Sampling-Free Uncertainty Quantification via Hidden State Dynamics in Language Models

Yixin Bu[1, 2], Guanyun Zou[1, 2], Renzhi Wang[1, 2], Runze Xia[1, 2], Cunjun Wang[3], Hongliang Dai[1, 2], Xiaoqing Ma[3], Piji Li[1, 2*]

[1]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China
[2]The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing, 211106, China
[3]COMAC Shanghai Aircraft Design and Research Institute, Shanghai, 201210, China
{sx2416068, zouguanyun, rzhwang, xiarunze, pjli}@nuaa.edu.cn, {wangcunjun, maxiaoqing}@comac.cc

Abstract

Large language models (LLMs) demonstrate remarkable capabilities in various complex language tasks, yet they face significant reliability challenges, including factual inaccuracies and generated biases (Yao et al. 2024; Zhang et al. 2022; Hong et al. 2024). Uncertainty quantification (UQ) plays a pivotal role in assessing model trustworthiness, particularly for high-stakes applications. However, current UQ methods for LLMs encounter computational efficiency bottlenecks due to their reliance on extensive sampling or external model invocations. In this work, we introduce a novel, sampling-free uncertainty quantification framework centered on hidden layer representation analysis. Our method facilitates real-time uncertainty quantification by modeling hierarchical internal semantic dynamics during the generation process. Through comprehensive experiments on multiple QA datasets and diverse model scales, we show that our approach consistently outperforms existing uncertainty quantification techniques in distinguishing correct from incorrect generations. Our results reveal that analyzing the dynamic evolution of hidden states provides a potent and computationally efficient signal for uncertainty quantification, directly from the model's internal workings, surpassing methods that depend solely on output probabilities or approximations via multiple samples.

1 Introduction

Large Language Models (LLMs) (Brown et al. 2020; Chowdhery et al. 2023; Touvron et al. 2023; Chung et al. 2024), despite their inherent remarkable capabilities, exhibit significant reliability issues such as factual inaccuracies and generated biases (Yao et al. 2024; Zhang et al. 2022; Hong et al. 2024). Robust Uncertainty Quantification (UQ) is thus crucial for assessing their trustworthiness and ensuring safer deployment in high-stakes applications (Lin et al. 2023; Sharma et al. 2023).

UQ gauges model confidence, distinct from "hallucination"(Manakul, Liusie, and Gales 2023) (plausible yet false outputs), and helps predict such errors.Its inherently effective UQ for Natural Language Generation (NLG) is challenging due to the vast output space (Sai, Mohankumar, and Khapra 2022). While semantic clustering approaches (Kuhn, Gal, and Farquhar 2023) offer a valuable direction, many existing UQ methods remain bottle-necked by the computationally intensive nature of requisite Monte Carlo sampling. Furthermore, the reliance on invoking external model APIs for auxiliary tasks like semantic assessment (Aichberger et al. 2024; Chen and Mueller 2023) not only escalates computational overhead but critically introduces the risk of error accumulation from these systems.

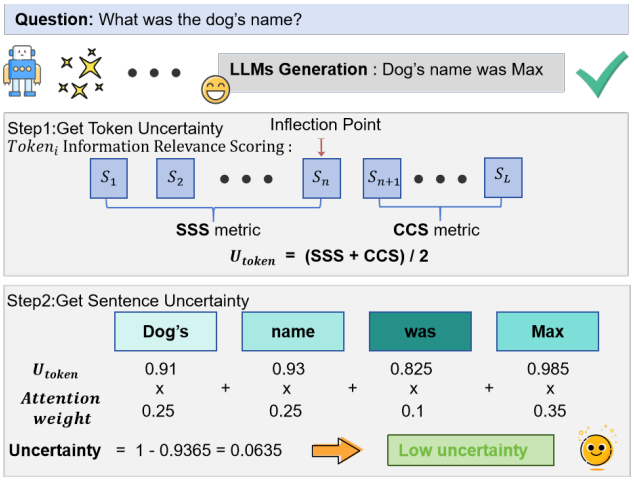

Figure 1: Our proposed framework for sampling-free uncertainty quantification. The model's internal hidden state dynamics are first translated into token-level scores by assessing semantic stability (SSS) and convergence (CCS). These scores are then intelligently aggregated using an attention mechanism to produce a final, sequence-level confidence score, enabling a reliable assessment of the generation's trustworthiness.

Our empirical investigations, leveraging interpretability techniques like the "logit lens" (nostalgebraist 2020; Belrose et al. 2023) , reveal a key phenomenon: tokens corresponding to well-established concepts or high-confidence predictions tend to emerge and stabilize in the vocabulary space decoded from earlier hidden layers. Conversely, less certain or more nuanced tokens often manifest clearly only in deeper layers (visualized in Figure 2). This observation leads us to hypothesize that these nuanced layer-wise evolutionary pathways—specifically, the timing of a token's emergence and the stability of its semantic representation across the

*Corresponding author.

network's depth—are not merely correlational but serve as strong, reliable, and direct indicators of the model's underlying certainty for that token.

Building on this hypothesis, our work operationalizes these internal state signals, encoding the model's generative deliberation, into a novel, sampling-free UQ framework. As detailed in our empirical study, we systematically track the trajectory of semantic similarity—a robust proxy for evolving semantic consensus—between each hidden layer's representation and the final output layer's representation. We find that the degree of semantic flux and the characteristics of convergence within these trajectories significantly differ between uncertain and confident generations. Our method then precisely quantifies these distinct dynamic patterns—specifically the emergent representational stability during an initial exploratory phase and the ultimate convergence decisiveness towards the final output—to derive actionable uncertainty scores. This direct interrogation of internal model states offers an efficient and more insightful approach to UQ, avoiding multiple forward passes or external model dependencies.

Our main contributions are threefold:

- A novel sampling-free UQ framework analyzing hierarchical semantic dynamics via hidden layer representations, eliminating multiple forward passes and external model calls.
- Two distinct metrics, Semantic Stability Score for exploration robustness and Convergence Confidence Score for decision certainty, providing a holistic uncertainty view.
- Extensive experiments show that our approach significantly outperforms existing methods in distinguishing correct/incorrect generations with high computational efficiency.

2 Empirical Study

2.1 Token Prediction Dynamics Across Transformer Layers

Leveraging interpretability techniques like the "logit lens" (nostalgebraist 2020; Belrose et al. 2023), we investigated token prediction evolution across transformer layers. This method decodes hidden states into vocabulary space, revealing how predictions mature and their associated confidence levels. Our analysis (visualized in Figure 2) shows that while early layers yield ambiguous predictions, later layers become increasingly precise. Notably, well-established concepts (e.g., named entities, common phrases) often emerge with high confidence in earlier layers. This suggests that the timing and stability of these early predictions may reflect the model's certainty and internalization of specific knowledge, a hypothesis central to our work.

2.2 Semantic Evolution Through Layer-wise Similarity Analysis

To explore this hypothesis, we initially examined Kullback-Leibler (KL) divergence between layer-wise probability distributions and the final output distribution. While KL divergence generally decreases with depth (Figure 4), per-token

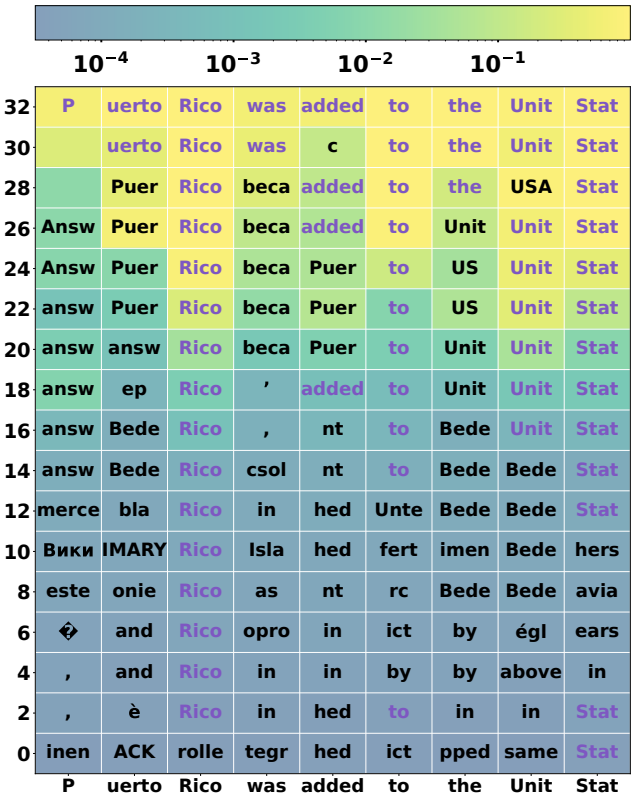

Figure 2: Token prediction dynamics across transformer layers. The heatmap shows how predictions evolve through hidden layers (vertical) during generation (horizontal). Confident generations exhibit early and accurate token predictions (highlighted regions), motivating our layer-wise uncertainty quantification approach.

trajectories can be noisy due to representational mismatches (Belrose et al. 2023); solutions like tuned probes add computational overhead.

We thus adopted a more direct and efficient measure: cosine similarity between each hidden layer's embedding and the final layer's output embedding. This approach effectively captures semantic relatedness (Church 2017; Vaswani et al. 2017; Mikolov et al. 2013). Analyzing CoQA (Reddy, Chen, and Manning 2019) samples (classified as correct/incorrect and confident/uncertain per (Kuhn, Gal, and Farquhar 2023)), we observed distinct evolutionary phases. As shown in Figure 3 (a-d), uncertain generations exhibit prolonged semantic fluctuations in early layers and slower convergence. In stark contrast, however, these confident generations show early stability and rapid, decisive convergence.

Further, analyzing inter-layer cosine similarity—the similarity between adjacent layer embeddings—reveals a notable pattern. Early layers rapidly achieve high similarity that subsequently stabilizes, indicating consistent semantic evolution (as shown in Figure 5). Crucially, this inter-layer similarity often decreases across the final few layers. We posit that this decrease signifies the model focusing its representational capacity along specific hidden dimensions to finalize the prediction for the target token. These dynamic character-

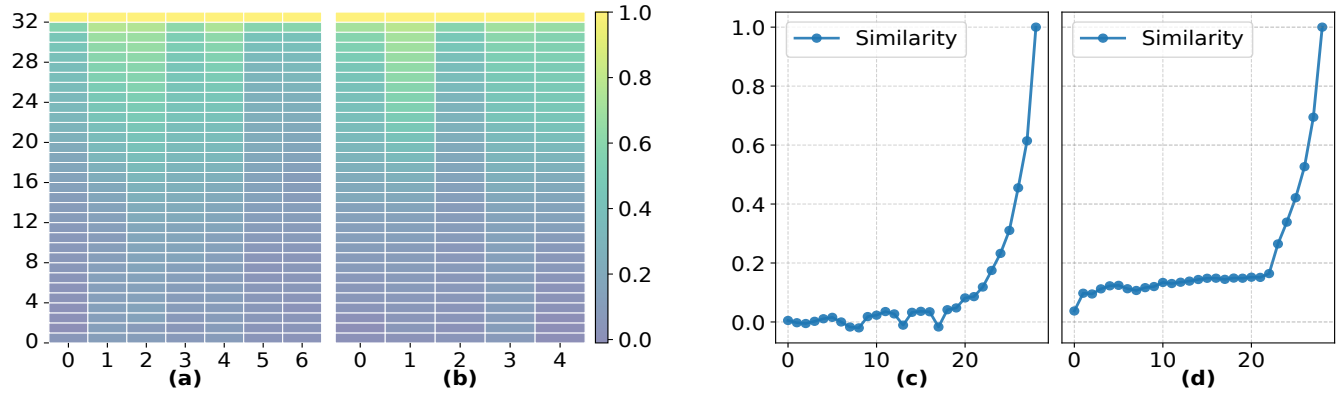

Figure 3: (a) and (b) provide token-level heatmaps comparing uncertain and certain model outputs. Notably, uncertain predictions demonstrate gradual color transitions across processing layers, whereas certain predictions converge with greater rapidity. (c) and (d) further present layer-wise similarity curves, underscoring increased early-stage semantic fluctuations and the protracted convergence specifically for uncertain cases.

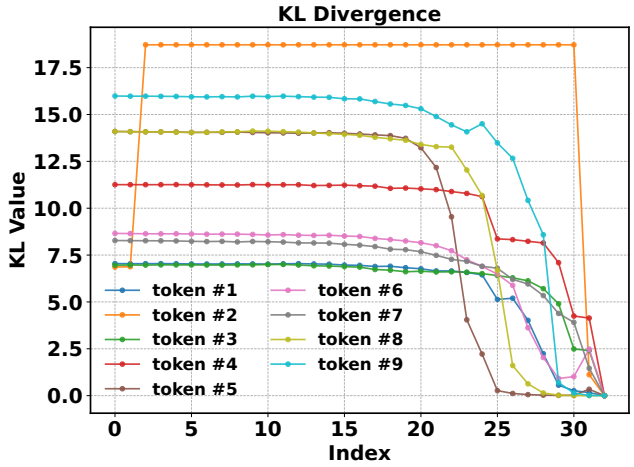

Figure 4: Illustration of the layer-wise evolution of KL divergence for individual tokens, with each trajectory highlighted in a distinct colour. The resulting curves are highly irregular, indicating that the per-token KL dynamics do not follow a consistent convergence pattern.

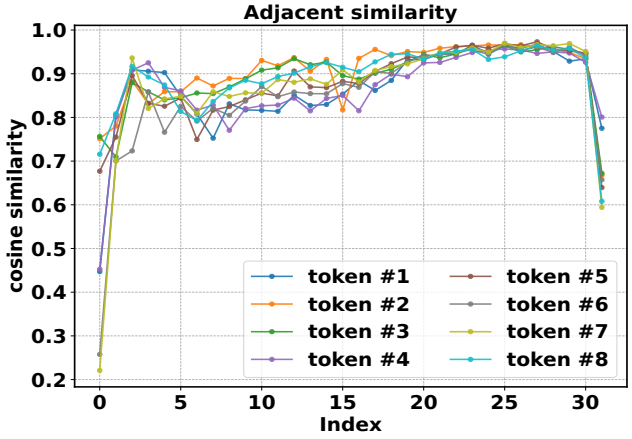

Figure 5: Adjacent layer cosine similarity for multiple tokens. Most tokens exhibit rapid increases in similarity at early layers, followed by stable high similarity, indicating consistent semantic evolution across layers.

istics of inter-layer similarity further underscore that internal semantic dynamics provide strong and often overlooked signals for uncertainty.

3 Method

Our method quantifies uncertainty by analyzing the layer-wise evolution of semantic relevance within Transformer hidden states during text generation. The process involves three main steps: (1) scoring the semantic relevance of each token across all layers, (2) identifying a critical inflection point in its semantic trajectory using our Curvature-driven Dynamic Inflection Detection (CDID) algorithm, and (3) computing uncertainty metrics based on the token's behavior before and after this point.

3.1 Information Relevance Scoring

For each token t in a generated sequence of length T, its hidden state tensor across L layers is $H_t \in \mathbb{R}^{L \times d}$. We define the Layer-wise Semantic Relevance Score (LSR-Score), denoted by $s_t^{(l)}$, for layer l as its cosine similarity to the final layer's hidden state $\boldsymbol{h_t^{(L)}}$:

$$s_t^{(l)} = \frac{\boldsymbol{h_t^{(l)}} \cdot \boldsymbol{h_t^{(L)}}}{||\boldsymbol{h_t^{(l)}}|| \cdot ||\boldsymbol{h_t^{(L)}}||} \quad (1)$$

This yields a relevance trajectory $\{s_t^{(1)}, ..., s_t^{(L)}\}$ for each token, capturing its semantic evolution towards its final representation.

3.2 Dynamic Inflection Detection

To identify critical transitions in these relevance trajectories, we employ the Curvature-driven Dynamic Inflection Detection (CDID) algorithm. We conceptualize CDID as a princi-

pled tool for locating the "knee" in the semantic trajectory, which we posit corresponds to the layer where a token's representation shifts from initial exploration to final refinement.

First, to mitigate noise in the raw relevance trajectories, which can be erratic, we apply a Savitzky-Golay smoothing filter (window width $w = 5$):

$$\hat{s}^{(l)} = \frac{1}{2w+1} \sum_{k=-w}^{w} a_k s^{(l+k)} \tag{2}$$

Subsequently, we analyze the geometry of the smoothed sequence $\hat{s}$ by computing its first and second derivatives using finite differences to understand its rate of change and concavity:

$$\begin{aligned} \hat{s}'^{(l)} &= \hat{s}^{(l+1)} - \hat{s}^{(l)} \\ \hat{s}''^{(l)} &= \hat{s}'^{(l+1)} - \hat{s}'^{(l)} \end{aligned} \tag{3}$$

The dynamic inflection point τ, signifying a critical juncture where semantic relevance stabilizes, is identified as the layer with maximum normalized curvature $\kappa^{(l)}$. Curvature is a natural choice for this task as it robustly identifies points of sharpest transition in a trajectory, independent of scale.

$$\kappa^{(l)} = \frac{|\hat{s}''^{(l)}|}{(1 + [\hat{s}'^{(l)}]^2 + \epsilon)^{3/2}} \tag{4}$$

Here, $\epsilon = 10^{-6}$ is a small constant for numerical stability. The inflection point τ is then:

$$\tau = \arg\max_{l \in [\delta, L]} \kappa^{(l)} \tag{5}$$

The search for τ is constrained to layers $l \in [\delta, L]$, where $\delta = 10$ is a heuristic threshold to filter out transient fluctuations often observed in the initial, shallower layers, ensuring the detected point reflects a meaningful semantic shift.

3.3 Uncertainty Quantification Metrics

Based on the identified inflection point τ, we derive two distinct uncertainty metrics, SSS and CCS. Their design is motivated by the empirical observation that when a model is confident about a token, its semantic relevance trajectory exhibits low variance in the early layers followed by a swift, monotonic convergence in the later layers. Conversely, uncertainty manifests as high initial fluctuation and a hesitant, slow, or non-monotonic convergence post-inflection. Our metrics are crafted to quantify these two phenomena directly.

The **Semantic Stability Score (SSS)** quantifies the stability of a token's representation during the initial exploratory phase (from layer 1 to τ_t). It is defined as one minus the variance of the relevance scores in this pre-inflection phase. A high SSS (approaching 1) indicates minimal variance and thus high confidence in the initial semantic interpretation.

$$\text{SSS}_t = 1 - \frac{1}{\tau_t} \sum_{l=1}^{\tau_t} \left(s_t^{(l)} - \frac{1}{\tau_t} \sum_{i=1}^{\tau_t} s_t^{(i)} \right)^2 \tag{6}$$

The **Convergence Confidence Score (CCS)** measures the decisiveness of semantic convergence in the layers subsequent to the inflection point τ_t. It is defined as the average slope of the relevance trajectory from the inflection point to the final layer. A high positive CCS indicates a strong and direct progression towards the final representation, signifying confidence.

$$\text{CCS}_t = \frac{1}{L - \tau_t} \sum_{l=\tau_t+1}^{L} \left(s_t^{(l)} - s_t^{(l-1)} \right) \tag{7}$$

3.4 Uncertainty Aggregation

Token-level Uncertainty ($\mathcal{U}_{\text{token}}$) is a linear combination of the two proposed metrics. As high SSS and CCS both indicate confidence, we formulate uncertainty as:

$$\mathcal{U}_{\text{token}}^{(t)} = \alpha \cdot (\text{SSS}_t) + \beta \cdot (\text{CCS}_t) \tag{8}$$

Here, α and β are not learnable parameters but rather hyperparameters that balance the two aspects of uncertainty. In our experiments, we use a default equal weighting ($\alpha = \beta = 0.5$), but these can be tuned on a validation set for task-specific optimization. This makes the framework transparent and largely unsupervised at its core.

For Sequence-level Uncertainty ($\mathcal{U}_{\text{seq}}$), we use an attention-guided aggregation of token uncertainties, positing that tokens with higher attention weights are more influential to the sequence's overall meaning and uncertainty.

$$\mathcal{U}_{\text{seq}} = \sum_{t=1}^{T} \omega_t \cdot \mathcal{U}_{\text{token}}^{(t)} \tag{9}$$

The token weights ω_t are derived from the attention distribution of a pre-selected head h^*. To clarify the notation, let $a_{T,t'}^{(h)}$ be the attention weight that head h in the final layer places on token t' when generating the final token of the sequence. The weights are these attention scores directly, with no further transformation applied: $\omega_{t'} = a_{T,t'}^{(h^*)}$.

The selection of this head, h^*, is crucial. We hypothesize that heads with high-entropy distributions, which attend broadly across the sequence, are better proxies for global context than low-entropy heads that focus narrowly. This data-driven selection allows our method to adapt to different models or tasks, which may use different heads to capture holistic dependencies. Therefore, h^* is chosen as the head that exhibits the maximum average entropy $\bar{H}^{(h)}$ over a representative dataset D (e.g., the training or validation set of the target task).

$$h^* = \arg\max_h \bar{H}^{(h)}, \quad \text{where} \quad \bar{H}^{(h)} = \frac{1}{|D|} \sum_{x \in D} H(A_{T_x}^{(h)}) \tag{10}$$

The entropy for head h on a sequence of length T is calculated as $H(A_T^{(h)}) = -\sum_{t'=1}^{T} a_{T,t'}^{(h)} \log a_{T,t'}^{(h)}$. By selecting the highest-entropy head, we leverage an attention mechanism that provides a holistic view of token interdependencies for weighting uncertainty contributions.

4 Experimental Setup

Our experimental setup involves a curated selection of pre-trained large language models, a diverse set of datasets, and

established evaluation metrics and baselines to assess model performance and uncertainty quantification techniques.

Models. We employ a diverse selection of Qwen(Bai et al. 2023) and Llama(Touvron et al. 2023) models, ranging from 1B to 8B parameters. These models are used in their pre-trained state, without task-specific fine-tuning. Full architectural specifics are detailed in Appendix.

Datasets. Evaluation is performed on a diverse suite of datasets including CoQA, NewsQA, SQuAD, Natural Questions (NQ), GSM8K, and Multi30K, spanning question answering, reasoning, and translation. We randomly sampled 1,000 instances per dataset, balancing cost with statistical significance, for our experiments. Refer to Appendix for dataset characteristics and sampling procedures.

Metrics. Generation correctness is assessed using ROUGE-L, Exact Match (EM), and BERTScore-F1. For robustness, ROUGE-L and BERTScore-F1 are evaluated with a primary threshold, with further threshold analyses presented in Appendix. Uncertainty quantification is primarily evaluated using Area Under the Receiver Operating Characteristic curve (AUROC) and Prediction Reliability Rate (PRR). AUROC measures the ability to distinguish between correct and incorrect generations, while PRR quantifies the reliability of predictions at various confidence levels. More complete details on these metrics and their justification are in Appendix.

Baselines. Our method is compared against several uncertainty quantification techniques, primarily focusing on white-box methods. These include Predictive Entropy (PE), Length-Normalized PE (LN-PE), Lexical Similarity (LS), Semantic Entropy (SE), Number of Semantic Clusters (NSC), Shifting Attention to Relevance (SAR). Detailed descriptions of these baselines are available in Appendix.

5 Result and Analysis

5.1 Main Results

Our comprehensive evaluation (Table 1) demonstrates that our layer-wise UQ method ("Ours") achieves marked superiority across all tested models, sizes, and datasets for question answering. It consistently outperforms baselines on two critical metrics: Percentage of Rejected incorrect answers (PRR) and AUROC. For instance, on the NQ dataset with the Llama 8B model, our method achieves a PRR of 32.18% and an AUROC of 83.67%, significantly surpassing strong baselines like Shifting Attention to Relevance (SAR) and Semantic Entropy (SE) by a considerable margin. The substantial improvement in PRR further highlights our method's practical utility in filtering out unreliable answers.

Furthermore, the superiority of our approach is by no means confined to QA tasks. As shown in Table 2, our method also achieves state-of-the-art performance on diverse domains such as mathematical reasoning (GSM8K) and machine translation (Multi-30k). For example, on GSM8K with the Llama-8B model, our method achieves an AUROC of 71.43, substantially outperforming SAR (65.41) and SE (60.31). This success across these highly varied domains underscores the general applicability and robustness of our uncertainty quantification technique.

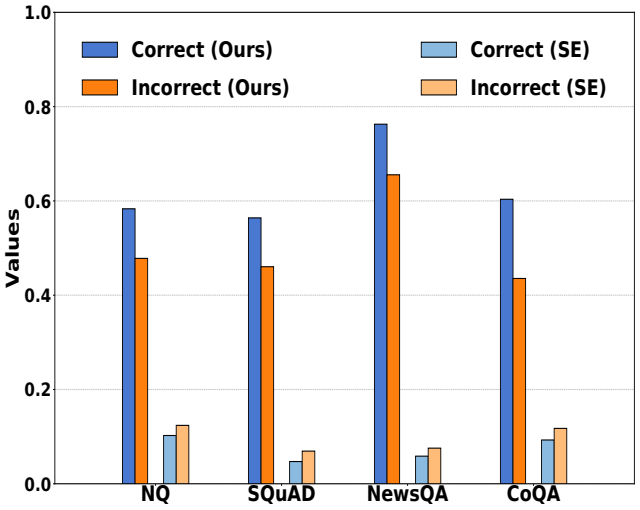

Figure 6: Normalized average uncertainty scores (Ours vs. SE) for correct/incorrect predictions (Qwen-7B, four datasets). Higher "Ours"/lower "SE" scores mean more confidence. Our method shows better separation.

The superior performance of our method in uncertainty estimation stems from its nuanced assessment of model certainty. We achieve this by interrogating the layer-by-layer evolution of hidden states, capturing subtle representational shifts that signal the model's confidence or doubt. This granular internal analysis explains its consistent outperformance against strong baselines (SE, LNPE, SAR), securing a 3 to 8 point advantage in AUROC and a substantial margin in Precision-Recall Rate (PRR)—a crucial metric for imbalanced settings.

Furthermore, the method's robustness and generalization are rigorously validated in Appendix. We demonstrate its superiority across a spectrum of correctness criteria, from the stringent exact_match to semantic similarity metrics like Rouge-L and bertscore_f1. These findings establish that analyzing internal state dynamics is an effective, principled approach to reliable uncertainty quantification. Crucially, our method achieves this in a single forward pass, offering a computationally efficient alternative to prevalent ensemble-based techniques.

5.2 Analysis

Computational Efficiency. A critical advantage of our methodology is its computational efficiency. As illustrated in Figure 7 (and further detailed in Appendix for Llama models), our single-pass approach offers substantial reductions in execution time compared to representative sampling-based baselines. This speedup is primarily because we avoid the iterative process of baselines, which often require multiple generation passes (e.g., five per instance). This foundational N-fold cost is then compounded by other processing overhead within the baseline's script, leading to the 25-fold advantage observed in our comparison of total script runtimes. For instance, with the Qwen 7B model, our method reduced computation time from approximately 25 hours to about 1 hour on datasets like CoQA and NewsQA. This sig-

Model	Size	Dataset	Ours	SE	SAR	LNPE	PE	LX	SE_num
Qwen	7B	CoQA	**60.19/76.66**	5.70/73.43	13.70/76.03	5.61/72.04	16.51/65.32	31.31/66.75	5.51/68.45
		NewsQA	**59.86/74.32**	25.20/70.05	25.20/69.36	25.61/71.83	18.22/68.45	23.60/65.43	25.71/67.32
		SQuAD	**75.95/71.85**	44.38/65.37	44.35/68.52	44.41/67.52	56.21/67.43	16.31/63.65	24.62/64.38
		NQ	**35.09/78.41**	6.24/75.61	10.52/70.10	6.46/74.32	5.05/74.38	12.47/65.43	6.35/68.37
	3B	CoQA	**49.41/66.95**	33.99/63.74	33.96/61.24	33.99/63.75	41.64/60.13	26.63/59.62	30.11/60.13
		NewsQA	**43.09/70.94**	41.26/64.43	35.42/68.74	33.69/65.36	39.77/62.45	29.54/60.32	30.39/61.56
		SQuAD	**50.79/70.06**	32.22/64.38	40.92/65.44	32.06/64.32	44.74/61.43	24.99/61.11	30.19/62.38
		NQ	**33.58/77.85**	33.01/70.15	28.99/75.05	12.48/69.43	12.25/65.78	23.19/60.74	32.00/63.43
	1.5B	CoQA	**49.48/64.83**	41.23/61.33	48.68/62.31	41.25/61.45	39.81/60.32	31.23/55.74	40.01/59.12
		NewsQA	**37.68/64.35**	33.31/62.38	30.73/62.45	34.08/62.33	36.99/60.15	25.73/58.35	25.33/60.18
		SQuAD	**50.12/68.41**	25.31/63.47	31.52/60.31	30.19/64.33	33.41/60.11	15.96/59.91	22.00/61.07
		NQ	**32.04/71.66**	23.44/65.62	18.35/66.32	23.45/65.08	25.66/61.45	22.33/60.75	19.83/55.43
Llama	8B	CoQA	**56.34/73.23**	9.30/70.14	10.35/69.03	13.56/71.03	29.88/67.91	5.31/65.09	9.31/60.03
		NewsQA	**33.49/70.01**	32.19/66.35	30.52/67.45	29.33/63.37	25.14/65.43	16.38/63.21	30.15/60.53
		SQuAD	**59.64/72.97**	45.31/67.42	46.37/70.31	38.73/68.35	44.38/67.91	25.17/62.04	33.59/66.66
		NQ	**32.18/83.67**	23.88/75.09	19.38/77.48	22.03/75.61	25.96/73.31	9.45/68.07	15.69/64.09
	3B	CoQA	**52.95/70.03**	45.38/63.05	46.35/65.42	45.38/64.71	44.31/60.04	25.07/55.97	35.95/64.67
		NewsQA	**29.82/67.21**	25.38/64.76	17.35/66.75	24.52/63.04	22.21/60.15	9.37/62.32	23.17/63.79
		SQuAD	**49.65/71.67**	32.76/65.03	30.43/64.03	33.15/66.43	45.31/62.46	22.73/63.05	25.65/61.40
		NQ	**28.98/76.91**	9.33/69.93	10.38/70.15	8.59/70.35	12.33/65.80	25.73/60.08	22.15/66.91
	1B	CoQA	**21.53/65.88**	9.65/61.12	9.65/63.31	7.38/61.11	9.38/60.04	28.77/55.70	14.88/57.03
		NewsQA	**29.68/65.94**	15.23/61.15	16.53/60.15	15.98/62.03	16.99/59.93	9.38/56.37	20.35/54.23
		SQuAD	**22.96/62.31**	20.33/59.39	18.95/55.43	19.73/60.91	22.33/57.70	8.99/56.13	5.33/50.03
		NQ	**25.03/64.19**	15.87/61.33	20.37/60.03	15.88/61.35	20.39/60.04	16.93/57.90	16.99/52.41

Table 1: Comparison of uncertainty quantification methods across different models and datasets. The values are presented as PRR (%) / AUROC (%). For both metrics, higher is better.

Model	Dataset	Ours	SAR	SE
Qwen-7B	GSM8K	**70.56**	60.41	57.83
	Multi-30k	**83.41**	75.38	73.41
Llama-8B	GSM8K	**71.43**	65.41	60.31
	Multi-30k	**85.83**	78.66	79.05

Table 2: AUROC (%) performance comparison on mathematical reasoning (GSM8K) and machine translation (Multi-30k) tasks. Higher values are better.

nificant efficiency gain, achieved by eliminating the need for multiple generation samples, makes our method more practical for a wider range of real-world applications, especially those demanding low latency, high throughput, and efficient resource utilization.

Discriminative Power of Uncertainty Metrics.A critical advantage of our methodology is its computational efficiency. As illustrated in Figure 7 (and further detailed in Appendix for Llama models), our single-pass approach offers substantial reductions in execution time compared to representative sampling-based baselines. This speedup is primarily because we avoid the iterative process of baselines, which often require multiple generation passes (e.g., five per instance). This foundational N-fold cost is then compounded

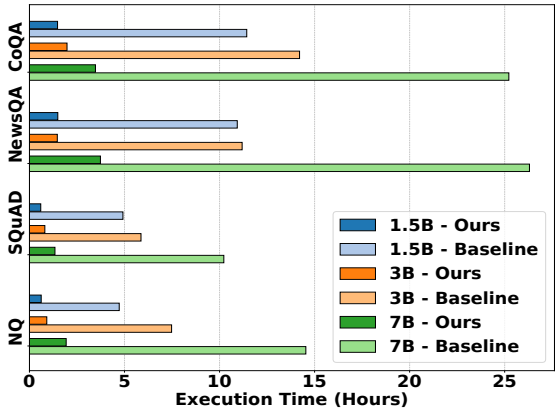

Figure 7: Qwen model execution time comparison: Our method ("Ours") demonstrates significant speedup over the sampling-based baseline ("Baseline").

by other processing overhead within the baseline's script, leading to the 25-fold advantage observed in our comparison of total script runtimes. For instance, with the Qwen 7B model, our method reduced computation time from approximately 25 hours to about 1 hour on datasets like CoQA and NewsQA. This significant efficiency gain, achieved by

Model	Size	CoQA	NewsQA	SQuAD	NQ
Qwen	7B	**76.66** / 51.14 / 47.43	**74.32** / 44.79 / 50.37	**71.85** / 41.00 / 49.77	**78.41** / 59.51 / 50.71
	3B	**66.95** / 48.82 / 61.66	**70.94** / 47.34 / 61.03	**70.06** / 48.44 / 59.78	**77.85** / 65.24 / 50.27
	1.5B	**64.83** / 48.07 / 58.57	**64.35** / 45.8 / 58.55	**68.41** / 56.00 / 51.46	**71.66** / 45.62 / 58.28
Llama	8B	**73.23** / 52.47 / 52.59	**70.01** / 49.33 / 50.09	**72.97** / 59.70 / 50.46	**83.67** / 70.80 / 57.40
	3B	**70.03** / 56.83 / 57.47	**67.21** / 52.53 / 56.42	**71.67** / 52.21 / 61.93	**76.91** / 62.48 / 67.83
	1B	**65.88** / 54.62 / 53.45	**65.94** / 50.76 / 52.34	**62.31** / 45.69 / 52.57	**64.19** / 56.09 / 60.80

Table 3: Ablation study: AUROC (%) of using only SSS, only CCS, and their combination (Ours) on different models and datasets. Cell values are reported as: Ours / SSS / CCS.

Model	Size	CoQA	NewsQA	SQuAD	NQ
Qwen	7B	**76.66** / 74.54 / 74.54	**74.32** / 72.06 / 72.06	**71.85** / 67.53 / 67.53	**78.41** / 68.41 / 68.41
	3B	**66.95** / 64.83 / 64.85	**70.94** / 66.04 / 66.06	**70.06** / 67.83 / 67.87	**77.85** / 73.97 / 74.07
	1.5B	**64.83** / 62.54 / 62.55	**64.35** / 61.62 / 61.62	**68.41** / 65.87 / 65.93	**71.66** / 66.58 / 66.66
Llama	8B	**73.23** / 70.52 / 70.53	**70.01** / 62.42 / 62.42	**72.97** / 66.93 / 66.93	**83.67** / 68.45 / 68.46
	3B	**70.03** / 65.55 / 65.56	**67.21** / 63.26 / 63.27	**71.67** / 67.89 / 67.89	**76.91** / 76.25 / 76.24
	1B	**65.88** / 55.71 / 55.71	**65.94** / 63.41 / 63.41	**62.31** / 60.05 / 60.05	**64.19** / 62.65 / 62.66

Table 4: Ablation study: AUROC (%) of different sequence-level aggregation strategies (Ours, AVG, Dynamic) on different models and datasets. Cell values are reported as: Ours / AVG / Dynamic.

eliminating the need for multiple generation samples, makes our method more practical for a wider range of real-world applications, especially those demanding low latency, high throughput, and efficient resource utilization.

Orthogonality to Semantic Entropy. We computed Pearson correlation coefficients between our metric and Semantic Entropy (SE) across datasets and models. Table 5 shows consistently low values near zero (e.g., 0.088 for Qwen-7B on NewsQA, 0.035 for Llama-8B on SQuAD), indicating our method and SE capture different uncertainty aspects. While SE relies on sampled output diversity, our approach analyzes hidden states' internal dynamics during a single generation. This difference enables our method to identify uncertainty signals orthogonal to SE. Our metric's superior discriminative power, evidenced by AUROC scores, shows that analyzing internal dynamics provides a more effective view of model confidence, complementing sampling-based entropy measures.

Model	CoQA	NewsQA	SQuAD	NQ
Qwen 7B	0.083	0.088	0.075	0.033
Qwen 3B	0.035	0.047	0.075	0.084
Qwen 1.5B	0.043	0.087	0.093	0.054
Llama 8B	0.042	0.035	0.014	0.064
Llama 3B	0.133	0.057	0.072	0.107
Llama 1B	0.056	0.075	0.135	0.067

Table 5: Pearson correlation between our metric and SE across models/datasets. Low values suggest weak linear relationship and differing uncertainty aspects.

5.3 Ablation Study

Ablation studies (Tables 3 and 4) evaluated our UQ framework's components: the Semantic Stability Score (SSS), Convergence Confidence Score (CCS), and sequence-level aggregation.

SSS and CCS Complementarity. As shown in Table 3, we evaluate the contributions of Semantic Stability Score (SSS) and Convergence Confidence Score (CCS). SSS assesses representational stability during the initial exploratory phase, while CCS measures the decisiveness of final convergence. While effective individually, results confirm that their combination ("Ours") yields superior AUROC scores. This highlights their complementarity: SSS captures early-stage semantic fluctuations, whereas CCS reflects the model's ultimate commitment, offering a comprehensive and robust uncertainty quantification across diverse architectures.

Aggregation Strategy Effectiveness. Table 4 examines sequence-level aggregation mechanisms for transforming token-level scores. We compare our attention-guided strategy ("Ours") against uniform averaging ("AVG") and "Dynamic" head selection. Our method, using a consistent high-entropy attention head, significantly outperforms both baselines. This confirms that a globally informed, stable weighting scheme effectively emphasizes salient token uncertainties while mitigating noise, thereby enhancing the overall accuracy of sequence-level uncertainty quantification.

6 Conclusion

We introduced a novel, sampling-free UQ framework that exploits hidden state dynamics, offering a computationally lean method that consistently outperformed baselines in distinguishing correct/incorrect generations. By eliminating the latency associated with standard sampling techniques, our method proves highly efficient for real-time applications. Our approach, based on synergistic SSS and CCS metrics and rigorously validated by ablation studies, confirms internal dynamics as a potent UQ signal, providing a solid foundation for fostering safer, more trustworthy AI systems.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.62476127), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Basic Research Program of the Bureau of Science and Technology (ILF24001), the Meituan Research Fund (No.PO250624101698), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

Aichberger, L.; Schweighofer, K.; Ielanskyi, M.; and Hochreiter, S. 2024. Semantically diverse language generation for uncertainty estimation in language models. *arXiv preprint arXiv:2406.04306*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. arXiv:2309.16609.

Belrose, N.; Furman, Z.; Smith, L.; Halawi, D.; Ostrovsky, I.; McKinney, L.; Biderman, S.; and Steinhardt, J. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; and Mueller, J. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Church, K. W. 2017. Word2Vec. *Natural Language Engineering*, 23(1): 155–162.

Hong, J.; Duan, J.; Zhang, C.; Li, Z.; Xie, C.; Lieberman, K.; Diffenderfer, J.; Bartoldson, B.; Jaiswal, A.; Xu, K.; et al. 2024. Decoding compressed trust: Scrutinizing the trustworthiness of efficient llms under compression. *arXiv preprint arXiv:2403.15447*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Lin, B.; Bouneffouf, D.; Cecchi, G.; and Varshney, K. R. 2023. Towards healthy AI: large language models need therapists too. *arXiv preprint arXiv:2304.00416*.

Manakul, P.; Liusie, A.; and Gales, M. J. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

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

Sai, A. B.; Mohankumar, A. K.; and Khapra, M. M. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2): 1–39.

Sharma, A.; Lin, I. W.; Miner, A. S.; Atkins, D. C.; and Althoff, T. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1): 46–57.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.