Sampling Less, Knowing More: Contradiction-aware Adaptive Spectral and Entropic Uncertainty Quantification for LLMs

Jiaming Zhang[1], Heng-yang Lu[2], Xinan Chen[1], Junyuan Zheng[1]
Yuncheng Jiang[1], Ying Zhu[1,*], Chenyou Fan[1,*]
[1]South China Normal University, Guangdong, China
[2]Jiangnan University, Jiangsu, China
zhang_jia_ming@m.scnu.edu.cn, {zhuy, fanchenyou}@scnu.edu.cn

Abstract

Uncertainty quantification (UQ) is essential for improving the reliability of large language models (LLMs). However, multi-sample UQ methods typically employ a fixed sampling budget for all queries, leading to computational redundancy. In addition, multi-sample UQ methods often fail in the presence of consistent errors. To address these limitations, we propose CASE (**C**ontradiction-aware **A**daptive **S**pectral and **E**ntropic UQ), a UQ framework that jointly models semantic disagreement and token-level uncertainty under an adaptive sampling paradigm. Specifically, CASE constructs a contradiction graph over sampled responses and quantifies semantic disagreement using normalized spectral norms of the graph representation. The resulting uncertainty signals are further leveraged for adaptive sampling, enabling query-dependent budget allocation while reducing redundant computation. In addition, CASE incorporates token-level predictive entropy as a complementary uncertainty signal to mitigate the failure of graph structures under consistent errors. Experimental results demonstrate that CASE outperforms existing multi-sample UQ methods across factual question answering datasets while reducing sampling costs by over 30% on average.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks (OpenAI et al., 2024; Grattafiori et al., 2024; Yang et al., 2025; Singh et al., 2026). However, the pervasive issue of hallucination of LLMs poses a major roadblock to their application in high-stakes domains such as healthcare and law (Huang et al., 2025; Anh-Hoang et al., 2025; Zhu et al., 2025; Simhi et al., 2025). Consequently, robust uncertainty quantification (UQ) becomes indispensable for calibrating model outputs and mitigating potential risks (Geng et al., 2024; Liu et al., 2025).

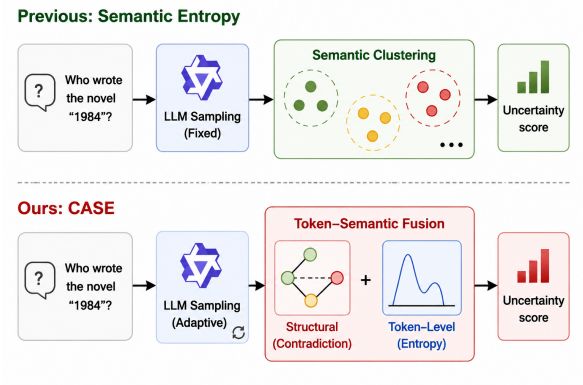


Figure 1: Comparison of UQ paradigms. Existing semantic-level approaches (top) rely on fixed-budget sampling and semantic clustering over generated responses. In contrast, CASE (bottom) performs adaptive sampling based on the evolving contradiction structure and combines spectral semantic uncertainty with token-level predictive entropy for more robust UQ.

Most existing multi-sample UQ methods for LLMs rely on a fixed sampling budget, estimating uncertainty from semantic diversity across sampled responses regardless of query difficulty (Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025; Li et al., 2025) (Figure 1, top). However, the degree of semantic disagreement differs substantially across queries. For many queries, sampled responses exhibit highly similar semantic patterns, causing additional samples to provide limited new information (Farquhar et al., 2024; Ciosek et al., 2025; McCabe et al., 2026; Hamidieh et al., 2026). Existing multi-sample UQ methods nevertheless employ a fixed sampling budget across all queries, resulting in substantial computational redundancy.

Beyond computational inefficiency, multi-sample UQ methods depend on semantic disagreement among sampled responses as the primary uncertainty signal (Manakul et al., 2023). Semantic-level uncertainty estimation is effective when incorrect generations exhibit substantial semantic variation, but becomes unreliable when models repeat-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 25092–25106
October 24-29, 2026 ©2026 Association for Computational Linguistics

edly generate semantically consistent yet factually incorrect responses (Farquhar et al., 2024). Strong semantic agreement under consistent errors can produce spuriously low uncertainty estimates, making reliable separation between correct predictions and confidently incorrect generations difficult.

To address these limitations, we propose CASE (**C**ontradiction-aware **A**daptive **S**pectral and **E**ntropic UQ), a UQ framework that jointly models semantic disagreement and token-level uncertainty under an adaptive sampling paradigm (Figure 1, bottom). By evaluating the evolving contradiction patterns among sampled responses, this framework dynamically determines whether newly generated samples contribute incremental uncertainty information. Specifically, CASE constructs a contradiction graph over sampled answers, where nodes correspond to candidate responses and edges encode pairwise contradictions. To ensure comparable uncertainty estimates across varying sampling budgets, the framework quantifies contradiction intensity using the normalized spectral norms of the graph representation.

Leveraging this spectral metric, CASE expands the sampling budget only when newly generated responses substantially alter the graph topology, and terminates sampling once the structure stabilizes, thereby reducing redundant computation for stable queries while preserving exploration for uncertain ones. To ensure robustness against consistent errors where spectral measures fail, CASE further incorporates token-level predictive entropy as a complementary uncertainty signal.

In summary, our contributions are as follows:

- We propose CASE, a contradiction-aware UQ framework that adaptively allocates sampling budgets according to the evolving contradiction structures among sampled responses.
- CASE introduces a spectral UQ strategy based on normalized spectral norms, enabling comparable UQ across varying sampling budgets.
- CASE incorporates token-level predictive entropy to mitigate failures of spectral measures under consistent errors.
- Experiments on multiple factual question answering datasets with open-source LLMs demonstrate that CASE outperforms existing multi-sample UQ methods while reducing sampling costs by more than 30%.

2 Related Work

Early approaches to UQ in LLMs primarily rely on single-sample internal model signals such as token-level predictive entropy and sequence likelihood (Desai and Durrett, 2020; Ovadia et al., 2019; Kadavath et al., 2022; Lin et al., 2022). However, single-sample uncertainty signals are often affected by superficial surface-form variations and fail to reliably capture uncertainty. Recent work has therefore shifted toward multi-sample UQ methods, which estimate uncertainty through semantic consistency across multiple generated responses (Kojima et al., 2022; Wang et al., 2023; Manakul et al., 2023). Representative approaches include information-theoretic uncertainty estimation methods based on semantic entropy and its variants, including kernel language entropy and semantic nearest neighbor entropy (Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025). In addition, several other methods have been proposed to further enhance estimation reliability (Shi et al., 2025; Li et al., 2025; Sun et al., 2026). Despite the efficacy of multi-sample UQ methods, they generally rely on rigid, fixed-budget sampling, leading to substantial computational redundancy across queries of varying difficulty. Furthermore, prevailing multi-sample UQ methods typically quantify uncertainty through local semantic clustering or pairwise comparisons, limiting the capacity of conventional multi-sample UQ metrics to capture global contradiction structures and failing in the presence of consistent errors. Consequently, these limitations pose a major obstacle to the practical deployment of multi-sample UQ methods in real-time, resource-constrained scenarios where both efficiency and robustness are critical.

3 Preliminaries: Algebraic Graph Theory

We briefly introduce the algebraic graph theory concepts underlying our UQ method (Chung, 1997).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with k nodes and a symmetric adjacency matrix $H \in \mathbb{R}^{k \times k}$. The spectrum of H consists of its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. The spectral radius of H is defined as $\rho(H) = \max_i |\lambda_i|$.

For a symmetric matrix, the spectral norm induced by the L_2 vector norm satisfies

$$\|H\|_2 = \rho(H). \tag{1}$$

Moreover, if H is non-negative, the Perron–Frobenius theorem (Horn and Johnson, 2012) guar-

antees that the principal eigenvalue is real and non-negative, yielding

$$\|H\|_2 = \rho(H) = \lambda_{\max}(H). \tag{2}$$

A brief derivation is provided in Appendix A. The spectral norm quantifies the overall connectivity strength within the graph, capturing how strongly information is aggregated across nodes through edge connections. Larger values correspond to more densely and strongly connected structures, thereby indicating tighter coupling and higher global interaction intensity within the graph. In Section 4, we leverage this structural property as an uncertainty signal in our UQ framework.

4 Methodology

We propose CASE, a spectral-semantic UQ framework (illustrated in Figure 2). At the semantic level, CASE models response contradictions as a sparsified graph and quantifies uncertainty via its spectral norm, utilizing adaptive sampling for computational efficiency. At the token level, it integrates predictive entropy to improve robustness against semantically consistent errors. Ultimately, CASE synergizes these structural and localized signals to provide a highly reliable uncertainty score.

4.1 Semantic-Level Uncertainty

4.1.1 Sparsified Contradiction Graph

Given an input query x and the LLM $\mathcal{M}$, we perform adaptive sampling to obtain a set of candidate responses $\mathcal{A}_K = \{a_1, a_2, \ldots, a_K\}$, where the sample size grows from $k = 2$ to a final budget K. At each intermediate step k, we construct a contradiction graph $\mathcal{G}_k = (\mathcal{A}_k, \mathcal{E}_k)$ to model semantic contradictions among the current candidates. In this graph, nodes correspond to generated responses, and edges encode the intensity of pairwise semantic contradictions.

Following prior works, we employ a pre-trained Natural language inference (NLI) model to evaluate directed semantic relations for each pair (a_i, a_j) (Farquhar et al., 2024; Nikitin et al., 2024). Let $P_C(a_i, a_j)$ and $P_E(a_i, a_j)$ denote the predicted probabilities of contradiction and entailment, respectively. NLI outputs often induce dense graphs, as sentence pairs with strong entailment relations may still receive non-zero contradiction probabilities. To mitigate this issue, we apply a non-linear sparsification function:

$$W_{i,j} = \max\Big(0, P_C(a_i, a_j) - P_E(a_i, a_j)\Big), \tag{3}$$

where $W_{i,j}$ denotes the directed edge weight representing the contradiction intensity from premise a_i to hypothesis a_j. Under this formulation, $W_{i,j} > 0$ only when the contradiction probability exceeds the entailment probability. This induces a sparse graph that highlights semantic contradictions while suppressing noisy edges.

Furthermore, NLI relations are asymmetric, which can introduce spurious unidirectional contradictions. To focus on contradictions, we symmetrize the directed matrix W into an undirected matrix $H^{(k)} \in \mathbb{R}^{k\times k}$ using the geometric mean:

$$H^{(k)}_{i,j} = \begin{cases} \sqrt{W_{i,j}W_{j,i}}, & \text{if } i \neq j, \\ 0, & \text{if } i = j. \end{cases} \tag{4}$$

where $H^{(k)}_{i,j}$ denotes the symmetric contradiction weight between responses a_i and a_j, and a non-zero value indicates the presence of bidirectional contradiction. This construction retains only mutual contradictions and filters out asymmetric relations that do not reflect genuine semantic disagreement.

4.1.2 Spectral Contradiction Score

Having constructed the sparsified symmetric contradiction matrix $H^{(k)} \in \mathbb{R}^{k\times k}$, we quantify the global semantic uncertainty by analyzing its spectral norm $\|H^{(k)}\|_2$. Drawing on the algebraic graph theory concepts introduced in Section 3, the spectral norm reflects the overall concentration of mutually contradictory interactions. As semantically conflicting responses typically indicate high uncertainty in the generative process, a larger $\|H^{(k)}\|_2$ signifies stronger global semantic inconsistency among the sampled responses.

However, $\|H^{(k)}\|_2$ is inherently scale-dependent, as its magnitude generally increases with the sample size k. To establish a comparable metric across different queries, we define the normalized spectral uncertainty score at step k as follows:

$$U^{(k)}_{\text{spec}} = \frac{\|H^{(k)}\|_2}{k-1}. \tag{5}$$

This specific choice of denominator is not arbitrary, as it represents the theoretical upper bound of the spectral norm. Consequently, we can demonstrate that this normalization naturally bounds the proposed uncertainty score within $[0, 1]$. Since $P_C, P_E \in [0, 1]$, it follows that the sparsified edge weights satisfy $W_{i,j} \in [0, 1]$, and thus $H^{(k)}_{i,j} \in [0, 1]$. For such a non-negative symmetric matrix, its spectral norm is bounded by the maximum

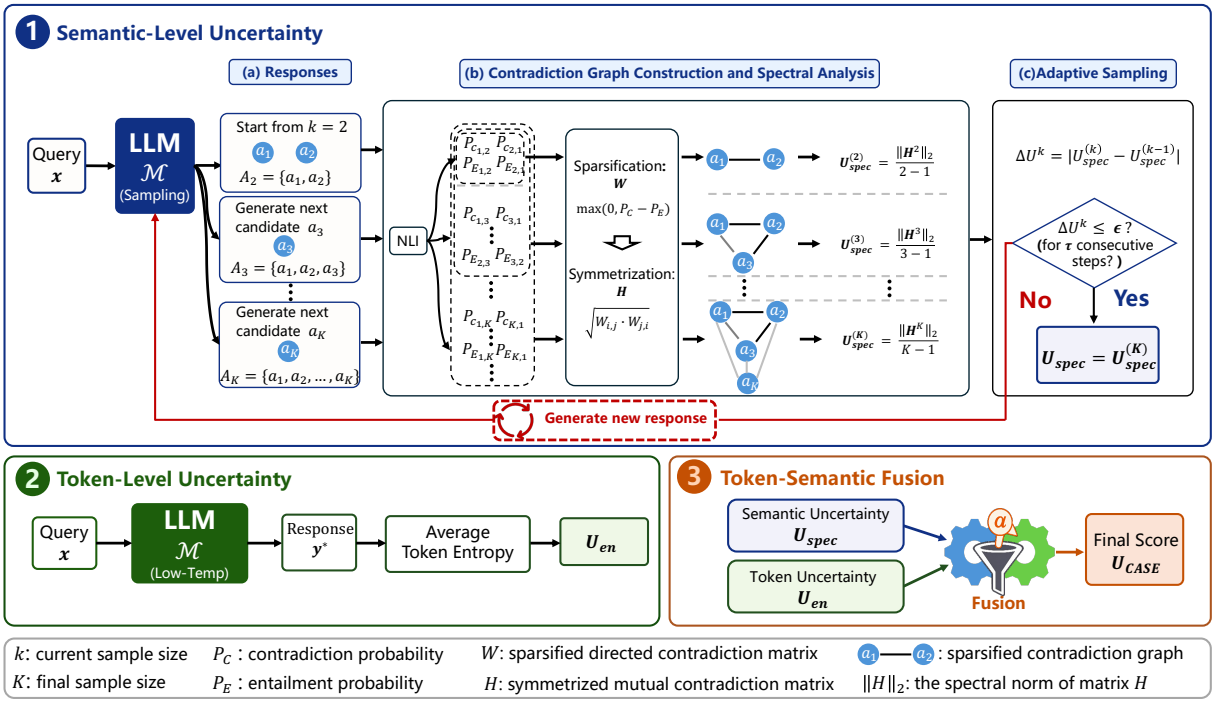


Figure 2: **Overview of the CASE framework.** The proposed architecture operates through three main components: (1) Semantic-Level Uncertainty quantifies the semantic uncertainty by adaptively sampling responses and calculating the spectral norm of their contradiction graph; (2) Token-Level Uncertainty derives the token uncertainty via the average token entropy of a low-temperature base response; and (3) Token-Semantic Fusion integrates both signals via a scaling factor α to produce the final uncertainty score U_{CASE}.

weighted node degree $d_{\max}$, yielding

$$\|H^{(k)}\|_2 \leq d_{\max} \leq k - 1. \tag{6}$$

Therefore, $U_{\text{spec}}^{(k)} \in [0, 1]$ is strictly guaranteed (a detailed proof is provided in Appendix B).

At the upper extreme, $U_{\text{spec}}^{(k)} = 1$ corresponds to an idealized, fully connected contradiction graph with maximal contradiction weights, indicating maximal semantic inconsistency. Conversely, values close to 0 reflect a high degree of agreement among generated responses, suggesting low semantic uncertainty.

4.1.3 Adaptive Sampling

A major limitation of existing multi-sample UQ frameworks is their reliance on a fixed sampling budget, which introduces unnecessary computational overhead for straightforward queries. To enable dynamic early stopping, a scale-invariant metric is desirable. In contrast to traditional metrics such as semantic entropy (SE) and kernel language entropy (KLE), which exhibit intrinsic scale dependence with increasing sample size (see Appendix C), our normalized score $U_{\text{spec}}^{(k)}$ is designed to be scale-invariant. This ensures that successive estimates remain comparable across different sampling cardinalities, providing a stable basis for a stopping criterion based on marginal changes.

Specifically, we employ a dynamic generation process coupled with a patience-based early stopping mechanism to prevent premature termination. Starting with an initial sample set $\mathcal{A}_2 = \{a_1, a_2\}$, at each subsequent step k, we query the LLM to generate a new candidate response a_k. We then incrementally update the sparsified contradiction matrix $H^{(k)}$ by evaluating the NLI-derived contradiction probabilities between a_k and all previously generated responses in $\mathcal{A}_{k-1}$. Based on this updated topology, we calculate the marginal spectral change:

$$\Delta U^{(k)} = \left| U_{\text{spec}}^{(k)} - U_{\text{spec}}^{(k-1)} \right|. \tag{7}$$

The sampling process terminates if $\Delta U^{(k)} \leq \epsilon$ for τ consecutive steps, where $\epsilon > 0$ is a tolerance threshold and τ is the patience parameter. The patience mechanism mitigates premature stopping caused by transient fluctuations in individual generations, thereby improving the stability. Upon termination at step K, we obtain the final semantic

uncertainty estimate:

$$U_{\text{spec}} = U_{\text{spec}}^{(K)} = \frac{\|H^{(K)}\|_2}{K-1}. \quad (8)$$

Fundamentally, a stabilized $\Delta U^{(k)}$ indicates that newly generated samples induce bounded incremental perturbations to the spectral structure of the contradiction matrix via rank-2 updates. Therefore, the stopping condition corresponds to a stable spectral regime in which further sampling yields only marginal changes in the normalized spectral uncertainty estimate. By terminating sampling upon this criterion, CASE balances estimation stability and computational efficiency across inputs of varying complexity (see Appendices D and E).

4.2 Token-Level Uncertainty

While adaptive sampling effectively captures contradiction structures within the generated response set, semantic uncertainty measures remain unreliable under consistent errors. In such cases, the contradiction graph degenerates, yielding $U_{\text{spec}}^{(K)} \approx 0$ and causing semantic-based signals to collapse.

To overcome this limitation, we directly incorporate token-level predictive uncertainty. Even when the sampled set $\mathcal{A}_K$ exhibits strong semantic consistency, the model's inherent predictive distributions during autoregressive generation often reveal latent uncertainty. To access the model's latent predictive uncertainty, we generate a single base response $y^* = (y_1^*, \ldots, y_L^*)$ using low-temperature decoding, which serves as the final candidate for correctness verification. We then compute the average token-level entropy of the model's predictive distribution across the generated tokens:

$$U_{\text{en}}(x) = -\frac{1}{L}\sum_{t=1}^{L}\sum_{v\in\mathcal{V}} P(v \mid x, y_{<t}^*) \log P(v \mid x, y_{<t}^*), \quad (9)$$

where L is the length of the low-temperature response y^*, and $\mathcal{V}$ denotes the model's vocabulary.

4.3 Token-Semantic Fusion

The final CASE uncertainty score is defined as:

$$U_{\text{CASE}}(x) = U_{\text{spec}}(x) + \alpha\, U_{\text{en}}(x), \quad (10)$$

where α is a scaling hyperparameter.

By coupling contradiction-driven semantic signals with token-level predictive entropy, this formulation mitigates degeneracy under consistent errors and provides more robust uncertainty estimates.

5 Experiments

5.1 Experimental Setup

Datasets and Models. We evaluate CASE on four factual QA datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and PopQA (Mallen et al., 2023). Following prior work SE (Farquhar et al., 2024) and KLE (Nikitin et al., 2024), we randomly sample 1,000 queries from the validation or test split of each dataset. We conduct experiments on three LLMs: LLaMA-3-8B-Instruct (Meta AI, 2024), Qwen3-8B (Qwen Team, 2025b), and Qwen3-14B (Qwen Team, 2025a).

Baselines. We compare CASE against UQ baselines. Single-sample baselines include predictive entropy (Malinin and Gales, 2021), perplexity (PPL) (Ren et al., 2023), and P(true) (Kadavath et al., 2022). Multi-sample baselines include SE (Farquhar et al., 2024), KLE (Nikitin et al., 2024), and SNNE (Nguyen et al., 2025). For SE, KLE, and CASE, we use `nli-deberta-v3-large` (Cross-Encoder Team, 2025), whereas SNNE uses `Qwen3-0.6B-embedding` (Qwen Team, 2025c). All multi-sample baselines sample exactly 10 answers per query.

Evaluation Metrics. Following prior work (Nikitin et al., 2024; Nguyen et al., 2025), we report the area under the receiver operating characteristic curve (AUROC) and the area under the accuracy–rejection curve (AUARC). Detailed definitions are provided in Appendix F.

Implementation Details. For each query, we first generate a deterministic answer (temperature $= 0$), which is used to compute token-level entropy U_{en} and evaluate correctness. We then construct the candidate set $\mathcal{A}$ using stochastic sampling with temperature $= 1.0$, top-$p = 0.9$, and top-$k = 50$. The samples are used to build the contradiction graph and compute the normalized spectral norm U_{spec}. Answer correctness is evaluated using Qwen-Flash under an LLM-as-a-judge protocol. For adaptive sampling, we use $K_{\min} = 2$, $K_{\max} = 10$, stopping tolerance $\epsilon = 0.01$, and patience $\tau = 2$. The token-semantic fusion weight is set to $\alpha = 1$. Additional implementation details in Appendix G.

Table 1: Main experimental results across four QA datasets. AUROC and AUARC scores are reported as percentages. The Samp. indicates the average number of sampled responses per query. Best results for each model are highlighted in **bold**, and second-best results are underlined.

Methods	PopQA			HotpotQA			NQ			TriviaQA		
	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓
LLaMA-3-8B-Instruct												
PPL	74.81	67.68	-	73.50	67.58	-	70.91	55.32	-	77.80	51.66	-
P(true)	78.51	71.41	-	63.43	59.64	-	78.16	60.36	-	79.82	51.81	-
Entropy	78.29	71.76	-	76.12	70.55	-	72.92	77.86	-	80.36	86.08	-
SE	77.32	69.43	10.0	75.58	70.25	10.0	71.42	76.66	10.0	81.06	85.60	10.0
KLE	82.30	72.35	10.0	78.61	71.05	10.0	79.77	79.97	10.0	84.96	87.48	10.0
SNNE	68.12	65.68	10.0	71.10	67.06	10.0	72.05	76.50	10.0	75.88	83.37	10.0
CASE (ours)	**83.22**	**73.94**	7.50	**81.34**	**73.45**	6.57	**80.74**	**81.42**	6.69	**85.28**	**88.03**	6.40
Qwen3-8B												
PPL	71.66	66.19	-	76.12	70.56	-	73.86	65.20	-	76.80	65.17	-
P(true)	61.71	59.81	-	67.25	65.12	-	72.32	66.62	-	75.46	67.56	-
Entropy	73.96	68.14	-	77.86	69.84	-	75.40	72.49	-	79.06	75.52	-
SE	75.68	69.37	10.0	79.01	69.40	10.0	76.86	71.48	10.0	79.56	74.23	10.0
KLE	80.35	71.54	10.0	81.08	69.96	10.0	79.61	71.95	10.0	80.52	73.84	10.0
SNNE	75.76	68.94	10.0	72.01	66.16	10.0	72.40	68.85	10.0	75.27	71.21	10.0
CASE (ours)	**81.43**	**72.15**	7.27	**81.65**	**71.49**	6.57	**80.63**	**74.41**	6.40	**82.45**	**76.67**	6.41
Qwen3-14B												
PPL	65.65	63.74	-	75.46	68.90	-	73.08	60.12	-	75.52	56.31	-
P(true)	66.67	64.84	-	78.28	70.49	-	76.18	61.31	-	**84.28**	62.34	-
Entropy	69.07	61.42	-	77.60	71.47	-	74.90	75.97	-	77.96	81.60	-
SE	70.24	61.67	10.0	78.37	67.36	10.0	72.84	73.57	10.0	78.30	80.15	10.0
KLE	76.16	63.60	10.0	80.99	70.42	10.0	76.76	74.91	10.0	79.36	78.24	10.0
SNNE	68.86	61.60	10.0	70.10	65.17	10.0	72.00	74.23	10.0	73.20	78.06	10.0
CASE (ours)	**77.00**	**65.55**	6.98	**82.39**	**73.16**	6.34	**79.37**	**77.49**	6.14	81.65	**82.53**	5.88

5.2 Main Results

Table 1 presents the main results on four QA datasets using three representative LLMs. Overall, CASE consistently achieves the best UQ performance in terms of both AUROC and AUARC across different datasets and model scales. Compared with multi-sample baselines such as SE and KLE, CASE not only improves estimation accuracy but also substantially reduces computational cost through adaptive sampling. Specifically, while SE and KLE require a fixed budget of 10 samples per query, CASE uses only 5.88–7.50 samples on average, reducing inference cost by approximately 25%–40%. Despite the lower sampling budget, CASE consistently outperforms all baselines in nearly all settings, indicating its ability to identify informative generations while avoiding redundant sampling. Moreover, the gains remain stable across different model scales, demonstrating the robustness and scalability of the proposed framework. Interestingly, as model capability improves, the average sampling budget required by CASE further decreases, suggesting that the framework can adaptively allocate computational resources according to the uncertainty characteristics of model outputs.

5.3 Ablation and Robustness Analysis

Effect of Token-level Entropy and Adaptive Sampling. Table 2 reports the ablation results of CASE on Qwen3-8B. Removing the token-level signal U_{en} consistently degrades performance across all datasets. This empirical drop validates our hypothesis that purely semantic metrics (U_{spec}) become unreliable under semantically consistent errors, whereas token-level predictive entropy effectively compensates by capturing the model's latent uncertainty (Section 4.2). Furthermore, we evaluate our adaptive sampling mechanism. While experiencing a marginal drop in AUROC and AUARC scores, the adaptive approach maintains highly competitive performance while requiring only 6.40–7.27 samples on average. This substantial 27%–36% reduction in generation overhead firmly validates our spectral stopping criterion: once the contradiction graph stabilizes, further sampling merely introduces redundant computational costs without

Table 2: Ablation study of CASE on the Qwen3-8B model. AUROC and AUARC scores are reported as percentages. Best results are highlighted in **bold**, and second-best results are underlined.

Method	PopQA			HotpotQA			NQ			TriviaQA		
	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓	AUROC⇑	AUARC⇑	Samp.⇓
CASE (ours)	81.43	72.15	7.27	81.65	71.49	6.57	80.63	74.41	6.40	82.45	76.67	6.41
- *w/o* U_{en}	80.50	70.52	7.27	77.62	65.37	6.57	77.81	70.00	6.40	78.28	72.60	6.41
- *w/o adaptive sampling*	**82.64**	**73.07**	10.0	**82.08**	**71.75**	10.0	**81.55**	**75.15**	10.0	**83.29**	**77.25**	10.0

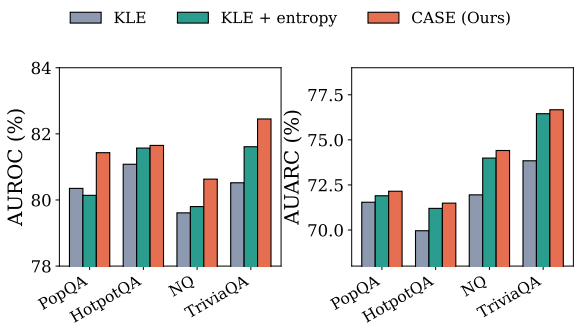


Figure 3: Performance comparison of CASE against KLE, with and without token-level entropy integration. The number of samples is set to 10 for all methods on the Qwen3-8B model. CASE consistently achieves the best performance across all datasets.

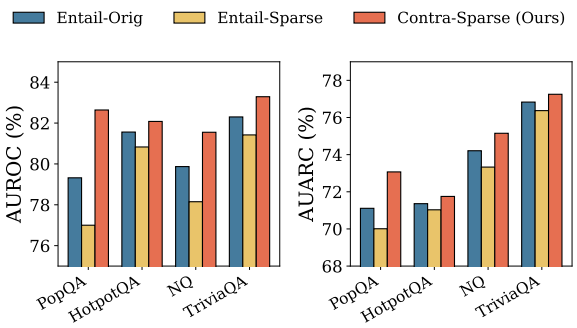


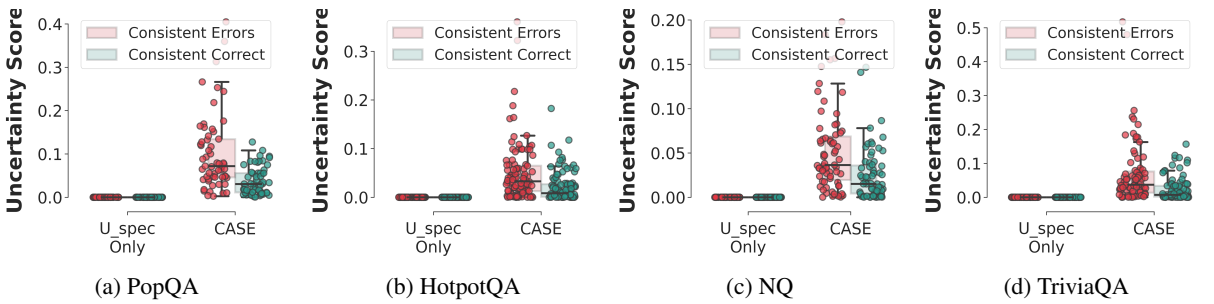
Figure 4: Ablation on graph construction strategies (Qwen3-8B). We compare our proposed contradiction-based sparse graph (Contra-Sparse) against dense and sparse entailment-based graphs (Entail-Orig and Entail-Sparse).

yielding novel semantic insights.

Generalizability of Token-level Entropy. We further examine whether token-level entropy can generally improve semantic-based uncertainty estimation beyond the proposed CASE framework. To this end, we augment KLE, an entailment graph-based uncertainty estimator, with token-level entropy signals and compare it with CASE. As shown in Figure 3, incorporating entropy consistently improves KLE across all datasets, with particularly noticeable gains in AUARC. This suggests that token-level uncertainty provides complementary information that is not fully captured by semantic consistency alone, and can be effectively integrated into graph-based uncertainty estimators. However, despite these improvements, the enhanced KLE variant remains inferior to CASE. This underscores that the success of CASE stems not merely from the naive inclusion of token-level entropy, but from the synergistic design of our framework: the sparsified contradiction graph and its spectral normalization inherently offer a superior topological representation, allowing the token-level signals to be aggregated far more effectively.

Effect of Graph Construction Strategies. To isolate the effect of graph construction, we compare three formulations under a fixed budget of 10. As shown in Figure 4, the standard Entail-Orig used in KLE achieves moderate performance, while its sparsified variant (Entail-Sparse) degrades AUROC and AUARC. This suggests that naive pruning of entailment graphs may remove informative agreement structure required for reliable uncertainty aggregation. In contrast, Contra-Sparse consistently outperforms both entailment-based variants. This indicates that modeling semantic disagreement via a contradiction-centric topology provides a more discriminative signal for hallucination detection, compared to aggregation over entailment relations. By explicitly emphasizing conflict structure rather than global agreement, CASE yields improved separation between uncertain and certain predictions.

Token-level Entropy Mitigates Consistent Errors. Figure 5 visualizes uncertainty distributions for consistent outputs across four QA datasets, where all sampled generations agree. In this setting, U_spec alone assigns uniformly low uncertainty to both correct outputs and consistent errors, indicating that structural signals become ineffective when contradiction patterns disappear. After incorporating token-level predictive entropy, CASE produces clearly separated distributions, assigning



(a) PopQA (b) HotpotQA (c) NQ (d) TriviaQA

Figure 5: Uncertainty distributions for consistent outputs across four QA datasets (Qwen3-8B). Spectral norm alone (U_spec Only) fails to distinguish consistent errors from correct answers. By integrating token-level predictive entropy, CASE significantly improves the discriminability between the two distributions, effectively mitigating the vulnerability to semantically consistent errors.

higher uncertainty to consistent errors while maintaining low uncertainty for correct outputs. These results suggest that token-level predictive entropy provides complementary uncertainty information even in the absence of structural variation, thereby mitigating consistent errors.

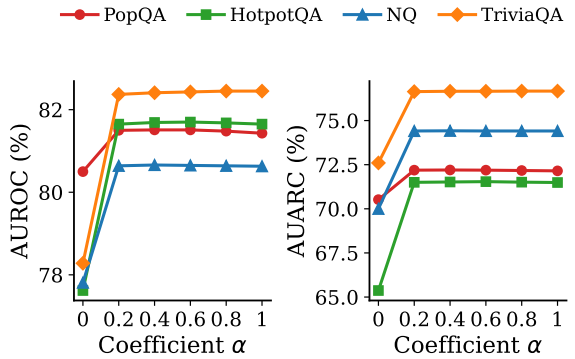


Figure 6: Hyperparameter sensitivity analysis of the coefficient (α) on the Qwen3-8B model.

Sensitivity to the Fusion Coefficient α. Figure 6 shows the impact of α in Eq. (10). Relying solely on the purely semantic metric ($\alpha = 0$) yields suboptimal results due to its vulnerability to consistent errors. However, introducing token-level entropy ($\alpha > 0$) substantially boosts both AUROC and AUARC. Crucially, performance plateaus and remains stable for $\alpha \geq 0.2$. This demonstrates that CASE is highly robust to hyperparameter variations, requiring no meticulous tuning to deliver reliable uncertainty estimates.

Hyperparameters of Adaptive Sampling. Figure 7 illustrates the trade-off between UQ performance and sampling efficiency, controlled by the stopping tolerance ϵ and patience τ. Stricter early-stopping settings (smaller ϵ or larger τ) enforce more exhaustive exploration of the semantic space, typically improving AUROC at the cost of increased generation overhead. With $\epsilon = 0.01$ and $\tau = 2$, CASE achieves a favorable balance, yielding strong hallucination detection performance while reducing redundant sampling costs. Full results are provided in Appendix H.

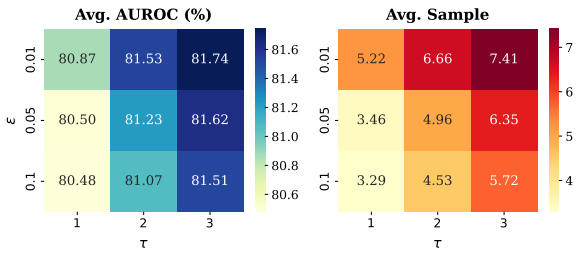


Figure 7: Performance–efficiency trade-off of adaptive sampling under different hyperparameter configurations on Qwen3-8B, reported as averaged performance across datasets.

Table 3: UQ results on GSM8K using Qwen3-8B. The best results are in **bold**.

Method	AUROC (%)	AUARC (%)
SE	60.91	53.65
KLE	70.77	54.15
CASE	**74.85**	**60.61**

Evaluation on Reasoning Tasks. Although our main evaluation focuses on factual question answering, we further evaluate CASE on GSM8K to examine whether the proposed uncertainty estimation generalizes to reasoning-oriented tasks. Using Qwen3-8B, CASE achieves an AUROC of 74.85 and an AUARC of 60.61, outperforming both SE (60.91/53.65) and KLE (70.77/54.15). These results suggest that contradiction-aware spectral uncertainty remains effective beyond factual QA.

Table 4: Robustness of CASE to different NLI model capacities. Results are reported on four factual QA datasets using Qwen3-8B.

NLI Model	PopQA		TriviaQA		NQ		HotpotQA	
	AUROC (%)	AUARC (%)	AUROC (%)	AUARC (%)	AUROC (%)	AUARC (%)	AUROC (%)	AUARC (%)
`deberta-v3-large`	81.43	72.15	82.45	76.67	80.63	74.41	81.65	71.49
`deberta-v3-base`	81.22	72.08	82.16	76.55	79.93	74.06	81.90	71.56
`deberta-v3-small`	81.19	72.06	82.33	76.62	79.76	73.96	81.76	71.55

Table 5: Premature early stopping statistics across four factual QA datasets. The count is computed over 1,000 evaluated instances per dataset. A premature stop is defined as terminating with fewer than five responses while continuing sampling to ten responses would reveal semantic contradictions.

Dataset	Count	Proportion (%)
HotpotQA	74 / 1000	7.4
PopQA	65 / 1000	6.5
NQ	86 / 1000	8.6
TriviaQA	67 / 1000	6.7

Robustness to NLI Model Capacity. Since CASE relies on an external NLI model to identify contradiction relations, we examine its sensitivity to NLI model capacity. We replace the default `nli-deberta-v3-large` with the corresponding base and small variants while keeping all other components unchanged. As shown in Table 4, CASE remains stable across NLI capacities, with only marginal variations in average AUROC and AUARC over the four factual QA datasets. This indicates that the effectiveness of CASE is not strongly dependent on a particular NLI model size.

Reliability of Adaptive Sampling. We further investigate whether the stabilization-based stopping criterion can lead to premature termination. Specifically, we identify cases where adaptive sampling stops with fewer than five responses, while continuing sampling up to ten responses would reveal semantic contradictions ($U_{\text{spec}} > 0$). As shown in Table 5, premature stopping occurs in only 6.5%–8.6% of the evaluated instances across the four datasets, corresponding to 65–86 cases out of 1,000. Moreover, removing adaptive sampling yields less than one percentage point improvement in the average UQ metrics while increasing the number of generated responses by more than 30%. These results demonstrate that CASE can reliably terminate sampling without sacrificing uncertainty estimation quality, thereby reducing unnecessary generation cost.

Table 6: Wall-clock inference time on NQ using Qwen3-8B. Reduction indicates the relative time saved by CASE over each baseline.

Method	Time (s)	Reduction (%)
SE	8,402.1	30.6
KLE	7,719.4	24.5
CASE	**5,826.7**	–

Wall-clock Efficiency. Beyond the number of generated responses, we further measure the actual wall-clock inference time to assess whether the sampling reduction of CASE translates into practical efficiency gains. As shown in Table 6, CASE requires only 5,826.7 seconds on NQ with Qwen3-8B, compared with 8,402.1 seconds for SE and 7,719.4 seconds for KLE. Thus, CASE reduces wall-clock inference time by 30.6% and 24.5% relative to SE and KLE, respectively. This reduction is achieved through CASE's adaptive sampling strategy, which terminates generation once the estimated uncertainty becomes stable instead of continuing to generate a fixed number of responses. The results confirm that CASE not only improves uncertainty estimation while reducing the average sampling cost, but also converts this reduction into substantial end-to-end inference-time savings.

6 Conclusion

In this paper, we propose CASE, a UQ framework that jointly models semantic disagreement and token-level predictive uncertainty. CASE constructs a contradiction graph over multiple responses and uses the normalized spectral norm to capture semantic uncertainty in a scale-invariant manner. This enables adaptive sampling that stops generation once the graph stabilizes, reducing redundant computation. To address failures of semantic metrics under semantically consistent but factually incorrect responses, CASE further incorporates token-level predictive entropy as a complementary signal. Experiments on four factual QA datasets show that CASE improves UQ performance while reducing sampling cost by over 30%.

Limitations

While CASE achieves promising results, several limitations remain. The current adaptive sampling strategy and entropy formulation could be further improved by incorporating richer graph-level signals and more selective uncertainty mechanisms. CASE also relies on an external NLI evaluator for contradiction detection; although our experiments show robustness across different NLI model capacities, its reliability may still degrade for long-form, multilingual, or highly compositional responses. Moreover, the uncertainty score does not directly identify the underlying source of uncertainty, such as ambiguity, reasoning difficulty, inconsistency, or hallucination. Broader evaluation on complex tasks and complementary signals for failure-type identification could therefore further improve CASE, while human oversight remains important for high-risk applications.

Acknowledgments

This project is supported by the Guangdong Basic and Applied Basic Research Foundation (2024A1515140109, 2024A1515011650) and Guangdong Special Support Plan Science and Technology Innovation Young Top-notch Talents (2023TQ07X621).

References

Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Fan RK Chung. 1997. *Spectral graph theory*, volume 92. American Mathematical Soc.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. 2025. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *Transactions on Machine Learning Research*.

Cross-Encoder Team. 2025. nli-deberta-v3-large. `https://huggingface.co/cross-encoder/nli-deberta-v3-large`.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302, Online. Association for Computational Linguistics.

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

Jiahui Geng, Fengyu Cai, and Others. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kimia Hamidieh, Veronika Thost, and Others. 2026. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*.

Roger A Horn and Charles R Johnson. 2012. *Matrix analysis*. Cambridge university press.

Lei Huang, Weijiang Yu, and Others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Mandar Joshi, Eunsol Choi, and Others. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, and Others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Takeshi Kojima, Shixiang Shane Gu, and Others. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Tom Kwiatkowski, Jennimaria Palomaki, and Others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhaoye Li, Siyuan Shen, and Others. 2025. Enhancing uncertainty quantification in large language models through semantic graph density. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, UAI '25. JMLR.org.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xiaoou Liu, Tiejin Chen, and Others. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, page 6107–6117, New York, NY, USA. Association for Computing Machinery.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *The Ninth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, and Others. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Lucas Hurley McCabe, Rimon Melamed, and Others. 2026. Estimating semantic alphabet size for LLM uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*.

Meta AI. 2024. Meta-llama-3-8b-instruct. https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. 2025. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4530–4540, Vienna, Austria. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, and Others. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pages 8901–8929. Curran Associates, Inc.

OpenAI, Josh Achiam, and Others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yaniv Ovadia, Emily Fertig, and Others. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32.

Qwen Team. 2025a. Qwen3-14B. https://huggingface.co/Qwen/Qwen3-14B.

Qwen Team. 2025b. Qwen3-8B. https://huggingface.co/Qwen/Qwen3-8B.

Qwen Team. 2025c. Qwen3-Embedding-0.6B. https://huggingface.co/Qwen/Qwen3-Embedding-0.6B.

Jie Ren, Jiaming Luo, and Others. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Zhenning Shi, Yijia Zhu, and Others. 2025. Reasoning under uncertainty: Efficient LLM inference via unsupervised confidence dilution and convergent adaptive sampling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32204–32218, Suzhou, China. Association for Computational Linguistics.

Adi Simhi, Itay Itzhak, and Others. 2025. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14665–14688, Suzhou, China. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, and Others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Qiyao Sun, Xingming Li, and Others. 2026. Efficient hallucination detection: Adaptive bayesian estimation of semantic entropy with guided semantic exploration. *Preprint*, arXiv:2603.22812.

Xuezhi Wang, Jason Wei, and Others. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

An Yang, Anfeng Li, and Others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhilin Yang, Peng Qi, and Others. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhihong Zhu, Yunyan Zhang, and Others. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

Appendix

A Spectral Norm Characterization

Let $H \in \mathbb{R}^{k\times k}$ denote the symmetric non-negative contradiction matrix.

Since H is symmetric, the spectral theorem gives

$$H = Q\Lambda Q^\top,$$

where Q is orthogonal and

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

contains the eigenvalues of H.

For symmetric matrices, the induced spectral norm satisfies

$$\|H\|_2 = \max_i |\lambda_i|.$$

Moreover, because H has non-negative entries, the Perron–Frobenius theorem implies that the dominant eigenvalue is real, non-negative, and equal to the spectral radius:

$$\lambda_{\max}(H) = \max_i |\lambda_i|.$$

Therefore,

$$\|H\|_2 = \lambda_{\max}(H).$$

B Upper Bound of the Spectral Uncertainty Score

Let $H^{(k)} \in \mathbb{R}^{k\times k}$ denote a symmetric, entrywise non-negative mutual contradiction matrix at stage k, where

$$H^{(k)}_{i,j} \in [0,1], \quad H^{(k)}_{i,i} = 0.$$

We analyze the spectral norm $\|H^{(k)}\|_2$. Since $H^{(k)}$ is real symmetric, its spectral norm equals its largest eigenvalue magnitude, i.e.,

$$\|H^{(k)}\|_2 = \lambda_{\max}(H^{(k)}). \tag{11}$$

For any matrix, the spectral norm is upper-bounded by the maximum row sum (induced ℓ_∞ norm bound):

$$\|H^{(k)}\|_2 \le \max_i \sum_{j=1}^{k} |H^{(k)}_{i,j}|. \tag{12}$$

Since $H^{(k)}$ is entrywise non-negative and $H^{(k)}_{i,i} = 0$, we obtain

$$\|H^{(k)}\|_2 \le \max_i \sum_{j\neq i} H^{(k)}_{i,j}. \tag{13}$$

Because each off-diagonal entry satisfies $H^{(k)}_{i,j} \le 1$, each row sum is bounded by

$$\sum_{j\neq i} H^{(k)}_{i,j} \le k-1, \quad \forall i. \tag{14}$$

Combining Eqs. (12)–(14), we obtain

$$\|H^{(k)}\|_2 \le k-1. \tag{15}$$

This bound is tight and is achieved when

$$H^{(k)}_{i,j} = 1, \quad \forall i \neq j,$$

i.e.,

$$H^{(k)} = \mathbf{1}\mathbf{1}^\top - I,$$

whose spectral norm is exactly $k-1$.

Accordingly, the spectral uncertainty score is defined as

$$U^{(k)}_{\mathrm{spec}} = \frac{\|H^{(k)}\|_2}{k-1}. \tag{16}$$

Therefore,

$$U^{(k)}_{\mathrm{spec}} \in [0,1].$$

C Why SE and KLE Are Less Suitable for Adaptive Sampling

Adaptive sampling requires the uncertainty estimates to be comparable across different sampling stages k. To this end, we define the stage-wise variation of the spectral uncertainty score as

$$\Delta U^{(k)} = \left| U^{(k)}_{\mathrm{spec}} - U^{(k-1)}_{\mathrm{spec}} \right|. \tag{17}$$

However, both Semantic Entropy (SE) and Kernel Language Entropy (KLE) exhibit intrinsic scale dependence during sequential sampling.

For SE, the empirical semantic distribution continuously evolves as additional samples introduce new semantic clusters and redistribute cluster probabilities. Consequently, the entropy estimate changes not only due to genuine uncertainty variation, but also due to support expansion and empirical distributional drift.

Similarly, KLE depends on the eigenspectrum of a semantic kernel constructed from sampled generations. As new samples are added, the semantic graph, kernel dimensionality, and eigenspectrum all evolve, causing additional spectral redistribution even when newly generated responses are semantically redundant.

Importantly, these scale dependencies cannot generally be removed through simple normalization by the sample count. Since both SE and KLE

are already defined over normalized probability or spectral distributions, further normalization lacks a principled statistical interpretation and may artificially collapse the estimator magnitude without eliminating estimator evolution.

As a result, uncertainty estimates produced by SE and KLE are not directly comparable across different sampling stages, making it difficult to design stable adaptive stopping criteria based on marginal uncertainty variation.

D Spectral Stability Under Incremental Sampling

We analyze the stability of the spectral uncertainty score under incremental sampling and justify the use of the marginal variation

$$\Delta U^{(k)} = \left| U_{\text{spec}}^{(k)} - U_{\text{spec}}^{(k-1)} \right|$$

as a stopping signal in adaptive sampling.

Incremental Matrix Perturbation. When a new sample a_k is added, the contradiction matrix is updated as

$$H^{(k)} = \begin{bmatrix} H^{(k-1)} & v \\ v^\top & 0 \end{bmatrix}, \tag{18}$$

where $v \in \mathbb{R}^{k-1}$ contains contradiction scores between a_k and previous samples. All previously computed entries remain unchanged, implying that sampling introduces a rank-2 perturbation.

Define the zero-padded matrix

$$H_{\text{pad}}^{(k-1)} = \begin{bmatrix} H^{(k-1)} & 0 \\ 0 & 0 \end{bmatrix},$$

so that

$$H^{(k)} = H_{\text{pad}}^{(k-1)} + E^{(k)}, \quad E^{(k)} = \begin{bmatrix} 0 & v \\ v^\top & 0 \end{bmatrix}.$$

Since $H_{i,j}^{(k)} \in [0,1]$, we have $\|v\|_2 \le \sqrt{k-1}$, and thus $\|E^{(k)}\|_2 \le \sqrt{k-1}$.

Spectral Sensitivity. Let $\|H^{(k)}\|_2$ denote the spectral norm. By Weyl's inequality for symmetric matrices,

$$\left| \|H^{(k)}\|_2 - \|H^{(k-1)}\|_2 \right| \le \|E^{(k)}\|_2.$$

Since

$$U_{\text{spec}}^{(k)} = \frac{\|H^{(k)}\|_2}{k-1},$$

the change in normalized spectral uncertainty satisfies

$$\begin{aligned} \Delta U^{(k)} \le & \frac{\left| \|H^{(k)}\|_2 - \|H^{(k-1)}\|_2 \right|}{k-1} \\ & + \|H^{(k-1)}\|_2 \left| \frac{1}{k-1} - \frac{1}{k-2} \right|. \end{aligned} \tag{19}$$

Using $\|H^{(k-1)}\|_2 \le k-2$, we obtain the finite-sample bound

$$\Delta U^{(k)} \le \frac{\|E^{(k)}\|_2}{k-1} + \frac{1}{k-1}. \tag{20}$$

This shows that the marginal change is controlled by the magnitude of the newly introduced contradiction edges and a diminishing normalization correction term.

Interpretation for Adaptive Stopping. Importantly, the stopping criterion is not based on the absolute value of $U_{\text{spec}}^{(k)}$, but on its stabilization behavior across consecutive sampling steps. When newly generated samples produce only small perturbations in the contradiction structure, the induced change in spectral uncertainty becomes negligible.

Therefore, monitoring $\Delta U^{(k)}$ provides a principled and computationally efficient criterion for detecting stabilization of the underlying contradiction structure, which is used as the basis for early stopping in adaptive sampling.

E Detailed Algorithms

This appendix provides the detailed pseudo-code for the adaptive sampling mechanism within the CASE framework.

F Evaluation Metrics

AUROC. Let $s(x)$ denote the uncertainty score for input x, and $y \in \{0,1\}$ indicate whether the prediction is correct. AUROC measures the probability that a randomly chosen correct prediction is ranked lower in uncertainty than a randomly chosen incorrect one:

$$\text{AUROC} = \mathbb{P}(s(x^+) < s(x^-)), \tag{21}$$

where x^+ and x^- denote correct and incorrect samples, respectively.

AUARC. Let θ be a rejection threshold, and consider the subset of samples satisfying $s(x) \le \theta$. The accuracy–rejection curve plots the accuracy

Algorithm 1 Adaptive Sampling via Spectral Contradiction Score

Require: Query x, LLM $\mathcal{M}$, NLI Model $\mathcal{N}$, Bounds $[K_{\min}, K_{\max}]$, Tolerance ϵ, Patience p
Ensure: Generated candidate set $\mathcal{A}$, Contradiction density U_{spec}

Initialize candidate set $\mathcal{A} \leftarrow \emptyset$, sparsified graph $H \leftarrow$ Empty Matrix
Initialize $streak \leftarrow 0$, $U_{spec}^{(0)} \leftarrow 0$
for $k = 1$ **to** $K_{\max}$ **do**
  Generate new candidate: $a_k \sim \mathcal{M}(x)$
  Update generated set: $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_k\}$
  if $k \geq 2$ **then**
    for $j = 1$ **to** $k - 1$ **do**
      Predict NLI probs: $P_C, P_E = \mathcal{N}(a_k, a_j)$ and $\mathcal{N}(a_j, a_k)$
      $W_{k,j} = \max\big(0, P_C(a_k, a_j) - P_E(a_k, a_j)\big)$
      $W_{j,k} = \max\big(0, P_C(a_j, a_k) - P_E(a_j, a_k)\big)$
      $H_{k,j} = H_{j,k} = \sqrt{W_{k,j} W_{j,k}}$
    end for
    $U_{spec}^{(k)} = \frac{\|H^{(k)}\|_2}{k-1}$
  end if
  if $k \geq N_{\min}$ **then**
    $\Delta U^{(k)} = \left| U_{spec}^{(k)} - U_{spec}^{(k-1)} \right|$
    if $\Delta U^{(k)} \leq \epsilon$ **then**
      $streak \leftarrow streak + 1$
    else
      $streak \leftarrow 0$
    end if
    if $streak \geq \tau$ **then**
      break
    end if
  end if
end for
return $\mathcal{A}$, $U_{spec}^{(k)}$

over the retained samples as a function of the rejection rate. AUARC is defined as:

$$\text{AUARC} = \int_0^1 \text{Acc}(r)\, dr, \tag{22}$$

where r is the rejection rate and $\text{Acc}(r)$ denotes the accuracy at rejection rate r.

G Dataset, Resources, and Reproducibility

Dataset Characteristics. The evaluated datasets cover diverse factual QA settings, including open-domain factual retrieval, multi-hop reasoning, and long-tail factual knowledge. Specifically, Natural Questions and TriviaQA focus on open-domain QA, HotpotQA emphasizes multi-hop reasoning, and PopQA evaluates long-tail entity-centric factual recall. All experiments are conducted in English.

Computational Resources. We evaluate three open-source LLMs ranging from 8B to 14B parameters, including LLaMA-3-8B-Instruct, Qwen3-8B, and Qwen3-14B. All experiments are conducted on 2 NVIDIA 4090 GPUs. Since CASE is an inference-time uncertainty estimation framework, no additional model training is required. The computational overhead mainly comes from stochastic answer sampling at inference time, whereas the additional cost introduced by pairwise contradiction estimation is negligible compared with generation.

H Detailed Hyperparameter Search for CASE

In the main text, we introduced two key hyperparameters in our adaptive sampling mechanism: the stopping tolerance ϵ and the patience parameter τ. To investigate their impact on the effectiveness-efficiency trade-off, we conducted a comprehen-

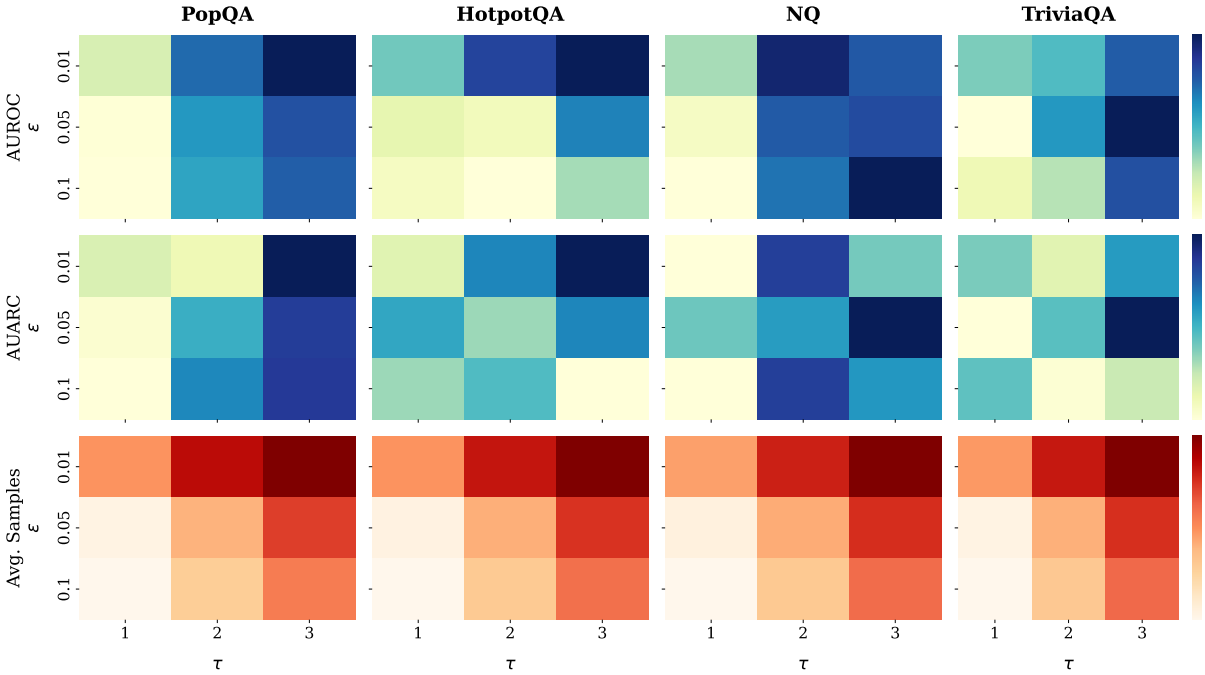


Figure 8: Complete hyperparameter search results for the stopping tolerance ϵ and patience τ across four QA datasets. The top and middle rows illustrate the uncertainty detection performance in terms of AUROC and AUARC, respectively (darker blue indicates higher performance). The bottom row displays the corresponding inference cost, measured by the average number of sampled responses per query (darker red indicates higher sampling cost). The configuration ($\epsilon = 0.01, \tau = 2$) effectively balances high detection performance with moderate sampling overhead.

sive grid search across all four evaluation datasets (PopQA, HotpotQA, NQ, and TriviaQA). The search space included $\epsilon \in \{0.01, 0.05, 0.1\}$ and $\tau \in \{1, 2, 3\}$. The complete results are visualized in Figure 8.

Effectiveness (AUROC and AUARC). The top two rows of Figure 8 illustrate the detection performance. As expected, adopting a more conservative generation strategy—either by lowering the threshold ϵ (making it harder to accept a response) or increasing the patience τ (requiring more consecutive consistent generations)—consistently yields higher AUROC and AUARC scores, indicated by the darker blue regions. Specifically, the configuration with the strictest conditions ($\epsilon = 0.01, \tau = 3$) achieves the highest theoretical performance across most datasets.

Efficiency (Sampling Cost). The bottom row of Figure 8 displays the corresponding inference overhead, measured by the average number of sampled responses per query. The heatmaps (in red) clearly show that the performance gains obtained from stricter parameters come at a steep cost. Setting $\tau = 3$ forces the model to generate near the maximum budget frequently, resulting in the darkest red regions and significantly degrading inference efficiency.

Optimal Trade-off. When analyzing the performance gains relative to the cost, the configuration of $\epsilon = 0.01$ and $\tau = 2$ emerges as the optimal operating point. At $\epsilon = 0.01$, increasing τ from 1 to 2 provides a substantial boost in detection performance. However, further increasing τ from 2 to 3 yields only marginal improvements in AUROC and AUARC while disproportionately maximizing the sampling overhead. Therefore, setting $\epsilon = 0.01$ and $\tau = 2$ effectively captures the majority of the performance benefits while maintaining a reasonable computational budget.