Estimating the Black-box LLM Uncertainty with Distribution-Aligned Adversarial Distillation

Huizi Cui[1], Huan Ma[1], Qilin Wang[2], Yuhang Gao[3], Changqing Zhang[4]*

[1] School of Computer Science and Technology, Tianjin University, China
[2] School of Future Technology, Tianjin University, China
[3] Georgia Tech Shenzhen Institute, Tianjin University, China
[4] School of Artificial Intelligence, Tianjin University, China
{huizicui, mahuan520, wangqilin, yuhang_gao, zhangchangqing}@tju.edu.cn

Abstract

Large language models (LLMs) have progressed rapidly in complex reasoning and question answering, yet LLM hallucination remains a central bottleneck that hinders practical deployment, especially for commercial black-box LLMs accessible only via APIs. Existing uncertainty quantification methods typically depend on computationally expensive multiple sampling or internal parameters, which prevents real-time estimation and fails to capture information implicit in the black-box reasoning process. To address this issue, we propose Distribution-Aligned Adversarial Distillation (DisAAD), which introduces a generation-discrimination architecture to guide a lightweight proxy model to learn the high-quality regions of the output distribution of the black-box LLM, thus effectively endowing it with the ability to "know whether the black-box LLM knows or not". Subsequently, we use the proxy model to reproduce the specific responses of the black-box LLM and estimate the corresponding uncertainty based on evidence learning. Extensive experiments have verified the effectiveness and promise of our proposed method, indicating that a proxy model even one that only accounts for 1% of the target LLM's size can achieve reliable uncertainty quantification. Our model and related resources are released at https://github.com/huizi-Cui/DisAAD.

1 Introduction

Large language models (LLMs) have made significant progress in recent years, demonstrating outstanding performance in complex reasoning and text generation tasks (Kadavath et al., 2022; Rawte et al., 2023; Zhang et al., 2025). Despite these remarkable achievements, LLMs are prone to generate seemingly reasonable responses with non-factual or unfaithful information, a phenomenon widely known as LLM hallucinations (Shah, 2024; Banerjee et al., 2024; Tonmoy et al., 2024). Moreover, recent works have further indicated that larger and more instructive LLMs usually tend to deceive by pretending to understand, creating a false sense of confidence that makes users easily believe their responses (Abbasi Yadkori et al., 2024; Zhou et al., 2024; Huang et al., 2025). Therefore, hallucination issues present a significant barrier to the widespread application of LLMs, particularly in safety-critical fields (Chen et al., 2025; Perković et al., 2024).

Uncertainty quantification has emerged as a promising way to mitigate the limitations arising from hallucinations by enabling LLMs to express doubt when generating potentially unreliable responses (Huang et al., 2024; Zhang et al., 2023). High uncertainty indicates that users need to be cautious, since the LLM may be influenced by hallucinations and offer unreliable responses. Depending on computational cost, existing uncertainty quantification methods can be broadly categorized into self-evaluation methods, multi-sample methods and single-sample methods (Xiong et al., 2024).

Self-evaluation methods allow an LLM to assess the confidence of its own generated response through internal mechanisms or with the help of additional advanced models (Kadavath et al., 2022; Kapoor et al., 2024). However, these methods often fail to produce credible estimation results, and some specific fine-tuning interventions are necessary. Multi-sample methods perceive the diversity within the possible answer space from multiple reference calls, leveraging statistical patterns across generations to identify areas where the model exhibits hesitation or inconsistency (Lakshminarayanan et al., 2017; Farquhar et al., 2024). For example, Semantic Entropy quantifies uncertainty in LLMs by measuring consistency across multiple responses generated for the same prompt, and further identifies those inconsistent outputs

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42721–42740

as potentially unreliable information sources (Farquhar et al., 2024). Although multi-sample methods are theoretically well-founded, they face several significant issues: (1) fail to estimate the uncertainty of single response; (2) inefficient in practical applications due to the need for multiple sampling iterations; (3) miss inherent uncertainty when models consistently generate incorrect answers due to knowledge gaps.

In view of the above issues, single-sample methods are developed to estimate the uncertainty of individual sentences by accessing the internal information derived from the LLM (e.g., the next token probability distribution) to estimate the real-time uncertainty (Fadeeva et al., 2024). LogTokU is a representative method for quantifying token-level uncertainty by treating logits as parameters of the Dirichlet distribution (Ma et al., 2025). It provides mathematical evidence that logits offer more accurate uncertainty representations than maximum probability or entropy measurements. However, these methods are not applicable to closed-source LLMs such as GPT-4 and Claude-3, which still dominate in current practical applications (Sriramanan et al., 2024). Since these LLMs do not provide complete access to their internal mechanisms and parameter states, a fundamental issue arises: **How well can we predict the real-time uncertainty of black-box LLM only based on the single response?**

Recent research demonstrates that small LLMs often refuse to provide answers to difficult issues, reflecting a better awareness of their knowledge limitations. In contrast, larger and more instructive LLMs (such as GPT-4) tend to give seemingly reasonable but actually incorrect responses more frequently, making them easily overlooked by users (Zhou et al., 2024; Steyvers et al., 2025). Since simple LLMs are more reliable, a natural idea emerges: estimating the uncertainty of black-box LLM by leveraging a smaller LLM. To achieve this, we propose a novel Distribution-Aligned Adversarial Distillation (DisAAD), which introduces a small proxy model to learn how to "know whether the black-box LLM knows or not" and enables it to guide uncertainty quantification for the target LLM in downstream tasks. Specifically, we first systematically collect the outputs from the target black-box LLM across diverse prompts and create a comprehensive distillation dataset (Zeng et al., 2024). Then, the proxy model is specifically optimized within a generation-discrimination architecture to approximate the high-probability regions of the target output distribution. Benefiting from adversarial distillation, we further utilize the distilled proxy model to reproduce the responses of the black-box LLM and estimate real-time uncertainty via evidential deep learning (Sensoy et al., 2018). Extensive experiments verify the effectiveness of the proposed method in various question-answer tasks, a distilled proxy model even with only 1% of the target model size can achieve superior response reliability estimation performance.

The main contributions of our work are summarized as follows: (1) We propose a new paradigm for estimating the uncertainty of black-box LLMs, which not only eliminates the need for accessing model states but also obviates the requirement for multiple response sampling. (2) We propose a new method that enables proxy models to approximate the high-probability regions of the target output distribution, thereby characterizing the uncertainty of black-box LLMs. (3) Through extensive experiments and theoretical analysis, we validate the effectiveness of our proposed method in the detection of hallucinations of LLM, outperforming the strongest baselines in black-box setting with an average improvement of 18.2% in AUROC and 22.9% in AUPR.

2 Related work

2.1 Multi-sample Methods

Multi-sample methods evaluate uncertainty by measuring semantic consistency across multiple responses to the same prompt. For instance, Semantic Entropy computes the entropy over a distribution of semantic clusters formed by grouping the sampled responses based on semantic equivalence (Farquhar et al., 2024). EigV quantifies uncertainty using a graph-based calculation to estimate how many distinct groups of similar answers exist, which allows it to effectively identify different semantic clusters (Lin et al., 2023). Furthermore, recent works advance this by integrating the model's internal confidence. CoCoA calculates LLM uncertainty by multiplying the model's confidence in a specific response with the average semantic inconsistency compared to other samples (Vashurin et al., 2025). Similarly, SAR creates a hybrid measure by combining sentence-level semantic relevance with token-level probability adjustments for a more fine-grained balance of uncertainty (Duan et al., 2024).

2.2 Single-sample Methods

Single-sample methods usually achieve LLM uncertainty quantification by leveraging token probabilities, logits or hidden layer activations without requiring additional sampling. The related works include perplexity, negative sequence probability, and mean token entropy (Fomicheva et al., 2020), along with more advanced techniques that account for the semantic importance. For example, CCP isolates factual uncertainty by analyzing the semantic relationships within the candidate token distribution at each step (Fadeeva et al., 2024). Focus uses a proxy model to re-weight token-level uncertainty based on semantic properties like keyword importance and entity type (Zhang et al., 2023). Recent research suggests that logits provide more direct insight into model confidence than normalized probabilities, leading to approaches like LogTokU (Ma et al., 2025), which treats logits as Dirichlet distribution parameters (Abdar et al., 2021).

3 Method

3.1 Notations

Given a white-box LLM $\mathcal{M}_{\mathrm{W}}$ with a vocabulary $\mathcal{V}$, we formalize the next-token prediction process. An input prompt is tokenized as sequence $\boldsymbol{x} = (x_1, \ldots, x_L)$. The model autoregressively generates a response $\boldsymbol{y} = (y_1, \ldots, y_T)$. At each step t, $\mathcal{M}_{\mathrm{W}}$ processes the context, which comprises the prompt $\boldsymbol{x}$ and previously generated tokens $\boldsymbol{y}_{<t} = (y_1, \ldots, y_{t-1})$ to produce logit vector $\boldsymbol{z}_t \in \mathbb{R}^{|\mathcal{V}|}$. The probability of generating token $v_k \in \mathcal{V}$ as the next token y_t is:

$$P(y_t = v_k | \boldsymbol{x}, \boldsymbol{y}_{<t}; \mathcal{M}_{\mathrm{W}}) = \frac{\exp(\boldsymbol{z}_{t,k})}{\sum_{j=1}^{|\mathcal{V}|} \exp(\boldsymbol{z}_{t,j})}, \tag{1}$$

where $\boldsymbol{z}_{t,k}$ represents the k-th element of the logit vector $\boldsymbol{z}_t$, corresponding to the token v_k. The next token y_t would be sampled from the distribution $y_t \sim P(\cdot | \boldsymbol{x}, \boldsymbol{y}_{<t}; \mathcal{M}_{\mathrm{W}})$.

The uncertainty of the white-box LLM $\mathcal{M}_{\mathrm{W}}$ can be directly evaluated according to the intermediate outputs including probability distribution, logits, etc. In contrast, we consider black-box LLM settings, which are increasingly prevalent in real-world applications. For a given input prompt $\boldsymbol{x}$, the black-box LLM $\mathcal{M}_{\mathrm{B}}$ simply returns a final response sequence $\boldsymbol{y}_{\mathrm{B}}$. The objective of our work is to quantify real-time uncertainty by only relying on the single input-response pair $(\boldsymbol{x}, \boldsymbol{y}_{\mathrm{B}})$ derived from $\mathcal{M}_{\mathrm{B}}$.

3.2 Distribution-Aligned Adversarial Distillation

As proven by recent work, larger and more instructive LLMs like GPT-4 often exhibit overconfidence, while smaller models tend to be better calibrated. Motivated by this, we propose to estimate the black-box uncertainty by leveraging the specifically optimized proxy model with small size. The proposed work is a two-stage method for black-box uncertainty quantification, where the first stage is distribution-aligned adversarial distillation for obtaining the proxy model, and the second stage is proxy-guided LLM uncertainty quantification based on evidential deep learning.

3.2.1 Distillation Data Collection

To align the distribution of our proxy model $\mathcal{M}_{\mathrm{p}}$ and the target model $\mathcal{M}_{\mathrm{B}}$, we first construct a small distillation dataset $\mathcal{D}_{\text{distill}}$. This process begins by creating a diverse set of prompts $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, collated from both large-scale conversational dataset (open-domain) and task-specific evaluation dataset (in-domain). For each prompt $\boldsymbol{x}^{(i)}$, we query $\mathcal{M}_{\mathrm{B}}$ multiple times to generate a candidate pool of responses $D_{\mathrm{B}}^{(i)}$. It constitutes an empirical sampling of the model's true conditional output distribution $P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x}^{(i)})$, which is often characterized by a long-tailed nature. To ensure that the distillation dataset represents the most characteristic outputs of the target LLM, we select the top M responses from each $D_{\mathrm{B}}^{(i)}$ that exhibit the highest mutual semantic consistency. This strategy effectively isolates the high-probability regions of the output distribution, and the selected prompt-response pairs $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\mathrm{B}}^{(i,j)})\}_{i=1,j=1}^{N,M}$ form our final distillation dataset $\mathcal{D}_{\text{distill}}$.

3.2.2 Proxy Model Training

Based on the collected distillation dataset $\mathcal{D}_{\text{distill}}$, we establish an adversarial training between the proxy model (generator) and the discriminator. As shown in Fig. 1, the proxy model learns to generate the responses consistent with those of the target LLM, while the discriminator learns to distinguish the responses derived from the proxy model and black-box LLM. Our goal is to optimize the proxy model until its generated responses are statistically indistinguishable from those of the target LLM,

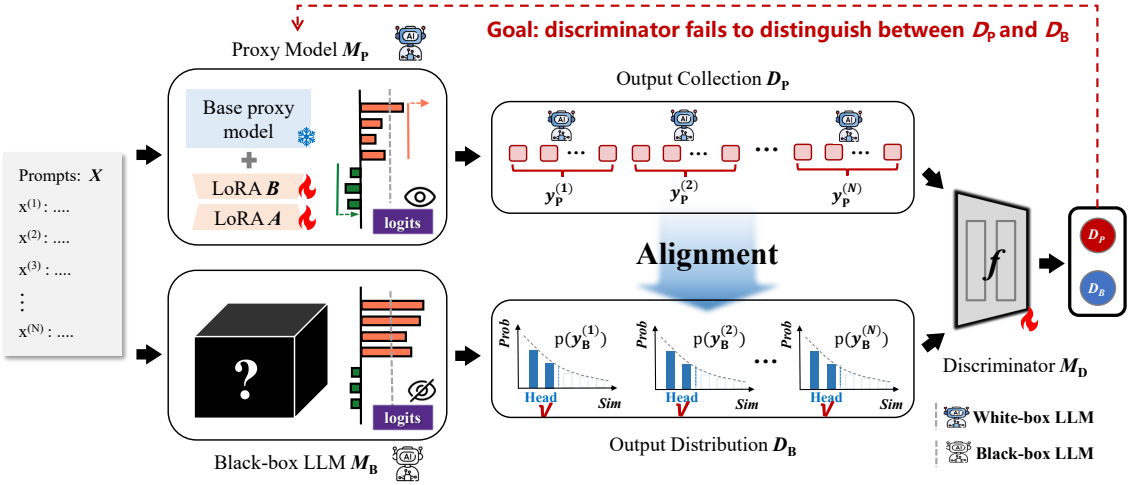


Figure 1: Overview of Distribution-Aligned Adversarial Distillation (DisAAD) framework. For a given small-size prompt set, we first collect the distillation dataset based on the target black-box LLM. Then, the LoRA-based proxy model is trained with the collected distillation dataset under the generation-discrimination architecture. The adversarial training objective is to enable the discriminator to learn to distinguish whether the responses generated by the proxy model are aligned with the high-probability regions from the target output distributions. The adversarial distillation terminates when the discriminator is unable to effectively distinguish the responses of the proxy model and the target black-box LLM.

reaching the point where the discriminator cannot distinguish them.

To efficiently fine-tune the proxy model, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022). Instead of fine-tuning all of the model's parameters, LoRA freezes the original weights (W_0) and injects trainable low-rank matrices (B and A) into each layer of the model. The effective weights are represented as:

$$W = W_0 + BA, \tag{2}$$

where $W_0 \in \mathbb{R}^{d\times d}$ represents the pre-trained weight matrix, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ are trainable low-rank matrices with rank $r \ll d$. We denote the base proxy model as $\mathcal{M}_{\text{p}}$, while LoRA-enhanced proxy model as $\mathcal{M}_{\text{p}}$, parameterized by its trainable LoRA weights θ.

In our proposed DisAAD framework, we aim to train a proxy model $\mathcal{M}_{\text{p}}$ to approximate the output behavior of a black-box LLM, denoted as $\mathcal{M}_{\text{B}}$. To this end, we define the training objective of the proxy model as minimizing the following loss:

$$\min_{\theta} \mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda \mathcal{L}_{\text{reg}}(\theta), \tag{3}$$

where θ denotes the parameters of the proxy model $\mathcal{M}_{\text{p}}$, and $\lambda > 0$ is a regularization coefficient. The loss consists of two parts:

- **Task Loss** $\mathcal{L}_{\text{task}}$: A standard distillation loss that aligns the output of the proxy model with the black-box model at the token level.
- **Regularization Loss** $\mathcal{L}_{\text{reg}}$: A sequence-level alignment constraint that enforces the proxy model's output to be indistinguishable from the black-box model by a discriminator.

Token-Level Distillation Loss (updating $\mathcal{M}_{\text{p}}$). We adopt a next-token prediction loss to encourage the proxy model to imitate the output of the black-box LLM. Given a dataset of N prompts $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, each associated with M sampled responses from the black-box model $\{\boldsymbol{y}_{\text{B}}^{(i,j)}\}_{j=1}^{M}$, the task loss is defined as:

$$\mathcal{L}_{\text{task}}(\theta) = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=l(\boldsymbol{x}^{(i)})+1}^{l(\boldsymbol{x}^{(i)})+l(\boldsymbol{y}_{\text{B}}^{(i,j)})} \log P_\theta(y_t \mid y_{<t}), \tag{4}$$

where $l(\boldsymbol{x}^{(i)})$ denotes the length of the prompt and $l(\boldsymbol{y}_{\text{B}}^{(i,j)})$ is the length of the target response. Following common instruction tuning practices (Zeng et al., 2024), we mask gradients from prompt tokens to prevent interference during learning.

Discriminator-Based Regularization (updating $\mathcal{M}_{\text{p}}$). To further align the proxy model with the target model at the sequence level, we introduce a discriminator $\mathcal{M}_{\text{D}}$ parameterized by ϕ. The discriminator receives a prompt-response pair and attempts

to distinguish whether the response is from the black-box model or the proxy model. The regularization loss encourages the proxy model to generate outputs that the discriminator cannot confidently classify as fake:

$$\mathcal{L}_{\text{reg}}(\theta) = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\log \mathcal{M}_{\text{D}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\text{P}}^{(i,j)}; \phi), \tag{5}$$

where $\boldsymbol{y}_{\text{P}}^{(i,j)}$ denotes the response generated by the current proxy model $\mathcal{M}_{\text{p}}$ for prompt $\boldsymbol{x}^{(i)}$, and $\mathcal{M}_{\text{D}}(\cdot)$ is the discriminator's estimated probability that the generated response is from the target model.

Discriminator Loss (updating $\mathcal{M}_{\text{D}}$). The discriminator is trained to classify whether a response is generated by the black-box model or by the proxy. Its objective is to maximize classification accuracy over real (target) and fake (proxy) samples:

$$\begin{aligned}\mathcal{L}_{\text{D}}(\phi) = &-\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\log \mathcal{M}_{\text{D}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\text{B}}^{(i,j)}; \phi) \\ &-\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\log\left(1 - \mathcal{M}_{\text{D}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\text{P}}^{(i,j)}; \phi)\right).\end{aligned} \tag{6}$$

Training Procedure. The proxy model and discriminator are trained in an alternating fashion:

1. Fix the discriminator, update the proxy model θ by minimizing $\mathcal{L}(\theta)$.

2. Fix the proxy model, update the discriminator ϕ by minimizing $\mathcal{L}_{\text{D}}(\phi)$.

This iterative training allows the proxy model to gradually learn to produce responses that are both locally (token-wise) and globally (sequence-wise) aligned with the black-box LLM. The discriminator acts as an adaptive regularizer, improving the semantic fidelity of the proxy model's generations. For more details please refer to the Appendix C.2.

3.3 Proxy-guided Uncertainty Quantification

As shown in Fig. 2, we first utilize the proxy model to reproduce the responses of black-box LLMs, then extract the corresponding logits and estimate the real-time uncertainty based on evidential deep learning (Sensoy et al., 2018; Han et al., 2022).

3.3.1 Logits as Evidence

Probability-based uncertainty estimation methods perform poorly because softmax normalization eliminates absolute evidence scale. Logits, however, can more flexibly capture both epistemic uncertainty and aleatoric uncertainty (Ma et al., 2025). Given a prompt $\boldsymbol{x}$ and response $\boldsymbol{y}_{\text{B}}$, we process the pair through our proxy model $\mathcal{M}_{\text{p}}$ to extract token-level logits $\boldsymbol{z}_t \in \mathbb{R}^{|\mathcal{V}|}$. To reduce noise from extremely low logits, we select only the top K token candidates with the largest logits to model a Dirichlet distribution (Tang et al., 2024):

$$\alpha_k = f(\boldsymbol{z}_{t,k}), \quad \alpha_0 = \sum_{k=1}^{K}\alpha_k \tag{7}$$

where $\boldsymbol{z}_{t,k} \in \mathbb{R}$ denotes the logit value for the k-th token candidate at decoding step t, $f(\cdot)$ is the ReLU activation function that converts logits to evidence parameters, and α_0 represents the total evidence strength.

3.3.2 Aleatoric Uncertainty (AU)

Aleatoric uncertainty, also known as data uncertainty, reflects the peak characteristic of output distributions. A lower AU indicates a peaked distribution where the probability mass is mainly concentrated on a single class, while the higher AU reflects the mass distributed uniformly across different classes. The specific definition is listed as follows:

$$\text{AU}(u_t) = -\sum_{k=1}^{K}\frac{\alpha_k}{\alpha_0}(\psi(\alpha_k{+}1) - \psi(\alpha_0{+}1)), \tag{8}$$

where $\psi(\cdot)$ denotes the digamma function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, $u_t = \{\alpha_1, \alpha_2, ..., \alpha_K\}$ is the set of evidence parameters at position t which represents the Dirichlet distribution parameters derived from logits.

3.3.3 Epistemic Uncertainty (EU)

Epistemic uncertainty, also known as model uncertainty, reflects the model's overall confidence in its prediction regardless of which specific token is selected. It is measured by:

$$\text{EU}(u_t) = \frac{K}{\sum_{k=1}^{K}(\alpha_k + 1)}, \tag{9}$$

where $\sum_{k=1}^{K}(\alpha_k + 1)$ is a smoothed measure of total evidence strength. A lower EU corresponds

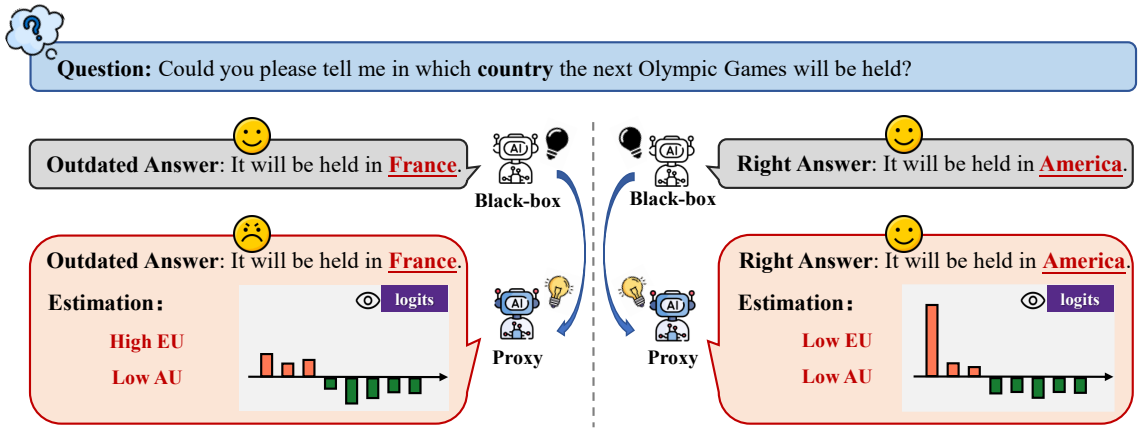


Figure 2: Overview of proxy-guided uncertainty quantification. For a given response from the target LLM, we first utilize the proxy model to reproduce the responses, then extract the corresponding logits and estimate the decoupled uncertainty. For an incorrect or outdated answer (e.g., "France"), the proxy model would identify this scenario as High EU and Low AU, which indicates that the response of the target model is unreliable. Conversely, for a correct answer (e.g., "America"), the proxy model would identify this scenario as Low EU and Low AU, which indicates that the response of the target model is reliable.

to high confidence prediction, while higher EU indicates knowledge gaps.

4 Experiments

4.1 Response Reliability

Consistent with concurrent work (Duan et al., 2024; Wang et al., 2025), we determine sentence reliability by focusing on the most uncertain tokens. The reliability of a given token u_t is defined as:

$$R(u_t) = -\mathrm{AU}(u_t) \cdot \mathrm{EU}(u_t), \tag{10}$$

where $R(u_t)$ captures the synergistic impact between aleatoric and epistemic uncertainty. Then the overall response reliability R_{response} is calculated by averaging the reliability values of the K^* least reliable tokens:

$$R_{\text{response}} = \frac{1}{K^*} \sum_{u_t \in T_{K^*}} R(u_t), \tag{11}$$

where T_{K^*} represents the set of K^* tokens with the lowest reliability values.

4.2 Evaluation Metrics and Datasets

The evaluation is formulated as a binary classification task, and conducted on question-answering benchmarks from different domains, including life sciences (BioASQ (Tsatsaronis et al., 2015)), truthfulness (TruthfulQA (Lin et al., 2021)) and knowledge seeking (TriviaQA (Joshi et al., 2017)). The responses generated by LLM with "BLEURT>0.5" or "LLM-Judge=1" are considered the correct answers (Xiong et al., 2024). The performance is quantified by AUROC, AUPR and ECE. For further details, please refer to the Appendix D.1.

4.2.1 Baseline Methods

We compare the proposed work with several SOTA black-box LLM uncertainty quantification methods, including Semantic Entropy (SE) and Discrete Semantic Entropy (DSE) (Farquhar et al., 2024), LN-Entropy (LNE) (Malinin and Gales, 2020), Lexical Similarity (LeS) (Lin et al., 2023) and EigV (Zhou et al., 2024). In addition, to show that our work can offer competitive performance without requiring access to the LLM's internal states, some white-box methods including probability-based method, entropy-based method and LogTokU (Ma et al., 2025; Xiong et al., 2024) are also used for comparative analysis.

4.3 Distillation Dataset Collection

We construct the distillation set by mixing half from in-domain evaluation prompts and half from out-of-domain conversational prompts to balance domain relevance and generalization. For each prompt, we draw 10 high-quality candidates from the target LLM. DisAAD yields an effective and computationally efficient proxy model using only $1K$ distillation samples (100 prompts). For additional information, please refer to Appendices C.1 and D.5.

4.4 Model Training

For the generator (proxy model), we employ LLaMA series models (Zheng et al., 2024) fine-tuned via Low-Rank Adaptation (LoRA) with rank=32, alpha=64, and dropout=0.1. Following the LLaMA architecture, we target all attention and feed-forward projections ("q_proj", "v_proj", "k_proj", "o_proj", "gate_proj", "down_proj", "up_proj"). The model is optimized via AdamW with learning rate 1×10^{-4} during training. For the discriminator, we utilize a GPT-2 encoder with the final three transformer layers unfrozen, optimized via AdamW at learning rate 1×10^{-5}. Throughout the adversarial distillation process, the discriminator serves a dual purpose by monitoring training sufficiency and providing quantitative evaluation metrics, effectively determining when the proxy model has successfully captured the characteristics of the target LLM's distribution. For more details, please refer to Appendix D.3.

4.5 Main Results

4.5.1 Results in Black-box Settings

Table 1 presents our main comparative results in black-box settings. The findings demonstrate that the proposed DisAAD achieves the best performance in response reliability estimation in most cases. Besides, a key advantage is its ability to perform real-time reliability estimation on a single response, which avoids the significant computational overhead of the multi-sample generation required by the other comparison methods. Specifically, when GPT-4 serves as the target LLM (Achiam et al., 2023), DisAAD achieves an average AUROC of 0.7321 and AUPR of 0.9134, substantially outperforming the best-performing baselines by 18.2% and 22.9%, respectively. In particular, this is achieved using a proxy model with only 1% of the target size. Similar performance is observed with other target LLMs. These results fully establish that DisAAD offers a more effective paradigm for LLM uncertainty quantification. For more details please refer to Appendix E.

4.5.2 Results in White-box Settings

To further investigate the performance of our proposed method, some representative white-box methods are used for comparison analysis. The results shown in Table 1 indicate that our method not only performs comparably to these methods, but often outperforms them in most cases, despite the fact that the proposed DisAAD does not require internal information derived from the target LLM. Specifically, with LLaMA2-70B as target LLM, DisAAD improves the average AUROC by 6.7% over the strongest baseline LogTokU. Compared to white-box methods that directly inherit the potentially flawed signal from the target's internal states, DisAAD achieves the uncertainty estimation by leveraging a smaller distilled proxy model, which appears to provide a more reliable result.

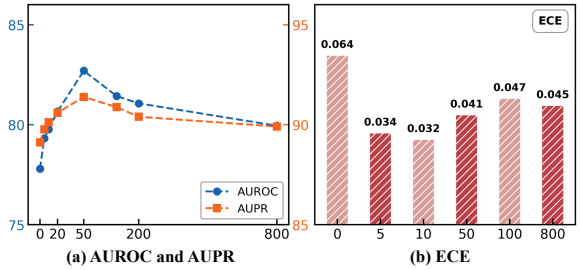


Figure 3: Reliability estimation performance from our distilled proxy model with different numbers of distillation prompts. Left: AUROC and AUPR results. Right: ECE results.

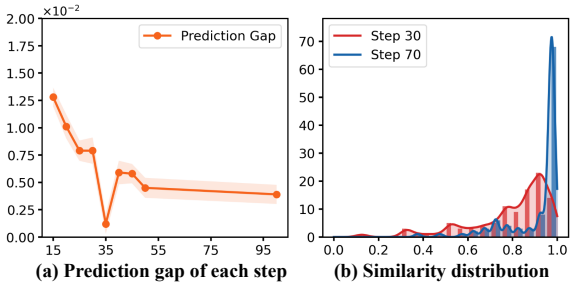
4.6 Further analysis

4.6.1 Efficiency of the Small-Size Distillation Dataset and Discriminator

To demonstrate the training efficiency of DisAAD, we analyze the effectiveness of small-size distillation datasets and our generation-discrimination architecture (LLaMA3-3B and GPT-2). Fig. 3(a) shows both AUROC and AUPR rapidly improve and stabilize after approximately 50 distillation prompts, with a slight pre-convergence decrease suggesting the proxy model captures essential uncertainty patterns before minor overfitting. The consistent ECE across different dataset sizes in Fig. 3(b) further support this finding. Fig. 4(a) illustrates the prediction gap (discriminator confidence difference between proxy and target outputs) narrowing to approximately 0.0050, indicating the discriminator can no longer effectively distinguish between the models' outputs. This serves as an auxiliary indicator that the proxy model has been adequately trained. Additionally, Fig. 4(b) visualizes the semantic similarity distribution for validation samples, showing a clear shift toward similarity scores exceeding 0.9, confirming high consistency between the proxy and target models. For additional theoretical proof and experimental results, refer to Appendices A and E.1.

Target	Method	$O(1)$	TruthfulQA		BioASQ		TriviaQA	
			AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
Black-box setting								
LLaMA2-70B	LEN	✗	$49.19^{\pm 4.12}$	$46.52^{\pm 3.50}$	$54.66^{\pm 4.18}$	$48.11^{\pm 3.33}$	$64.24^{\pm 3.90}$	$84.21^{\pm 4.52}$
	SE	✗	$53.93^{\pm 3.98}$	$46.73^{\pm 4.45}$	$59.94^{\pm 3.01}$	$48.25^{\pm 4.29}$	$\mathbf{65.80}^{\pm 4.82}$	$82.91^{\pm 5.60}$
	DSE	✗	$52.52^{\pm 3.05}$	$45.04^{\pm 4.61}$	$59.83^{\pm 3.03}$	$44.43^{\pm 3.58}$	$65.28^{\pm 4.85}$	$81.62^{\pm 6.69}$
	LES	✗	$64.19^{\pm 4.80}$	$76.39^{\pm 5.95}$	$52.85^{\pm 4.15}$	$45.73^{\pm 3.51}$	$46.22^{\pm 4.40}$	$\mathbf{90.26}^{\pm 3.05}$
	EigV	✗	$66.21^{\pm 3.75}$	$77.03^{\pm 4.88}$	$53.15^{\pm 5.10}$	$46.22^{\pm 4.40}$	$55.88^{\pm 3.51}$	$86.95^{\pm 4.12}$
	DisAAD	✓	$\mathbf{80.15}^{\pm 1.12}$	$\mathbf{78.07}^{\pm 1.25}$	$\mathbf{70.46}^{\pm 2.40}$	$\mathbf{78.74}^{\pm 1.19}$	$65.03^{\pm 3.86}$	$88.52^{\pm 2.20}$
GPT-4	LEN	✗	$53.57^{\pm 3.10}$	$67.83^{\pm 4.01}$	$60.90^{\pm 3.95}$	$48.27^{\pm 4.22}$	$50.89^{\pm 3.20}$	$89.16^{\pm 4.15}$
	SE	✗	$49.62^{\pm 4.23}$	$63.54^{\pm 3.15}$	$65.84^{\pm 4.70}$	$57.80^{\pm 5.98}$	$57.56^{\pm 4.01}$	$91.54^{\pm 6.99}$
	DSE	✗	$51.29^{\pm 3.18}$	$64.66^{\pm 5.09}$	$60.51^{\pm 3.98}$	$47.89^{\pm 4.25}$	$54.89^{\pm 3.11}$	$90.98^{\pm 4.03}$
	LES	✗	$65.10^{\pm 4.75}$	$77.20^{\pm 5.90}$	$53.15^{\pm 3.10}$	$46.05^{\pm 3.40}$	$47.33^{\pm 4.30}$	$91.10^{\pm 3.00}$
	EigV	✗	$66.05^{\pm 3.72}$	$76.81^{\pm 4.90}$	$63.90^{\pm 3.08}$	$56.01^{\pm 6.48}$	$55.91^{\pm 3.45}$	$90.15^{\pm 4.08}$
	DisAAD	✓	$\mathbf{80.78}^{\pm 2.83}$	$\mathbf{90.79}^{\pm 1.77}$	$\mathbf{69.93}^{\pm 3.42}$	$\mathbf{87.42}^{\pm 4.90}$	$\mathbf{68.93}^{\pm 1.60}$	$\mathbf{95.80}^{\pm 2.50}$
White-box setting								
LLaMA2-70B	Probability	✓	$65.89^{\pm 4.70}$	$59.90^{\pm 3.20}$	$57.07^{\pm 4.05}$	$67.20^{\pm 3.66}$	$60.06^{\pm 4.99}$	$85.89^{\pm 3.40}$
	Entropy	✓	$67.32^{\pm 3.65}$	$61.21^{\pm 4.15}$	$59.67^{\pm 3.99}$	$69.24^{\pm 4.59}$	$59.32^{\pm 3.01}$	$85.56^{\pm 4.42}$
	LogTokU	✓	$72.17^{\pm 4.40}$	$68.74^{\pm 3.80}$	$58.98^{\pm 4.00}$	$69.43^{\pm 3.58}$	$64.40^{\pm 4.88}$	$86.62^{\pm 3.35}$
	DisAAD	✓	$\mathbf{80.15}^{\pm 1.12}$	$\mathbf{78.07}^{\pm 1.25}$	$\mathbf{70.46}^{\pm 2.40}$	$\mathbf{78.74}^{\pm 1.19}$	$\mathbf{65.03}^{\pm 3.86}$	$\mathbf{88.52}^{\pm 2.20}$
Qwen3-32B	Probability	✓	$65.70^{\pm 4.72}$	$73.10^{\pm 3.66}$	$63.93^{\pm 4.85}$	$83.33^{\pm 3.20}$	$63.78^{\pm 3.90}$	$79.12^{\pm 4.80}$
	Entropy	✓	$65.90^{\pm 3.70}$	$74.07^{\pm 4.60}$	$64.67^{\pm 3.81}$	$83.86^{\pm 4.15}$	$66.58^{\pm 4.77}$	$80.57^{\pm 3.75}$
	LogTokU	✓	$66.02^{\pm 1.69}$	$74.97^{\pm 2.55}$	$64.47^{\pm 4.82}$	$83.41^{\pm 3.18}$	$\mathbf{75.22}^{\pm 3.45}$	$\mathbf{86.78}^{\pm 4.33}$
	DisAAD	✓	$\mathbf{74.78}^{\pm 2.30}$	$\mathbf{83.10}^{\pm 3.10}$	$\mathbf{65.71}^{\pm 2.75}$	$\mathbf{84.80}^{\pm 1.05}$	$69.35^{\pm 1.65}$	$81.25^{\pm 2.70}$

Table 1: Reliability estimation performance in the QA tasks of different domains. Response correctness for TruthfulQA is determined by "BLEURT>0.5", while the others are based on "LLM-Judge=1". $O(1)$ reflects the complexity of the response sampling process (multi-sample or single-sample).



(a) Prediction gap of each step (b) Similarity distribution

Figure 4: The performance of the discriminator in our DisAAD. Left: The prediction gap between the proxy model and the target LLM of each step. Right: The comparison semantic similarity distribution of Epoch 1.

4.6.2 Effectiveness of Distilled Proxy Model

To verify the effectiveness of the proposed adversarial distillation, we compare the reliability estimation performance of the base proxy model and our distilled proxy model on TruthfulQA. As demonstrated in Table 2, the distilled proxy model consistently shows better performance in AUROC and AUPR than the base proxy model. Furthermore, the ECE results indicate that the improvement of the discriminative power does not always come at the cost of calibration, e.g., DisAAD achieves a better ECE for both the Qwen3-32B (Yang et al., 2025) and GPT-4. The results verify that the adversarial distillation process can significantly enhance the ability of the proxy model to accurately capture the output characteristics of the target black-box LLM, so that our method can provide more accurate uncertainty quantification results.

Target	Proxy	TruthfulQA		
		AUROC↑	AUPR↑	ECE↓
LLaMA2-70B	Base*	0.7801	0.7608	0.0779
	DisAAD	**0.8015**	**0.7807**	**0.0741**
Qwen3-32B	Base*	0.7228	0.7793	0.0749
	DisAAD	**0.7478**	**0.8310**	**0.0664**
GPT-4	Base*	0.7780	0.8912	**0.0636**
	DisAAD	**0.8078**	**0.9079**	0.0685

Table 2: Performance comparison of base and distilled proxy models on the TruthfulQA dataset. "Base*" refers to the vanilla LLaMA3-3B, while "DisAAD" refers to the same architecture trained with our proposed method.

4.6.3 Flexibility in the Choice of Proxy Model

We evaluated LLaMA3 models of varying sizes (1B, 3B, and 8B) as distilled proxy models for reliability estimation. Table 3 reveals a non-monotonic

relationship between model size and performance, with the 3B model achieving optimal results. This model shows a 3.2% AUROC improvement over the 1B variant, which lacks sufficient capacity for complex uncertainty modeling. Conversely, the 8B model demonstrates a 2.7% AUROC reduction compared to the 3B model, likely because larger LLMs tend to exhibit overconfidence even when uncertain (Zhou et al., 2024; Steyvers et al., 2025). These findings suggest that the ideal proxy model requires moderate scale to effectively capture uncertainty patterns without inheriting the overconfidence issues of larger LLMs.

Target	Proxy	TruthfulQA		
		AUROC↑	AUPR↑	ECE↓
LLaMA2-70B	LLaMA3-1B	0.7768	0.7596	**0.0461**
	LLaMA3-3B	**0.8015**	**0.7807**	0.0741
	LLaMA3-8B	0.7797	0.7367	0.0723

Table 3: Reliability estimation performance of the distilled proxy model with different scales.

5 Conclusion

In this work, we introduce Distribution-Aligned Adversarial Distillation (DisAAD), a novel framework for estimating the uncertainty of black-box LLMs using a lightweight proxy model. Unlike existing methods that require multiple queries or access to internal model parameters, our approach enables real-time uncertainty quantification solely based on a single input-response pair. Through adversarial training and distribution alignment, the proxy model effectively learns to approximate the high-probability output regions of the target black-box LLM, thus acquiring the ability to discern response reliability. Extensive experiments on multiple question-answering benchmarks demonstrate that DisAAD achieves state-of-the-art performance in black-box LLM uncertainty estimation, paving the way for a safer and more trustworthy deployment in real-world applications.

Limitations

In this work, we mainly focus on token-level uncertainty estimation of black-box LLMs, which may overlook higher-level semantic or contextual inconsistencies in the LLM's full responses that are not captured by isolated token analysis. Besides, the proposed work relies on a prior adversarial distillation step to fine-tune the proxy model, ensuring that the proxy model can align with the target black-box LLM's output distribution before uncertainty assessment. Though the step of adversarial distillation would incur additional costs, both experimental results and theoretical analysis demonstrate that the proposed work can train an effective proxy model using only limited distillation samples.

Ethical Considerations

In this work, we propose the DisAAD framework to estimate the uncertainty of black-box LLM via a lightweight proxy model. The datasets we sourced from are publicly available, ensuring transparency and reproducibility of the distillation dataset construction. We do not expect any direct ethical concern from our work, as the framework solely aims to enhance the reliability of black-box LLM outputs by quantifying uncertainty, rather than modifying or exploiting the target LLMs in harmful ways.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2025YFF0515600) and the National Natural Science Foundation of China (62376193). We thank Zongbo Han (BUPT), Jingdong Chen (TJU), Yitao He (TJU) and Haiyun Yao (TJU) for their helpful discussion about the project and comments on the manuscript. The authors appreciate the valuable feedback from anonymous reviewers.

References

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Advances in Neural Information Processing Systems*, 37:58077–58117.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, and 1 others. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. LLMs will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Zheng Feng. 2025. Enhancing uncertainty modeling with semantic graph for hallucination detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23586–23594.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 5050–5063.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. A survey of uncertainty estimation in LLMs: Theory meets practice. *arXiv preprint arXiv:2410.15326*.

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

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

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

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

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

Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. 2025. Estimating LLM uncertainty with logits. *arXiv preprint arXiv:2502.00290*.

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

Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in LLMs: Understanding and addressing challenges. In *MIPRO ICT and electronics convention (MIPRO)*, pages 2084–2088. IEEE.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Savyasachi V Shah. 2024. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953–e2425953.

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

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. 2024. Top-nσ: Not all logits are you need. *arXiv preprint arXiv:2411.07641*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*.

Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. 2024. Efficient and effective uncertainty quantification for LLMs. In *Neurips Safe Generative AI Workshop*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Cong Zeng, Shengkun Tang, Xianjun Yang, Yuanzhou Chen, Yiyou Sun, Zhiqiang Xu, Yao Li, Haifeng Chen, Wei Cheng, and Dongkuan DK Xu. 2024. DALD: Improving logits-based detector without logits from black-box LLMs. *Advances in Neural Information Processing Systems*, 37:54947–54973.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and Ji Heng. 2025. The law of knowledge overshadowing: Towards understanding, predicting and preventing LLM hallucination. In *Findings of the Association for Computational Linguistics: ACL*, pages 23340–23358.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

Appendix

A Theoretical Analysis

A.1 Proof of Lipschitz Continuity

Theorem 1 (Lipschitz Continuity)**.** *For the distilled model $G : \mathcal{X} \to \mathcal{Y}$, there exists a constant $L > 0$ such that for any $x_1, x_2 \in \mathcal{X}$: $\|G(x_1) - G(x_2)\| \leq L\|x_1 - x_2\|$.*

Proof. Let W denote the proxy model's weight matrix. It employs low-rank adaptation (LoRA) on top of base model, with weight matrix:

$$W = W_0 + BA, \tag{12}$$

where W_0 represents the base model weights, and BA represents the low-rank adaptation with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$. Then, the output difference between the proxy model and black-box LLM under the given inputs x_1, x_2 can be expressed as:

$$\begin{aligned}\|G(x_1) - G(x_2)\| &= \|(W_0 + BA)(x_1 - x_2)\| \\ &\leq \|W_0(x_1 - x_2)\| + \|BA(x_1 - x_2)\|.\end{aligned} \tag{13}$$

For the LoRA component, it satisfies the submultiplicative property of matrix norms:

$$\|G(x_1) - G(x_2)\| \leq (\|W_0\| + \|B\| \cdot \|A\|) \|x_1 - x_2\|. \tag{14}$$

Let α and r denote the LoRA scaling factor and the adaptation rank, respectively. Given the $\|BA\| \leq \frac{\alpha}{r}$, $\|W_0\| + \|B\| \cdot \|A\|$ is a constant, thus we obtain:

$$\|G(x_1) - G(x_2)\| \leq L\|x_1 - x_2\|, \tag{15}$$

where $L = \|W_0\| + \|B\| \cdot \|A\|$ serves as our Lipschitz constant.

□

A.2 Proof of Adversarial Distillation with Limited Samples

First, we offer the definitions of missing mass and Zipf distribution for LLMs.

Definition 1 (Missing Mass)**.** *Given the input prompt x, the corresponding true output distribution $P_B(\boldsymbol{y}|\boldsymbol{x})$ of a black-box LLM and k samples $\{\boldsymbol{y}_B^{(i)}\}_{i=1}^k$ drawn from $P_B(\boldsymbol{y}|\boldsymbol{x})$, the missing mass U_k is defined as:*

$$U_k = \sum_{\boldsymbol{y}\in\mathcal{Y}} P_B(\boldsymbol{y}|\boldsymbol{x}) \cdot \mathbb{I}\{\boldsymbol{y} \notin \{\boldsymbol{y}_B^{(1)}, \ldots, \boldsymbol{y}_B^{(k)}\}\}, \tag{16}$$

where U_k represents the total probability of responses not observed in the k samples, $\boldsymbol{y}$ represents a possible model response, $\mathcal{Y}$ is the set of all possible responses.

Definition 2 (Zipf Distribution and Concentration Function)**.** *The output distribution of one LLM typically follows a Zipf-like power law where the probability of the i-th most likely response is proportional to $i^{-\alpha}$ for some $\alpha > 1$:*

$$P_B(\boldsymbol{y}_i|\boldsymbol{x}) \propto i^{-\alpha}. \tag{17}$$

The concentration function $H(v)$ for a distribution is defined as:

$$H(v) = \sum_{P_B(\boldsymbol{y}|\boldsymbol{x})\leq v} P_B(\boldsymbol{y}|\boldsymbol{x}). \tag{18}$$

To prove our main theorem, we need the following lemmas.

Lemma 1. *For distributions following Zipf's law with parameter $\alpha > 1$, the parameter β defined as:*

$$\beta = \liminf_{v \to 0} \frac{\ln H(v)}{\ln v}, \tag{19}$$

where the parameter β satisfies $\beta \geq \frac{\alpha-1}{\alpha}$. Furthermore, for any $\varepsilon > 0$, there exists k_0 such that for all $k > k_0$:

$$\mathbb{E}[U_k] \leq k^{-(\beta-\varepsilon)}. \tag{20}$$

Proof. For distributions following Zipf's law, the concentration function exhibits specific scaling properties near zero. Using the theory of regular variation, for a Zipf distribution with parameter $\alpha > 1$, it can be shown that $H(v) \sim v^{(\alpha-1)/\alpha}$ as $v \to 0$. This directly leads to $\beta \geq \frac{\alpha-1}{\alpha}$.

The bound on $\mathbb{E}[U_k]$ follows from established results in the concentration function theory for discrete distributions with heavy tails. Specifically, the expected missing mass decays polynomially with the sample size at a rate determined by the parameter β. □

Lemma 2. *Using Hoeffding's inequality, with probability at least $1 - \delta/2$:*

$$|U_k - \mathbb{E}[U_k]| \leq \sqrt{\frac{\ln(2/\delta)}{2k}}. \tag{21}$$

Proof. The missing mass U_k can be viewed as a function of k independent samples from $P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})$. This function satisfies a bounded differences condition: changing any single sample can change the value of U_k by at most $\frac{1}{k}$. Applying Hoeffding's inequality for such functions yields the stated result. □

Lemma 3. *For the empirical distribution $\hat{P}_B$ constructed from k samples:*

$$\mathcal{D}_{KL}(P_B \parallel \hat{P}_B) \leq -\ln(1 - U_k) + \frac{1}{1 - U_k} \sum_{i=1}^{k} \frac{|f_i - p_i|}{p_i}, \tag{22}$$

where $p_i = P_B(\boldsymbol{y}_B^{(i)}|\boldsymbol{x})$ and f_i is the empirical frequency of $\boldsymbol{y}_B^{(i)}$.

Proof. We decompose the KL divergence between the true distribution P_{B} and the empirical distribution $\hat{P}_{\mathrm{B}}$:

$$\mathcal{D}_{\mathrm{KL}}(P_{\mathrm{B}} \parallel \hat{P}_{\mathrm{B}}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x}) \ln \frac{P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})}{\hat{P}_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})} \tag{23}$$

$$= \sum_{\boldsymbol{y} \notin \{\boldsymbol{y}_{\mathrm{B}}^{(i)}\}_{i=1}^{k}} P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x}) \ln \frac{P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})}{\hat{P}_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})} + \sum_{i=1}^{k} P_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x}) \ln \frac{P_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x})}{\hat{P}_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x})}. \tag{24}$$

For the first term, note that $\hat{P}_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x}) = 0$ for all $\boldsymbol{y} \notin \{\boldsymbol{y}_{\mathrm{B}}^{(i)}\}_{i=1}^{k}$, a standard approach involves using a smoothed version of the empirical distribution to handle the undefined logarithm, which leads to a bound related to the missing mass U_k. This yields:

$$\sum_{\boldsymbol{y} \notin \{\boldsymbol{y}_{\mathrm{B}}^{(i)}\}_{i=1}^{k}} P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x}) \ln \frac{P_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})}{\hat{P}_{\mathrm{B}}(\boldsymbol{y}|\boldsymbol{x})} \leq -\ln(1 - U_k). \tag{25}$$

For the second term, by applying convexity arguments and properties of logarithms, one can derive the following bound for the observed samples:

$$\sum_{i=1}^{k} P_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x}) \ln \frac{P_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x})}{\hat{P}_{\mathrm{B}}(\boldsymbol{y}_{\mathrm{B}}^{(i)}|\boldsymbol{x})} \leq \frac{1}{1 - U_k} \sum_{i=1}^{k} \frac{|f_i - p_i|}{p_i}. \tag{26}$$

Combining these bounds gives us the desired result. □

Lemma 4. *With probability at least* $1 - \delta/2$*:*

$$\sum_{i=1}^{k} \frac{|f_i - p_i|}{p_i} \leq C_2 \sqrt{\frac{\ln(2/\delta)}{k}}, \tag{27}$$

where C_2 *is a constant.*

Proof. This follows from standard concentration inequalities for multinomial distributions. The empirical frequencies f_i are unbiased estimators of the true probabilities p_i, and their deviations can be bounded using results from statistical learning theory. □

Lemma 5. *Using the bound on* U_k *and the inequality* $-\ln(1-z) \leq \frac{z}{1-z}$ *for* $z \in [0, 1)$*:*

$$-\ln(1 - U_k) \leq \frac{U_k}{1 - U_k} \leq \frac{C_1}{k^\gamma}, \tag{28}$$

where $\gamma = \beta - \varepsilon$ *and* C_1 *is a constant.*

Proof. From Lemma 1 and Lemma 2, we have with probability at least $1 - \delta/2$:

$$U_k \leq \mathbb{E}[U_k] + \sqrt{\frac{\ln(2/\delta)}{2k}} \leq k^{-(\beta-\varepsilon)} + \sqrt{\frac{\ln(2/\delta)}{2k}}. \tag{29}$$

For a typical Zipf distribution in LLM, $\beta > 1/2$, making the polynomial term $k^{-(\beta-\varepsilon)}$ the dominant one for sufficiently large k. Thus, we can bound $U_k \leq C' k^{-(\beta-\varepsilon)}$ for some constant C'. Applying the inequality $-\ln(1-z) \leq \frac{z}{1-z}$, we get:

$$-\ln(1 - U_k) \leq \frac{U_k}{1 - U_k} \leq \frac{C' k^{-(\beta-\varepsilon)}}{1 - C' k^{-(\beta-\varepsilon)}} \leq \frac{C_1}{k^\gamma}, \tag{30}$$

where $\gamma = \beta - \varepsilon$ and C_1 is a constant that depends on C'. □

Now we are ready to present the proof of our main theorem.

Theorem 2 (Sample-Finited Adversarial Distillation)**.** *Let* $P_B(\boldsymbol{y}|\boldsymbol{x})$ *be the true output distribution of a black-box LLM* $\mathcal{M}_B$ *for input prompt* $\boldsymbol{x}$*. For* k *samples* $\{\boldsymbol{y}_B^{(i)}\}_{i=1}^k$ *drawn from* $P_B(\boldsymbol{y}|\boldsymbol{x})$*, the empirical distribution* $\hat{P}_B(\boldsymbol{y}|\boldsymbol{x})$ *constructed from these samples satisfies, with probability at least* $1 - \delta$*:*

$$\mathcal{D}_{KL}(P_B(\boldsymbol{y}|\boldsymbol{x}) \parallel \hat{P}_B(\boldsymbol{y}|\boldsymbol{x})) \leq \frac{C_1}{k^\gamma} + C_2 \sqrt{\frac{\ln(1/\delta)}{k}}, \tag{31}$$

where $\gamma > 0$ *depends on the power-law characteristics of* P_B*, and* C_1, C_2 *are constants.*

Proof. By combining Lemma 3 with Lemmas 4 and 5, and using the union bound, we get with probability at least $1 - \delta$:

$$\mathcal{D}_{\mathrm{KL}}(P_{\mathrm{B}} \parallel \hat{P}_{\mathrm{B}}) \leq -\ln(1 - U_k) + \frac{1}{1 - U_k} \sum_{i=1}^{k} \frac{|f_i - p_i|}{p_i} \tag{32}$$

$$\leq \frac{C_1}{k^\gamma} + C_2 \sqrt{\frac{\ln(2/\delta)}{k}}, \tag{33}$$

where $\gamma = \beta - \varepsilon \geq \frac{\alpha-1}{\alpha} - \varepsilon$, and C_1, C_2 are constants. Since $U_k \to 0$ as $k \to \infty$, the term $1/(1 - U_k)$ approaches 1 and can be absorbed into a new constant. After adjusting constants to reflect the total probability bound of $1 - \delta$, we arrive at the final result shown in the theorem.

□

B Prompt Templates

We provide the specific prompt templates used in our experiments for both response generation and evaluation.

B.1 Prompts for Response Generation and Reliability Estimation

Following the previous works, we utilize distinct prompt structures tailored to the specific formatting requirements of each for the LLaMa2 series, LLaMa3 series, Qwen3 series and GPT-4, respectively. The {question} placeholder is dynamically replaced with the input question from the question-answer datasets.

Prompt for Response Reliability Estimation

LLaMa2 Series Prompt

```
Answer the question concisely.
    Q: {question} A:
```

LLaMa3 Series Prompt

```
<|eot_id|>
<|start_header_id|>user<|
   end_header_id|>
Answer the question concisely.
    Q: {question} A:<|eot_id|>
```

GPT-4 Prompt

```
{"role": "user", "content": "
   Answer the question
   concisely. Q: {question} A
   :"}
```

Qwen3 Series Prompt

```
<|im_start|>user
Answer the question concisely.
Q: {question} A:<|im_end|>
<|im_start|>assistant
```

B.2 System Prompt for LLM-as-a-Judge

To automatically evaluate the correctness of generated responses for datasets like BioASQ and TriviaQA, we employ an LLM-as-a-Judge approach. The following system prompts are used to instruct the judge model (e.g., GPT-4) to act as an impartial evaluator. The prompt provides clear instructions, examples of both correct (1) and incorrect (0) evaluations, and the final template used for batch processing.

System Prompt for LLM-as-Judge

System: Your task is to determine if the provided answer is true or false based solely on the ground truth answers given to you in the format ['answer 1', 'answer 2', ...]. DO NOT rely on your memory; only use the information provided after this instruction. Respond with 1 if the predicted answer is correct, which means semantically consistent with any of the ground truth answers, otherwise respond with 0. Respond with just 0 or 1, and DO NOT include anything else in your response. This is the only instruction you need to follow.
User: Input: Who is elected as the vice president of india in 2017?
Ground Truth: ['Venkaiah Naidu', 'Muppavarapu Venkaiah Naidu']
Provided Answer: M. Venkaiah Naidu
Assistant:1
User: Input: who sings you are a magnet and i am steel?
Ground Truth: ['Walter Egan']
Provided Answer: The song 'You Are a Magnet and I Am Steel' is performed by the band The 1975.
Assistant:0

User: Input: {Question}
Ground Truth: {Your Ground Truth List}
Provided Answer: {The Answer to be Judged}
Assistant:

C Method Description

C.1 Distillation Dataset Generation

This section provides a detailed description of our proposed methodology. Algorithm 1 outlines the process for constructing our high-quality distillation dataset. This process begins by creating a diverse set of prompts, collated from both large-scale conversational WildChat (open-domain) and task-specific evaluation dataset (in-domain). For each prompt, we leverage the black-box LLM to generate a set of candidate responses using a dual-temperature sampling strategy to capture both pre-

cision and diversity. These responses then undergo a rigorous filtering process to ensure that each prompt in the final distillation dataset is paired with a set of representative responses.

Algorithm 1: Distillation Dataset Generation

Input: A collection of N prompts from mixed multi-source datasets

Model: Black-box LLM $\mathcal{M}_{\mathrm{B}}$ with specified generation parameters (e.g., temperature, top-M sampling)

Output: Distillation dataset derived from target LLM

Sample $N/2$ prompts from WildChat and $N/2$ from the evaluation dataset
Merge and shuffle to form the mixed prompt set
for each prompt **do**
    Generate one response with low temperature ($T \approx 0$) using the black-box LLM
    Generate multiple responses with high temperature ($T > 0.5$) using the black-box LLM
end for
for each prompt and its response set **do**
    Discard high-temp responses that are too short, repetitive, or high in perplexity
    Keep 1 low-temp + up to $M-1$ high-quality high-temp responses
    if valid response count $< M$ **then**
        Discard the sample
    end if
end for
return

C.2 Proxy Model Optimization

Algorithm 2 describes the adversarial distillation process for optimizing the proxy model $\mathcal{M}_{\mathrm{p}}$ using the proposed DisAAD framework. It establishes an adversarial training dynamic between the LoRA-based proxy model (the generator) and the discriminator $\mathcal{M}_{\mathrm{D}}$. The training objective is to align the proxy model's responses with the high-probability region of the output distribution of the target black-box LLM.

To maintain a stable and effective adversarial process, the training proceeds in a carefully balanced alternating fashion: for every single update to the proxy model, the discriminator is first updated twice, which ensures that the discriminator can provide a robust and informative learning signal by refining its ability to distinguish between responses from the black-box model's distribution (real) and those from the proxy model's distribution (fake). Following this, the proxy model's trainable parameters are updated based on a composite loss function, which combines a standard token-level task loss with a sequence-level adversarial loss provided by the discriminator. This iterative process drives the proxy model to generate responses that are increasingly indistinguishable from those of the target black-box LLM.

Algorithm 2: The Optimized Proxy Model using DisAAD

Input: Supervised distillation dataset $\{(\boldsymbol{x}^{(i)}, \{\boldsymbol{y}_{\mathrm{B}}^{(i,j)}\}_{j=1}^{M})\}_{i=1}^{N}$

Model: Proxy model $\mathcal{M}_{\mathrm{p}}$ with LoRA parameters θ; Discriminator $\mathcal{M}_{\mathrm{D}}$ with parameters ϕ

Output: The optimized proxy model $\mathcal{M}_{\mathrm{p}}$ with optimized parameters θ

Divide $\mathcal{D}_{\text{distill}}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$
for each training iteration **do**
    Sample a batch of prompts $\{\boldsymbol{x}^{(i)}\}_{i\in\mathcal{B}}$ from $\mathcal{D}_{\text{train}}$, and retrieve all corresponding target responses $\{\{\boldsymbol{y}_{\mathrm{B}}^{(i,j)}\}_{j=1}^{M}\}_{i\in\mathcal{B}}$
    Generate M responses for each prompt: $\{\{\boldsymbol{y}_{\mathrm{p}}^{(i,j)}\}_{j=1}^{M}\}_{i\in\mathcal{B}} \leftarrow \mathcal{M}_{\mathrm{p}}(\{\boldsymbol{x}^{(i)}\}_{i\in\mathcal{B}}; \theta)$
    // Update Discriminator $\mathcal{M}_D$
    Define real pairs $\mathcal{P}_{\text{real}} \leftarrow \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\mathrm{B}}^{(i,j)}) \mid i \in \mathcal{B}, j \in [1, M]\}$
    Define fake pairs $\mathcal{P}_{\text{fake}} \leftarrow \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\mathrm{p}}^{(i,j)}) \mid i \in \mathcal{B}, j \in [1, M]\}$
    Compute discriminator loss $\mathcal{L}_{\mathrm{D}}(\phi)$ over all pairs in $\mathcal{P}_{\text{real}}$ and $\mathcal{P}_{\text{fake}}$
    Update discriminator parameters: $\phi \leftarrow \phi - \eta_D \nabla_\phi \mathcal{L}_{\mathrm{D}}(\phi)$
    // Update Proxy Model $\mathcal{M}_p$
    Compute task loss $\mathcal{L}_{\text{task}}(\theta)$ over the set of real pairs $\mathcal{P}_{\text{real}}$
    Compute regularization (adversarial) loss $\mathcal{L}_{\text{reg}}(\theta)$ over the set of fake pairs $\mathcal{P}_{\text{fake}}$
    Aggregate total loss $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{task}}(\theta) + \lambda \mathcal{L}_{\text{reg}}(\theta)$
    Update proxy model's LoRA parameters: $\theta \leftarrow \theta - \eta_p \nabla_\theta \mathcal{L}(\theta)$
    Periodically validate $\mathcal{M}_{\mathrm{p}}$ on $\mathcal{D}_{\text{val}}$ and save the best-performing checkpoint
end for
return

C.3 Uncertainty Description

Here, we also give more explanation about the decoupled uncertainty based on evidential deep learning. As shown in Fig. 2 and Fig. 5, the proposed work can effectively quantify the LLM uncertainty even when there are multiple correct responses. Four different scenarios considered for black-box LLM uncertainty are listed as follows:

- High AU, High EU: the black-box LLM lacks domain knowledge and produces uncertain predictions;
- Low AU, High EU: it is an overconfidence scenario where the target LLM generates confident prediction despite knowledge gaps;
- Low AU, Low EU: it is an optimal reliability scenario with both strong knowledge and high prediction confidence;
- High AU, Low EU: the black-box LLM knows more than one reasonable answer.

D More Experimental Details

D.1 Dataset Description

D.1.1 TruthfulQA

TruthfulQA is a benchmark designed to evaluate the truthfulness of language model outputs. It contains questions that are adversarially selected to elicit false or misleading answers from models. The dataset spans multiple categories such as health, law, and finance, emphasizing factual consistency over plausibility. Each question has a reference answer annotated for truthfulness, making it suitable for both uncertainty and calibration evaluation.

D.1.2 BioASQ

BioASQ is a biomedical question answering benchmark comprising expert-annotated questions based on PubMed articles. We use the factoid subset of the dataset, which consists of questions with short, factual answers. BioASQ is particularly challenging due to domain-specific terminology and the requirement for precise biomedical knowledge. Each question is paired with a gold-standard answer list, enabling both exact match and ranking-based evaluation.

D.1.3 TriviaQA

TriviaQA is an open-domain question answering dataset containing over $650K$ question-answer pairs sourced from trivia websites and verified using evidence documents from Wikipedia. The questions are naturally complex and often require multi-hop reasoning. We use the unfiltered open-domain version, which pairs each question with multiple evidence documents, making it suitable for evaluating answer faithfulness and uncertainty under broader context exposure.

D.2 Evaluation Metrics

The evaluation is formulated as a binary classification task, and conducted on question-answering benchmarks from different domains, including life sciences (BioASQ (Tsatsaronis et al., 2015)), truthfulness (TruthfulQA (Lin et al., 2021)) and knowledge seeking (TriviaQA (Joshi et al., 2017)). The responses generated by LLM with "BLEURT$>$0.5" or "LLM-Judge=1" are considered the correct answers (Xiong et al., 2024). During the testing phase, for each dataset, we randomly selected 800 samples to evaluate the performance of different methods. The performance is quantified by the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR) assess its discriminative power, while the Expected Calibration Error (ECE) evaluates its calibration accuracy.

D.3 Target and Proxy Models

We employ Llama2-70B-Instruct, Qwen3-32B and GPT-4 (i.e., GPT-4-0613) as the target LLMs, and utilize Llama3-1B-Instruct, Llama3-3B-Instruct, Llama3-8B-Instruct as proxy models. Although our method is adaptable to various open-source proxy models, we standardize on the LLaMA3 series to ensure a controlled and consistent basis for our comparative analysis. We run all the experiments on 2-4 NVIDIA GeForce RTX 4090 GPUs with parallel processing. It is worth noting that we did not carefully select the hyperparameters, we believe that by making careful adjustments, better results can be achieved.

D.4 Baseline Methods

We demonstrate the response reliability estimation performance of our proposed framework in both black-box and white-box settings. For the multi-sample methods applicable in the black-box setting, including LN-Entropy (LNE), Semantic Entropy

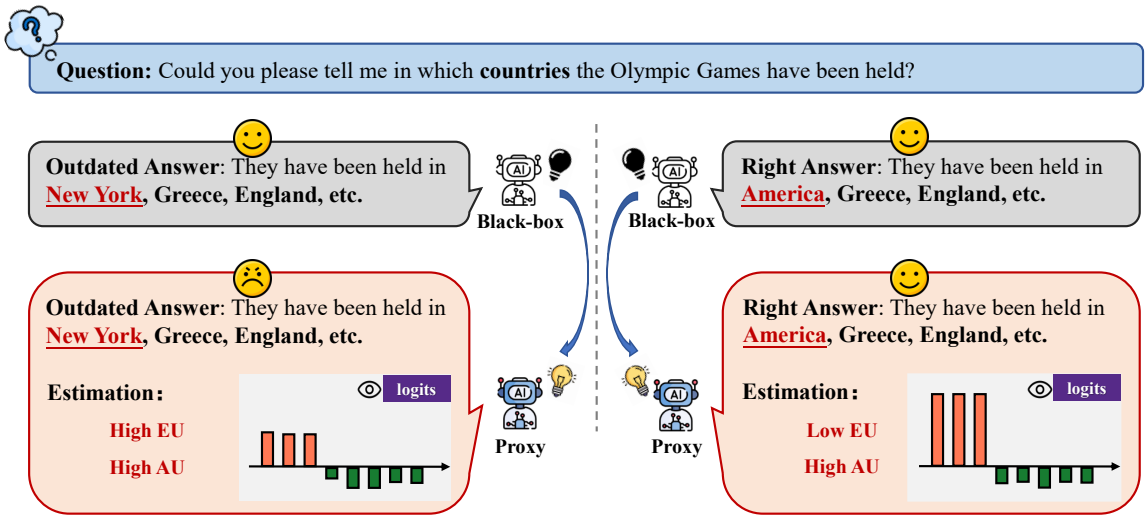


Figure 5: Overview of proxy-guided uncertainty quantification. For a given response from the target LLM, we first utilize the proxy model to reproduce the responses, then extract the corresponding logits and estimate the decoupled uncertainty. For an incorrect answer (e.g., "New York"), the proxy model identifies this scenario as High EU and High AU, since the LLM mistakenly identifies "New York" as one of the countries that have held the Olympic Games. Conversely, for a correct answer (e.g., "America"), the proxy model identifies this scenario as Low EU and High AU. This combination suggests that the target model is confident in its knowledge (Low EU) while also recognizing that the question has multiple correct answers (High AU), thus indicating the response is reliable.

(SE), Discrete Semantic Entropy (DSE), and Lexical Similarity (LeS), we follow the default setting in the original paper by generating 10 candidate responses with a temperature of 0.5 to derive the uncertainty scores. For LeS, Rouge-L is utilized as the similarity metric, whereas for SE and DSE, Deberta-Large-MNLI model is employed to form semantic clusters. Furthermore, for the single-sample methods applicable in the white-box setting, such as probability, entropy and LogTokU, we produce the single most probable response via greedy decoding strategy, then retain the token-level logits and compute the uncertainty scores.

D.5 Distillation Dataset Collection

To construct a distillation dataset that is both diverse and domain-specific, we strategically combine prompts from in-domain and out-of-domain datasets. Half are drawn from a general dataset rich in real-world interactions (e.g., WildChat), while the other half originate from the relevant evaluation dataset. This hybrid approach balances domain relevance with generalizability, enabling the proxy model to align closely with the target LLM's output distribution. For each input prompt, we generate 15 candidate responses: one response with low temperature ($T \approx 0.01$) and 14 diverse responses with high temperature ($T > 0.5$). These candidates then undergo a systematic filtering process. First, we discard high-temperature responses that fail baseline quality checks, specifically insufficient length (e.g., fewer than 15 words) or excessive sequence repetition. Following this, we compute the cosine similarity of each valid high-temperature response to its corresponding low-temperature response. We then retain the low-temperature response and the top nine high-temperature responses based on this similarity ranking. An input prompt is discarded entirely if it fails to yield this complete set of 10 valid responses after this procedure. This procedure is designed to capture the characteristic output distribution of the black-box model, rather than to filter exclusively for factual correctness. The final dataset consists of these newly generated responses paired with their corresponding original prompts, forming a collection of prompt-response pairs. Our experiments demonstrate that an effective proxy model can be obtained based on the proposed DisAAD with merely $1K$ samples, making our method not only powerful but also computationally efficient.

E More Experimental Results

E.1 Efficiency of the Small-Size Distillation Dataset and Discriminator

To demonstrate the training efficiency of our proposed DisAAD framework, we analyze the per-

Method	$K=1$		$K=5$		$K=10$		$K=15$		$K=20$		$K=25$		$K=ALL$	
	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
Probability	0.5503	0.6734	0.5739	0.6785	0.5611	0.6751	0.5548	0.6747	0.5533	0.6746	0.5533	0.6727	0.5707	0.6720
Entropy	0.5849	0.6946	0.6045	0.7010	0.5907	0.6945	0.5798	0.6908	0.5757	0.6914	0.5738	0.6899	0.5967	0.6924
Logtoku	0.5898	0.6943	0.6320	0.7185	0.6370	0.7178	0.6351	0.7189	0.6326	0.7177	0.6324	0.7199	0.6256	0.7253
DisAAD	**0.7046**	**0.7848**	**0.7105**	**0.7868**	**0.7038**	**0.7808**	**0.6946**	**0.7756**	**0.6905**	**0.7764**	**0.6866**	**0.7765**	**0.7046**	**0.7848**

Table 4: The comparison reliability estimation performance of different numbers of tokens with high uncertainty on BioASQ dataset. The proxy model and the target LLM are LLaMA3-3B and LLaMA2-70B, respectively. "K=ALL" denotes all the generated tokens (128 tokens) are used for estimation.

formance of small-size distillation dataset and the generation-discrimination architecture. As shown in Fig. 3(a), both AUROC and AUPR metrics rapidly improve with initial samples and stabilize after approximately 50 distillation prompts (equivalent to $0.5K$ distillation data). Notably, performance slightly decreases before final convergence, suggesting that the proxy model first captures essential uncertainty patterns from limited data, then experiences minor overfitting to noise present in larger datasets. The calibration results in Fig. 3(b) further support this finding, as the ECE values remain consistent across different distillation dataset sizes, indicating that additional data beyond the optimal point does not meaningfully improve calibration quality.

Furthermore, the effectiveness of the generation-discrimination architecture is demonstrated in Fig. 4, where we employ LLaMA3-3B as the generator and GPT-4 as the discriminator. To quantitatively measure the performance of the distilled proxy model, we define the prediction gap as the difference in discriminator confidence when classifying outputs from the proxy model versus the target LLM. As shown in Fig. 4(a), this gap progressively narrows throughout the training process, eventually stabilizing at approximately 0.0050. This convergence indicates the discriminator can no longer effectively distinguish between the outputs of both models, confirming that the proxy model has successfully learned to mimic the target LLM's response patterns. The semantic similarity analysis in Fig. 4(b) further supports this finding, illustrating the distribution of similarity scores for a number of 100 validation samples at different training steps. The clear shift toward similarity scores exceeding 0.9 between steps 30 and 70 demonstrates that the proxy model achieves high consistency with the target model's output distribution. Overall, empirical results demonstrate that the proposed DisAAD framework requires only a small distillation dataset, with the proxy model effectively approximating the target LLM's performance during the early stages of adversarial training, thus substantially reducing the query cost associated with distillation.

E.2 Discussion of the Choice of Top-K Tokens with High Uncertainty

Table 4 presents a comparative analysis of reliability estimation results using the Top-K tokens with the highest uncertainty. It is observed that for all tested choices of K, our DisAAD method maintains a substantial performance margin over all baseline methods. In addition, the reliability estimation achieved using a small subset of tokens (e.g., $K=5$ or $K=10$) is not only comparable to but even slightly superior to that achieved using the entire sequence of generated tokens ($K=$ ALL). This observation aligns with recent research suggesting that during an LLM's inference, only a small portion of tokens are important for estimating reliability. Therefore, focusing on these key tokens with higher uncertainty can provide more focused and clearer signals, while considering the entire token sequence as a whole might potentially affect the accuracy of the estimation. In this paper, we uniformly set K as 20% of the total token count for the generated response for all datasets.

E.3 Effectiveness of the Logits-based Uncertainty Estimation

As shown in Table 5, the proposed DisAAD outperforms both probability-based method and entropy-based method in estimating the response reliability in most cases. This superiority is evident from two key observations. First, DisAAD consistently achieves the highest average AUROC and AUPR scores across different types of target LLMs (LLaMA2-70B, Qwen3-32B, and GPT-4), indicating its robust and generalizable performance. For instance, with LLaMA2-70B as the target LLM, DisAAD's average AUROC of 0.7188 significantly surpasses the 0.6699 from entropy and 0.6390 from probability. Theoretically, this advantage stems from the nature of logits as raw and unnormal-

Target	Method	$O(1)$	TruthfulQA		BioASQ		TriviaQA		Average	
			AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑	AUROC↑	AUPR↑
LLaMA2-70B	Probability	✓	0.7261	0.6825	0.6482	0.7426	0.5427	0.8175	0.6390	0.7475
	Entropy	✓	0.7527	0.7166	0.6763	0.7622	0.5808	0.8465	0.6699	0.7751
	DisAAD	✓	**0.8015**	**0.7807**	**0.7046**	**0.7874**	**0.6503**	**0.8852**	**0.7188**	**0.8178**
Qwen3-32B	Probability	✓	0.6608	0.7369	0.6599	0.8396	0.6852	0.7905	0.6686	0.7890
	Entropy	✓	0.6724	0.7710	**0.6867**	**0.8522**	**0.7127**	**0.8215**	0.6906	0.8149
	DisAAD	✓	**0.7478**	**0.8310**	0.6571	0.8480	0.6935	0.8125	**0.6995**	**0.8305**
GPT-4	Probability	✓	0.6803	0.8297	0.6620	0.8701	0.6627	0.9522	0.6683	0.8840
	Entropy	✓	0.7370	0.8729	0.6862	0.8758	0.6838	0.9540	0.7023	0.9009
	DisAAD	✓	**0.8078**	**0.9079**	**0.6993**	**0.8742**	**0.6893**	**0.9580**	**0.7321**	**0.9134**

Table 5: Comparison of reliability estimation performance based on probability, entropy, and logits-based DisAAD.

ized scores from the model's final layer. Unlike probabilities, which are normalized by the softmax function and thus only reflect the relative differences between scores, logits retain crucial information about the absolute magnitude of the model's conviction. By operating directly on these richer, uncompressed logit representations, DisAAD can access a more fine-grained and reliable signal of the model's true internal state of confidence, ultimately enabling a more accurate distinction between reliable and unreliable answers.

E.4 Potential Risks

AI safety is closely related to the reduction of LLM hallucination. These non-factual but seemingly reasonable outputs pose risks for safety-critical applications. Our DisAAD framework estimates the uncertainty of black-box LLMs through a streamlined proxy model, which helps detect illusion phenomena and thereby enhances the security of deployment. However, our proxy model relies on the quality of the dataset, which can lead to deviations in uncertainty estimation and result in the omission of illusion phenomena in critical scenarios. Additionally, the bias of discriminator may also cause the proxy model to overestimate or underestimate the response uncertainty of the black-box LLM. Therefore, users who adopt the proposed method need to be cautious when conducting adversarial distillation of the proxy model.