Intrinsic Cognition and jointly reasoning with output uncertainty, our two-dimensional framework provides a principled approach to uncertainty calibration via uncertainty post-processing, which is training-free, black-box friendly, and explicitly boundary-aware.

3 Method

3.1 Problem Formulation and Analysis

Let θ denote a frozen LLM. For an input question x, the LLM autoregressively generates an answer $y = (y_1, \ldots, y_t)$. Our goal is to produce a calibrated hallucination score for the pair (x, y) without updating θ and without any labelled data.

We introduce two random variables:

1. Epistemic uncertainty $I(x) \in \{0, 1\}$. $I(x) = 1$ signifies that x lies inside the LLM's cognitive boundary; $I(x) = 0$ signifies that the LLM lacks the knowledge required for x.

2. Output uncertainty $U(y \mid x) \in \mathbb{R}$. This statistic is directly computable from the LLM output, *e.g.*, perplexity or energy.

Hallucination risk is defined as the conditional uncertainty given that the LLM believes it knows:

$$\eta(x, y) \triangleq E\big(U(y \mid x) \mid I(x) = 1\big). \tag{1}$$

Applying the law of total probability (Ash and Doléans-Dade, 2000) to $E(U(y \mid x))$ gives

$$E\big(U(y \mid x)\big) = E_1\pi_1 + E_0\pi_0, \tag{2}$$

where $E_* = E\big(U(y \mid x) \mid I(x) = *\big)$ and $\pi_* = P\big(I(x) = *\big)$. Solving (2) for $\eta(x, y)$ yields

$$\eta(x, y) = \frac{E\big(U(y \mid x)\big) - E_0\pi_0}{\pi_1}. \tag{3}$$

The term $E_0\pi_0$ captures the contribution to uncertainty arising from the model's lack of knowledge regarding the query condition. When the model is entirely unaware of the relevant context, uncertainty is not merely unquantifiable; the concept itself ceases to apply. Hence $E_0\pi_0$ is not a bias to be bounded but a nuisance term inherent to standard uncertainty formulations, and its presence partly explains the inaccuracy of conventional uncertainty estimates.

We therefore discard this term and retain only the epistemically valid component:

$$\eta(x, y) \approx \frac{E(U(y \mid x))}{\pi_1}, \tag{4}$$

which explicitly recalibrates any one-dimensional uncertainty $U(y \mid x)$ by the epistemic certainty π_1. Section 3.2 describes an unsupervised estimator of π_1, and Section 3.3 instantiates the resulting two-dimensional score. An overview of the pipeline is given in Figure 1.

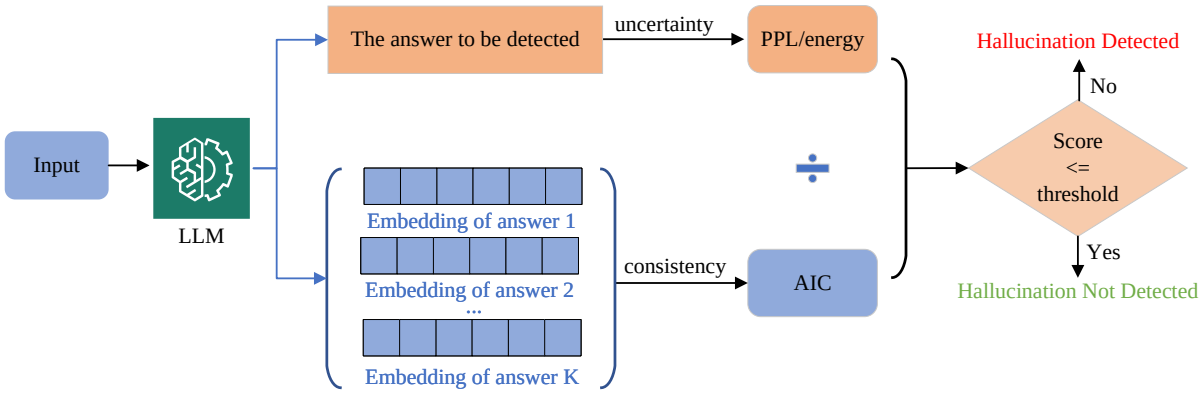


Figure 1: Two-dimensional uncertainty framework for hallucination detection.

3.2 Estimating Intrinsic Cognition

Because $I(x)$ is not observable, we approximate $\pi_1 = P(I(x) = 1)$ by analysing the semantic concentration of multiple stochastic decodings. The key intuition is that answers sampled inside the cognitive boundary scatter around a dominant direction, whereas outside-boundary queries yield either chaotic or misleadingly consistent but incorrect output (Chen et al., 2024; Lin et al., 2023; Kuhn et al., 2023). We therefore quantify concentration via the leading explained-variance ratio of an empirical covariance matrix.

Concretely, given x we sample n answers

$$\mathcal{Y} = \{y^{(i)} \sim \pi_\theta(\cdot \mid x; T)\}_{i=1}^{n} \tag{5}$$

with temperature $T > 0$. Each answer is mapped to a vector $\mathbf{h}_i \in \mathbb{R}^d$:

1. **White-box:** feed $[x; y^{(i)}]$ into θ and extract the last-token hidden state from a selected layer;
2. **Black-box:** encode $y^{(i)}$ with a fixed sentence encoder (*e.g.*, Sentence-BERT).

Let $\tilde{\mathbf{H}} \in \mathbb{R}^{d \times n}$ denote the mean-centered matrix. The empirical covariance

$$\mathbf{\Sigma} = \frac{1}{n-1}\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top \in \mathbb{R}^{d \times d} \tag{6}$$

is regularized as $\mathbf{\Sigma}_\alpha = \mathbf{\Sigma} + \alpha\mathbf{I}$ with $\alpha = 10^{-3}$ to ensure numerical stability. Eigen-decomposition yields

$$\begin{aligned} \mathbf{\Sigma}_\alpha &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \\ \mathbf{\Lambda} &= \mathrm{diag}(\lambda_1 \geq \cdots \geq \lambda_d). \end{aligned} \tag{7}$$

Normalizing the eigenvalues by total variance gives the explained-variance ratios

$$\tilde{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}. \tag{8}$$

We define the AIC score as

$$\hat{\pi}_1 = \tilde{\lambda}_1, \tag{9}$$

which is large when samples align along a single semantic direction and small when they are scattered. Algorithm 1 summarizes the procedure.

Algorithm 1 Estimating $\hat{\pi}_1$ (AIC)

Require: Question x, model π_θ, sample count n, temperature T, encoder $\mathsf{enc}(\cdot)$, regularizer $\alpha = 10^{-3}$

$\mathcal{Y} \leftarrow \{y^{(i)} \sim \pi_\theta(\cdot \mid x; T)\}_{i=1}^{n}$
for $i = 1$ **to** n **do**
  $\mathbf{h}_i \leftarrow \mathsf{enc}(x, y^{(i)})$
end for
$\mathbf{H} \leftarrow [\mathbf{h}_1, \ldots, \mathbf{h}_n] \in \mathbb{R}^{d \times n}$; mean-center to obtain $\tilde{\mathbf{H}}$
$\mathbf{\Sigma} \leftarrow \frac{1}{n-1}\tilde{\mathbf{H}}\tilde{\mathbf{H}}^\top$; $\mathbf{\Sigma}_\alpha \leftarrow \mathbf{\Sigma} + \alpha\mathbf{I}$
Eigendecompose $\mathbf{\Sigma}_\alpha = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ and extract λ_1
$\tilde{\lambda}_1 \leftarrow \lambda_1 / \sum_{j=1}^{d} \lambda_j$
return $\hat{\pi}_1 = \tilde{\lambda}_1$

3.3 Two-dimensional Uncertainty Score

Given any one-dimensional uncertainty statistic $\hat{\mathcal{U}}(y \mid x)$, we produce a calibrated hallucination score by

$$\mathcal{S}(x, y) = \frac{\hat{\mathcal{U}}(y \mid x)}{\hat{\pi}_1}, \tag{10}$$

where $\hat{\pi}_1$ is the AIC estimated in Algorithm 1. Equation (10) is plug-and-play: $\hat{\mathcal{U}}$ can be perplexity or energy.

$\mathcal{S}$ behaves as an odds-like ratio: numerator quantifies "how uncertain the LLM sounds," while denominator quantifies "how confident the LLM is that it possesses the knowledge." Thus, a small $\hat{\pi}_1$ inflates $\mathcal{S}$, flagging high-confidence but incorrect outputs—the dominant failure mode of one-dimensional detectors. Conversely, when $\hat{\pi}_1$ is large, $\mathcal{S}$ reduces to the original uncertainty, preserving sensitivity to genuine ambiguity.

Under the generative model described in §3.1, $\mathcal{S}$ is asymptotically unbiased for $\eta(x, y)$ up to an additive constant that cancels during thresholding. Formally, as $n \to \infty$ and d fixed, $\hat{\pi}_1 \xrightarrow{p} \pi_1$; hence ranking by $\mathcal{S}$ is consistent with ranking by the true conditional uncertainty η.

Because $\hat{\pi}_1$ is computed solely from π_θ's own samples, no human-annotated hallucination labels are required. We pick the threshold τ as the elbow of the empirical cumulative distribution function (CDF) of $\mathcal{S}$; if ground-truth labels become available, the same τ can be plotted on the receiver operating characteristic (ROC) curve and evaluated via the area under the ROC curve (AUROC) (Satopaa et al., 2011). Algorithm 2 summarizes the three-line procedure.

Algorithm 2 Two-dimensional uncertainty hallucination detector

Require: Question x, answer y, model π_θ, hyper-parameters $(n, T, \texttt{type})$

1: $\hat{\pi}_1 \leftarrow$ Algorithm 1(x) ▷ AIC estimation
2: $\hat{\mathcal{U}} \leftarrow$ one-dimensional uncertainty of type $\texttt{type}$ on (x, y)
3: **return** $\mathcal{S}(x, y) = \hat{\mathcal{U}}/\hat{\pi}_1$

Relation to consistency-based detectors Existing methods such as SelfCheckGPT or EigenScore treat high answer agreement as a sufficient signal of truthfulness (Manakul et al., 2023; Chen et al., 2024). In contrast, AIC regards agreement as necessary but not sufficient for falling inside the cognitive boundary: outside-boundary queries may also collapse to a misleadingly consistent but incorrect answer. By using the leading explained-variance ratio $\tilde{\lambda}_1$ as an explicit estimate of knowability π_1 and dividing the original uncertainty by this estimate, we convert consistency from a binary verdict into a continuous calibration factor; Table 2 summarizes the conceptual difference.

Table 2: Comparison of uncertainty-based detectors. "Cal. low-H" = calibrates high-confidence hallucinations outside cognitive boundary.

Method	Train-free	Cal. low-H
One-dimensional	✓	✗
Consistency	✓	✗
Two-dimensional (ours)	✓	✓

4 Experiments

4.1 Experimental Setup

Datasets We conduct closed-book evaluation on three standard open-domain QA benchmarks; models must answer solely from parametric memory. For each corpus, we fix the first 1,000 examples of the official test set to ensure deterministic and reproducible uncertainty estimation (Niu et al., 2025). **TriviaQA** (Joshi et al., 2017) — trivia-style questions requiring encyclopedic knowledge. **Natural Questions (NQ)** (Kwiatkowski et al., 2019) — real user queries with Wikipedia gold spans; we retain single-short answers. **HaluEval-QA** (Li et al., 2023; Zhu et al., 2024) — adversarially constructed QA pairs where 50% of fluent answers are factually incorrect, enabling direct assessment of low-entropy hallucinations.

Models and Hyper-parameters We cover both scale and architectural diversity: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B-Instruct, and Qwen-2.5-14B-Instruct (Qwen et al., 2025). All parameters remain frozen; only decoding hyper-parameters are adjusted. Following consistency-based baselines (Chen et al., 2024), a hybrid greedy–stochastic strategy is adopted: one greedy sequence for answer generation and nine stochastic samples (T=0.5, top-k=5, top-p=0.99, repetition penalty 1.2) for covariance estimation.*

Evaluation Metrics Hallucination detection is cast as a binary classification task. We report AUROC, which measures the probability that a randomly drawn positive example is assigned a higher uncertainty score than a negative example; higher AUROC indicates better ranking quality (Azaria and Mitchell, 2023). To mitigate surface-form bias, two complementary correctness labels are generated for every answer:

1. **Lexical (AUROC$_r$):** ROUGE-L F-measure ≥ 0.5 against any reference, following Lin (2004). ROUGE-L quantifies the longest common subsequence between the generated text and the ground-truth span; we use the F_1 score to balance precision and recall.

2. **Semantic (AUROC$_s$):** cosine similarity ≥ 0.9 between the generated sentence and the ground-truth answer, using 768-dimensional

*White-box hidden states are extracted from the *middle* layer of the model (layer $\frac{L}{2}$).

Table 3: Main results across three public open-domain QA benchmarks (NQ, TriviaQA, HaluEval-QA) and three model scales (8B, 7B, 14B). Each cell reports **AUROC$_r$ / AUROC$_s$** (%, higher is better) under lexical and semantic correctness criteria, respectively. Best score in **bold**; second-best underlined. Suffix "-AIC" denotes coupling the corresponding one-dimensional uncertainty with our proposed Answer-level Intrinsic Cognition to form a two-dimensional hallucination detector. All improvements are absolute percentage points over the baseline method in the same row. The consistent gains validate that AIC calibration suppresses low-entropy hallucinations across domains, model sizes, and uncertainty backbones.

Model	Method	NQ		TriviaQA		HaluEval-QA	
		AUROC$_r$	AUROC$_s$	AUROC$_r$	AUROC$_s$	AUROC$_r$	AUROC$_s$
Llama-3.1-8B	EigenScore	76.8	76.2	80.0	78.8	74.3	73.8
	LNE	70.9	73.2	80.6	78.8	77.7	77.2
	SAR	75.9	75.7	80.4	79.3	79.4	79.3
	SE	76.7	76.3	79.3	78.1	76.3	76.6
	AIC	73.0	71.6	75.2	74.3	72.4	72.1
	PPL	73.0	76.9	80.5	80.5	76.7	77.3
	PPL-AIC	76.7	79.4	82.8	82.7	78.8	79.2
	Energy	63.0	65.5	74.5	74.1	73.5	74.8
	Energy-AIC	73.7	73.0	79.0	78.2	75.8	75.9
	CCP	79.3	81.6	82.9	82.8	80.5	82.3
	CCP-AIC	**80.9**	**82.4**	**83.5**	**82.9**	**81.8**	**83.2**
Qwen2.5-7B	EigenScore	74.7	75.2	77.8	76.8	74.4	73.5
	LNE	74.1	73.5	81.5	80.7	79.4	78.1
	SAR	63.0	63.7	70.5	71.9	70.4	71.2
	SE	64.7	65.9	70.5	71.9	70.4	71.2
	AIC	75.9	**77.2**	77.4	77.0	77.0	76.2
	PPL	75.6	75.0	84.9	84.4	82.3	81.7
	PPL-AIC	**76.7**	76.4	**85.5**	**84.9**	**82.7**	**82.1**
	Energy	66.0	66.9	79.4	80.3	75.2	75.9
	Energy-AIC	72.0	73.6	82.8	83.1	78.3	78.4
	CCP	68.4	67.8	70.5	71.9	70.4	71.2
	CCP-AIC	71.2	70.7	74.7	74.7	72.7	73.2
Qwen2.5-14B	EigenScore	75.2	73.8	77.0	76.6	74.1	74.6
	LNE	76.0	74.3	80.8	79.6	81.9	82.1
	SAR	57.9	55.2	65.2	64.8	61.8	61.6
	SE	62.4	62.0	66.8	66.7	65.5	66.0
	AIC	76.2	74.4	77.5	76.4	78.6	79.2
	PPL	78.5	76.7	84.1	83.8	86.4	86.7
	PPL-AIC	**79.3**	**77.4**	**84.8**	**84.3**	**86.9**	**87.1**
	Energy	69.8	66.8	78.4	79.0	78.8	78.9
	Energy-AIC	74.2	71.4	82.7	82.6	81.6	82.1
	CCP	58.8	58.4	61.6	62.6	60.0	60.2
	CCP-AIC	60.6	60.0	63.6	64.6	61.8	62.0

embeddings from `nli-roberta-large`[†]. The threshold 0.9 corresponds to strong semantic equivalence under the NLI-trained space (Reimers and Gurevych, 2019).

For each (x, y) pair, detectors output a scalar uncertainty; AUROC is computed once with respect to both label sets.

Baselines We compare against seven mainstream uncertainty estimators. **PPL** (Ren et al., 2022) computes sequence-level uncertainty as mean token-level perplexity. **Energy** (Liu et al., 2020) derives an energy-based score from the last-layer logits. **LN-Entropy** (Malinin and Gales, 2020) measures length-normalized predictive entropy across multiple samples. **EigenScore** (Chen et al., 2024) leverages the explained variance of the leading principal component to assess multi-answer consistency. **Semantic Entropy (SE)** (Kuhn et al., 2023) clusters answers by semantic equivalence. **CCP** (Fadeeva et al., 2024) provides token-level conformal prediction with coverage guarantees. **SAR** (Duan et al., 2024) detects context misalignment via attention-relevance divergence.

Notably, PPL, Energy, and CCP are inherently scalar indicators; we extend them to two-dimensional features by introducing the answer-level dimension and denote the resulting variants

[†] https://huggingface.co/sentence-transformers/nli-roberta-large

with the suffix "-AIC". In contrast, EigenScore, LN-Entropy, SE, and SAR already capture consistency across multiple answers and thus do not require such extension. AIC can be coupled with any one-dimensional uncertainty measure to form complementary features without additional training. All experiments were conducted on a local GPU workstation with an AMD EPYC 7742 64-core CPU and an NVIDIA DGX A800 (80 GB).

4.2 Main Results

Effectiveness Table 3 shows that coupling *AIC* with one-dimensional indicators to form two-dimensional features yields consistent and significant AUROC gains across all baselines, supporting the generality of the proposed calibration. For Llama-3.1-8B on NQ, *PPL-AIC* improves AUROC$_r$ from 73.0 to 76.7 ($\Delta \uparrow +3.7$) and AUROC$_s$ from 76.9 to 79.4 ($\Delta \uparrow +2.5$); *Energy-AIC* delivers an absolute AUROC$_r$ gain of 10.7. Similar trends hold on TriviaQA and HaluEval-QA and remain stable across model scales, indicating that the gains are not domain-specific.

Scale effect on hallucination rate As parameter count increases, the raw hallucination rate decreases monotonically: on HaluEval-QA, Qwen2.5-14B achieves 86.4/86.7 with *PPL*, outperforming the 7B counterpart by 4.1 points. This observation aligns with the theoretical expectation that $\hat{P}(I(x) = 1)$ approaches 1 for larger models, thereby reducing the room for calibration.

Diminishing calibration gains Although two-dimensional calibration remains beneficial, its marginal improvement shrinks as model scale grows. On HaluEval-QA (AUROC$_r$), PPL-AIC yields an absolute gain of 0.4 percentage points for Qwen2.5-7B and 0.5 points for 14B; Energy-AIC drops from 3.1 points (7B) to 2.8 points (14B). The same decay pattern holds for AUROC$_s$ and for NQ/TriviaQA, where the absolute gains likewise decrease with model scale, consistent with Eq. (10): as $\hat{\pi}_1 \to 1$, $\mathcal{S}(x, y) \to \mathcal{U}(y \mid x)$, leaving less room for correction.

4.3 Ablation Studies

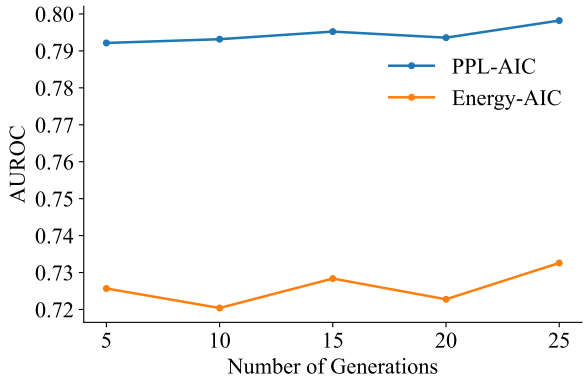
Impact of automatic correctness criteria Designing QA correctness metrics that align with human judgment is non-trivial; the chosen threshold therefore constitutes a major source of variance in hallucination detection. On Llama-3.1-8B + NQ, we sweep ROUGE-L and Sentence-Similarity thresholds from 0.3 to 0.9. As shown in Table 4, tightening the criterion re-labels partially correct answers as negatives, shifting class balance toward negatives. Because "*-AIC" assigns higher uncertainty to these high-confidence but incorrect borderline samples, they migrate from the false-positive to the true-negative region of the ROC curve, simultaneously increasing the true-positive rate (TPR) and decreasing the false-positive rate (FPR), hence yielding higher AUROC. The effect is most pronounced at the stringent threshold 0.9, where one-dimensional baselines suffer from confident nonsense while our two-dimensional score remains robust.

Table 4: Correctness measures vs. threshold (ROUGE-L / Sentence Similarity).

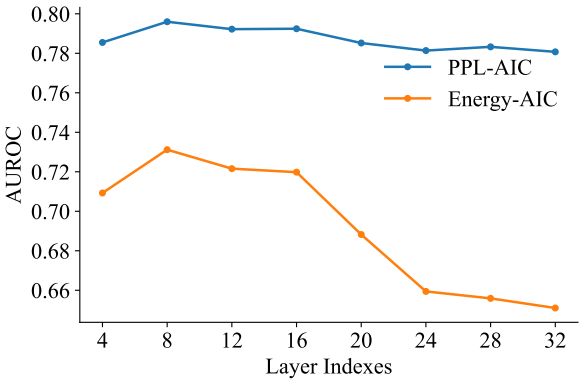
Method	ROUGE-L			Sentence Similarity		
	0.3	0.5	0.7	0.7	0.8	0.9
EigenScore	75.8	76.8	76.1	71.7	74.1	76.2
AIC	72.4	73.0	72.1	66.8	69.7	71.6
PPL	72.4	73.0	75.7	65.4	71.1	76.9
PPL-AIC	75.9	76.7	78.5	68.5	73.8	79.4
Energy	64.6	65.5	65.3	58.5	61.6	65.5
Energy-AIC	73.0	73.6	73.5	66.8	70.1	73.0
CCP	76.6	79.3	80.7	71.5	77.8	81.6
CCP-AIC	**78.2**	**80.9**	**81.8**	**72.5**	**78.9**	**82.4**

Sample budget analysis All ablation runs in this subsection adopt the most stringent correctness criterion (Sentence-Similarity > 0.9) to focus on high-confidence errors. Figure 2a reports AUROC versus the number of stochastic answers $K \in \{5, 10, 15, 20, 25\}$ (Llama-3.1-8B, NQ). Performance increases with K and saturates around $K = 15$; beyond this point, the curve plateaus, indicating diminishing returns. Thus, $K = 15$ offers a practical trade-off between accuracy and computational cost.

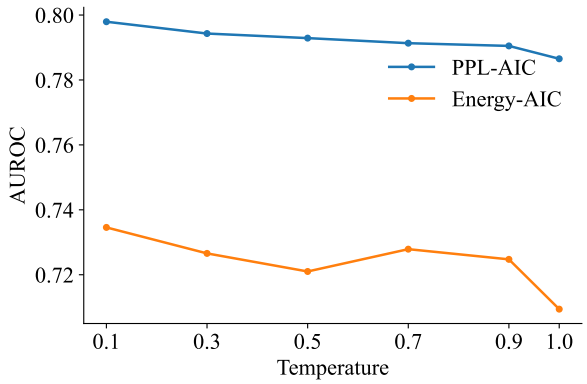
Layer sensitivity for white-box embeddings In the main experiments, we use the embedding of the last token from an intermediate layer as the sentence representation. Here we further study the impact of using different layers. Figure 2b shows hallucination detection performance across layers. We observe: (1) intermediate-layer sentence vectors (especially near the middle) achieve the best performance, potentially because intermediate layers better capture semantic information; (2) deeper-



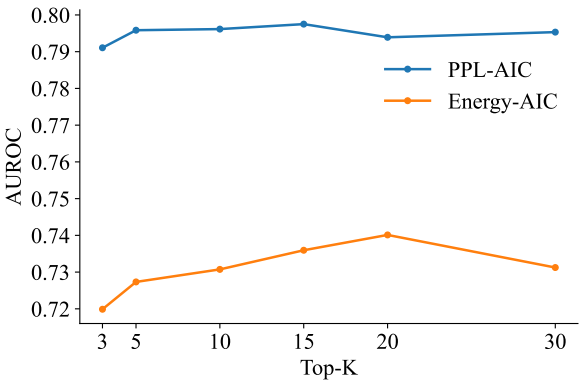
(a) AUROC vs. number of sampled answers.



(b) AUROC vs. representation layer.



(c) AUROC vs. temperature.



(d) AUROC vs. top-k.

Figure 2: Sensitivity analyses: (a) number of stochastic answers; (b) representation layer; (c) temperature; (d) top-k.

layer sentence vectors lead to a substantial drop in performance, suggesting that these layers may not capture key semantic information effectively.

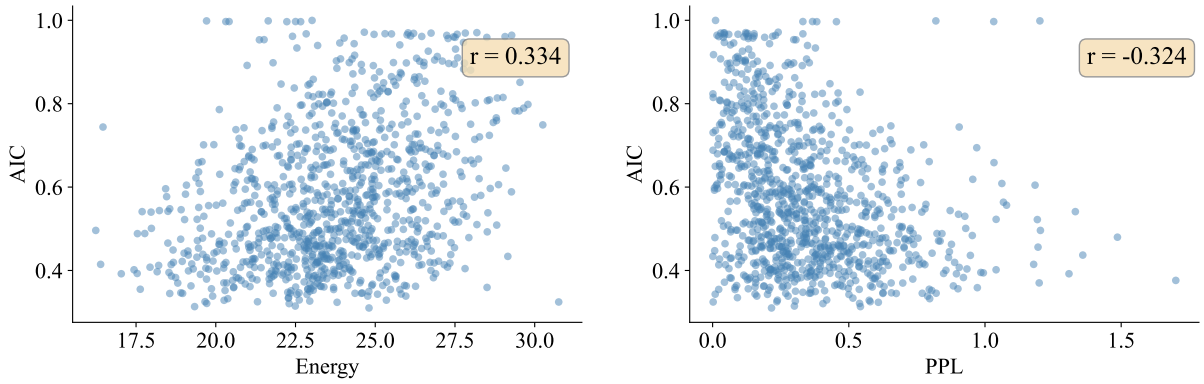
Hyperparameter sensitivity Decoding hyperparameters such as temperature, top-k, and top-p control output diversity. We provide sensitivity analyses in Figure 2c and Figure 2d. Temperature has a strong impact, while top-k has a smaller effect. Temperature (Figure 2c): as temperature increases from 0.3 to 1.0, performance decreases, most notably for Energy-AIC at $T = 1.0$, indicating that excessive diversity may introduce more irrelevant or incorrect content and degrade hallucination detection.

Top-k (Figure 2d): varying top-k from 3 to 30 has limited impact across all three methods, suggesting that these methods are relatively insensitive to top-k. Overall, temperature is the key hyperparameter and should be tuned carefully to avoid overly increasing generation diversity, while top-k can be set more flexibly.

Feasibility of AIC Calibration To validate that AIC effectively complements one-dimensional uncertainty measures, we analyze its Pearson correlation with representative scalar indicators PPL and Energy. As shown in Figure 3, AIC exhibits weak correlations with both PPL and Energy, with absolute values below 0.35 in both cases. This empirical evidence suggests that AIC captures largely independent information from these scalar baselines: PPL and Energy assess token-level or sequence-level model confidence, whereas AIC measures answer-level consistency across multiple generations. The low correlation indicates limited redundancy between the two dimensions, making them suitable candidates for feature concatenation. Consequently, coupling AIC with any one-dimensional uncertainty measure yields complementary two-dimensional features without requiring additional training or architectural modifications.

5 Conclusion

We introduced a training-free, two-dimensional uncertainty framework that couples output uncertainty



(a) Energy vs. AIC ($r = 0.334$)

(b) PPL vs. AIC ($r = -0.324$)

Figure 3: Pearson correlation analysis between AIC and one-dimensional uncertainty measures. Weak correlations ($|r| < 0.35$) indicate that AIC captures complementary information to scalar baselines, validating its use for feature calibration.

with Answer-level Intrinsic Cognition. By explicitly modeling the model's own cognitive-boundary stability, the approach recalibrates any off-the-shelf uncertainty score without gradients, labels, or external knowledge. Consistent improvements across multiple benchmarks and model scales indicate that the paradigm effectively suppresses low-entropy hallucinations that remain invisible to one-dimensional detectors. Theoretically, the decomposition bounds the bias induced by out-of-boundary queries and confines the hallucination signal to the region where "the model believes it knows yet answers incorrectly." Beyond empirical gains, the work shifts hallucination detection from single scalar statistics to joint cognition–uncertainty modeling, offering a plug-in module for safety-critical deployments under black-box constraints. Future directions include reducing sampling cost and extending the principle to long-form generation and retrieval-augmented dual verification.

6 Limitations

The framework relies on answer-level consistency across stochastic decodings as an indirect proxy for the model's cognitive boundary. This design choice incurs extra forward passes, yields a single global estimate for the entire answer, and becomes less informative under low-temperature sampling, thereby propagating latency, granularity, and sensitivity issues into downstream deployment. Future work that obtains the boundary through more direct and efficient signals could lift these constraints without altering the core two-dimensional calibration principle.

AI Usage Statement

Large language models were used exclusively for grammar checking, wording refinement, and LaTeX formatting. All scientific content, claims, and final decisions were made by the human authors.

References

Robert B Ash and Catherine A Doléans-Dade. 2000. *Probability and measure theory*. Academic press.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *Preprint*, arXiv:2402.03744.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and

Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

X Hu and 1 others. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jieran Li, Xiuyuan Hu, Yang Zhao, Shengyao Zhuang, and Hao Zhang. 2025. Leveraging reference documents for zero-shot ranking via large language models. *Preprint*, arXiv:2506.11452.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Shuliang Liu, Hongyi Liu, Aiwei Liu, Duan Bingchen, Zheng Qi, Yibo Yan, He Geng, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. A survey on proactive defense strategies against misinformation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18144–18155.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. 2019. Advances in neural information processing systems. *Advances in neural information processing systems*, 32.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In

Proceedings of the 2023 conference on empirical methods in natural language processing, pages 9004–9017.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. 2025. Robust hallucination detection in llms via adaptive token selection. *arXiv preprint arXiv:2504.07863*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Derui Zhu, Dingfan Chen, Qing Li, Zongxiong Chen, Lei Ma, Jens Grossklags, and Mario Fritz. 2024. Pollmgraph: Unraveling hallucinations in large language models via state transition dynamics. *arXiv preprint arXiv:2404.04722*.