LCIC: A Unified Hallucination Detection-and-Mitigation Method Using LLM's Tail-Pooled Internal Signals

Jiahang Li*, Jiafei Niu*, Yidong Ding, Ping Yi†
School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China
Email: {williamlee2002, jiafei_niu, ydding2001, yiping}@sjtu.edu.cn

***Abstract*—Hallucination remains a critical reliability bottleneck for large language model (LLM) applications, where models may generate fluent yet unverifiable or incorrect statements when internal evidence is insufficient. Most governance pipelines treat hallucination handling as two loosely coupled stages: detection, which estimates the reliability of a generated response, and mitigation, which suppresses hallucinations through inference-time control or training-time alignment. Such designs often incur substantial overhead: external evidence pipelines increase system complexity and latency, while sampling-based consistency approaches raise inference cost proportionally to the number of generations.**

In this paper, we propose LCIC, a unified detect–mitigate method that enables single forward-pass hallucination detection and provides a unified risk signal for both detection and mitigation, without relying on external tools or additional generations. LCIC derives a continuous hallucination risk score, denoted as S_{LCIC}, by combining two complementary cues: (i) tail-focused internal consistency, which aggregates token-level confidence and attention focus to capture localized instability, and (ii) layer contrast, which measures distributional disagreement between intermediate and final layers to reflect internal inconsistency. The same score is further used as a reward signal for group-relative policy optimization and stabilizing alignment with a Best-of-N(BoN) stablizing regularizer.

Experiments on four datasets and multiple open LLMs demonstrate that LCIC achieves strong unsupervised detection performance and delivers consistent mitigation gains over representative decoding-time and self-refinement baselines[1].

***Index Terms*—Large Language Models, Hallucination Detection, Mitigation, Internal Signals, Layer Contrast, Reinforcement Learning.**

I. Introduction

Large language models (LLMs) have become the foundation of modern question answering and assistant systems. Despite strong generative abilities, their practical reliability is constrained by hallucinations—outputs that are fluent and confident but unsupported by verifiable facts (**factual hallucination**) or by the input context (**faithfulness hallucination**) [1]–[4]. These failure modes demand reliable detection and mitigation mechanisms.

Existing detectors face a fundamental tension between latency and granularity. Many state-of-the-art methods are not single forward-pass: external verification pipelines validate claims using retrieval or tools [5], and consistency-based detectors rely on multiple sampled outputs or multi-round self-checking [6], [7]. Their computational cost grows with sample count, impeding large-scale deployment.

To reduce latency, recent work turns to internal signals—token probabilities, attention statistics, or hidden states [8]–[12]. However, these methods typically summarize a generation via **global aggregation**, which dilutes localized failures. Hallucinations are empirically sparse and span-concentrated—confined to a few critical tokens such as entities or numbers [5], [13], [14]. Averaging over the full sequence overwhelms these concentrated signals, weakening discriminability and calibration.

This raises a central question: can we characterize hallucination risk in a single forward-pass while explicitly prioritizing high-risk spans to prevent signal dilution?

We propose LCIC, a unified detect–mitigate method. At its core is a tail-aware, single forward-pass score S_{LCIC} prioritizing localized failures and serving as a shared signal for both detection and mitigation.

- **Tail-Event Risk Score**: S_{LCIC} fuses (i) CVaR-style low-tail-pooled [15] internal consistency (token confidence and attention focus) with (ii) high-tail-pooled layer contrast (Jensen–Shannon divergence between mid- and final-layer distributions), emphasizing localized high-risk spans.
- **Unified Detect–Mitigate Framework**: the same scalar S_{LCIC} drives threshold-based online detection and GRPO-based training-time alignment with BoN regularization [16], [17], coupling both stages under a consistent criterion.
- **Empirical Robustness**: evaluations across four QA datasets and two open-source LLM families demonstrate strong detection and mitigation performance, with ablations validating each component.

II. Related Work

A. Hallucination Detection

1) External and Consistency-Based Detection: Early hallucination detection relied on external knowledge verification. The "extract-and-verify" paradigm (e.g., FactScore [5])

*These authors contributed equally to this work.
†Corresponding author.
[1]The data and code for this work will be made publicly available upon acceptance.

decomposes generated text into atomic claims and validates each against retrieved evidence or structured knowledge bases. While effective for factual accuracy, these approaches introduce substantial system complexity and inference latency. Self-consistency methods such as Chain-of-Verification (CoVe) [6] and SelfCheckGPT [7] leverage multiple sampled outputs from the same model to detect contradictions, reducing reliance on external tools but incurring cost proportional to sample count. More recent approaches prompt stronger LLMs to assess generation quality (LLM-as-judge) [18], though these remain limited by coarse supervision and significant inference overhead.

2) Internal-Signal Detection: Recent work seeks to detect hallucinations without external queries by leveraging the model's internal activations. Token-level confidence (probability of generated tokens) provides a basic uncertainty measure [1], while attention statistics capture where the model focuses during generation [8], [9]. Hidden state-based approaches such as EigenScore [11] and LLM-Check [12] exploit distributional properties of intermediate representations. However, these methods typically aggregate signals via arithmetic averaging over the entire sequence, which dilutes localized failure signals. Our approach differs by applying CVaR-style tail pooling [15] to explicitly prioritize high-risk spans.

3) Faithfulness Detection: For context-grounded generation, faithfulness hallucination detection verifies alignment between outputs and provided sources. Natural Language Inference (NLI) classifiers determine entailment relationships between generated text and source documents [19], while QA-consistency loops generate questions from model outputs and verify answers against source text [4]. These methods excel at identifying deviations from source material but require task-specific supervision or multiple inference passes.

B. Hallucination Mitigation

Mitigation strategies operate at three stages of the LLM pipeline. **Data-level** approaches include filtering noisy training samples [20], [21] and augmenting generation with retrieved evidence via RAG [22], though the latter introduces retriever latency. **Training-level** methods align model behavior through supervised fine-tuning on conservative responses [23] or RLHF with human preference data; however, these typically decouple risk estimation from optimization, requiring external judgment signals. **Inference-level** strategies such as DoLa [24] contrast layer-wise distributions to suppress unreliable tokens, while post-generation revision methods [25], [26] iteratively refine outputs using self-feedback or external verification.

Our work bridges the gap between detection and mitigation by reusing a single internal-signal risk score for both tasks. Unlike prior approaches that treat detection and mitigation as independent stages with potentially inconsistent criteria, LCIC unifies both under the same S_{LCIC} signal, enabling coherent optimization without additional inference overhead at deployment.

III. Methodology

A. Preliminaries

Consider an L-layer Transformer with H attention heads. For query $q = (x_1, \ldots, x_{T_x})$ and generated response $y = (y_1, \ldots, y_{T_y})$, let $z = [q; y]$ and $\mathcal{T}_y = \{T_x + 1, \ldots, T_x + T_y\}$. For each token $t \in \mathcal{T}_y$, we access internal signals $\mathcal{S}_t = \{p_t, \mathbf{A}_t, \mathbf{H}_t\}$: conditional probability $p_t = p(y_{t-T_x} \mid z_{<t})$, attention weights $\mathbf{A}_t$, and hidden states $\mathbf{H}_t$.

Sparse Hallucination Hypothesis. Hallucination risk is not uniformly distributed: empirically, errors concentrate in a small number of critical spans (entities, quantities) [5], [13], [14]. Thus $\mathcal{T}_{\text{halluc}} \subset \mathcal{T}_y$ is sparse with $|\mathcal{T}_{\text{halluc}}| \ll |\mathcal{T}_y|$. Global averaging $\bar{R} = \frac{1}{|\mathcal{T}_y|}\sum_t r_t$ dilutes these localized signals as $|\mathcal{T}_y|$ grows, motivating tail-aware pooling [15].

B. The LCIC Risk Score

The LCIC pipeline (Fig. 1) derives a scalar risk score from a single forward-pass.

1) Internal Consistency via Low-Tail Pooling: **Token confidence.** For $t \in \mathcal{T}_y$, token-level generation confidence is $p_t = p(z_t = y_{t-T_x} \mid z_{<t}) \in (0, 1]$. Low p_t indicates local reliability risk when the model lacks internal support.

Attention focus. We quantify evidence localization via layer-averaged top-1 attention:

$$s_t = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{H}\sum_{h=1}^{H}\max_i A_{t,i}^{(l,h)}. \tag{1}$$

Low s_t reflects diffuse attention and weaker evidence localization [9], [10].

Low-tail pooling. To capture the most fragile spans under the sparse hypothesis, we apply CVaR-style pooling [15] over the lowest q_{stab} quantile index sets $\mathcal{J}_p, \mathcal{J}_s$:

$$\text{CVaR}_p = \frac{1}{|\mathcal{J}_p|}\sum_{t\in\mathcal{J}_p} p_t, \quad \text{CVaR}_s = \frac{1}{|\mathcal{J}_s|}\sum_{t\in\mathcal{J}_s} s_t, \tag{2}$$

$$R_{\text{stab}} = \sqrt{\text{CVaR}_p \cdot \text{CVaR}_s}. \tag{3}$$

R_{stab} is high only when the model remains confident and focused on the frailest tokens.

2) Layer Contrast via High-Tail Pooling: We quantify internal disagreement by contrasting token predictive distributions from intermediate layers L_{mid} (the middle third of $\{1, \ldots, L-1\}$) with the final layer L [11], [24]. For hidden state $h_t^{(l)}$, applying the shared LM head yields logits $z_t^{(l)} = W_{\text{LM}} h_t^{(l)} + b_{\text{LM}}$ and distribution $p_t^{(l)} = \text{Softmax}(z_t^{(l)})$. We average logits over L_{mid} before softmax: $\bar{z}_t^{(\text{mid})} = \frac{1}{|L_{\text{mid}}|}\sum_{l\in L_{\text{mid}}} z_t^{(l)}$, giving $p_t^{(\text{mid})} = \text{Softmax}(\bar{z}_t^{(\text{mid})})$.

Token-wise divergence. We compute the Jensen–Shannon(JS) divergence between mid- and final-layer distributions:

$$M_t = \tfrac{1}{2}(p_t^{(\text{mid})} + p_t^{(L)}) \tag{4}$$

$$\text{JS}_t = \tfrac{1}{2}\big(\text{KL}(p_t^{(\text{mid})} \| M_t) + \text{KL}(p_t^{(L)} \| M_t)\big). \tag{5}$$

High-tail pooling. Let $\mathcal{J}_{\text{JS}}$ be the highest q_{lc} quantile by JS_t:

$$D_{\text{lc}} = \frac{1}{|\mathcal{J}_{\text{JS}}|}\sum_{t\in\mathcal{J}_{\text{JS}}} \text{JS}_t. \tag{6}$$

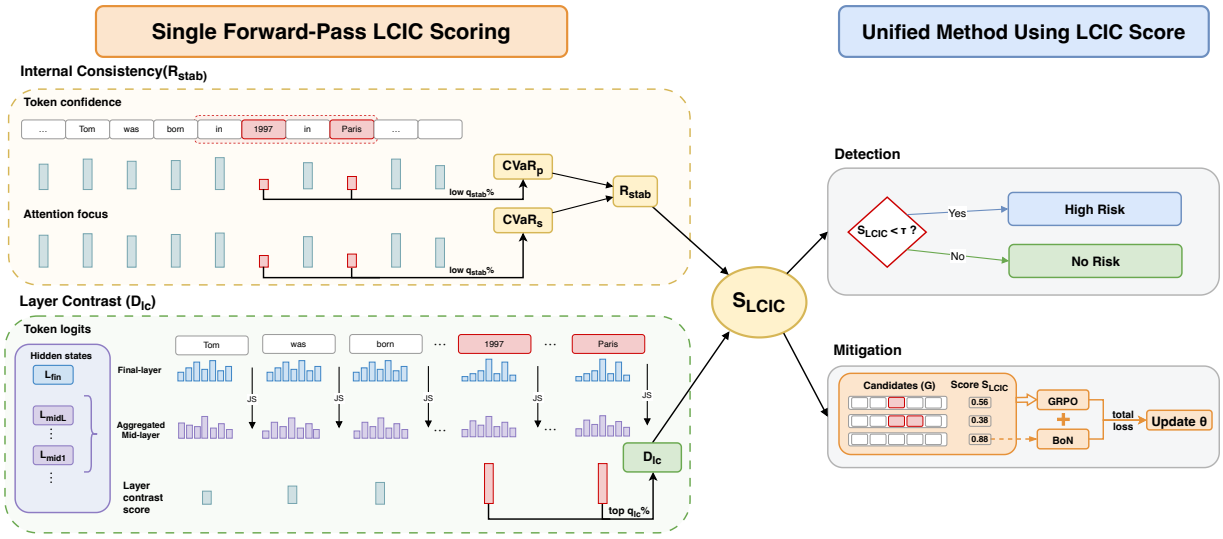

Fig. 1. **Unified LCIC pipeline.** A single forward-pass produces tail-pooled components R_{stab} (internal consistency) and D_{lc} (layer contrast), fused into S_{LCIC}. The same score supports online detection (thresholding) and training-time alignment (GRPO with Best-of-N regularization).

This highlights localized semantic conflicts often concentrated in entities and numbers.

C. Unified Detection–Mitigation Method

Synthesizing both signals, the LCIC risk score is:

$$S_{\text{LCIC}}(q, y) = R_{\text{stab}}(q, y) \cdot \exp(-\lambda_{\text{lc}} D_{\text{lc}}(q, y)), \quad \lambda_{\text{lc}} > 0. \tag{7}$$

1) Online Detection: We perform detection by thresholding: $\hat{h}(q, y) = \mathbb{I}[S_{\text{LCIC}}(q, y) < \tau]$, where τ is an application-dependent operating point and $\hat{h} = 1$ indicates high-risk output, feedable to an intervention policy without additional queries.

2) Training-Time Mitigation: We reuse S_{LCIC} as the reward signal for alignment, ensuring mitigation is optimized under the same criterion as detection (Fig. 1, right branch).

GRPO with LCIC-guided rewards. We adopt GRPO [16]: for each prompt q, sample G responses $\{y_i\}$, compute $s_i = S_{\text{LCIC}}(q, y_i)$, and convert to group-relative rewards $r_i = (s_i - \mu_q)/(\sigma_q + \delta)$ where μ_q, σ_q are the group mean and std. This yields stable policy gradient updates with KL regularization toward a reference policy.

Best-of-N trajectory regularization. To stabilize training, we anchor on the highest-scoring candidate $y^* = \arg\max_i s_i$ via NLL loss:

$$L_{\text{BoN}}(\theta) = -\mathbb{E}_q \left[\sum_{t=1}^{T_{y^*}} \log \pi_\theta(y_t^* \mid q, y_{<t}^*) \right]. \tag{8}$$

The BoN loss encourages the model to replicate decision patterns along the highest-S_{LCIC} trajectory, providing stable gradients independent of advantage estimation.

The final objective is:

$$L_{\text{total}}(\theta) = -\mathcal{J}_{\text{GRPO}}(\theta) + \lambda_{\text{BoN}} L_{\text{BoN}}(\theta). \tag{9}$$

where $\mathcal{J}_{\text{GRPO}}$ denotes the standard clipped GRPO objective (with group-relative rewards from S_{LCIC} and KL regularization toward a reference policy).

IV. Experiments

We evaluate LCIC on two dimensions: (i) **detection**: separability and calibration of S_{LCIC} with frozen base models; and (ii) **mitigation**: test-set improvements from S_{LCIC}-guided alignment.

A. Experimental Setup

1) Datasets and Models: We evaluate on four QA datasets: **TriviaQA** [27] and **Chinese SimpleQA** [28] for no-context factuality; **CoQA** [29] and **DRCD** [30] for context faithfulness (bilingual). Models: **Llama-3.1-8B-Instruct** and **Qwen3-8B**.

2) Detection and Mitigation Settings:

a) Detection: Base model frozen; S_{LCIC} computed from greedy decoding. Default hyperparameters: q_{stab}=0.1, q_{lc}=0.3, λ_{lc}=1.0.

b) Mitigation: 80/20 train–test split; GRPO with Best-of-N (λ_{BoN}=0.5) following Sec. III-C2, sampling G=8 candidates per prompt.

3) Evaluation Metrics: We derive binary correctness labels from three criteria: **Exact Match (EM)** [31] for strict token-level consistency; **ROUGE-L** with threshold 0.5 for lexical overlap; and **semantic similarity** (cosine similarity >0.9 via BAAI/bge-m3 embeddings) for paraphrase equivalence. For detection, we report AUROC and Pearson correlation (PCC) between S_{LCIC} and quality scores to assess separability and calibration. For mitigation, we report test-set accuracy.

4) Baselines:

a) Detection: Perplexity [32], Energy Score [33], LN-Entropy [34], SelfCheckBERTScore [7], EigenScore [11], and LLM-Check [12].

TABLE I
HALLUCINATION DETECTION PERFORMANCE (%). BEST IN BOLD.

Model	Method	TriviaQA						CoQA						Chinese SimpleQA						DRCD					
		ROUGE		EM[a]		SemSim[b]		ROUGE		EM		SemSim		ROUGE		EM		SemSim		ROUGE		EM		SemSim	
		AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC	AUC	PCC
Qwen3-8B	Perplexity	78.10	43.37	74.38	47.01	78.04	54.96	50.27	1.22	60.08	12.16	52.22	3.61	63.87	17.18	63.11	11.66	64.01	12.55	72.11	33.21	69.48	32.58	70.59	30.88
	Energy Score	77.98	47.14	76.30	44.69	76.19	53.42	57.22	18.80	61.99	19.29	58.78	19.36	58.23	13.93	59.99	10.07	54.76	10.31	58.50	12.16	56.46	10.04	56.73	11.29
	LN-Entropy	79.24	44.58	78.65	41.96	77.31	44.43	58.65	14.18	58.58	9.46	60.86	10.85	54.64	5.76	54.39	4.00	55.28	5.65	69.96	25.47	68.49	27.36	68.58	27.67
	SelfCheckBERTScore	78.21	51.43	76.07	45.26	74.67	50.67	73.78	47.63	73.41	37.54	73.53	45.73	77.67	37.00	78.00	24.29	78.10	41.30	63.30	32.36	64.41	28.48	63.64	28.97
	EigenScore	76.99	50.14	75.42	47.46	74.62	51.70	71.45	40.57	74.05	36.72	71.27	39.08	76.70	35.60	77.09	26.78	75.49	39.46	65.21	**34.19**	63.26	33.61	64.85	37.49
	LLM-Check (Hidden)	67.59	27.49	68.89	30.10	66.34	26.51	79.04	49.12	**90.37**	**51.73**	80.60	51.10	78.86	25.73	**87.60**	18.40	78.20	27.76	57.54	15.16	62.14	18.52	61.10	17.61
	LCIC(ours)	**85.62**	**61.91**	**85.00**	**60.52**	**83.37**	**62.50**	**80.84**	**53.48**	89.97	49.21	**82.92**	**54.55**	**85.10**	**48.90**	86.10	**37.00**	**85.10**	**53.50**	**73.81**	32.54	**73.81**	**38.98**	**75.06**	**46.60**
Llama3-8B	Perplexity	80.30	52.74	80.02	52.86	80.68	56.30	57.59	11.90	52.73	5.98	54.62	0.55	67.49	25.31	72.94	17.69	69.81	18.32	78.02	46.57	75.50	38.88	75.25	45.74
	Energy Score	65.02	27.78	61.50	22.06	65.81	32.31	54.56	12.36	60.31	15.93	57.34	22.93	50.22	5.03	52.05	0.76	51.46	4.07	60.74	17.43	54.32	8.25	55.43	15.06
	LN-Entropy	81.72	50.06	82.28	48.06	81.19	44.61	59.50	15.35	61.52	17.63	58.20	8.73	64.37	17.64	74.90	17.23	67.05	7.53	78.22	43.35	78.35	39.77	77.41	39.57
	SelfCheckBERTScore	81.64	58.97	81.07	52.96	81.56	55.38	75.79	47.86	72.54	35.86	72.71	42.95	74.24	38.10	**80.59**	23.23	78.48	37.93	79.95	52.51	77.19	40.64	77.39	49.15
	EigenScore	79.43	50.37	79.80	50.22	79.38	48.12	61.89	21.43	65.42	22.99	61.32	21.31	59.09	9.43	62.26	11.52	61.47	18.38	76.28	42.01	77.43	47.25	76.77	42.94
	LLM-Check (Hidden)	62.55	19.59	63.73	23.13	60.91	23.51	76.44	39.25	84.41	30.62	64.55	36.61	74.70	7.66	74.72	7.63	67.89	24.27	64.59	22.13	70.01	35.58	69.54	25.47
	LCIC(ours)	**85.70**	**62.53**	**83.96**	**55.79**	**85.06**	**58.27**	**80.14**	**56.41**	**88.33**	**53.13**	**81.57**	**47.22**	**79.49**	**40.00**	79.85	**29.62**	**81.46**	**50.20**	**82.44**	**54.27**	**80.93**	**49.54**	**79.32**	**51.12**

[a]EM: Exact Match. [b]SemSim: cosine similarity via BAAI/bge-m3 embeddings.

TABLE II
HALLUCINATION MITIGATION PERFORMANCE (ACCURACY %). BEST IN BOLD.

Model	Method	TriviaQA			CoQA			Chinese SimpleQA			DRCD		
		ROUGE	**EM**	**SemSim**	**ROUGE**	**EM**	**SemSim**	**ROUGE**	**EM**	**SemSim**	**ROUGE**	**EM**	**SemSim**
Qwen3-8B	Greedy Decoding	61.91	56.90	55.22	72.39	64.92	61.36	27.79	16.81	22.30	91.14	78.27	86.39
	DoLa	65.16	64.16	60.71	78.38	70.20	69.71	38.37	24.71	23.92	94.17	85.92	90.52
	Self-Reflection	66.75	58.92	57.32	77.24	68.03	65.39	31.64	21.79	25.35	94.21	83.63	88.62
	Self-Refine	63.19	58.07	57.44	70.09	62.47	60.70	29.98	22.84	29.49	91.19	79.71	85.76
	Integrative Decoding	**72.27**	**67.91**	66.91	82.68	74.35	**74.80**	37.19	24.71	32.61	95.11	84.16	90.25
	LCIC(ours)	70.45	66.75	**70.59**	**87.52**	**78.21**	73.93	**42.50**	**26.60**	**33.79**	**96.77**	**88.21**	**92.75**
Llama3-8B	Greedy Decoding	56.03	50.54	47.31	70.17	61.53	60.85	13.14	6.66	10.15	89.60	80.84	85.42
	DoLa	65.54	54.17	52.25	73.08	68.50	71.09	18.28	9.66	14.20	92.35	84.05	91.66
	Self-Reflection	64.38	56.87	55.63	75.87	66.07	67.76	17.19	8.51	13.99	90.20	82.39	89.32
	Self-Refine	57.34	52.83	48.52	69.16	60.86	61.97	15.24	7.22	11.17	88.10	81.78	86.01
	Integrative Decoding	**70.57**	58.36	51.08	80.32	69.54	74.83	19.64	10.77	16.90	94.13	87.38	93.97
	LCIC(ours)	69.85	**63.21**	**58.06**	**83.65**	**72.28**	**77.12**	**22.38**	**15.96**	**18.92**	**94.92**	**91.06**	**95.20**

b) Mitigation: Greedy decoding, DoLa [24], Self-Reflection [26], Self-Refine [25], and Integrative Decoding [35].

B. Main Results

a) Hallucination Detection: Table I demonstrates that LCIC consistently achieves superior discriminative performance across all datasets and evaluation criteria. On TriviaQA, LCIC attains 85.62 ROUGE-AUROC with Qwen3-8B, surpassing Perplexity (78.10) by +7.52 points and SelfCheckBERTScore (78.21) by +7.41 points. This advantage remains pronounced under the stricter EM criterion (85.00 vs. 74.38 for Perplexity), indicating that LCIC maintains robust separability even when penalizing minor lexical variations.

On Chinese SimpleQA with Llama3-8B, LCIC achieves 79.49 ROUGE-AUROC, significantly outperforming uncertainty-based proxies such as Perplexity (67.49) and LN-Entropy (64.37). Notably, LCIC achieves higher PCC than all baselines on TriviaQA (62.53 vs. 58.97 for SelfCheckBERTScore), demonstrating that the score correlates not only with binary correctness but also with graded answer quality, which is crucial for threshold-based deployment.

b) Hallucination Mitigation: Table II shows that LCIC-guided alignment consistently delivers the strongest mitigation performance across datasets. On CoQA with Qwen3-8B, LCIC achieves 87.52 ROUGE accuracy and 78.21 EM accuracy, substantially surpassing inference-time baselines including Self-Reflection (77.24; 68.03) and even Integrative Decoding (82.68), which requires additional decoding-time computation. The gains are most pronounced on context-grounded datasets: on DRCD, LCIC improves EM accuracy by +9.94 points over greedy decoding for Qwen3-8B, suggesting that optimizing toward internal stability encourages the model to better align with passage evidence.

On TriviaQA, Integrative Decoding retains a slight ROUGE advantage (72.27 vs. 70.45), suggesting that decoding heuristics can occasionally recover surface-level lexical matches on open-domain factual questions. However, LCIC leads on semantic similarity (70.59), indicating superior meaning preservation, while avoiding the inference overhead of multi-pass decoding strategies.

C. Ablation Study

1) LCIC Components and Aggregation Strategy: Table III isolates the contribution of each S_{LCIC} component on Qwen3-4B-Instruct (EM correctness). Single-signal variants are con-

sistently suboptimal: confidence-only achieves 75.60 AUC on TriviaQA, while attention-only reaches only 69.31. Removing layer contrast from the full configuration causes the largest degradation (-5.83 AUC on TriviaQA; -9.22 on Chinese SimpleQA), confirming that inter-layer predictive disagreement provides critical cues beyond confidence- and attention-based fragility.

Table IV compares aggregation strategies. Global averaging performs worst (-10.54/-11.79 AUC), validating that hallucination risk is indeed localized and diluted by full-sequence pooling. Local maximum improves over global averaging but remains sensitive to stochastic noise from single-token outliers. Sliding-window pooling provides moderate gains but is constrained by fixed granularity. CVaR pooling achieves the highest AUC (87.09; 87.98), demonstrating that tail-focused aggregation optimally balances sensitivity to high-risk spans with statistical stability through multi-token averaging.

TABLE III
ABLATION: COMPONENT CONTRIBUTIONS TO LCIC (AUC).

Components			TriviaQA		ChSimpleQA	
Conf.	**Focus**	**Contrast**	**AUC**	**Δ**	**AUC**	**Δ**
✓			75.60	−11.49	78.00	−9.98
	✓		69.31	−17.78	72.05	−15.93
		✓	74.12	−12.97	80.31	−7.67
✓	✓		81.26	−5.83	78.76	−9.22
✓	✓	✓	**87.09**	–	**87.98**	–

Δ: change vs. full LCIC; ✓: component included.

TABLE IV
ABLATION: AGGREGATION STRATEGY COMPARISON (AUC).

Aggregation	TriviaQA		ChSimpleQA	
	AUC	**Δ**	**AUC**	**Δ**
Global Mean	76.55	−10.54	76.19	−11.79
Sliding Window	78.63	−8.46	81.26	−6.72
Local Maximum	82.08	−5.01	84.71	−3.27
CVaR (ours)	**87.09**	–	**87.98**	–

Δ: change vs. CVaR.

2) Mitigation Framework and Reward Signal: Table V decomposes the contribution of training objectives. BoN-only SFT improves TriviaQA accuracy from 38.48% to 42.06% by fitting the highest-S_{LCIC} sample, but has limited exploration. GRPO-only reaches 46.71% by leveraging group-relative comparisons to discover better trajectories, but is sensitive to sampling variance. The joint objective achieves 49.93%, demonstrating that GRPO-driven exploration benefits from BoN anchoring as a stabilizing regularizer.

Table VI validates whether S_{LCIC} specifically provides superior training signal compared to generic unsupervised rewards. Probability-based rewards (Perplexity: 36.92%) perform worst, as optimizing for low uncertainty risks encouraging overconfident but incorrect outputs. Sample-consistency rewards (SelfCheckBERTScore: 45.28%) improve over probability signals but remain vulnerable to consistent hallucinations where multiple samples converge to the same erroneous answer. Against the strongest white-box baseline LLM-Check (45.63%), LCIC achieves a +4.30% accuracy gain (49.93%). This confirms that combining inter-layer divergence with internal consistency provides finer-grained supervision than coarse hidden-state statistics, validating the unified design of using S_{LCIC} for both detection and training-time alignment.

TABLE V
ABLATION: TRAINING STRATEGY FOR MITIGATION (ACC.%).

Method	GRPO loss	BoN loss	TriviaQA		ChSimpleQA	
			Acc.	**Δ**	**Acc.**	**Δ**
Baseline (no training)			38.48	−11.45	18.80	−10.50
SFT (best sample)		✓	42.06	−7.87	23.97	−5.33
RL only (no BoN)	✓		46.71	−3.22	25.16	−4.14
Full (joint)	✓	✓	**49.93**	–	**29.30**	–

✓: term included; Δ: vs. full method.

TABLE VI
ABLATION: REWARD SIGNAL FOR MITIGATION (ACC.%).

Reward Signal	Src.	TriviaQA		ChSimpleQA	
		Acc.	**Δ**	**Acc.**	**Δ**
Perplexity	Prob.	36.92	−13.01	20.08	−9.22
Energy Score	Prob.	35.13	−14.80	16.31	−12.99
LN-Entropy	Prob.	40.51	−9.42	22.59	−6.71
LexicalSim	Cons.	43.76	−6.17	26.85	−2.45
SelfCheckBERTScore	Cons.	45.28	−4.65	24.17	−5.13
LLM-Check (Attention)	Int.	41.37	−8.56	23.82	−5.48
LLM-Check (Hidden)	Int.	45.63	−4.30	25.63	−3.67
LCIC (Ours)	Int.	**49.93**	–	**29.30**	–

Src.: Prob.=probability, Cons.=consistency, Int.=internal states. Δ: vs. LCIC.

V. CONCLUSIONS

We propose LCIC, a unified tail-pooled internal-signal method for LLM hallucination governance that supports both online detection and training-time mitigation. By combining CVaR-style low-tail-pooled internal consistency (token confidence and attention focus) with high-tail layer-contrast divergence (mid–final JS divergence), LCIC derives a single forward-pass risk score S_{LCIC} that captures localized hallucination risk without external tools or repeated sampling, and reuses it as a GRPO+BoN reward for training-time alignment. Extensive experiments across four QA datasets and two open LLM families demonstrate strong unsupervised detection capability and consistent mitigation gains over decoding-time and self-refinement baselines, with ablations validating each component and the superiority of CVaR tail pooling. A current limitation is the reliance on white-box access and future work will explore surrogate strategies to extend LCIC to black-box settings.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under Grant 2023YFB3106501.

REFERENCES

[1] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[2] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[3] E. Durmus, H. He, and M. Diab, "Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization," *arXiv preprint arXiv:2005.03754*, 2020.

[4] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "Qafacteval: Improved qa-based factual consistency evaluation for summarization," in *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2022, pp. 2587–2601.

[5] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 076–12 100.

[6] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," in *Findings of the association for computational linguistics: ACL 2024*, 2024, pp. 3563–3578.

[7] P. Manakul, A. Liusie, and M. Gales, "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*, 2023, pp. 9004–9017.

[8] E. Quevedo, J. Y. Salazar, R. Koerner, P. Rivas, and T. Cerny, "Detecting hallucinations in large language model generation: A token probability approach," in *World Congress in Computer Science, Computer Engineering & Applied Computing*. Springer, 2024, pp. 154–173.

[9] G. Attanasio, D. Nozza, D. Hovy, and E. Baralis, "Entropy-based attention regularization frees unintended bias mitigation from lists," in *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics, 2022, pp. 1105–1119.

[10] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu, "Unsupervised real-time hallucination detection based on the internal states of large language models," in *Findings of the Association for Computational Linguistics, ACL*, 2024, pp. 14 379–14 391.

[11] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye, "INSIDE: llms' internal states retain the power of hallucination detection," in *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

[12] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "Llm-check: Investigating detection of hallucinations in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34 188–34 216, 2024.

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

[14] A. Goel, D. Schwartz, and Y. Qi, "Zero-knowledge llm hallucination detection and mitigation through fine-grained cross-model consistency," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025, pp. 1982–1999.

[15] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[16] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024," *URL https://arxiv. org/abs/2402.03300*, vol. 2, no. 3, p. 5, 2024.

[17] P. G. Sessa, R. Dadashi, L. Hussenot, J. Ferret, N. Vieillard, A. Ramé, B. Shariari, S. Perrin, A. Friesen, G. Cideron *et al.*, "Bond: Aligning llms with best-of-n distillation, 2024," *URL https://arxiv. org/abs/2407.14622*.

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[19] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," *Advances in neural information processing systems*, vol. 34, pp. 27 263–27 277, 2021.

[20] J. He, Z. Fan, S. Kuang, L. Xiaoqing, K. Song, Y. Zhou, and X. Qiu, "Fine: Filtering and improving noisy data elaborately with large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 8686–8707.

[21] S. Si, H. Zhao, G. Chen, C. Gao, Y. Bai, Z. Wang, K. An, K. Luo, C. Qian, F. Qi *et al.*, "Aligning large language models to follow instructions and hallucinate less via effective data filtering," *arXiv preprint arXiv:2502.07340*, 2025.

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[23] H. Zhang, S. Diao, Y. Lin, Y. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, and T. Zhang, "R-tuning: Instructing large language models to say 'i don't know'," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 7106–7132.

[24] Y. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," in *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

[25] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 534–46 594, 2023.

[26] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.

[27] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, 2017, pp. 1601–1611.

[28] Y. He, S. Li, J. Liu, Y. Tan, W. Wang, H. Huang, X. Bu, H. Guo, C. Hu, B. Zheng *et al.*, "Chinese simpleqa: A chinese factuality evaluation for large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 19 182–19 208.

[29] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[30] C. C. Shao, T. Liu, Y. Lai, Y. Tseng, and S. Tsai, "Drcd: a chinese machine reading comprehension dataset," *arXiv preprint arXiv:1806.00920*, 2018.

[31] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

[32] J. Ren, J. Luo, Y. Zhao, K. Krishna, M. Saleh, B. Lakshminarayanan, and P. J. Liu, "Out-of-distribution detection and selective generation for conditional language models," in *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

[33] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.

[34] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," *arXiv preprint arXiv:2002.07650*, 2020.

[35] Y. Cheng, X. Liang, Y. Gong, W. Xiao, S. Wang, Y. Zhang, W. Hou, K. Xu, W. Liu, W. Li *et al.*, "Integrative decoding: Improve factuality via implicit self-consistency, 2024," *URL: https://arxiv. org/abs/2410.01556. doi*, vol. 10.