

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

Jiaming Zhang¹ Heng-yang Lu² Xinan Chen¹ Junyuan Zheng¹ Yuncheng Jiang¹ Ying Zhu^{1,*} Chenyou Fan^{1,*}
¹South China Normal University ²Jiangnan University
 zhang_jia_ming@m.scnu.edu.cn {zhuy, fanchenyou}@scnu.edu.cn



Motivation

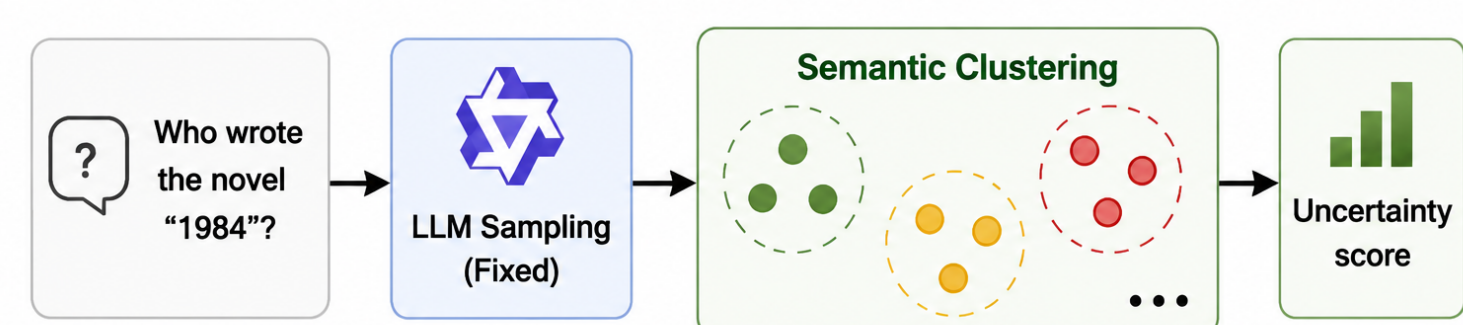
Multi-sample UQ methods commonly use a fixed sampling budget, leading to:

- **Redundant sampling:** easy queries may not need the full sampling budget.
- **Consistent errors:** semantically consistent but incorrect generations can yield spuriously low uncertainty.

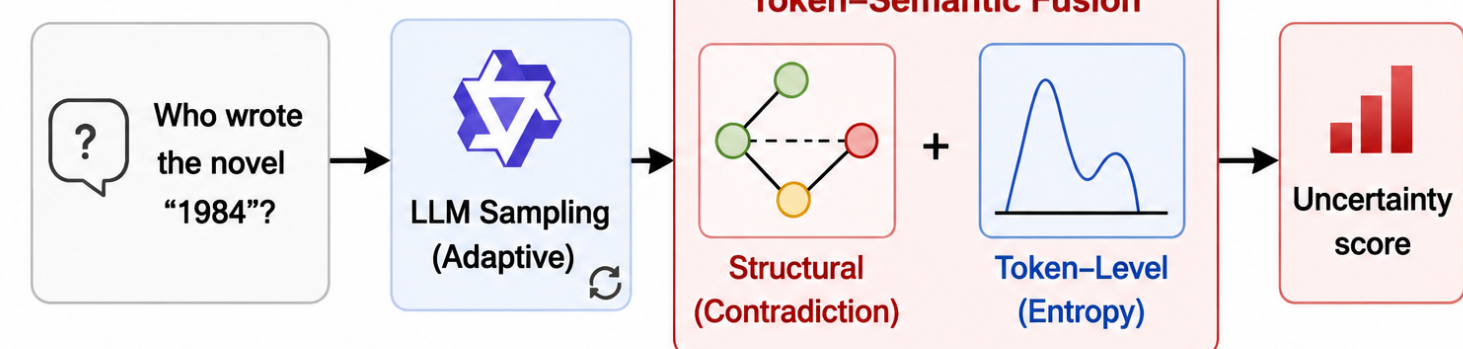
► Key Insight

Sampling should stop when semantic uncertainty stabilizes, while a complementary token-level signal is needed to detect consistent errors.

Previous: Semantic Entropy

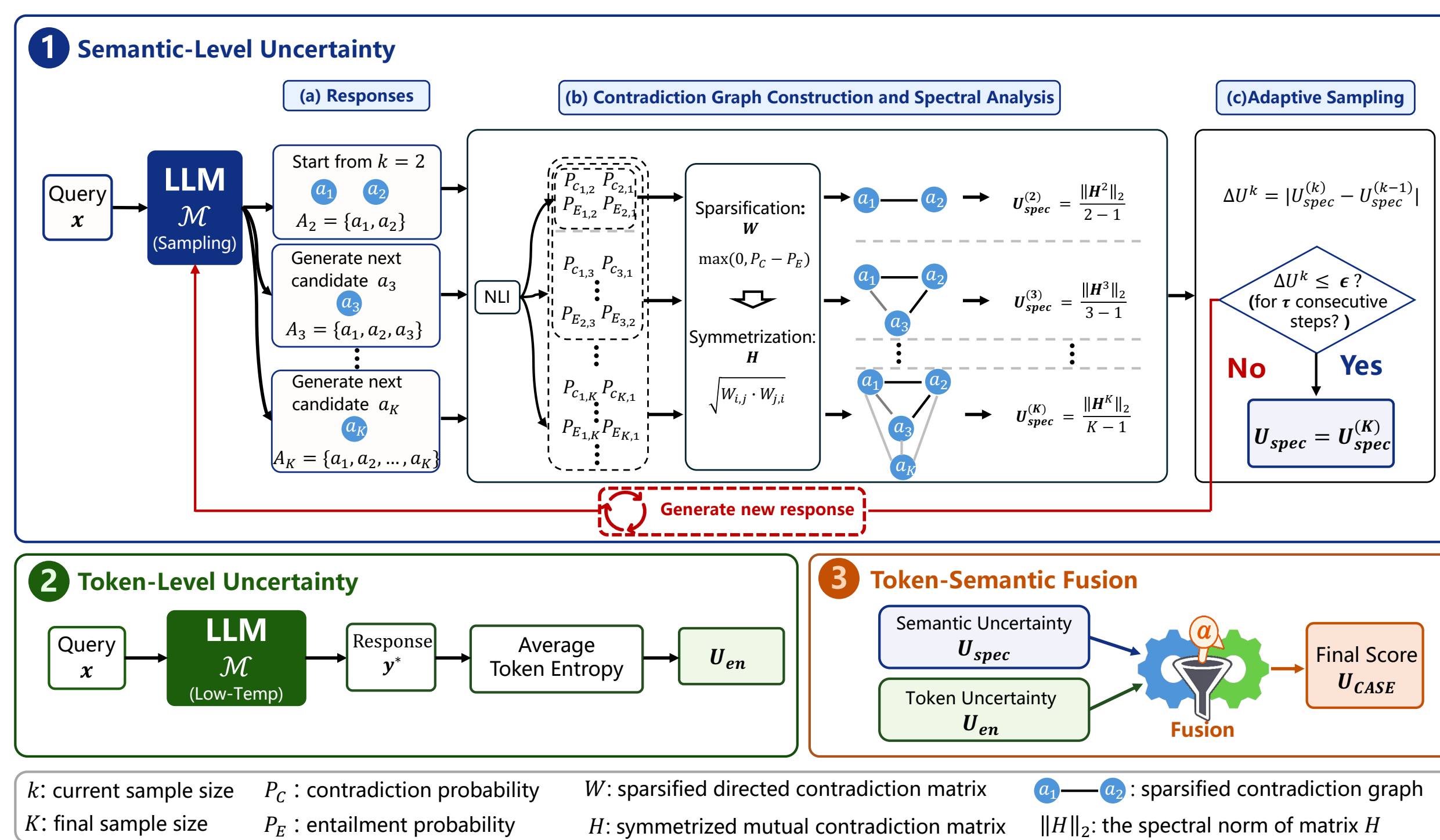


Ours: CASE



Comparison of UQ paradigms. Existing semantic-level approaches (top). In contrast, CASE (bottom).

CASE Framework



Overview of the CASE framework. The proposed architecture operates through three main components: (1) Semantic-Level Uncertainty; (2) Token-Level Uncertainty ; and (3) Token-Semantic Fusion.

Token-Level Uncertainty

Semantic signals can collapse when all sampled responses agree, including cases of consistent errors. CASE therefore generates one low-temperature base response y^* and computes average token entropy:

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

The final score is

$$U_{\text{CASE}}(x) = U_{\text{spec}}(x) + \alpha U_{\text{en}}(x).$$

Experimental Setup

Datasets: NQ, TriviaQA, HotpotQA, PopQA.

LLMs: LLaMA-3-8B-Instruct, Qwen3-8B, Qwen3-14B.

Baselines: PPL, P(true), Entropy, SE, KLE, SNNE.

Metrics: AUROC, AUARC.

Adaptive sampling: $K_{\min} = 2$, $K_{\max} = 10$, $\epsilon = 0.01$, $\tau = 2$.

Main Results

Methods	PopQA			HotpotQA			NQ			TriviaQA		
	AUROC $\uparrow$	AUARC $\uparrow$	Samp. $\downarrow$	AUROC $\uparrow$	AUARC $\uparrow$	Samp. $\downarrow$	AUROC $\uparrow$	AUARC $\uparrow$	Samp. $\downarrow$	AUROC $\uparrow$	AUARC $\uparrow$	Samp. $\downarrow$
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	<u>82.30</u>	<u>72.35</u>	10.0	<u>78.61</u>	<u>71.05</u>	10.0	<u>79.77</u>	<u>79.97</u>	10.0	<u>84.96</u>	<u>87.48</u>	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	<u>70.56</u>	-	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	<u>72.49</u>	-	79.06	<u>75.52</u>	-
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	<u>80.35</u>	<u>71.54</u>	10.0	<u>81.08</u>	69.96	10.0	<u>79.61</u>	71.95	10.0	<u>80.52</u>	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	<u>64.84</u>	-	78.28	70.49	-	76.18	61.31	-	84.28	62.34	-
Entropy	69.07	61.42	-	77.60	<u>71.47</u>	-	74.90	<u>75.97</u>	-	77.96	<u>81.60</u>	-
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	<u>76.16</u>	63.60	10.0	<u>80.99</u>	70.42	10.0	<u>76.76</u>	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

Main results across four factual QA datasets.

Compared with fixed-budget multi-sample baselines:

- CASE uses **5.88–7.50** samples per query, versus 10 for fixed-budget SE/KLE.
- This corresponds to approximately **25–40%** fewer samples.

Ablation & Robustness

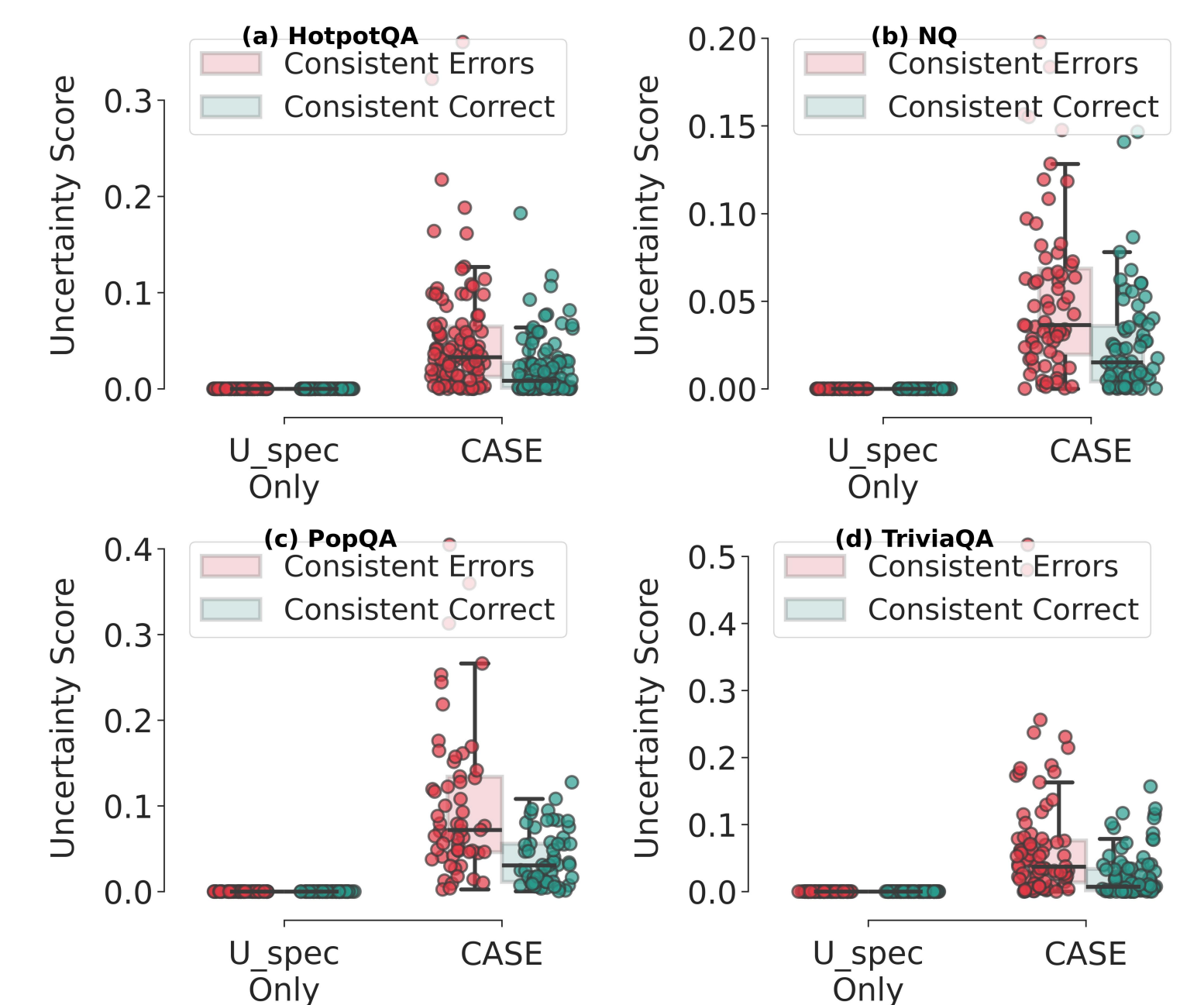
Ablation of CASE Components

CASE combines adaptive sampling and token-level entropy. We remove each component to analyze its contributions.

Method	AUROC	AUARC	Samp.
w/o Adaptive	82.39	74.30	10.0
w/o U_{en}	78.55	69.62	6.66
CASE	81.54	73.68	6.66

Token-level Entropy Mitigates Consistent Errors

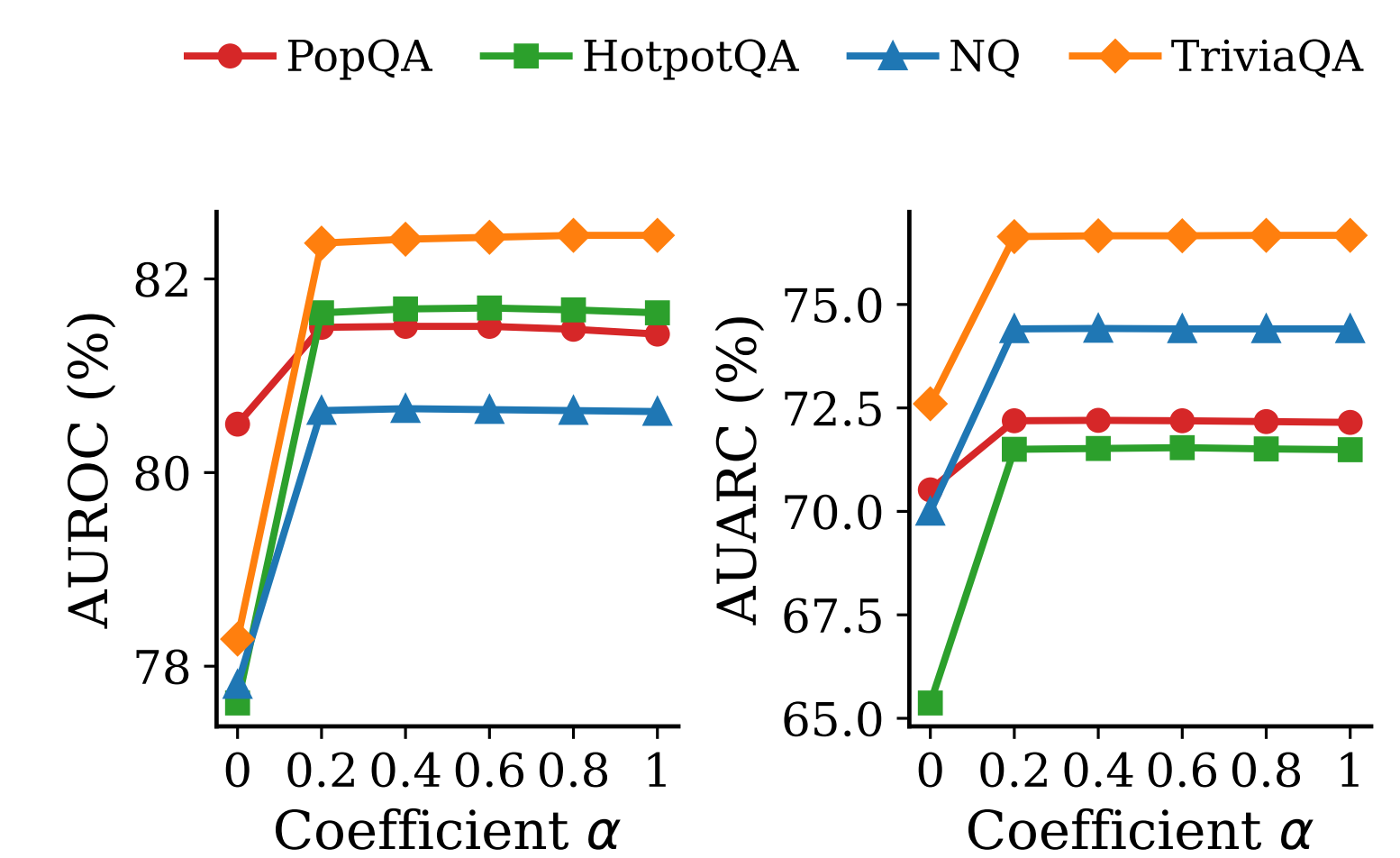
When sampled responses are semantically consistent, spectral uncertainty alone struggles to identify incorrect generations. Token-level entropy provides a complementary uncertainty signal.



Token-level entropy improves discrimination between correct answers and consistent errors.

Fusion Coefficient Sensitivity

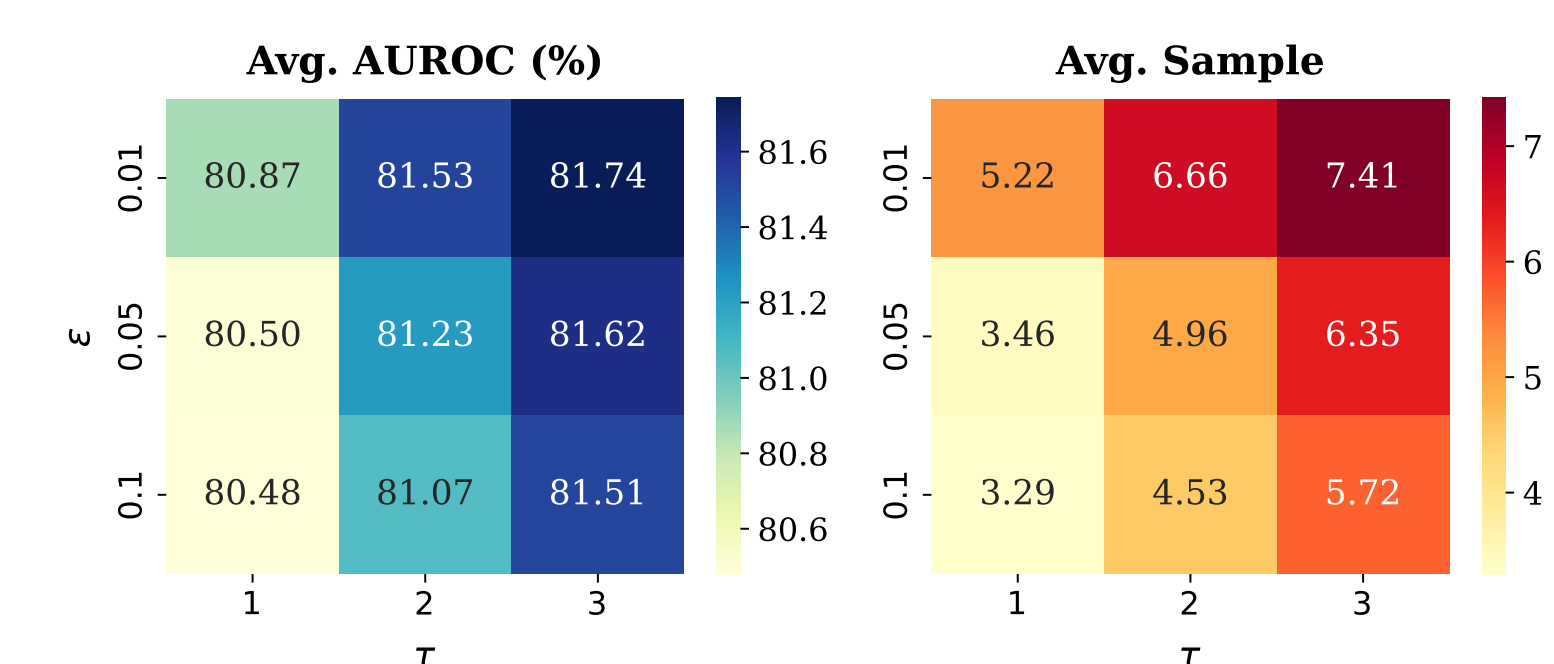
CASE remains robust across different fusion coefficients α . Introducing token entropy improves uncertainty estimation and stabilizes performance.



AUROC and AUARC under different fusion coefficients.

Adaptive Sampling Trade-off

The stopping tolerance ϵ and patience τ control the trade-off between uncertainty quality and sampling cost.



Performance–efficiency trade-off under different configurations.