LUMINA: Detecting Hallucinations in RAG System with Context–Knowledge Signals

Samuel Yeh[1,2], Sharon Li[1], Tanwi Mallick[2]
[1]Department of Computer Science, University of Wisconsin-Madison
{samuelyeh, sharonli}@cs.wisc.edu
[2]Argonne National Laboratory
tmallick@anl.gov

Abstract

Retrieval-Augmented Generation (RAG) aims to mitigate hallucinations in large language models (LLMs) by grounding responses in retrieved documents. Yet, RAG-based LLMs still hallucinate even when provided with correct and sufficient context. A growing line of work suggests that this stems from an imbalance between how models use external context and their internal knowledge, and several approaches have attempted to quantify these signals for hallucination detection. However, existing methods require extensive hyperparameter tuning, limiting their generalizability. We propose LUMINA, a novel framework that detects hallucinations in RAG systems through *context–knowledge signals*: external context utilization is quantified via distributional distance, while internal knowledge utilization is measured by tracking how predicted tokens evolve across transformer layers. We further introduce a framework for statistically validating these measurements. Experiments on common RAG hallucination benchmarks and four open-source LLMs show that LUMINA achieves consistently high AUROC and AUPRC scores, outperforming prior utilization-based methods by up to +13% AUROC on HalluRAG. Moreover, LUMINA remains robust under relaxed assumptions about retrieval quality and model matching, offering both effectiveness and practicality.

1 Introduction

Large language models (LLMs) are prone to hallucination, *i.e.*, producing responses that are factually incorrect, nonsensical, or not grounded in the input or available data, while still appearing fluent and plausible [1, 2, 3]. One commonly used strategy to mitigate hallucination is providing LLMs with relevant information retrieved from external knowledge bases, so-called Retrieval-Augmented Generation (RAG) [4, 5, 6]. However, despite having sufficient and relevant retrieved documents, RAG systems still have a chance to hallucinate and produce statements that are either unsupported or contradict the retrieved information [7, 8].

Recent work has shown that such failures often arise from conflicts between an LLM's internal knowledge and the retrieved external context [9]. In these cases, models tend to over-rely on internal knowledge regardless of correctness, undermining factual reliability [10, 11, 12, 13]. Inspired by this observation, recent approaches attempt to quantify hallucinations in RAG [14, 15, 16]. However, existing methods rely on mechanistic interpretability heuristics—such as selecting specific attention heads or transformer layers to achieve the optimal hallucination detection performance—which require heavy hyperparameter tuning and often fail to generalize across models and datasets.

To overcome these limitations, we propose LUMINA, a new framework for detecting hallucinations in RAG system through *context–knowledge signals*, namely the signals of external context utilization and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

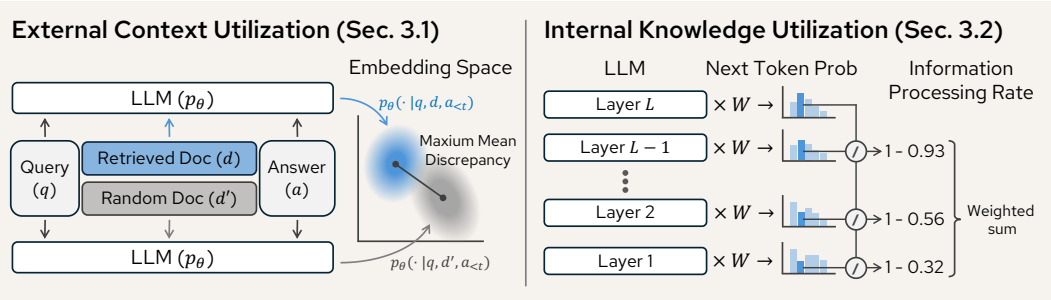

Figure 1: **The overview of LUMINA.** For external context utilization, we propose to measure the maximum mean discrepancy between two next token probability distributions conditioned on different documents. For internal knowledge utilization, we introduce the idea of information processing rate by looking at the ratio of the most probable output token's probability across transformer layers and use it to determine the amount of utilized internal knowledge when generating the next token.

internal knowledge utilization, as shown in Figure 1. Rather than targeting particular attention heads or layers, LUMINA measures these signals in a layer-agnostic manner, requiring less hyperparameter tuning. Specifically, for **external context utilization**, we measure the discrepancy between predictive distributions conditioned on retrieved documents *vs.* random documents. A larger discrepancy indicates that the LLM is more sensitive to semantic changes in documents when generating the answer, implying higher reliance on the external context. For **internal knowledge utilization**, we track how the model's internal states and token predictions evolve across layers: if the internal layers' predictions do not converge to the final output until later layers, it suggests more information is added during the layer-wise process, implying stronger reliance on internal knowledge. We further validate the soundness of our measurements through statistical hypothesis testing on verifiable implications, establishing a stronger link between the proposed scores and actual utilization.

We conduct extensive experiments on common RAG hallucination benchmarks and across four LLMs to evaluate the performance of LUMINA on hallucination detection. The results show that the hallucination score calculated with LUMINA outperforms existing methods by a significant margin. For example, LUMINA achieves more than 0.9 AUROC on the HalluRAG datasets across models, with improvements of up to **+13**% over prior state-of-the-art. Importantly, the decomposition into external context utilization and internal knowledge utilization provides interpretable insights: hallucinations are strongly associated with low external context scores and disproportionately high internal knowledge scores. We further demonstrate that LUMINA is robust across different retrieval settings. These results validate both the effectiveness and practicality of our framework.

Our key contributions are summarized as follows:

1. We propose LUMINA, a novel approach to quantify utilization of external context and internal knowledge for RAG-based hallucination detection.
2. We propose a framework to statistically validate LUMINA, showing that they align with the intended results.
3. We conduct extensive experiments and show that LUMINA outperforms both score-based and learning based methods in hallucination detection, establishing new *state-of-the-art*.

2 Preliminaries

2.1 Problem Formulation and Motivation

RAG systems aim to improve factuality by incorporating external documents into the generation process. However, hallucinations still occur when a model over-relies on its internal parametric knowledge and under-utilizes the retrieved external context. We provide a formal definition below.

Conjecture 1 (**External context vs. internal knowledge utilization**). *Let p_θ be an RAG-based LLM that takes a query q and retrieved documents d as inputs to generate a response a. Assume d is relevant to q and contains correct and sufficient information to respond to q. Denote $\mathcal{E}_{p_\theta}(a|q,d), \mathcal{I}_{p_\theta}(a|q,d) \in$*

$\mathbb{R}$ *be the signals of external context utilization and internal knowledge utilization of* p_θ*, respectively, when generating* a*. The response* a *is more likely to be hallucination if* $\mathcal{I}_{p_\theta}(a|q,d) \gg \mathcal{E}_{p_\theta}(a|q,d)$*.*

Definition 2.1 (**Hallucination in an RAG system**)**.** *Based on Conjecture 1, we define hallucination scores at both the token and response level. Specifically, for a generated answer* $a = (a_1, \ldots, a_T)$ *with* T *tokens, let* $\mathcal{E}_{p_\theta}(a_t|q,d,a_{<t}), \mathcal{I}_{p_\theta}(a_t|q,d,a_{<t}) \in \mathbb{R}$ *be the signals of external context utilization and internal knowledge utilization of* p_θ *when generating the token* a_t*, respectively. The token-level hallucination score of* a_t *is defined as*

$$\mathcal{H}_t(a_t|q,d,a_{<t}) := \lambda \cdot \mathcal{I}_{p_\theta}(a_t|q,d,a_{<t}) - (1-\lambda) \cdot \mathcal{E}_{p_\theta}(a_t|q,d,a_{<t}), \tag{1}$$

where λ *is a hyperparameter. Similarly, the response-level hallucination score of the response* a *is defined as the average of the token-level hallucination scores, i.e.,*

$$\mathcal{H}_r(a|q,d) := \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}_t(a_t|q,d,a_{<t}). \tag{2}$$

In this paper, we focus on the core question: ***How to quantify the utilization of external context and internal knowledge?***

2.2 Related Work

Prior works have attempted to quantify $\mathcal{E}_{p_\theta}(a_t|q,d,a_{<t})$ and $\mathcal{I}_{p_\theta}(a_t|q,d,a_{<t})$ using empirical metrics [14, 15]. For example, Sun et al. [14] proposed ReDeEP, which measures external context utilization through cosine similarity between the generated token and tokens in context that have high attention weights w.r.t. certain attention heads. For internal knowledge utilization, it measures the Jensen-Shannon (JS) divergence between the hidden states before/after the FFN layer of certain transformer layers. The success of ReDeEP on some RAG hallucination detection datasets validates the idea of Conjecture 1. Wang [15] combine the idea of ReDeEP with semantic entropy probes (SEP) [17]. They quantified external context utilization by measuring the semantic correlation between the semantic entropy of the generated token and attended tokens in the context. For internal knowledge utilization, they measured the absolute difference between the semantic entropy corresponding to hidden states before and after the FFN layer.

Although these approaches effectively detect hallucinations in the RAG system, they have two major limitations. First, these approaches require selecting specific attention heads and transformer layers to compute the external context score and internal knowledge score. However, the selection process is non-trivial and requires extensive hyperparameter tuning. In addition, these hyperparameters are dataset and model-specific, limiting the generalizability across different datasets and models. Another limitation is that although these works demonstrated the correlation between their proposed scores and hallucination, they did not validate whether the scores truly reflect the utilization of external context and internal knowledge.

3 Methodology

Overview. To overcome the limitations of prior empirical approaches, we introduce LUMINA, a new framework for quantifying both external context and internal knowledge utilization. In Section 3.1 and Section 3.2, we formalize the quantification of the two signals, which will be combined to compute the final hallucination score. In Section 3.3, we propose to validate the soundness of LUMINA through extensive hypothesis testing, addressing the challenges of score validation in previous works.

3.1 Quantifying External Context Utilization

To measure LLM's external context utilization, our key idea is to assess its sensitivity to semantic changes in the input documents. If the LLM effectively incorporates the external context to generate a response, then replacing relevant documents with random ones should noticeably change the token probability distribution. Formally, we propose the following measurement:

Measurement 1 (**External context utilization**)**.** *Let* a *be an LLM-generated answer to query* q *with retrieved documents* d *as input. Assume* d *is relevant to* q *and contains correct and sufficient*

information to respond to q. Let d' be a subset of random documents irrelevant to q. The model's predictive distribution over tokens induces two (approximated) distributions over embeddings:

$$P(E_v) = p_\theta(v \mid q, d, a_{<t}), \quad Q(E_v) = p_\theta(v \mid q, d', a_{<t}), \tag{3}$$

where each token $v \in \mathcal{V}$ in the vocabulary space is associated with an embedding $E_v \in \mathbb{R}^D$. Then, the degree to which the model uses external context for generating token a_t is reflected in the divergence between the two distributions conditioned on d versus d':

$$\mathcal{E}_{p_\theta}(a_t|q, d, a_{<t}) := \Delta(P, Q), \tag{4}$$

where $\Delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is a distance function between two probability distributions.

Note that we adopt $P(E_v)$ and $Q(E_v)$ as proxies to approximate the ground truth embedding distribution, as it is challenging to estimate it over the high-dimensional vector space. We instantiate Δ with Maximum Mean Discrepancy (MMD), which measures the distance of two probability distributions by mapping them into a Reproducing Kernel Hilbert Space.

Definition 3.1 (**Maximum Mean Discrepancy** [18]). *Given a positive semi-definite kernel function k, the squared MMD between two probability distributions P and Q is defined as*

$$\mathrm{MMD}_k^2(P, Q) := \mathbb{E}_{\boldsymbol{A},\boldsymbol{A}' \sim P}[k(\boldsymbol{A}, \boldsymbol{A}')] + \mathbb{E}_{\boldsymbol{B},\boldsymbol{B}' \sim Q}[k(\boldsymbol{B}, \boldsymbol{B}')] - 2\mathbb{E}_{\boldsymbol{A}\sim P,\boldsymbol{B}\sim Q}[k(\boldsymbol{A}, \boldsymbol{B})], \tag{5}$$

where $\boldsymbol{A}, \boldsymbol{A}'$ are i.i.d. vectors randomly sampled from P and $\boldsymbol{B}, \boldsymbol{B}'$ are sampled from Q.

This metric provides us with a non-parametric and LLM-agnostic way to quantify the utilization of external context, making it generalizable to different models and datasets.

By rewriting MMD with P and Q we defined in Eq. (3) over token embeddings, we obtain:

$$\begin{aligned}\mathcal{E}_{p_\theta}(a_t|q, d, a_{<t}) := &\sum_{u,v\in\mathcal{V}} P(E_u)P(E_v)k(E_u, E_v) + \sum_{u,v\in\mathcal{V}} Q(E_u)Q(E_v)k(E_u, E_v) \\ &- 2\sum_{u,v\in\mathcal{V}} P(E_u)Q(E_v)k(E_u, E_v).\end{aligned} \tag{6}$$

We adopt the cosine kernel:

$$k_{\cos}(E_u, E_v) := \frac{1}{2}\left(1 + \frac{E_u^T E_v}{\|E_u\|_2 \|E_v\|_2}\right). \tag{7}$$

Note that the cosine kernel acts equivalent to computing cosine similarity between two token embeddings, which is commonly used to measure the semantic similarity of two pieces of text. In Section 4.4, we experiment with alternative kernels such as the Gaussian kernel, and we show that our method is not sensitive to the choice of kernels.

3.2 Quantifying Internal Knowledge Utilization

To quantify the utilization of internal knowledge, we focus on the signals in internal states of an LLM. Specifically, a transformer-based autoregressive LLM has multiple layers, through which information is gradually added into a residual stream that flows from the input layer to the output layer, shaping the output token representation and probability distribution [19]. Studies have found that by projecting the hidden state of each layer to the token representation space, we can interpret what an LLM believes after the process of each layer [20]. In addition, via logit lens [20], studies have identified the saturation event in an LLM, *i.e.*, the top-k prediction of the LLM remains constant in all subsequent layers after a certain layer called the k-th saturation layer [19, 21].

Inspired by these observations, we propose a metric that quantifies how actively the model updates its predictions across layers. Formally, we define the rate of information processing below.

Definition 3.2 (**Information processing rate**). *Given an LLM p_θ with L layers, which takes $x_{<t}$ as the input and generate the next token x_t, we denote $x_{t,1} := \arg\max_v p_\theta(v|x_{<t})$ as the most probable next token and $h_{t,l} \in \mathbb{R}^D$ as the l-th layer hidden state when generating x_t. Let $f : \mathbb{R}^D \to \mathcal{P}$ be a projection from a hidden state to a probability distribution over the vocabulary $\mathcal{V}$. The information processing rate of p_θ conditioned on $x_{<t}$ is defined as*

$$\mathcal{R}_{p_\theta}(x_{<t}) := \frac{\sum_{l=1}^{L-1}\left(1 - \min\left\{\frac{[f(h_{t,l})]_{x_{t,1}}}{p_\theta(x_{t,1}|x_{<t})}, 1\right\}\right) \cdot l}{\sum_{l'=1}^{L-1} \frac{l'}{H(f(h_{t,l'}))}}, \tag{8}$$

where $H(\cdot)$ is the entropy function, and f is the logit lens [20] that projects the hidden state of each layer to logits using the LayerNorm and the unembedding matrix $\boldsymbol{W}$, *i.e.*,

$$\text{LogitLens}(h) := \text{LayerNorm}(h)\boldsymbol{W}, \quad f(\cdot) := \text{Softmax}(\text{LogitLens}(\cdot)). \tag{9}$$

Specifically, $\mathcal{R}_{p_\theta}(x_{<t})$ captures two key elements: (1) The numerator measures the extent to which each layer's prediction for the most probable token differs from the final output, weighted by layer depth to emphasize later-layer processing. When $\frac{[f(h_{t,l})]_{x_{t,1}}}{p_\theta(x_{t,1}|x_{<t})}$ is small, it indicates the layer has not yet converged to the final prediction, suggesting active information processing. (2) The denominator provides adaptive normalization based on each layer's prediction uncertainty (entropy), giving higher relative weight to layers that exhibit confident, decisive processing patterns. Given this definition, we attribute the utilization of internal knowledge to the 1st information processing rate and propose the following measurement:

Measurement 2 (Internal knowledge utilization). *An LLM is considered to be more heavily utilizing its internal knowledge to generate a_t when it exhibits a higher information processing rate. Specifically, we propose that the internal knowledge utilization of an LLM to generate a_t given q and d can be measured as*

$$\mathcal{I}_{p_\theta}(a_t|q, d, a_{<t}) := \mathcal{R}_{p_\theta}(q, d, a_{<t}). \tag{10}$$

3.3 Statistical Validation of the Measurement

In this section, we validate the soundness of our approach. Previous work such as Sun et al. [14] primarily verified whether their scores have a causal relationship with hallucination but failed to show the relationship between the scores and actual external context/internal knowledge utilization. To address this, we directly assess whether our measurements capture the intended notion of utilization. Specifically, we derive verifiable implications that must hold if our proposed measurements are valid. We then use the proposed score to verify these implications with statistical hypothesis testing. If the proposed score passes all tests, the score reflects the corresponding utilization.

External context utilization. To validate Measurement 1, we examine the following implications:

H1. If Measurement 1 is valid, then $\mathcal{E}_{p_\theta}(a_t|q, d, a_{<t}) > \mathcal{E}_{p_\theta}(a'_t|q, \varnothing, a'_{<t})$. That is, generations with retrieved documents have stronger external context utilization than generations without.

H2. If Measurement 1 is valid, then $\mathcal{E}_{p_\theta}(a_t|q_{\text{sum}}, d_{\text{sum}}, a_{<t}) > \mathcal{E}_{p_\theta}(a_t|q_{\text{QA}}, d_{\text{QA}}, a_{<t})$. That is, summarization tasks should exhibit higher external context utilization than question answering.

Internal knowledge utilization. To validate Measurement 2, we examine the following:

H3. If Measurement 2 is valid, then $\mathcal{R}^1_{p_\theta}(q, \varnothing, a_{<t}) > \mathcal{R}^1_{p_\theta}(q, d, a_{<t})$. That is, generating an answer without retrieved documents requires more internal knowledge than with retrieved documents.

H4. If Measurement 2 is valid, then $\mathcal{R}^1_{p_\theta}(q_{\text{D2T}}, d_{\text{D2T}}, a_{<t}) > \mathcal{R}^1_{p_\theta}(q_{\text{sum}}, d_{\text{sum}}, a_{<t})$. In other words, data-to-text generation requires more internal knowledge than summarization.

Table 1: **All the hypotheses pass the statistical tests.** For H1, H2, H4, we report one-tailed t-statistic; for H3, we report paired-sample one-tailed t-statistic. All four implications reject their null hypothesis, validating the soundness of LUMINA. Note that the tests are run with > 65k tokens and the magnitude of the t-statistic means how easy we can distinguish the two distributions. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

LLM	H1	H2	H3	H4
Llama2-7B	79.85***	27.67***	101.20***	15.36***
Llama2-13B	73.49***	20.51***	91.00***	7.71***
Llama3-8B	94.15***	6.35***	102.44***	15.85***
Mistral-7B	88.70***	6.21***	109.26***	9.69***

To examine **H1**, we utilize data in the QA set of RAGTruth [7]. We use the original data to compute $\mathcal{E}_{p_\theta}(a_t|q, d, a_{<t})$, and generate additional answers without providing retrieved documents as a' to compute $\mathcal{E}_{p_\theta}(a'_t|q, \varnothing, a'_{<t})$. For **H2**, we utilize the Summary and QA set of RAGTruth; for **H4**, the Summary and Data2Text set; and for **H3**, the entire RAGTruth dataset. We test the hypotheses

with four different instruction-tuned LLMs, including Llama2-{7B, 13B} [22], Llama3-8B [23], and Mistral-7B [24]. Results in Table 1 indicate that all four implications reject their null hypothesis, validating our measurements for external context utilization and internal knowledge utilization.

4 Experiments

4.1 Experimental Settings

Baselines. We compare LUMINA with baselines across 8 different hallucination detection strategies: (1) **Uncertainty-based**, which detects hallucination by estimating uncertainty via token-level probability or entropy. Baselines of this category include Perplexity [25], LN-Entropy [26], and Focus [27]. (2) **Cross-sample consistency**, which detects hallucination by sampling multiple responses for a query and measuring their (logic/semantic) consistency. Approaches include SelfCKGPT [28] and EigenScore [29]. (3) **Verbalization**, which detects hallucinations by prompting another LLM to score the correctness of the answer. Approaches include P(True) [30] and RefChecker [31]. (4) **Utilization of external context and internal knowledge**, which decouples these two signals via findings in the study of mechanistic interpretability. Baseline of this category is ReDeEP [14]. Details of each baseline are introduced in Appendix B.

LLMs. To demonstrate the generalizability of LUMINA, we conduct experiments with four open-sourced LLMs, including Llama2-{7B, 13B}, Llama3-8B, and Mistral-7B. Specifically, each LLM is used to detect hallucinations in responses generated by the same model. We also report the performance of proxy LLM setting, *i.e.*, using one LLM to detect hallucinations in responses generated by another model, in Sec. 4.3. All LLMs are the instruction-tuned version.

Datasets. Experiments are conducted on two representative RAG hallucination detection benchmarks: **RAGTruth** [7], the first high-quality RAG hallucination detection dataset, consisting of three types of RAG tasks, including question answering, data-to-text writing, and news summarization. **HalluRAG** [8], a dataset of free-form question answering in an RAG setting. Details of these datasets are introduced in Appendix C.

Evaluation metrics. We measure the performance with three metrics: **AUROC**, **AUPRC**, and **Pearson's correlation coefficient** (PCC). AUPRC captures precision-recall trade-offs, while AUROC evaluates the trade-offs between true and false positive rates. These metrics are threshold-agnostic and better suited for comparing scoring-based methods. We also report the optimal precision, recall, and F1 score ($\text{Prec}_{\text{Opt}}, \text{Recall}_{\text{Opt}}, \text{F1}_{\text{Opt}}$) in Appendix E.1, where F1_{Opt} is the optimal F1 score among all possible threshold and Prec_{Opt} and $\text{Recall}_{\text{Opt}}$ are corresponding Precision and Recall.

Implementation details. We adopt $\lambda = 0.5$ to compute Eq. (1) as ablations show that balancing the scores of external context and internal knowledge yields relatively strong performance (see Appendix E.3 for detailed ablations). Other implementation details and computational resources of LUMINA are reported in Appendix D and G, respectively.

4.2 Main Results

LUMINA achieves state-of-the-art performance. Table 2 summarizes the experimental comparison across methods. The results show that LUMINA has a consistently high performance across datasets and LLMs. In particular, it almost always outperforms ReDeEP, the previous attempt of measuring the utilization of external context and internal knowledge to detect hallucinations. The gap between them is particularly large on the HalluRAG dataset. Noticeably, LUMINA achieves more than 0.9 AUROC on the HalluRAG dataset across models, outperforming the baselines by a substantial margin. We further conduct an error analysis to see when and why LUMINA fails. Specifically, we sample 20 false-negative and false-positive cases from the RAGTruth dataset, respectively, and qualitatively analyze the reason of errors. The result reveals that most of the errors stem from incorrect labels and low-quality retrieved documents of the dataset, suggesting a potentially higher performance in a setting with high-quality data. The details of this analysis can be found in Appendix F.

Table 2: **LUMINA consistently achieves a high performance across datasets and LLMs.** The highest scores are set in **bold**. Note that HalluRAG dataset does not contain responses generated by Llama3-8B.

		RAGTruth			HalluRAG		
LLM	Approach	AUROC ↑	PCC ↑	AUPRC ↑	AUROC ↑	PCC ↑	AUPRC ↑
Llama2-7B	Perplexity	0.5103	-0.0118	0.4836	0.4610	-0.0673	0.2332
	LN-Entropy	0.6964	0.3318	0.6615	0.9102	0.5133	0.6812
	Focus	0.5633	0.0811	0.5386	0.5652	0.2415	0.3844
	SelfCKGPT	0.4787	-0.0279	0.4859	0.4669	-0.0070	0.2377
	EigenScore	0.5454	0.0717	0.5183	0.6720	0.2705	0.4470
	P(True)	0.5197	0.0404	0.5334	0.5847	0.1143	0.2976
	RefChecker	0.5869	0.1751	0.6827	0.4907	-0.0255	0.2750
	ReDeEP	0.7273	0.3859	0.6971	0.6771	0.1468	0.3378
	LUMINA	**0.7646**	**0.4546**	**0.7491**	**0.9153**	**0.6554**	**0.7572**
Llama2-13B	Perplexity	0.4539	-0.1020	0.3993	0.2548	-0.2366	0.0944
	LN-Entropy	0.7677	0.4446	0.6838	0.7826	0.3262	0.3567
	Focus	0.5451	0.0130	0.4603	0.6739	0.2563	0.3181
	SelfCKGPT	0.4545	-0.0835	0.4106	0.7729	0.2640	0.3029
	EigenScore	0.6329	0.2080	0.5202	0.7862	0.4250	0.4867
	P(True)	0.7543	0.3821	0.7418	0.6914	0.2480	0.2146
	RefChecker	0.6363	0.2723	0.6988	0.5670	0.1390	0.3169
	ReDeEP	0.8055	0.5195	0.7792	0.7645	0.2705	0.3001
	LUMINA	**0.8569**	**0.6041**	**0.8436**	**0.9166**	**0.6044**	**0.8497**
Llama3-8B	Perplexity	0.7130	0.3568	0.7183	-	-	-
	LN-Entropy	0.7072	0.3500	0.7109	-	-	-
	Focus	0.5258	0.0375	0.5380	-	-	-
	SelfCKGPT	0.5339	0.0491	0.5550	-	-	-
	EigenScore	0.6001	0.1774	0.5824	-	-	-
	P(True)	0.5407	0.0928	0.5502	-	-	-
	RefChecker	0.5718	0.1494	0.6874	-	-	-
	ReDeEP	**0.7495**	**0.4458**	0.7817	-	-	-
	LUMINA	0.7446	0.4236	**0.7874**	-	-	-
Mistral-7B	Perplexity	0.6200	0.1463	0.6106	0.5362	-0.0264	0.1261
	LN-Entropy	0.7607	0.4386	0.7377	0.9188	0.6076	0.7347
	Focus	**0.7803**	0.4188	0.7647	0.8565	0.4318	0.4219
	SelfCKGPT	0.5680	0.0812	0.5698	0.8275	0.5552	0.6098
	EigenScore	0.5642	0.1006	0.5637	0.8652	0.6411	0.7337
	P(True)	0.7530	0.4334	0.7494	0.5899	0.0886	0.1771
	RefChecker	0.6017	0.2047	0.7303	0.5065	0.0153	0.1784
	ReDeEP	0.7615	0.4613	**0.8133**	0.7870	0.2611	0.3516
	LUMINA	0.7685	**0.4623**	0.7942	**0.9899**	**0.7529**	**0.9431**

Comparison with supervised approach. We also compare LUMINA with SAPLMA [32], a supervised approach that trained a binary classifier on the last token hidden states to detect hallucination. Since our method is unsupervised in nature and does not rely on labeled data, the supervised baseline can be viewed as a performance upper bound. Results in Appendix E.2 show that LUMINA achieves a competitive performance against SAPLMA and even sometimes outperforms it, all without any training, highlighting both its supreme performance and ease of deployment.

4.3 Relaxing Assumptions

In Section 3, we implicitly make two assumptions: 1) ***perfect context assumption:*** we assume the retrieved documents d are correct, sufficient, and relevant to the query. 2) ***same LLM assumption:*** we assume the LLM used to compute the external context score and internal knowledge score is the same as the LLM used to generate responses. These two assumptions are usually introduced in other hallucination detection works as well [27, 14]. Unfortunately, they are often strong and have a significant impact on the performance, limiting the usability of these methods (such as for open-sourced model-generated responses only). In this section, we investigate the performance of LUMINA when relaxing these two assumptions, showing the robustness of LUMINA.

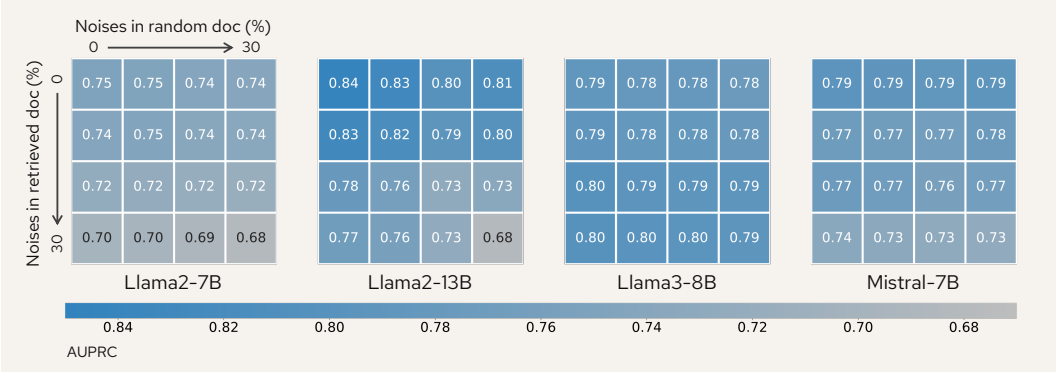

Figure 2: **Noises in context do not largely degrade the performance of LUMINA.** We add $0 \sim 30\%$ noises to the retrieved documents and random documents and evaluate the hallucination detection performance. The experiment is conducted on the RAGTruth dataset.

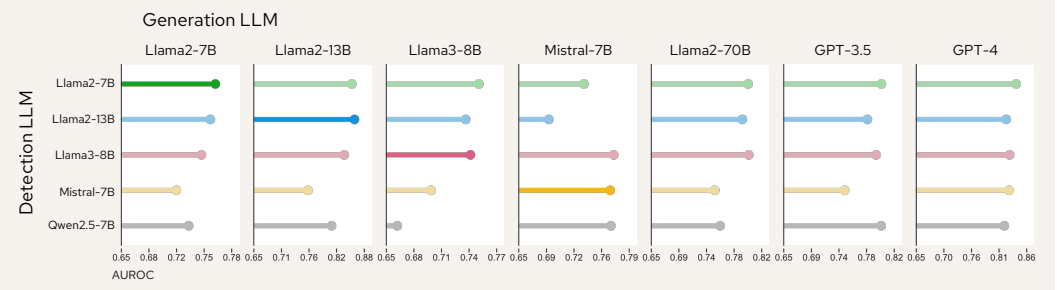

Figure 3: **The "same LLM" setting is not essential for LUMINA to achieve the optimal performance.** On the RAGTruth dataset, for each set of responses generated by the same LLM, we apply LUMINA with a different base LLM to detect hallucination. Bars in more saturated shades indicate settings where the same LLM is used for both generation and detection.

Relaxing perfect context assumption. We relax this assumption by gradually injecting noise into the retrieved documents d and random documents d'. Specifically, for the assumption on retrieved documents, we randomly remove $\{0\%, 10\%, 20\%, 30\%\}$ sentences from d. And for the assumption on the random documents, we randomly add $\{0\%, 10\%, 20\%, 30\%\}$ sentences from d to d'. Figure 2 shows the AUPRC of all noise injection combinations on the RAGTruth dataset. The result shows that except Llama2-13B, which has a > 0.1 performance drop after injecting noises, LUMINA with other LLMs yields stable performance. Furthermore, after removing sentences from retrieved documents, LUMINA with Llama3-8B even achieves a higher AUPRC. These results demonstrate the robustness of LUMINA against context noises.

Relaxing the same LLM assumption. We relax this assumption by using different LLMs to compute the scores for a response. Specifically, we use Llama2-7B, Llama2-13B, Llama3-8B, Mistral-7B, and Qwen2.5-7B [33] to detect hallucination on the RAGTruth dataset, which contains responses generated by Llama2-7B, Llama2-13B, Llama2-70B, Llama3-8B, Mistral-7B, GPT-3.5, and GPT-4. Figure 3 shows AUROC across different generator-detector LLM pairs.

The results show that the same model setting is not always necessary. Specifically, Llama2-7B achieves a comparable or higher AUROC than Llama3-8B on Llama3-8B responses. Moreover, LUMINA with Llama2-7B and Llama3-8B has stable performance across different generation LLMs. Overall, LUMINA demonstrates a plausible solution for generation LLM-agnostic hallucination detection, which is more practical in real-world scenarios.

4.4 Ablation Study

Impact of external context & internal knowledge. Our final hallucination score is the combination of the external context score and internal knowledge score. To obtain more insights into how each

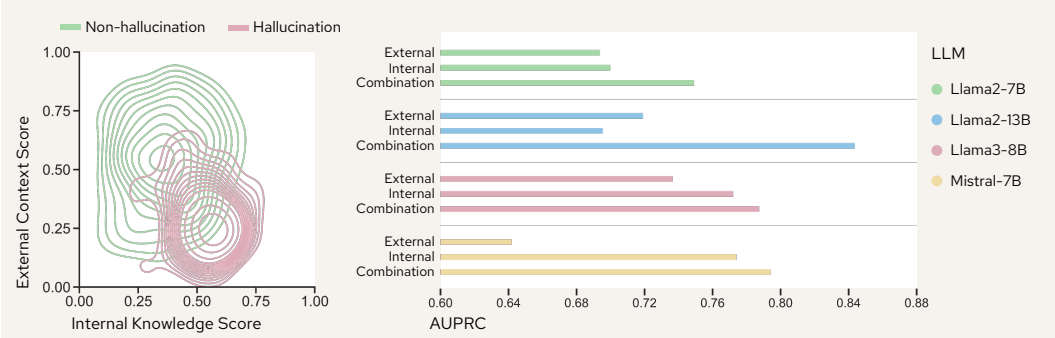

Figure 4: **Combining scores of external context and internal knowledge boosts the hallucination detection performance.** Left: 2D kernel density estimation (KDE) of the distribution of external context score and internal knowledge score of Llama2-13B responses on the RAGTruth dataset. Right: Hallucination detection performance with external/internal score only, as well as the performance of their combination.

component contributes to the final score, we ablate on the components by considering only the external context score and internal knowledge score. The right plot of Figure 4 shows that combining scores of external context and internal knowledge achieves the highest AUPRC on the RAGTruth dataset for every LLM. For example, on Llama2-13B, the combination leads to more than 10% improvement. This observation justifies the effectiveness of the hallucination score introduced in Definition 2.1. In addition, the left plot of Figure 4 shows that a response generated by Llama2-13B is more likely to be hallucination if it has a high internal knowledge score and a low external context score. This observation validates Conjecture 1 and suggests that Eq. (1) does not imply an objective function that forces LLM only using external context to answer questions. Instead, it suggests that the internal knowledge utilization should be grounded in an external context to achieve a reliable generation. Further experiments are shown in Appendix E.3.

Impact of kernel selection. We ablate on the selection of kernel $k \in \{\text{Cosine}, \text{RBF}_{0.5}, \text{RBF}_{0.7}, \text{RBF}_1, \text{RBF}_2, \text{RBF}_3\}$, where RBF_σ is a RBF kernel, *i.e.*, $\text{RBF}_\sigma(E_u, E_v) := \exp\left(-\frac{\|E_u - E_v\|_2^2}{2\sigma^2}\right)$. Figure 5 shows the AUPRC of different kernels on the RAGTruth dataset. The results show that the optimal setting of the RBF kernel has a similar performance to the cosine kernel, suggesting our external context score is insensitive to the kernel selection. We default to the cosine kernel as it is less dependent on hyperparameters, making it easy to use in practice.

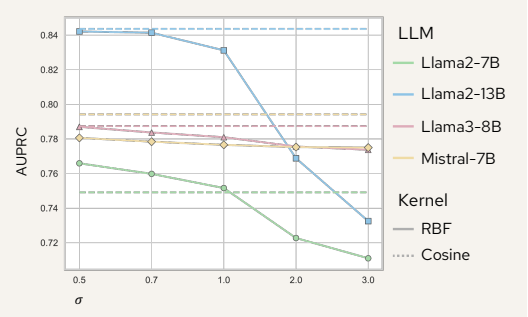

Figure 5: **MMD with cosine kernel performs similarly or better than with RBF kernel.**

5 Conclusion

In this paper, we introduce LUMINA, a novel approach to quantify the utilization of external context and internal knowledge. These context–knowledge signals provide a principled way to assess how LLMs balance retrieved evidence against their own parametric knowledge during generation. Experimental results on common benchmarks across four LLMs demonstrate that LUMINA has a consistently high performance on hallucination detection for RAG-based generations, outperforming prior attempts of quantifying external context and internal knowledge utilization, and being competitive with supervised hallucination detection models. Analyses also show that LUMINA is robust against noise in retrieved documents and can be generalized to the proxy LLM setting, demonstrating its usability in real-world scenarios.

Ethics Statement

This work introduces LUMINA, a novel way to estimate the utilization of external context and internal knowledge when an LLM generates responses with the RAG setup. LUMINA significantly improves the performance of hallucination detection, which will help increase the reliability of RAG systems in real-world deployments and reduce the risk of sharing misinformation. Through a deeper analysis of LUMINA in the future, researchers may better understand how LLMs utilize external context and internal knowledge to generate responses. Such findings will help the community design approaches to mitigate hallucinations and create a more reliable AI system.

Reproducibility Statement

We provide all details of the implementation of LUMINA in Appendix D, including the approximation of MMD, the selection of kernel, and the choice of random documents for measuring external context score, as well as the calibration of internal knowledge score. In Sec. 4.1, we illustrate the experimental settings, including baselines, datasets, LLMs, and evaluation metrics. The details of baselines and datasets are further provided in Appendix B and C, respectively. Furthermore, we provide the codebase of LUMINA at https://anonymous.4open.science/r/LUMINA-E71B. These comprehensive reports will help future studies easily reproduce our experiments.

Acknowledgement

We thank Chongdae Oh and Seongheon Park for their valuable feedback on the draft. This work was supported by Laboratory Directed Research and Development funding from Argonne National Laboratory, provided by the Office of Science, U.S. Department of Energy under Contract No. DE-AC02-06CH11357. Sharon Li acknowledges support from AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Schmidt Sciences Foundation, Open Philanthropy, and Alfred P. Sloan Fellowship. The funders had no role in study design, data collection, data analysis or interpretation, or manuscript preparation.

Bibliography

[1] Junliang Luo, Tianyu Li, Di Wu, Michael R. M. Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*, 2024.

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 2024. ISSN 1046-8188.

[3] Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer LLM latents for hallucination detection. In *Forty-second International Conference on Machine Learning*, 2025.

[4] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[5] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, 2024. ISBN 9798400704901.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

[7] 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[8] Fabian Ridder and Malte Schilling. The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm's internal states. *arXiv preprint arXiv:2412.17056*, 2025.

[9] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[10] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[11] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[12] Kaiser Sun, Fan Bai, and Mark Dredze. What is seen cannot be unseen: The disruptive effect of knowledge conflict on large language models. *arXiv preprint arXiv:2506.06485*, 2025.

[13] Khurram Yamin, Gaurav Ghosal, and Bryan Wilder. Llms struggle to perform counterfactual reasoning with parametric knowledge. *arXiv preprint arXiv:2506.15732*, 2025.

[14] ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025.

[15] Lei Wang. Seredeep: Hallucination detection in retrieval-augmented models via semantic entropy and context-parameter fusion. *arXiv preprint arXiv:2505.07528*, 2025.

[16] Yufei Tao, Adam Hiatt, Rahul Seetharaman, and Ameeta Agrawal. "lost-in-the-later": Framework for quantifying contextual grounding in large language models. *arXiv preprint arXiv:2507.05424*, 2025.

[17] Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[18] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. ISSN 1533-7928.

[19] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022.

[20] nostalgebraist. interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

[21] Daria Lioubashevski, Tomer Schlank, Gabriel Stanovsky, and Ariel Goldstein. Looking beyond the top-1: Transformers determine top tokens in order. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[22] AI @ Meta Llama Team. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[23] AI @ Meta Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[24] 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. *arXiv preprint arXiv:2310.06825*, 2023.

[25] 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. In *The Eleventh International Conference on Learning Representations*, 2023.

[26] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

[27] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[28] Potsawee Manakul, Adian Liusie, and Mark 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.

[29] 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. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] 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. *arXiv preprint arXiv:2207.05221*, 2022.

[31] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*, 2024.

[32] 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*, 2023.

[33] Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

APPENDIX

CONTENTS

A Broader Impacts **13**

B Details of Baselines **13**

C Details of Datasets **14**

D Implementation Details of LUMINA **14**

E Additional Experimental Results **15**

E.1 Evaluation with Other Metrics 15

E.2 Compare with supervised baselines 16

E.3 Performance with Hyperparameter Tuning 16

F Error Analysis **17**

G Computational Resources **17**

A Broader Impacts

Beyond hallucination detection, LUMINA has broader impacts in interpretability and LLM understanding. Specifically, our proposed score validation framework in Sec. 3.3 suggests a novel way to empirically validate the finding of mechanistic interpretability, which can be used to highlight the soundness of proposed hypotheses. In addition, our proposed information processing rate in Sec. 3.2 presents a new lens for examining the internal states of LLMs. Deeper investigation of this measure could help the community better characterize how LLMs reason and leverage internal knowledge, potentially leading to more reliable training and inference processes. While our experiments focus on using LUMINA for hallucination detection, its utility extends further. For instance, it could inform the design of new training objectives or decoding algorithms aimed at mitigating hallucinations, ultimately making LLMs more reliable and trustworthy.

B Details of Baselines

(1) Token-level uncertainty:

- **Perplexity:** This approach measured the perplexity of the generated response as uncertainty and to detect hallucinations.
- **LN-Entropy:** This approach measured sequence-level uncertainty with entropy normalized by sequence length. A higher entropy indicates greater uncertainty and a higher likelihood of hallucinations.
- **Focus:** This approach used entropy and token probability as a based score, and calibrated it by focusing only on key informative tokens and propagating the score according to the attention weight.

(2) Cross-sample consistency:

- **SelfCKGPT:** This approach sampled multiple responses and used an NLI model to check the logistic consistency between the target generation and additional samples. In our experiment, we follow the setting of Manakul et al. [28] to set the sample size as 20.

- **EigenScore:** Similar to SelfCKGPT, this approach sampled multiple responses and checked the semantic consistency between the additional samples and the target generation through measuring the eigenvalues of responses' covariance matrix. In our experiment, we set the sample size as 20.

(3) Verbalization:

- **P(True):** This approach prompted an LLM with the generated answer and asked whether the LLM think the answer is true. The approach then estimated the probability of the "Yes" generated by the LLM.
- **RefChecker:** This approach prompted an LLM to extract claims from generation, and prompted another LLM to verify the logical consistency between each claim and reference documents. In our experiment, we use `dongyru/Mistral-7B-Claim-Extractor`, the model finetuned by Hu et al. [31], to extract claims.

(4) Utilization of external context and internal knowledge:

- **ReDeEP:** For external context utilization, ReDeEP measured the cosine similarity between the generated token and topK attended tokens in retrieved documents. For internal knowledge utilization, it measured the JS divergence of the vocabulary distributions between logit lens outputs before and after FFN layers in a Transformer. At the end, it weighted summed the two scores to obtain a hallucination score.

C Details of Datasets

RAGTruth. The RAGTruth dataset is a human annotated hallucination detection dataset, containing 15,090 training data and 2,700 testing data. Each data point consists of a query, retrieved documents, LLM-generated answer, and span-level hallucination annotation. The dataset covers three tasks, including summarization, data to text generation, and question answering. For each query-and-documents pair, RAGTruth provides answers generated by six different LLMs, including Llama2-7B, Llama2-13B, Llama2-70B, Mistral-7B, GPT-3.5, and GPT-4. In our experiment, we also utilize the extended test set provided by Sun et al. [14], who curated and annotated Llama3-8B generated responses.

HalluRAG. HalluRAG is an LLM annotated hallucination detection dataset for question answering. Ridder and Schilling [8] prompted GPT-4o to generate question given sentences from Wikipedia, then used Llama2-7B, Llama2-13B, and Mistral-7B to generate answer for each question given the relevant Wikipedia article. The hallucination labels were assigned by GPT-4o with a Chain-of-Thought (CoT) prompt and verified by human. HalluRAG contains both answerable and unanswerable questions, while we only use the answerable instances for evaluation.

D Implementation Details of LUMINA

For external context utilization, we measure MMD with Eq. (6), which requires summing over the combinations of the entire vocabulary. In practice we approximate it with the top 100 tokens to reduce the computational cost. To obtain $p_{\text{ctx}'}$, in our experiment we treat the retrieved documents of another data point as the d' of the target data point. In a real-world RAG system, d' can be obtained by selecting random documents from the data store or retrieving less relevant documents of the query with a retrieval model.

For internal knowledge utilization, Eq. (10) computes the first information process rate of generating a_t based on the next token with the highest probability. However, due to the sampling process of generation, the generated token a_t is not always the highest probability token. Thus, the internal knowledge used during the generation process may not fully apply to a_t. To take this factor into account, we calibrate the internal knowledge score by the ratio of probability between the generated token and the highest probability token. In the end, the calibrated internal knowledge score of a_t is

Table 3: **LUMINA consistently achieves a balanced precision-recall trade-off and high F1 score across datasets and LLMs.** We report the score of $\mathrm{Prec}_{\mathrm{Opt}}$, $\mathrm{Recall}_{\mathrm{Opt}}$, and $\mathrm{F1}_{\mathrm{Opt}}$ for LUMINA and baselines on each dataset.

		RAGTruth			HalluRAG		
LLM	Approach	$\mathrm{Prec}_{\mathrm{Opt}} \uparrow$	$\mathrm{Recall}_{\mathrm{Opt}} \uparrow$	$\mathrm{F1}_{\mathrm{Opt}} \uparrow$	$\mathrm{Prec}_{\mathrm{Opt}} \uparrow$	$\mathrm{Recall}_{\mathrm{Opt}} \uparrow$	$\mathrm{F1}_{\mathrm{Opt}} \uparrow$
Llama2-7B	Perplexity	0.5080	0.9867	0.6707	0.2531	1.0000	0.4040
	LN-Entropy	0.6303	0.7920	0.7020	0.7143	0.7500	0.7317
	Focus	0.5276	0.9292	0.6731	0.3077	1.0000	0.4706
	SelfCKGPT	0.5125	1.0000	0.6777	0.2631	1.0000	0.4167
	EigenScore	0.5201	0.9735	0.6780	0.4333	0.6500	0.5200
	P(True)	0.5079	0.9956	0.6726	0.3065	0.9500	0.4634
	RefChecker	0.5022	1.0000	0.6686	0.2532	1.0000	0.4040
	ReDeEP	0.6898	0.7478	0.7176	0.4167	0.7500	0.5357
	LUMINA	0.7131	0.7699	0.7404	0.7826	0.9000	0.8372
Llama2-13B	Perplexity	0.4926	0.9662	0.6525	0.1519	1.0000	0.2637
	LN-Entropy	0.6602	0.8164	0.7300	0.5385	0.5833	0.5600
	Focus	0.4938	0.9565	0.6513	0.5556	0.4167	0.4762
	SelfCKGPT	0.4801	0.9903	0.6467	0.3056	0.9167	0.4583
	EigenScore	0.5389	0.9034	0.6751	0.5833	0.5833	0.5833
	P(True)	0.6890	0.6957	0.6923	0.2449	1.0000	0.3934
	RefChecker	0.4600	1.0000	0.6301	0.2727	0.2500	0.2609
	ReDeEP	0.7772	0.7246	0.7500	0.4706	0.6667	0.5517
	LUMINA	0.7816	0.7778	0.7797	1.0000	0.7500	0.8571
Llama3-8B	Perplexity	0.6369	0.8519	0.7289	-	-	-
	LN-Entropy	0.5852	0.9465	0.7233	-	-	-
	Focus	0.5571	0.9630	0.7059	-	-	-
	SelfCKGPT	0.5657	0.9918	0.7205	-	-	-
	EigenScore	0.5907	0.9383	0.7250	-	-	-
	P(True)	0.5718	0.9342	0.7094	-	-	-
	RefChecker	0.5400	1.0000	0.7013	-	-	-
	ReDeEP	0.6621	0.7901	0.7205	-	-	-
	LUMINA	0.6988	0.7449	0.7211	-	-	-
Mistral-7B	Perplexity	0.6187	0.9243	0.7412	0.1702	0.8000	0.2807
	LN-Entropy	0.6890	0.9040	0.7820	0.8571	0.6000	0.7059
	Focus	0.7175	0.9004	0.7986	0.7143	0.5000	0.5882
	SelfCKGPT	0.5914	0.9920	0.7411	0.5385	0.7000	0.6087
	EigenScore	0.5931	0.9522	0.7309	1.0000	0.5000	0.6667
	P(True)	0.7030	0.8486	0.7690	0.3333	0.3000	0.3158
	RefChecker	0.5578	1.0000	0.7161	0.1266	1.0000	0.2247
	ReDeEP	0.6506	0.8640	0.7423	0.6250	0.5000	0.5556
	LUMINA	0.6600	0.9320	0.7728	0.9000	0.9000	0.9000

defined as

$$\mathcal{I}_{p_\theta}(a_t|q,d,a_{<t}) := \frac{p_\theta(a_t|q,d,a_{<t})}{p_\theta(a_{t,1}|q,d,a_{<t})} \cdot \mathcal{R}^1_{p_\theta}(q,d,a_{<t}). \tag{11}$$

E Additional Experimental Results

E.1 Evaluation with Other Metrics

Table 3 shows the scores of $\mathrm{Prec}_{\mathrm{Opt}}$, $\mathrm{Recall}_{\mathrm{Opt}}$, and $\mathrm{F1}_{\mathrm{Opt}}$ on each dataset. The results show that LUMINA consistently has a balanced precision-recall trade-off, where the differences between $\mathrm{Prec}_{\mathrm{Opt}}$ and $\mathrm{Recall}_{\mathrm{Opt}}$ are smaller than other baselines. Specifically, it achieves $(\mathrm{Prec}_{\mathrm{Opt}}, \mathrm{Recall}_{\mathrm{Opt}}) = (0.9, 0.9)$ on HalluRAG with Mistral-7B. This suggests that LUMINA does not over-predict hallucinations to achieve a high $\mathrm{F1}_{\mathrm{Opt}}$ score.

Table 4: **LUMINA achieves a competitive performance against supervised approaches.** We report the score of AUROC (ROC), Pearson's correlation coefficient (PCC), and AUPRC (PRC) for LUMINA and baselines on each dataset. The highest scores are set in **bold**.

LLM	Approach	RAGTruth			HalluRAG		
		ROC ↑	PCC ↑	PRC ↑	ROC ↑	PCC ↑	PRC ↑
Llama2-7B	SAPLMA	0.6508	0.2530	0.6446	0.8813	**0.6710**	**0.8023**
	LUMINA	**0.7646**	**0.4546**	**0.7491**	**0.9153**	0.6554	0.7572
Llama2-13B	SAPLMA	0.8337	0.5623	**0.8466**	0.8925	**0.8249**	**0.8647**
	LUMINA	**0.8569**	**0.6041**	0.8436	**0.9166**	0.6044	0.8497
Mistral-7B	SAPLMA	**0.8073**	**0.5027**	**0.8164**	0.9667	**0.7920**	0.9088
	LUMINA	0.7685	0.4623	0.7942	**0.9899**	0.7529	**0.9431**

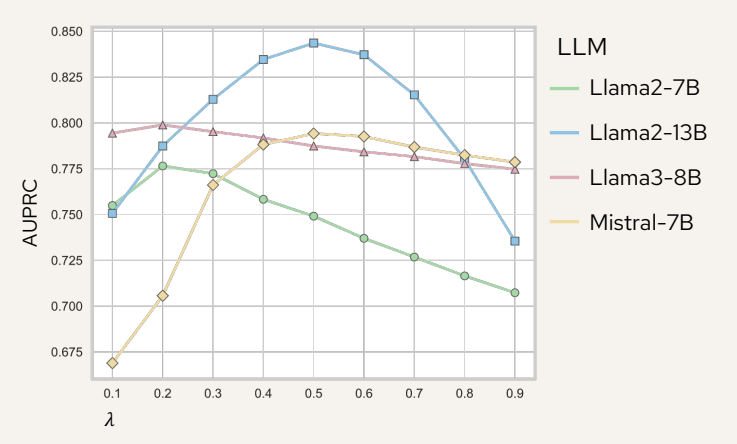

Figure 6: **A good performance of LUMINA happens with a medium λ value**. We alter λ in Eq. (1) to control the weight of internal knowledge score and external context score and evaluate the resulted hallucination detection performance. We conduct the experiment on the RAGTruth dataset and report the AUPRC score.

E.2 Compare with supervised baselines

We further compare LUMINA with SAPLMA [32], a supervised approach that trained a MLP model over the internal hidden states of the last generated token to classify whether the generation is hallucination or not. Following the original paper, we use hidden states at the 20th layer as input features of SAPLMA. Result in Table 4 shows that LUMINA has a competitive performance against SAPLMA and even sometimes outperforms it. Note that Table 4 doesn't show the result of Llama3-8B as the training set doesn't contain responses generated by Llama3-8B.

E.3 Performance with Hyperparameter Tuning

We evaluate the hallucination detection performance with $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Figure 6 shows the AUPRC of different λ on the RAGTruth dataset. The results show that the LUMINA achieves the optimal performance with varies λ across LLMs. For Llama2-13B and Mistral-7B, setting $\lambda = 0.5$, *i.e.*, the default setting, is the optimal. While for Llama2-7B and Llama3-8B, the optimal λ is 0.2. However, for these two models, their performance only drops less than 0.025 when setting $\lambda = 0.5$, suggesting that weighting internal knowledge and external context utilization equally is still a good practice.

F Error Analysis

To analyze the failure of LUMINA, we sample 20 cases from the RAGTruth dataset that are (1) hallucinated with high-external context and low-internal knowledge scores (*i.e.*, false negative) or (2) non-hallucinated with low-external context and high-internal knowledge scores (*i.e.*, false positive). We qualitatively analyze these cases and categorize them into three groups:

(1) Incorrect labels. Sometimes LLMs generate fabricated content that is not sourced from the retrieved document (*e.g.*, a detailed menu of a restaurant). However, these fabricated contents are sometimes not identified by human annotators. Also, human annotators sometimes misclassify semantically equivalent content as hallucination. In these cases, the provided labels are incorrect, and LUMINA indeed correctly detects hallucination.

(2) Generally low hallucination score for the summarization task. We observe that many false negative samples come from the summarization task. In these cases, the LLM does generate content that contradicts the retrieved documents and has a relatively high internal knowledge score. However, since most of the generated content is still grounded in the retrieved documents, they usually have a high external score as well, resulting in a relatively low hallucination score. This observation suggests that different tasks might have different distributions of hallucination scores. A better practice is to independently evaluate the hallucination detection performance on each task.

(3) Low quality of retrieved documents. For the false positive cases, we observe that many of them are due to the quality issue of the retrieved documents. These documents often contain only irrelevant information or are too vague to concretely answer the query. Thus, the LLM has to reason over them and respond with "unable to answer" or use its internal knowledge to generate answers with details and examples. This results in a relatively high internal knowledge score and a low external context score. To address this, a future direction can focus on assessing whether the utilization of internal knowledge is necessary and correct, and using that to calibrate the hallucination score.

G Computational Resources

LUMINA is a lightweight and efficient approach, which requires only two forward passes to obtain the necessary information to compute external context and internal knowledge scores. As LUMINA does not require generating multiple samples nor training, it is easy to scale up to a large amount of data. All the experiments of LUMINA are conducted on a single Nvidia H100 GPU. The execution time of computing both external context and internal knowledge scores varies depending on the length of the response. For responses around 150 tokens, the average computational time is less than 1 second.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The claims in Abstract and Introduction are aligned with the content in Section 3 and 4.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [No]
   Justification:
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Section 4 and Appendix D, G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: We provide the link to our released code in the Reproducibility Statement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No randomness in our approach.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on computing resources in Appendix G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics, and confirmed that our work does not deviate from it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix A

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite relevant works in Section 4 and Appendix B, C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The link to the released code is presented in the Reproducibility Statement.

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use LLM for core method development.

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.