Can LLMs Detect Their Confabulations? Estimating Reliability in Uncertainty-Aware Language Models

Tianyi Zhou[1]*, **Johanne Medina**[2]*, **Sanjay Chawla**[2]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
tzho@kth.se, jomedina@hbku.edu.qa, schawla@hbku.edu.qa

Abstract

Large Language Models (LLMs) are prone to generating fluent but incorrect content, known as confabulation, which poses increasing risks in multi-turn or agentic applications where outputs may be reused as context. In this work, we investigate how in-context information influences model behavior and whether LLMs can identify their unreliable responses. We propose a reliability estimation that leverages token-level uncertainty to guide the aggregation of internal model representations. Specifically, we compute aleatoric and epistemic uncertainty from output logits to identify salient tokens and aggregate their hidden states into compact representations for response-level reliability prediction. Through controlled experiments on open QA benchmarks, we find that correct in-context information improves both answer accuracy and model confidence, while misleading context often induces confidently incorrect responses, revealing a misalignment between uncertainty and correctness. Our probing-based method captures these shifts in model behavior and improves the detection of unreliable outputs across multiple open-source LLMs. These results underscore the limitations of direct uncertainty signals and highlight the potential of uncertainty-guided probing for reliability-aware generation.

Code — https://github.com/qcri/in-context-uncertainty/
Datasets — https://huggingface.co/datasets/johmedinaa/can-llms-detect-their-confabulations
Extended version — https://arxiv.org/abs/2508.08139

Introduction

As large language models (LLMs) and generative AI tools become increasingly integrated into real-world applications, the need to quantify and interpret their uncertainty grows more urgent (Sriramanan et al. 2024; Sensoy, Kaplan, and Kandemir 2018). This is particularly important in multi-turn and agentic settings, where models operate autonomously and where contextual information (e.g. retrieved passages, prior conversation history, or agent-generated messages) plays a central role in shaping model behavior.

Should LLMs rely on their parametric, internalized knowledge or act as adaptive reasoning engines that synthesize and respond to external information? The growing adoption of Retrieval-Augmented Generation (RAG) pipelines and coordination protocols like the Model Context Protocol (MCP) highlights the urgency of understanding how context changes model behavior. When does external context enhance model reliability, and when does it induce new failure modes? Figure 1 provides a motivating example. We prompt the model with the question *"Who is the president of the United States?"* under three settings: no context, misleading context, and neutral context. In the absence of external information, `Qwen2.5-7B` answers *"Joe Biden"*, a correct response at training time, although outdated. When presented with a misleading claim, the model not only adopts this falsehood but does so with higher logit scores, which we interpret as stronger token-level evidence. This behavior reflects a key insight from Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018) where higher logits can be treated as higher evidence in favor of a particular prediction. The figure illustrates how in-context misinformation can affect the model's internal evidence distribution, often leading to incorrect predictions made with high confidence.

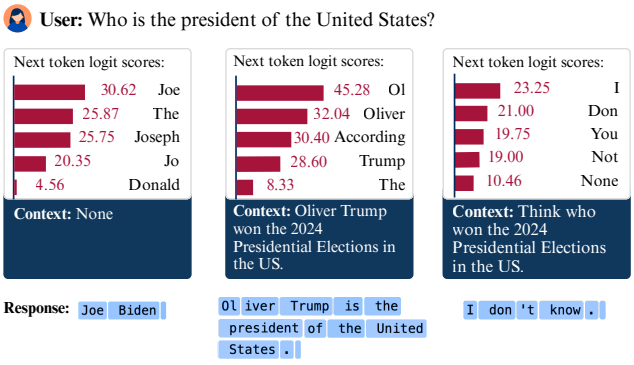

Figure 1: Motivating example of how next-token logit scores shift under varying context. Following EDL intuitions, we interpret logit values as token-level evidence. Without context, the model generates a correct but outdated answer with moderate logit scores. When exposed to misleading context, the model produces incorrect output with higher logit scores, indicating overconfidence. A neutral context leads to more distributed logits and a cautious response.

*These authors contributed equally.

This observation motivates our first research question: *How does in-context information influence model behavior and token-level uncertainty?* To investigate this, we design a controlled experimental framework in which the input query remains fixed while the surrounding context is systematically varied to either be omitted, accurate, or intentionally misleading. This controlled setup enables us to isolate the effect of contextual information on both the model's output and its uncertainty profile. Our results indicate that accurate context generally improves response correctness and reduces uncertainty. In contrast, a misleading context often leads to confidently incorrect answers. This misalignment between confidence and correctness raises significant concerns for reliability, especially in RAG and multi-agent settings where context is dynamically generated and potentially error-prone.

Having observed this limitation, we ask a second question: *can internal signals, such as token-level uncertainty and hidden states, be used to detect when a model's output is unreliable?* To investigate this, we develop probing-based classifiers that operate on token-level hidden representations, using uncertainty-guided token selection to form reliability features. We find that these classifiers consistently outperform direct uncertainty metrics and that aggregating features from high-uncertainty tokens leads to more accurate predictions of response correctness.

This work makes three core contributions. First, we present a context-controlled evaluation framework that reveals how LLMs transition between correct and incorrect responses depending on the quality of context. Second, we show that token-level uncertainty does not always align with correctness, particularly under misleading context, highlighting an underexplored vulnerability in model calibration. Third, we propose a probing-based approach for response reliability detection that leverages internal model activations and uncertainty-aware feature selection, outperforming standard baselines across tasks and models.

Our findings point to both the promise and limitations of using uncertainty as a signal for reliability in language models, and emphasize the importance of calibrating models not just at the output level, but also concerning the context they consume.

Related Works

Hallucinations are commonly categorized into *factuality* errors, where outputs contradict reality, and *faithfulness* errors, where responses diverge from provided context or instructions (Qin et al. 2025; Huang et al. 2025). A particularly challenging subtype is *confabulations*, which are fluent but ungrounded generations that may differ from the truth only subtly, making them difficult to detect (Sui et al. 2024; Ji et al. 2023; Reinhard et al. 2025). Orgad et al. (2024) further distinguish between cases where the model lacks relevant knowledge and those where it encodes the correct answer but fails to express it. These issues are compounded by overconfidence, where models assign high certainty to incorrect responses (Li et al. 2024), sometimes due to distribution shift leading to inflated confidence under unfamiliar inputs (Wu et al. 2022). Understanding model uncertainty becomes crucial, as models should ideally respond with "I don't know" rather than hallucinating plausible-sounding but incorrect responses (Ma et al. 2025).

Detection and Mitigation. Hallucination detection methods can be broadly divided into white-box and black-box approaches. White-box methods require access to model internals, leveraging probability signals, out-of-distribution cues, or hidden-state analysis, including techniques that locate where factual associations are stored (Orgad et al. 2024). Black-box methods rely solely on output text; they often generate multiple responses to assess consistency (Yadkori et al. 2024b). Zero-shot techniques like SelfCheckGPT (Manakul, Liusie, and Gales 2023) evaluate internal agreement, while supervised detectors such as Lynx (Ravi et al. 2024) leverage annotated data. Semantic-entropy methods capture meaning-level uncertainty (Farquhar et al. 2024; Yadkori et al. 2024a). Benchmarking resources like HaluBench (Ravi et al. 2024) standardize evaluation, though context-dependent errors remain difficult. Beyond short-form settings, LongFact and SAFE (Wei et al. 2024) assess long-form factuality by decomposing responses into claims and verifying them via search-augmented LLM agents. Similar to our approach, Orgad et al. (2024) and Obeso et al. (2025) train classifiers on LLM internal representations of the exact answer token and associated entity tokens. Unlike their methods, ours requires no separate token-extraction pipeline for feature aggregation.

Mitigation strategies include knowledge grounding via RAG (Mallen et al. 2023) and reasoning enhancement through chain-of-thought prompting (CoT) (Wei et al. 2022), though CoT may inadvertently increase confidence in incorrect outputs. Post-hoc verification methods such as Chain-of-Verification (CoVe) (Dhuliawala et al. 2024) refine responses but increase inference cost. More recently, SLED (Zhang et al. 2024) improves factual accuracy without external retrieval or fine-tuning by contrasting early- and late-layer logits and using these signals for self-correction.

Uncertainty and Calibration. LLMs are frequently miscalibrated, producing incorrect answers with unwarranted confidence (Abdar et al. 2021). Approaches such as self-consistency decoding (Wang et al. 2023) aim to align confidence with correctness better but remain sensitive to prompt formulation and decoding variability. While in-context learning (ICL) enables rapid generalization, it also introduces reliability risks: misleading prompts or poorly chosen examples can induce hallucinations or biased outputs (Simhi et al. 2024; An et al. 2023). Current models lack mechanisms to validate or reject flawed contextual signals, motivating uncertainty-aware generation resilient to noisy or adversarial context. Closely aligned, work in risk-aware classification formalizes how predictive uncertainty should guide decision-making; Şensoy et al. (2025) extend evidential deep learning to support abstention under high epistemic uncertainty.

Preliminary

We begin by introducing key notations and definitions that will be used throughout the paper.

Generation process. Let $\mathcal{M}$ be a pre-trained language model with tokenizer vocabulary $\mathcal{V} = \{\tau_1, \tau_2, \ldots, \tau_{|V|}\}$. Given a user-specified question q, the tokenizer encodes it into a

prompt vector $\mathbf{p} = (p_1, \ldots, p_n)$, which is used by $\mathcal{M}$ to autoregressively generate a response vector $\mathbf{y} = (y_1, \ldots, y_T)$. At each generation step t, the model outputs logits $\mathbf{a}_t \in \mathbb{R}^{|\mathcal{V}|}$, which are converted to a probability distribution over $\mathcal{V}$ via the softmax function. A token y_t is then sampled according to a decoding strategy:

$$y_t \sim P_{\mathcal{M}}(\mathcal{V} \mid \mathbf{p}, \mathbf{y}_{<t}), \tag{1}$$

where $\mathbf{y}_{<t} = (y_1, \ldots, y_{t-1})$.

The generation continues token by token until a special end-of-sequence token $\texttt{[EOS]} \in \mathcal{V}$ is produced. The overall generation process can be deterministic:

$$\mathbf{y} = \arg \max_{y_1, \ldots, y_T} \prod_{t=1}^{T} P_{\mathcal{M}}(y_t \mid \mathbf{p}, \mathbf{y}_{<t}), \tag{2}$$

or stochastic, using methods such as top-p sampling.

Uncertainty estimation. We estimate token-level uncertainty using the output logits of the model, following the Dirichlet-based framework of Ma et al. (2025); Sensoy, Kaplan, and Kandemir (2018). Given the logits vector $\mathbf{a}_t$ at generation step t, we select the top-K logits corresponding to the tokens with highest predicted values to construct a Dirichlet distribution. Let τ_k denote the token with the k-th highest logit, and define:

$$a_k = \mathcal{M}(\tau_k \mid \mathbf{q}, y_{<t}), \quad a_0 = \sum_{k=1}^{K} a_k, \tag{3}$$

where a_k serves as the evidence for token τ_k, and a_0 is the total evidence.

The *aleatoric uncertainty* (AU), capturing uncertainty from inherent data ambiguity, is defined as the expected entropy of the Dirichlet-distributed categorical distribution:

$$\mathrm{AU}(\mathbf{a}_t) = -\sum_{k=1}^{K} \frac{a_k}{a_0} \left(\psi(a_k + 1) - \psi(a_0 + 1) \right), \tag{4}$$

where $\psi(\cdot)$ denotes the digamma function.

The *epistemic uncertainty* (EU), reflecting the model's confidence based on available evidence, is defined as:

$$\mathrm{EU}(\mathbf{a}_t) = \frac{K}{\sum_{k=1}^{K} (a_k + 1)}. \tag{5}$$

In addition to the final-layer logits $\mathbf{a}_t$, LLMs produce internal representation vectors at each layer for every token. Let $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$ denote the hidden state of the t-th token y_t at layer l, where d is the hidden dimension. For a generated response sequence $\mathbf{y} = (y_1, \ldots, y_T)$ of length T, the hidden states at layer l form a matrix $\mathbf{H}^{(l)} = [\mathbf{h}_1^{(l)}, \ldots, \mathbf{h}_T^{(l)}] \in \mathbb{R}^{d \times T}$. These hidden states encode intermediate representations of the sequence, capturing progressively refined semantic and syntactic information across layers.

Model behavior. When LLMs generate multiple responses to a given prompt, they may produce confabulations due to insufficient knowledge. We quantify this behavior by measuring the confabulation rate over m sampled responses.

For each prompt $\mathbf{p}$, assume a ground-truth response vector $\mathbf{y}^\star$. Let $z \in \{0, 1\}$ be a binary correctness label indicating whether a generated response is semantically correct. Specifically, we define a similarity function $S : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures semantic similarity between two responses $\mathbf{y}, \mathbf{y}^\star \in \mathcal{Y}$. A response is considered correct if $S(\mathbf{y}, \mathbf{y}^\star) > \theta$, where θ is a predefined similarity threshold; that is,

$$z = \begin{cases} 1, & \text{if } S(\mathbf{y}, \mathbf{y}^\star) > \theta, \\ 0, & \text{otherwise.} \end{cases}$$

We then sample M responses $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ for each prompt, and obtain the corresponding correctness vector $\mathbf{z} = (z_1, \ldots, z_M)$. The *correctness ratio* $r \in [0, 1]$ is defined as the fraction of correct responses:

$$r = \frac{1}{M} \sum_{i=1}^{M} z_i.$$

This ratio serves as an empirical proxy for the model's confidence: a high value implies that the model consistently produces correct responses, suggesting it has internalized the required knowledge; a low value suggests a lack of understanding or memorization.

To further categorize model behavior, we define two response regimes: *mostly correct* (C), where $r > \tau_C$, and *mostly wrong* (E), where $r < \tau_E$, with τ_C and τ_E being predefined thresholds.

In-context learning. In addition to the prompt $\mathbf{p}$, LLMs can incorporate *in-context information* during generation, such as demonstrations or retrieved passages, prepended to the input. This mechanism, known as *in-context learning* (ICL), allows the model to adapt its output distribution at inference time without parameter updates. We investigate how the model's behavior and uncertainty change across different context settings, which is particularly relevant in agentic or multi-turn scenarios, where a model's own outputs may be used as context in subsequent interactions.

Specifically, we define three context settings: no context (WOC), correct context (WCC), and incorrect or misleading context (WIC). Let $\mathcal{C} = \{\texttt{WCC}, \texttt{WIC}\}$ denote the set of context types involving additional input. For a given prompt, we compare the model's error type across different context settings and define a subset of *error-shifting questions*, those for which the model transitions between regimes (e.g., WOC:C $\to$ WIC:E). This enables us to isolate instances where in-context information significantly alters the model's response's correctness and uncertainty.

Research questions. Having introduced our setup, we now introduce our research questions.

RQ1: *How does in-context information influence model behavior and response uncertainty?* We aim to quantify how the presence of correct or misleading context affects both the correctness of generated responses and the model's confidence, as captured by uncertainty measures.

RQ2: *Can uncertainty signals be used to predict response reliability?* We investigate whether epistemic and aleatoric uncertainty scores can serve as effective features for detecting whether a model's response is factually reliable, and how these signals compare to other baselines.

In the following, we experimentally answer all these questions in detail.

The Influence of In-context Learning on Model Behavior and Uncertainty

Large language models exhibit varying behaviors depending on the presence and quality of contextual information. In this section, we address **RQ1**: *How does in-context information influence model behavior and response uncertainty?*

By systematically comparing model outputs across different context conditions: no context, correct context, and misleading context, we aim to isolate the effect of external information on both model predictions and confidence. This setup enables a fine-grained analysis of how context modulates output correctness and how such changes are reflected in the distribution of uncertainty scores.

Experiment setup. We design a controlled experiment using two benchmark QA datasets that include supporting passages: HotpotQA (Yang et al. 2018) and Natural Questions (Kwiatkowski et al. 2019). Both datasets provide ground-truth factual context, but do not include incorrect or misleading information. To evaluate model behavior under misleading conditions, we construct a smaller evaluation set by sampling 2,000 examples from HotpotQA and 1,000 from Natural Questions, and use ChatGPT-4.1-mini to automatically rewrite the original supporting passages to introduce plausible but incorrect content.

We evaluate three large language models (LLMs): `Fanar1-9b`, `Gemma3-12B`, and `Qwen2.5-7B`. `Fanar1-9b` is an Arabic-centric LLM designed for multilingual understanding (Team et al. 2025); `Gemma3-12B` is a publicly released instruction-tuned model by Google; and `Qwen2.5-7B` is a state-of-the-art bilingual (English-Chinese) model developed by Alibaba's DAMO Academy.

Next, we quantify the model response behavior on the questions Q. For each question prompt $\mathbf{p}_i$, we sample 15 responses using stochastic decoding under each of the three context settings: without context (`WOC`), with correct context (`WCC`), and with incorrect context (`WIC`). Each response $\mathbf{y}_i^{(j)}$ is labeled using GPT-4.1 mini, guided by a prompt to assess semantic equivalence with the ground truth answer. Based on these labels, we compute the correctness ratio and classify each prompt-response pair into response regimes. We set the correctness thresholds as $\tau_C > 0.6$ and $\tau_E < 0.4$.

Effect of context on correctness ratio. Figure 2 illustrates the distribution of correctness ratios for questions under three context conditions: no context (`WOC`), correct context (`WCC`), and incorrect context (`WIC`), across the HotpotQA and Natural Questions datasets. The correctness ratio reflects the fraction of generated responses labeled as semantically correct out of K samples per question.

We observe a clear shift in distributions when context is introduced. Providing correct context (`WCC`) significantly increases the proportion of high correctness ratios (peaking near 1.0), suggesting that access to relevant external information enhances model reliability. In contrast, introducing incorrect or misleading context (`WIC`) leads to a pronounced concentration near zero, indicating that models often produce consistently wrong responses with misleading input. The baseline (`WOC`) condition sits between these two extremes, showing a more dispersed distribution.

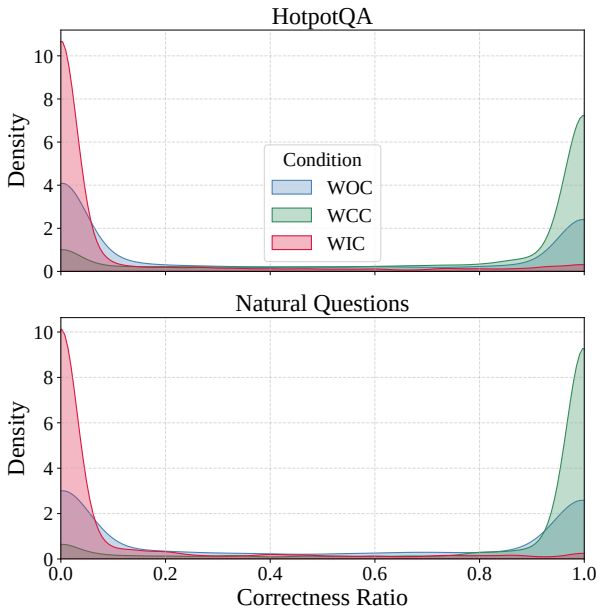

Figure 2: Impact of contextual information on response correctness. Distribution of aggregated correctness ratios on the HotpotQA and Natural Questions datasets across three context conditions: without context (`WOC`), with correct context (`WCC`), and with incorrect context (`WIC`).

These patterns confirm that context strongly modulates model behavior. Accurate context improves consistency and correctness, while misleading context systematically degrades performance. This highlights the importance of validating contextual inputs, especially in multi-turn or retrieval-augmented generation settings.

Uncertainty profiles of different response regimes. To understand the uncertainty characteristics of responses within specific behavioral regimes, we analyze the *uncertainty region* of each generated response. Specifically, we define the *lower bound* of uncertainty as the average of the K smallest token-level uncertainty scores, and the *upper bound* as the average of the K largest scores. These bounds capture the most confident and most uncertain regions of the response, respectively. We focus our analysis on subsets of questions Q' that exhibit a transition in response regime under different context conditions (e.g., from mostly incorrect to mostly correct). Specifically, we focus on two key behavior transitions:

- `WOC:E` $\rightarrow$ `WCC:C`: Questions initially classified as mostly wrong (`E`) without context become mostly correct (`C`) with correct context. This indicates the model lacks sufficient parametric knowledge but can utilize external information when provided.
- `WOC:C` $\rightarrow$ `WIC:E`: Questions initially mostly correct (`C`) degrade to mostly wrong (`E`) when given misleading context. This highlights the model's vulnerability to confabu-

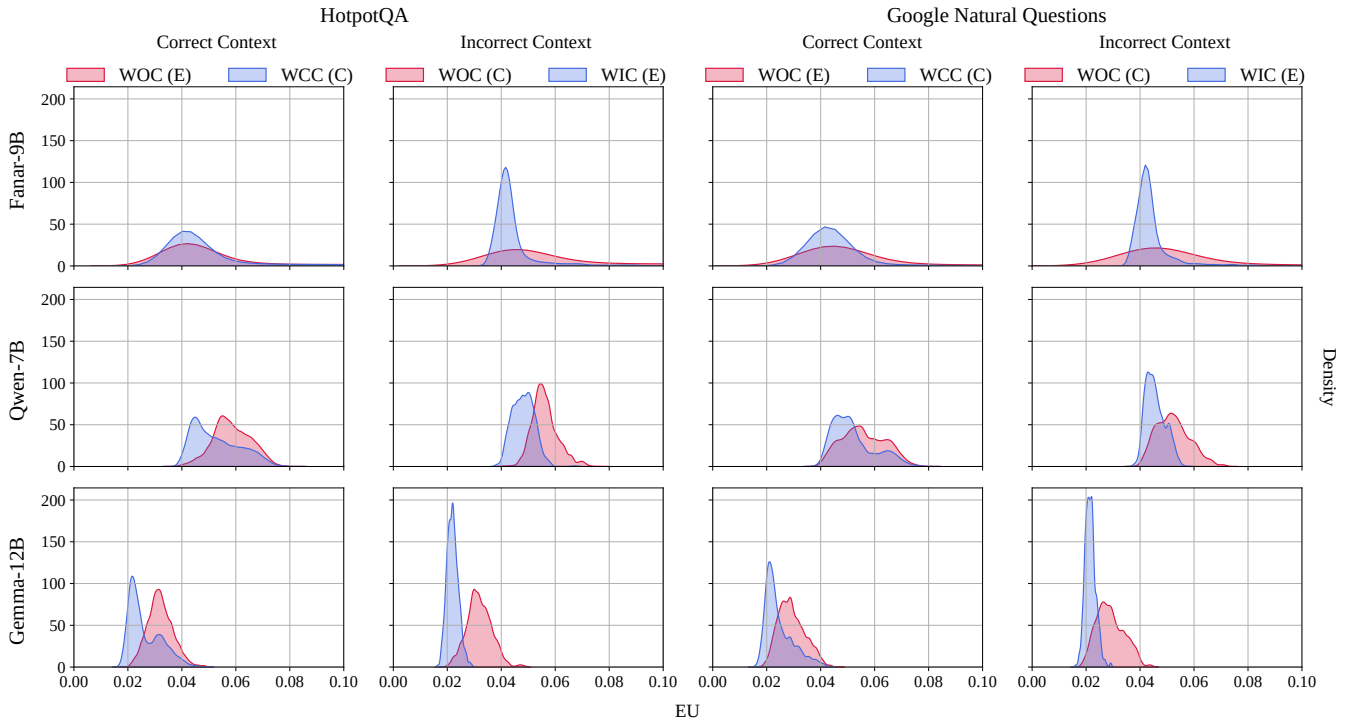

Figure 3: Model behavior transitions and epistemic uncertainty (EU) distribution shifts across HotpotQA and Natural Questions for three models (`Fanar1-9b`, `Qwen2.5-7B`, `Gemma3-12B`). Each subplot displays the distribution of lower-bound epistemic uncertainty scores for subsets of questions whose correctness regime changes between the no-context (`WOC`) and context-enhanced (`WCC` or `WIC`) settings. We focus on two key transitions: (1) `WOC:E` $\rightarrow$ `WCC:C`, where injecting correct context into previously incorrect responses leads to improved correctness and decreased uncertainty; and (2) `WOC:C` $\rightarrow$ `WIC:E`, where misleading context causes the model to produce incorrect responses with sustained low uncertainty. These shifts highlight how in-context information modulates both model predictions and confidence, revealing risks of overconfident confabulations in the presence of incorrect input.

lations triggered by incorrect external information, despite possessing sufficient internal knowledge.

Figure 3 visualizes the distribution of lower-bound epistemic uncertainty across these subsets using kernel density estimation (KDE), allowing for comparison of uncertainty profiles before and after the context shift. Results are shown for three models: `Fanar1-9b`, `Qwen2.5-7B`, and `Gemma3-12B`, on the HotpotQA and Natural Questions datasets.

Correct context reduces uncertainty. As expected, we observe a clear and consistent decrease in epistemic uncertainty in the transition from incorrect responses without context to correct responses with context (`WOC:E`$\rightarrow$`WCC:C`). Across all models, the KDE curves corresponding to the `WCC:C` setting shift leftward relative to those from the `WOC:E` setting, indicating that providing accurate contextual information not only improves answer correctness but also increases model confidence. This effect is particularly pronounced for `Qwen2.5-7B` and `Gemma3-12B`, where the uncertainty distributions in the `WCC:C` condition are sharply concentrated around low epistemic uncertainty values.

Misleading context induces confident errors. We analyze the setting where models transition from correct predictions without context (`WOC:C`) to incorrect predictions with misleading context (`WIC:E`). Ideally, such a transition should result in higher epistemic uncertainty, reflecting the model's recognition of ambiguity or conflict, visualized as broader, right-shifted distributions. However, all models instead show a contraction in their EU distributions, with `WIC:E` responses exhibiting sharper and more left-skewed profiles.

`Fanar1-9b`, despite appearing flat under correct context conditions, exhibits a notable increase in peakedness and reduced variance under misleading context, indicating an unjustified confidence in its wrong answers. This suggests that Fanar is responsive to misleading context and exhibits similar calibration issues as the other models, even if the mean EU shift is modest. `Qwen2.5-7B` also produces more confident predictions under misleading context, with `WIC:E` curves shifting left and becoming narrower relative to `WOC:C`. `Gemma3-12B` shows the most extreme behavior, with the narrowest and most left-shifted `WIC:E` distribution. This reflects strong contextual dependence but very poor calibration when that context is misleading.

These results reveal a dual role of contextual information in large language model behavior. When context is accurate, it reliably improves both correctness and model confidence.

However, misleading context can cause models to produce incorrect answers with high certainty. These findings align with expectations and emphasize the importance of robust uncertainty estimation in detecting context-induced confabulations. They motivate future research in reliability-aware generation and mechanisms for validating or filtering context in multi-turn or retrieval-augmented generation settings. In the following section, we investigate how to use uncertainty information to guide the response reliability detection.

Effectiveness of Uncertainty-Guided Probing for Reliability Detection

As shown in our analysis of **RQ1**, token-level uncertainty is not always aligned with correctness, particularly under in-context learning. In the presence of misleading information, models may produce confident yet incorrect responses. This phenomenon raises concerns in multi-turn or retrieval-augmented settings, where such confabulated outputs may be reused as context in future turns. This observation underscores the limitations of using uncertainty alone as a reliability signal when external context is present.

However, in scenarios where the model relies solely on its internal parameters (i.e., without additional context), uncertainty may still provide meaningful cues about response reliability. This motivates our investigation in **RQ2**: *Can token-level uncertainty, when combined with internal representations, be used to detect unreliable responses?*

We explore this question by training probing classifiers on token-level hidden states from various layers and positions, using both static and uncertainty-aware token selection strategies. Our goal is to assess whether internal signals, especially those grounded in model confidence, can serve as reliable indicators of output correctness.

Response reliability detection. We consider the following method from the related literature of uncertainty, reliability, and hallucination detection.

- **LogProb:** This method computes the mean of log-probability scores of the generated tokens (Yadkori et al. 2024a)

$$\frac{1}{T}\sum_{t=1}^{T}\log \mathbb{P}(y_t|\mathbf{p}, \mathbf{y}_{<t}).$$

- **P(True):** This method prompts the LLMs to judge whether their answer is correct. Our prompt followed the following template from Kadavath et al. (2022).
- **LogTokU:** This method computes the aggregated aleatoric and epistemic uncertainty to predict the response reliability. We follow the aggregation method from Ma et al. (2025).
- **Probing.** We train lightweight classifiers on token-level hidden states $\mathbf{h}_t^{(l)}$ to predict response-level reliability following previous work (Li et al. 2023). We consider several token selection strategies:
  - **Probe(EOS):** Uses the final generated token $\mathbf{h}_T^{(l)}$.
  - **Probe(Exact):** Selects tokens aligned with the exact answer span (Orgad et al. 2024).
  - **Probe(EU):** Selects the single token with either the highest or lowest epistemic uncertainty score.
  - **Probe(AVG):** Average hidden states across selected token subsets (e.g., top-k uncertain tokens or fixed positions) to form an aggregated feature representation.

Performance metric. We use the area under the receiver operating characteristic curve (AUROC) to evaluate the performance of reliability detectors. This metric summarizes the model's ability to distinguish between positive and negative cases across all classification thresholds, effectively balancing sensitivity (true positive rate) and specificity (false positive rate).

Model	Method	TruthfulQA	TriviaQA	Math
Fanar	**LogProb**	0.597	0.774	0.757
	P(true)	0.530	0.672	0.635
	LogTokU	0.541	0.683	0.666
	Prob(Exact)	<u>0.711</u>	**0.783**	0.827
	Probe(EOS)	0.706	0.739	0.790
	Probe(EU)	0.709	0.751	<u>0.794</u>
	Probe(AVG)	**0.734**	<u>0.765</u>	**0.833**
Qwen	**LogProb**	0.591	0.774	0.635
	P(true)	0.537	0.736	0.664
	LogTokU	0.642	0.773	0.565
	Prob(Exact)	0.758	0.781	0.627
	Probe(EOS)	**0.794**	**0.812**	**0.703**
	Probe(EU)	0.759	0.754	0.646
	Probe(AVG)	<u>0.761</u>	<u>0.786</u>	<u>0.699</u>
Gemma	**LogProb**	0.545	0.806	0.683
	P(true)	0.598	0.631	0.779
	LogTokU	**0.790**	0.611	<u>0.791</u>
	Prob(Exact)	0.728	0.796	0.773
	Probe(EOS)	0.728	<u>0.810</u>	**0.834**
	Probe(EU)	0.687	0.751	0.669
	Probe(AVG)	<u>0.733</u>	**0.818**	0.786

Table 1: Comparison of probing methods across `Fanar1-9b`, `Qwen2.5-7B`, and `Gemma3-12B` models on three datasets. We report AUROC scores (3-decimal precision). Bold indicates the best in each column; underlined indicates the second-best.

Reliability detection cross layers and tokens. Figure 4 presents the AUROC scores of probing classifiers trained on hidden states from the last 20 layers, using different token-level feature strategies under the epistemic uncertainty setup. Each heatmap column represents a token selection method ranging from single-token probing (e.g., using the token with highest or lowest uncertainty) to aggregated representations computed by averaging hidden states across multiple tokens.

We observe that individual token features (left columns) often yield weaker performance, especially in earlier layers. In contrast, aggregated features (right columns) consistently lead to better classification results. This trend holds across all models. In particular, strategies like `EU AVG (1-5) + EOS` achieve the highest AUROC scores, especially when features are extracted from middle to upper layers. These results

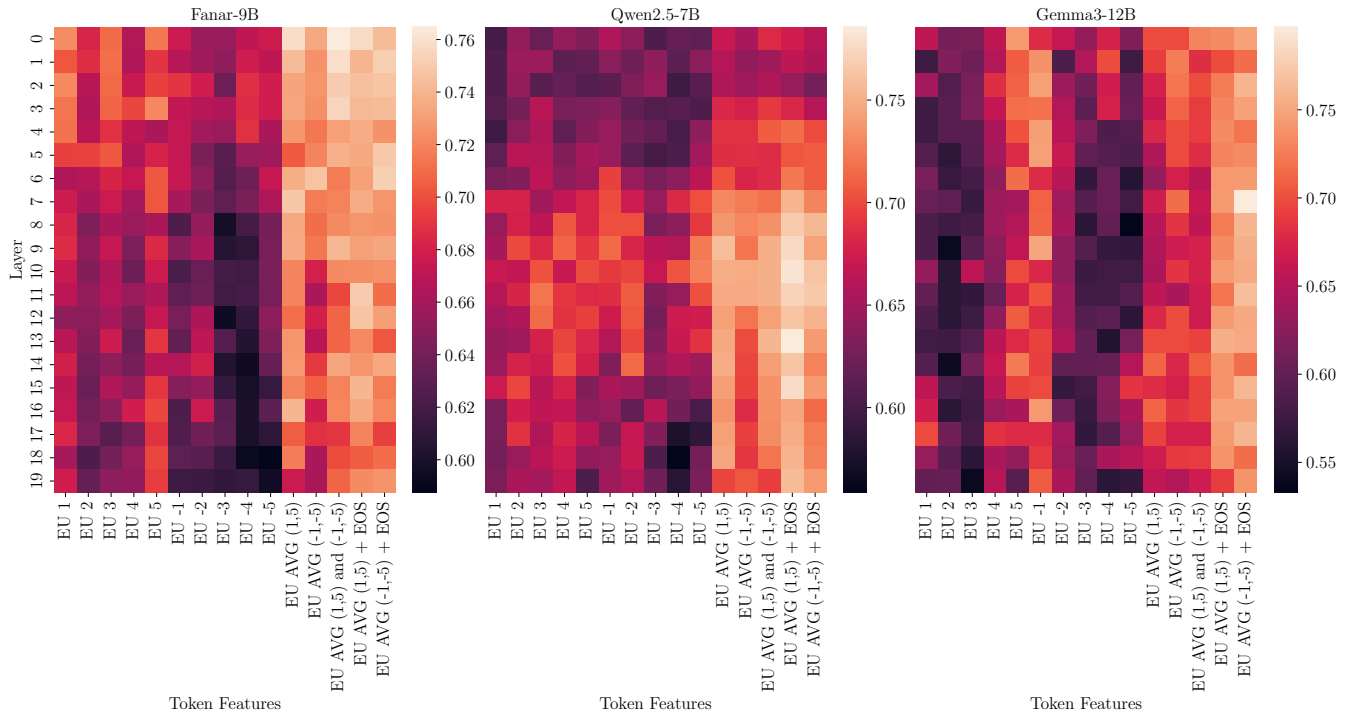

Figure 4: AUROC scores of probing classifiers across the last 20 layers using different token-level features, evaluated on the TriviaQA dataset for `Fanar1-9b`, `Qwen2.5-7B`, and `Gemma3-12B`. From left to right, columns correspond to probing with single tokens ranked by epistemic uncertainty: the k smallest (EU 1 to EU 5) and the k largest (EU -1 to EU -5). Aggregated features (EU AVG) formed by averaging hidden states across selected tokens yield the highest detection performance across all models.

suggest that combining multiple token-level signals enhances the robustness of response-level reliability detection.

Comparison with Uncertainty-Based Baselines. Next, we compare the reliability detection performance of different methods. Table 1 summarizes the AUROC performance of different methods across three LLMs (`Fanar1-9b`, `Qwen2.5-7B`, `Gemma3-12B`) and three QA datasets (TruthfulQA, TriviaQA, Math). Probing methods clearly outperform uncertainty-only baselines such as **LogProb** and **P(true)**, demonstrating the added value of internal model representations. Among all methods, **Probe(AVG)** yields the best overall performance, followed by **Probe(EOS)** and **Probe(EU)**. Although `Gemma3-12B` achieves strong performance with **LogTokU** on TruthfulQA, probing methods are more robust across tasks. Notably, performance is higher on TriviaQA and Math, indicating that response reliability is more predictable in factoid-style and structured QA than in open-ended questions.

These findings highlight the effectiveness of token-level probing for reliability detection. Aggregating hidden states over uncertain or boundary tokens provides a strong signal and consistently outperforms uncertainty-only baselines. This supports the utility of internal representations in enabling more reliable LLM-generated outputs.

Failure analysis. We observe that last-layer reliability probes are not optimal: the best-performing probes usually lie in middle layers, but this remains poorly understood and we lack a principled way to select the best layer.

Conclusion and Future Work

This work examines how LLMs respond to contextual inputs, focusing on failure modes. Accurate context improves both accuracy and confidence, whereas misleading context induces confidently incorrect answers, exposing a mismatch between uncertainty estimates and true correctness. This raises concerns for multi-turn and retrieval-augmented generation, where confabulated responses may propagate. To flag unreliable outputs, we propose a probing method that leverages token-level hidden states and uncertainty-guided token selection. Across multiple models and datasets, this approach outperforms direct uncertainty baselines; aggregating signals from multiple, especially high-uncertainty, tokens yields stronger reliability estimates. While our analysis focuses on question answering tasks, extending these techniques to open-ended generation and multi-turn dialogue remains an open challenge. Future work could explore incorporating reliability signals into generation-time decisions, combining probing-based methods with retrieval validation, and developing safeguards to limit the propagation of confabulated content in interactive applications.

Acknowledgements

This research was supported by the ERC Advanced Grant REBOUND [834862] and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Additionally, this research was partially supported by the Ministry of Transport, Qatar under grant number OVPR-SRO-MOT-2025-001.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion*, 76(C): 243–297.

An, S.; Zhang, Z.; Wang, Z.; Yuan, H.; Li, X. L.; and Yih, W.-t. 2023. Skill-Based Few-Shot Prompting for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578. Bangkok, Thailand: Association for Computational Linguistics.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 630(8017): 625–630.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).

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

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, K.; Patel, O.; Viégas, F. B.; Pfister, H.; and Wattenberg, M. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Li, L.; Chen, Z.; Chen, G.; Zhang, Y.; Su, Y.; Xing, E.; and Zhang, K. 2024. Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models.

Ma, H.; Chen, J.; Zhou, J. T.; Wang, G.; and Zhang, C. 2025. Estimating LLM Uncertainty with Evidence. *arXiv preprint arXiv:2502.00290*.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822. Toronto, Canada: Association for Computational Linguistics.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 9004–9017. Association for Computational Linguistics.

Obeso, O.; Arditi, A.; Ferrando, J.; Freeman, J.; Holmes, C.; and Nanda, N. 2025. Real-time Detection of Hallucinated Entities in Long-form Generation. *arXiv preprint arXiv:2509.03531*.

Orgad, H.; Toker, M.; Gekhman, Z.; Reichart, R.; Szpektor, I.; Kotek, H.; and Belinkov, Y. 2024. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. International Conference on Learning Representations (ICLR).

Qin, Y.; Li, S.; Nian, Y.; Yu, X. V.; Zhao, Y.; and Ma, X. 2025. Don't Let It Hallucinate: Premise Verification via Retrieval-Augmented Logical Reasoning. arXiv:2504.06438.

Ravi, S. S.; Mielczarek, B.; Kannappan, A.; Kiela, D.; and Qian, R. 2024. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*.

Reinhard, P.; Li, M. M.; Fina, M.; and Leimeister, J. M. 2025. Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 3183–3193. Red Hook, NY, USA: Curran Associates Inc.

Simhi, A.; Herzig, J.; Szpektor, I.; and Belinkov, Y. 2024. Constructing benchmarks and interventions for combating hallucinations in llms. *arXiv preprint arXiv:2404.09971*.

Sriramanan, G.; Bharti, S.; Sadasivan, V. S.; Saha, S.; Kattakinda, P.; and Feizi, S. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. *Advances in Neural Information Processing Systems*, 37: 34188–34216.

Sui, P.; Duede, E.; Wu, S.; and So, R. 2024. Confabulation: The Surprising Value of Large Language Model Hallucinations. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14274–14284. Bangkok, Thailand: Association for Computational Linguistics.

Team, F.; Abbas, U.; Ahmad, M. S.; Alam, F.; Altinisik, E.; Asgari, E.; Boshmaf, Y.; Boughorbel, S.; Chawla, S.; Chowdhury, S.; et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Wang, X.; Wei, J.; Schuurmans, D.; Bosma, M.; Chi, E.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.; Huang, J.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; Du, C.; and Le, Q. V. 2024. Long-form Factuality in Large Language Models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Wu, Y.; Zeng, Z.; He, K.; Mou, Y.; Wang, P.; and Xu, W. 2022. Distribution Calibration for Out-of-Domain Detection with Bayesian Approximation. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 608–615. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Yadkori, Y. A.; Kuzborskij, I.; György, A.; and Szepesvári, C. 2024a. To Believe or not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Yadkori, Y. A.; Kuzborskij, I.; Stutz, D.; György, A.; Fisch, A.; Doucet, A.; Beloshapka, I.; Weng, W.-H.; Yang, Y.-Y.; Szepesvári, C.; Cemgil, A. T.; and Tomasev, N. 2024b. Mitigating LLM Hallucinations via Conformal Abstention. (arXiv:2405.01563). ArXiv:2405.01563 [cs].

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2369–2380. Association for Computational Linguistics.

Zhang, J.; Juan, D.-C.; Rashtchian, C.; Ferng, C.-S.; Jiang, H.; and Chen, Y. 2024. SLED: self logits evolution decoding for improving factuality in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Şensoy, M.; Kaplan, L. M.; Julier, S.; Saleki, M.; and Cerutti, F. 2025. Risk-aware classification via uncertainty quantification. *Expert Systems with Applications*, 265: 125906.