CoSu-UQ: Uncertainty Quantification for Reasoning via Confidence and Semantic Support

Yinglun Feng[a], Qinhong Lin[a], Yuhao Zhang[a], Zhongliang Yang[b,a,*], Linna Zhou[b,a,*]
[a]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[b]QuanCheng Laboratory, Jinan, China
yangzl@bupt.edu.cn, zhoulinna@bupt.edu.cn

***Abstract*—Estimating uncertainty for long-form reasoning in large language models remains challenging, as existing methods based on token probabilities or semantic diversity often break down in multi-step reasoning scenarios. We propose CoSu-UQ, a multi-sample uncertainty quantification framework that jointly models generation confidence and cross-sample semantic support. CoSu-UQ aggregates token-level probabilities into response-level confidence and measures cross-sample semantic support via sentence-level entailment. These complementary signals are integrated through bidirectional entailment-based clustering over final answers to produce a robust uncertainty estimate. We conduct systematic evaluations on GSM8K, MATH, HotpotQA, 2WikiMultiHopQA, and MedQA across large language models of different scales. Results show that CoSu-UQ consistently outperforms representative baselines (e.g., Predictive Entropy, Semantic Entropy, LUQ) in terms of AUROC for distinguishing correct from incorrect generations. Ablation studies further confirm the complementarity between generation confidence and cross-sample semantic support, with particularly pronounced gains in long-form reasoning and multi-hop question answering scenarios.**

***Index Terms*—Uncertainty Quantification, Chain-of-Thought, Large Language Models, Multi-sample Reasoning, Semantic Support**

I. Introduction

Large language models (LLMs) have shown breakthrough capabilities across complex tasks, including dialogue systems, knowledge-intensive QA, and multi-step reasoning [1], [2], [3], [4]. Among them, Chain-of-Thought (CoT) prompting explicitly guides models to generate step-by-step reasoning, substantially improving performance on multi-turn dialogue, logical reasoning, and math reasoning tasks [5], [6], [7], [8]. CoT has therefore become a prominent paradigm for enhancing LLM reasoning.

However, stronger reasoning does not eliminate unreliability. A growing body of work shows that LLMs still hallucinate, producing fluent but incorrect reasoning or conclusions [9], [10]. This issue is especially critical in safety-sensitive settings such as medical decision making, scientific analysis, and multi-hop QA, where erroneous outputs can be costly [11]. Notably, CoT improves structure and interpretability, but fluent and self-consistent reasoning can also mask errors, making incorrect conclusions harder to detect.

To address this, uncertainty quantification (UQ) has been proposed as a tool for identifying when model outputs should be trusted [12]. Existing UQ for LLMs follows two main lines: probability-based measures that aggregate token-level likelihoods (e.g., Predictive Entropy and its variants) [13], [14], [15], and semantic methods that group multiple samples into semantically equivalent answers and estimate uncertainty from their frequency or distribution (e.g., semantic uncertainty/entropy) [15], [16]. Recent work further incorporates semantic similarity or natural language inference to capture cross-sample support structure [17], [18]. Despite their success, prior studies report that explicit long-form reasoning can lead to overconfidence in uncertainty estimates [19]. This is often attributed to an internal trust bias: models tend to over-trust both their final answers and their generated reasoning [20], [21]. In multi-sample reasoning settings, different reasoning paths can exhibit complex patterns of agreement and disagreement, which a single uncertainty signal cannot fully capture.

These observations highlight a central challenge for long-form, multi-path reasoning: how to jointly model generation confidence and cross-sample semantic support to better characterize uncertainty at the decision level.

Motivated by this, we propose **CoSu-UQ** (Confidence and Support-based Uncertainty Quantification), a multi-sample UQ framework for long-form reasoning. CoSu-UQ combines two complementary signals: (1) response-level generation confidence derived by aggregating token-level probabilities to capture local generation risk, and (2) cross-sample semantic support measured via sentence-level natural language inference to model agreement structure across samples. To reduce semantic noise introduced by long or redundant intermediate steps, CoSu-UQ performs semantic grouping and aggregation at the final-answer level, focusing uncertainty estimation on the decision itself.

Our contributions are summarized as follows:

1) Conceptually, we decompose uncertainty in long-form reasoning into complementary signals from generation-time confidence and cross-sample semantic support, enabling a more faithful characterization of uncertainty at the decision level.
2) Methodologically, we propose CoSu-UQ, a multi-sample uncertainty quantification framework that jointly models token-level generation confidence and cross-sample semantic support, and aggregates uncertainty at the final-answer level to mitigate noise from long intermediate reasoning.

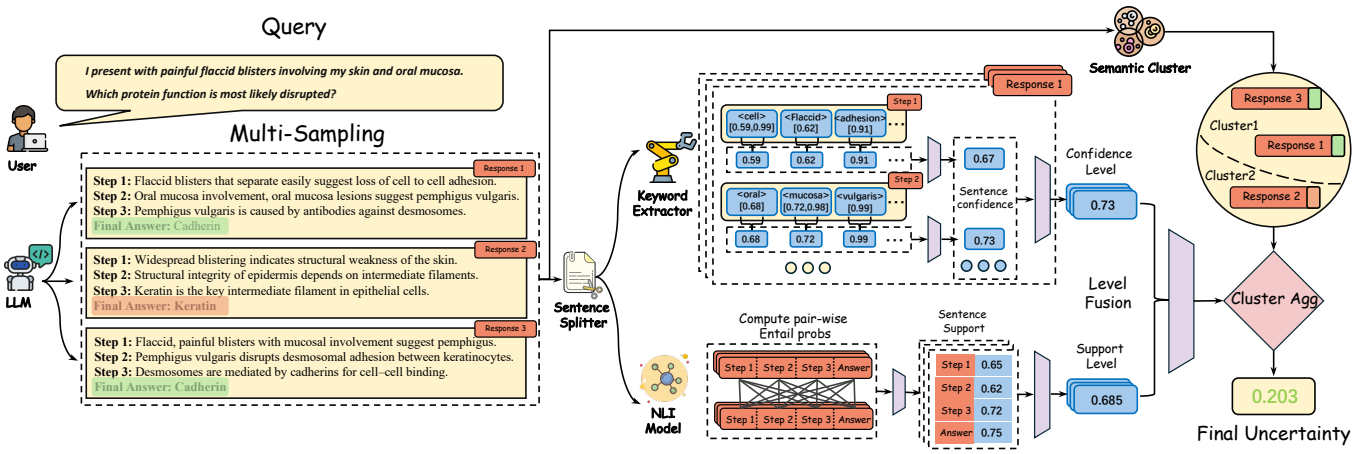


Fig. 1: Overview of the CoSu-UQ framework.

3) Empirically, we provide systematic evaluations on diverse reasoning benchmarks and model sizes, demonstrating that CoSu-UQ achieves more stable and consistent AUROC improvements compared to probability-based, semantic-based, and reasoning-aware baselines.

Our code and data are publicly available at https://github.com/interments/CoSu-UQ.

II. Related Works

Traditional uncertainty quantification for LLMs can be broadly grouped into probability-based methods and semantic aggregation over multiple samples. Probability-based approaches aggregate token-level generation probabilities to quantify uncertainty, such as Predictive Entropy (PE) and its length-normalized variant LN-PE [13], [14], [22]. Self-Consistency (SC) [23] measures agreement via majority voting over sampled answers, while Semantic Entropy performs semantic clustering and computes uncertainty from the resulting probability distribution over clusters [15], [16]. SentenceSAR and TokenSAR reweight uncertainty with sentence-level or token-level relevance [17].

For long-form reasoning, recent studies explicitly model uncertainty within or across reasoning trajectories. CoT-UQ analyzes keyword importance within a single chain to capture intra-chain uncertainty [24], whereas LUQ leverages sentence-level natural language inference to model cross-sample semantic support and agreement [18]. Other efforts adopt graph-based formulations, constructing semantic similarity graphs and combining them with probabilistic signals or learned calibration mechanisms for uncertainty estimation or confidence calibration [25], [26].

However, most prior work relies on either confidence signals or cross-sample consistency alone, limiting uncertainty estimation in long-form reasoning.

III. Methodology

We quantify uncertainty in the reasoning process by combining token-level confidence, cross-sample support, and bidirectional entailment clustering over multiple sampled chains of thought. Fig. 1 provides an overview of the proposed pipeline.

A. Multi-Sampling and Structured Reasoning

Given an input question, we prompt the model to produce step-by-step Chain-of-Thought (Step 1, Step 2, ...) ending with a *Final Answer*. To characterize uncertainty for the same question, we sample the model N times (default $N = 5$), obtaining a set of responses $\{R_1, R_2, \ldots, R_N\}$. Each response contains the full reasoning chain and final answer. We then split each response into sentence-level steps plus the final answer, so all subsequent metrics operate at a uniform sentence granularity.

B. Confidence Level

During generation, the model assigns a conditional probability $p(t \mid t_{<t}, x)$ to each token t in response R_i, where x is the input question; this is our lowest-level confidence signal. We then segment each response into sentences and extract content words from sentence s_i^a using spaCy POS tagging [27], denoted as $\mathcal{K}_i^a = \{k_1, k_2, \ldots, k_m\}$. Each keyword k can map to multiple tokens with probabilities $\mathcal{P}(k) = \{p_{k,1}, p_{k,2}, \ldots, p_{k,n_k}\}$. To capture the most uncertain part of a semantic unit, we use min aggregation [13], [28], [22]:

$$\text{Conf}(k) = \min_{p \in \mathcal{P}(k)} p. \tag{1}$$

For a sentence s_i^a, we average keyword confidence to obtain sentence-level confidence:

$$\text{Confidence}(s_i^a) = \frac{1}{|\mathcal{K}_i^a|} \sum_{k \in \mathcal{K}_i^a} \text{Conf}(k). \tag{2}$$

A full response consists of multiple steps, denoted $R_i = \{s_i^1, s_i^2, \ldots, s_i^{|R_i|}\}$. We use a product over steps to reflect serial dependence in multi-step reasoning:

$$\text{Confidence}(R_i) = \prod_{a=1}^{|R_i|} \text{Confidence}(s_i^a). \tag{3}$$

C. Support Level

Support captures cross-sample agreement at the sentence level across both reasoning steps and the final answer. We estimate sentence-level entailment with `potsawee/deberta-v3-large-mnli` and apply LUQ-style max-over-sentences aggregation to account for paraphrases and optional steps [18]. For a fixed sentence s_i^a, we take its maximum entailment probability p_{ent} over all sentences in R_j, so a step is supported if any sentence in R_j entails it. We then average these maxima over all sentences in R_i to obtain response-level similarity:

$$\text{Similarity}(R_i, R_j) = \frac{1}{|R_i|} \sum_{s_i^a \in R_i} \max_{s_j^b \in R_j} p_{\text{ent}}(s_i^a, s_j^b), \quad (4)$$

where $p_{\text{ent}}(s_i^a, s_j^b)$ denotes the probability that s_j^b entails s_i^a. Finally, we average similarities against all other responses to define support for R_i:

$$\text{Support}(R_i) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \text{Similarity}(R_i, R_j). \quad (5)$$

D. Level Fusion

We treat confidence and support as complementary signals and fuse them by arithmetic averaging:

$$\text{Reliability}(R_i) = \text{Mean}\big(\text{Support}(R_i), \text{Confidence}(R_i)\big). \quad (6)$$

This yields a reliability score for each response R_i, reflecting both internal generation confidence and cross-response consistency.

E. Clustering

To capture overall answer uncertainty at the multi-sample level, we cluster cleaned final answers using bidirectional entailment. Following semantic-entropy-style entailment clustering [16], we adopt an MNLI checkpoint, `microsoft/deberta-large-mnli`, built upon DeBERTa [29].

Concretely, given N sampled responses $\{R_1, \ldots, R_N\}$, we first deduplicate identical final answers. We then build an undirected graph over the remaining answers: an edge is added between two answers only when both directional entailment conditions hold ($R_i \Rightarrow R_j$ and $R_j \Rightarrow R_i$). Semantic clusters are defined as the connected components of this graph. Therefore, $|C| \leq N$, where $C = \{C_1, C_2, \ldots, C_{|C|}\}$ and $z_i \in \{1, \ldots, |C|\}$ denotes the cluster assignment of response R_i. This definition allows transitive connectivity through intermediate answers (e.g., A-B and B-C), without requiring every pair inside a cluster to be directly mutually entailed.

We define semantic uncertainty by aggregating reliability scores within each cluster:

$$H = -\frac{1}{|C|} \sum_{c=1}^{|C|} \log\left(\sum_{i: z_i = c} \text{Reliability}(R_i) \right). \quad (7)$$

We further normalize it into the final uncertainty score:

$$U = 1 - \frac{\exp(-H)}{N}. \quad (8)$$

Here, H is a cluster-mass aggregation quantity rather than a standard probability entropy; dividing by N in Eq. (8) is a scale normalization under a fixed sample budget ($N = 5$ in our experiments).

IV. Experiments

A. Experimental Setup

Models. We evaluate the proposed uncertainty measure on LLMs of different scales and architectures: Qwen3-4B, Qwen3-14B, Qwen3-32B [30], and Llama-3.1-8B [31]. All models use identical inference and sampling settings to ensure comparability across model sizes.

Datasets. We test on GSM8K [32], MATH [33], HotpotQA [34], and 2WikiMultiHopQA [35]. We additionally include MedQA [36] as a domain-specific benchmark. Dataset-specific preprocessing details are provided in Appendix A.

Sampling and Decoding Setup. We sample $N = 5$ CoT responses per question (temperature 1.0) and compute uncertainty from token-level confidence and semantic support across samples. Additional details are provided in Appendix B.

Baselines and Implementation. We compare CoSu-UQ with PE/LN-PE [13], [14], TokenSAR/SentenceSAR [17], SC [23], SE [16], LUQ [18], and CoT-UQ [24]. We apply only standard dataset-specific answer normalization and use default hyperparameters. For CoT-UQ, we average scores across sampled responses. All methods are evaluated under the same sampling budget. All experiments are conducted on a server with an Intel Xeon Platinum 8336C CPU @ 2.30GHz and two NVIDIA A100-SXM4-80GB GPUs.

Metric. Following standard practice in uncertainty quantification [15], we use AUROC as the primary metric for evaluating how well uncertainty scores separate correct from incorrect predictions [37]. In the multi-sample setting, we adopt the selective generation protocol commonly used in self-consistency reasoning [23]. Since the system ultimately outputs a single prediction corresponding to the dominant semantic answer cluster, uncertainty is assessed with respect to this final decision rather than individual samples. Concretely, we select the highest-probability response from the largest semantic cluster as the representative output. Correctness labels are obtained via majority voting among three independent LLM judges, which yields high agreement with human annotations; details are provided in Appendix C.

B. Main Results

Table I reports the AUROC results of CoSu-UQ compared with representative UQ baselines across reasoning benchmarks and model sizes. In general, CoSu-UQ achieves more consistent and robust performance by jointly capturing generation confidence and cross-sample semantic support.

Task dependence of single-signal uncertainty methods. Methods relying on a single uncertainty signal show clear task dependence. Probability-based methods are strong on short and structured reasoning tasks such as GSM8K (e.g., on GSM8K with Llama-3.1-8B, PE reaches 0.8314 while CoSu-UQ is 0.8261 in Table I), suggesting token-level likelihood already

TABLE I: Main results (AUROC) on reasoning datasets and MedQA. Bold indicates the best result; † indicates the second best. S-SAR and T-SAR denote SentenceSAR and TokenSAR, respectively; SC denotes Self-Consistency.

Models	Datasets	PE	LN-PE	S-SAR	T-SAR	SE	SC	LUQ	CoT-UQ	Ours
Llama3.1-8B	GSM8K	**0.8314**	0.8200	0.5698	0.8199	0.8201	0.7815	0.7953	0.7536	0.8261†
	MATH	0.8499	0.7805	0.7843	0.7805	0.7723	0.8867	0.8881†	0.7967	**0.9066**
	HotpotQA	0.7609	0.7776	0.7741	0.7776	0.7492	**0.8784**	0.8251	0.7522	0.8408†
	2WikiMultiHopQA	0.6811	0.7296	0.7471	0.7296	0.6120	**0.8589**	0.6710	0.7479	0.8004†
	MedQA	0.7022	0.6741	0.6370	0.6741	0.6719	0.8198†	0.7732	0.6498	**0.8209**
Qwen3-4B	GSM8K	0.8344†	0.8087	0.5664	0.8088	0.8091	0.7338	0.8174	0.7939	**0.8612**
	MATH	0.9174†	0.8578	0.6241	0.8579	0.8653	0.8733	0.8944	0.8469	**0.9580**
	HotpotQA	0.6852	0.6798	0.5612	0.6799	0.6947	0.8632†	0.8315	0.6343	**0.8653**
	2WikiMultiHopQA	0.6662	0.6826	0.4887	0.6823	0.6157	0.8146†	0.7726	0.7337	**0.8335**
	MedQA	0.6648	0.6233	0.5838	0.6234	0.6446	0.6576	0.6740†	0.5881	**0.7093**
Qwen3-14B	GSM8K	0.8534	0.8116	0.6791	0.8113	0.8113	0.6804	0.8717†	0.7474	**0.8921**
	MATH	0.9262†	0.8300	0.5877	0.8302	0.8296	0.9114	0.9232	0.8801	**0.9865**
	HotpotQA	0.6489	0.6748	0.6494	0.6747	0.6802	**0.8391**	0.8133	0.6316	0.8177†
	2WikiMultiHopQA	0.6941	0.7839	0.6584	0.7838	0.7656	0.8189†	0.7224	0.7881	**0.8247**
	MedQA	0.6875	0.7009	0.6747	0.7009	0.6992	0.7093†	0.6704	0.6614	**0.7407**
Qwen3-32B	GSM8K	0.8839	0.8858	0.7984	0.8864†	0.8864†	0.7620	0.8603	0.8569	**0.9465**
	MATH	0.9210†	0.8253	0.7193	0.8249	0.8211	0.8188	0.8593	0.8318	**0.9732**
	HotpotQA	0.6702	0.6874	0.6775	0.6875	0.6667	0.8606†	0.8423	0.6623	**0.8736**
	2WikiMultiHopQA	0.5826	0.7219	0.6443	0.7220	0.7505	0.8431†	0.8260	0.7323	**0.8452**
	MedQA	0.7223	0.7177	0.6829	0.7177	0.7189	0.8276†	0.7858	0.6777	**0.8587**

captures much of the uncertainty in these settings. By contrast, in multi-hop or long-form reasoning, probability signals are more affected by length bias and locally high-confidence errors (e.g., on 2WikiMultiHopQA with Qwen3-14B, PE improves from 0.6941 to 0.8247 with CoSu-UQ in Table I). Semantic relevance methods (SentenceSAR/TokenSAR) can provide additional discrimination in some cases but are overall less stable (e.g., on GSM8K with Qwen3-4B, SentenceSAR drops to 0.5664 compared to CoSu-UQ 0.8612 in Table I).

Limitations of semantic clustering under long-form reasoning. Semantic Entropy (SE) can be competitive in some multi-hop settings but often degrades in long-form reasoning due to redundancy in intermediate steps (e.g., on HotpotQA with Qwen3-14B, SE drops to 0.6802 while CoSu-UQ reaches 0.8177 in Table I). This suggests that SE clustering may reflect superficial similarity of reasoning traces rather than disagreement at the final-answer level, limiting its discriminative power in long-form reasoning.

Robust gains from jointly modeling confidence and semantic support. CoSu-UQ shows more consistent gains across tasks and model scales. Generation confidence captures locally unstable steps early in the reasoning process, while cross-sample semantic support explicitly models mutual corroboration among reasoning paths, reducing "high-confidence but wrong" cases. This advantage is especially clear in math and multi-hop reasoning (e.g., on MATH with Qwen3-14B, CoSu-UQ reaches 0.9865, exceeding PE 0.9262 and LUQ 0.9232 in Table I). CoSu-UQ also remains best on domain-specific medical QA across model sizes (e.g., on MedQA with Qwen3-32B, it achieves 0.8587, outperforming LUQ 0.7858 and PE 0.7223 in Table I).

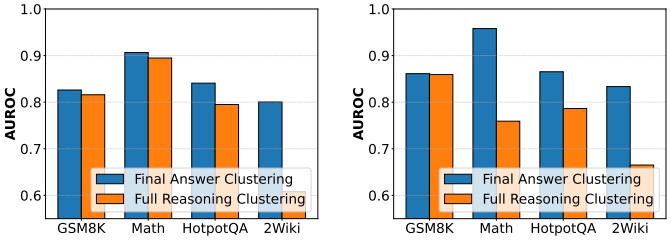
Majority agreement is informative but not sufficient. We additionally include a Self-Consistency (SC) baseline that measures uncertainty via the relative mass of the largest semantic cluster. SC performs strongly on some multi-hop QA settings, highlighting that majority answer agreement can serve as an informative uncertainty signal (e.g., on HotpotQA with Llama3.1-8B, SC reaches 0.8784 in Table I). However, SC is purely frequency-based and cannot detect cases where responses are highly consistent yet systematically wrong, leading to less stable behavior across tasks. In contrast, CoSu-UQ integrates both generation confidence and cross-sample semantic support, yielding more robust uncertainty estimation across diverse reasoning scenarios.

TABLE II: Ablation of confidence-only, support-only, and combined signals for uncertainty estimation (AUROC). Bold indicates the best result; † indicates the second best.

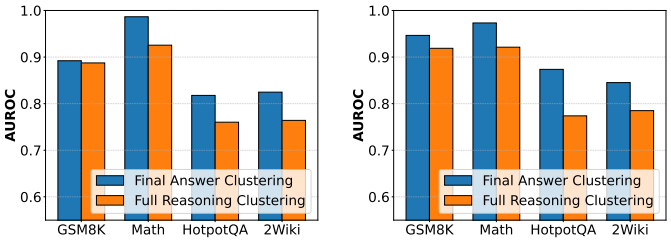
Run Setting	Dataset	Confidence-only	Support-only	Combined
Llama-3.1-8B	GSM8K	0.8238†	0.8215	**0.8261**
	MATH	0.8627	0.9046†	**0.9066**
	HotpotQA	0.7552	0.8400†	**0.8408**
	2WikiMultiHopQA	0.7942†	0.7730	**0.8004**
Qwen3-4B	GSM8K	0.8609†	0.8424	**0.8612**
	MATH	0.9469	0.9550†	**0.9580**
	HotpotQA	0.8027	0.8629†	**0.8653**
	2WikiMultiHopQA	0.8184†	0.8146	**0.8335**
Qwen3-14B	GSM8K	0.8440	0.8877†	**0.8921**
	MATH	0.9813	0.9825†	**0.9865**
	HotpotQA	0.7203	0.8153†	**0.8177**
	2WikiMultiHopQA	0.8152†	0.8063	**0.8247**
Qwen3-32B	GSM8K	0.9336	0.9462†	**0.9465**
	MATH	0.9689†	0.9571	**0.9732**
	HotpotQA	0.8308	0.8588†	**0.8736**
	2WikiMultiHopQA	0.8011	0.8403†	**0.8452**

C. Ablation Studies

Confidence vs. Support. Table II compares confidence-only and support-only variants of CoSu-UQ. Confidence captures



(a) Llama-3.1-8B (b) Qwen3-4B



(c) Qwen3-14B (d) Qwen3-32B

Fig. 2: Bidirectional entailment clustering on final answers vs. full reasoning.

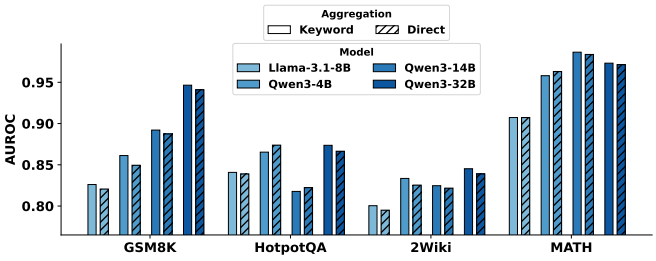


Fig. 3: Keyword-based vs. direct aggregation (AUROC) across models and datasets.

token-level reliability and remains strong and competitive in some GSM8K and MATH settings. Support captures cross-sample semantic agreement and is often more helpful in long multi-hop settings (e.g., HotpotQA and 2WikiMultiHopQA), where token-probability signals can be noisier. Notably, neither single signal dominates uniformly across all settings; performance depends on both dataset and model. Combining both signals consistently yields the best AUROC across all datasets, confirming their complementarity and motivating our final joint design.

Clustering Granularity. Fig. 2 studies how semantic clustering choices affect uncertainty estimation. We find that clustering final answers consistently outperforms clustering full reasoning traces. This is because intermediate chains often share superficial overlap even when the final conclusions diverge, which can blur disagreement signals. Entailment-based clustering further improves robustness by grouping answers that are logically consistent despite lexical variation. These results suggest that uncertainty should be anchored at the decision level rather than at noisy intermediate steps; accordingly, we adopt answer-level entailment clustering in

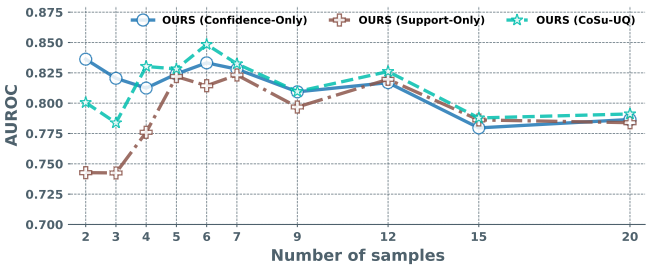


Fig. 4: Effect of sample count k on AUROC under different ablations.

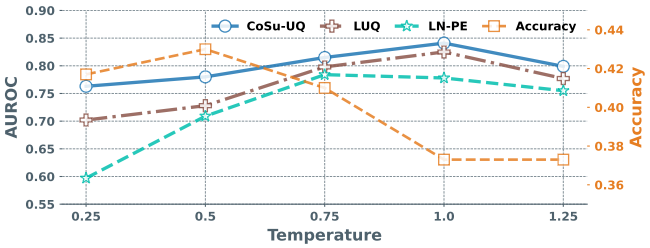


Fig. 5: Effect of sampling temperature on AUROC and accuracy.

CoSu-UQ. We report the full AUROC table in Appendix D.

Keyword Filtering. Fig. 3 evaluates whether keyword-based token aggregation improves confidence estimation. Overall, keyword filtering is beneficial in 12 of 16 model–dataset settings. Gains are consistent on GSM8K and 2WikiMultiHopQA (4/4 models each), while MATH and HotpotQA show mixed results (2/4 each). Therefore, we retain keyword-based aggregation as a lightweight refinement. From a computational perspective, the end-to-end runtime can be approximated as $T \approx T_{\text{sample}} + T_{\text{split}} + \max(T_{\text{conf}}, T_{\text{sup}}, T_{\text{cluster}}) + T_{\text{fusion}} + T_{\text{agg}}$. Under the same hardware setup and profiling protocol (300 sampled responses per dataset), our runtime measurements averaged across the four models show that $T_{\text{sup}} \approx 4$ h 32 min and $T_{\text{cluster}} \approx 3.54$ s. For the confidence branch, $T_{\text{conf}}^{\text{keyword}} \approx 4$ min 50 s, while $T_{\text{conf}}^{\text{direct}} \approx 10$ s. Since the critical path is dominated by T_{sup}, keyword filtering adds limited additional end-to-end wall-clock overhead in practice.

Sampling Size. We ablate aggregation uncertainty by varying sample count while keeping uncertainty sources fixed. We generate 20 responses per question and build subsets with $k \in \{2, 3, 4, 5, 6, 7, 9, 12, 15, 20\}$ using the first k samples, so differences are attributable only to sample size. We evaluate confidence-only, support-only, and the combined uncertainty. With small k, support estimates are noisier because each response has limited peer evidence. As Fig. 4 shows, performance improves with k and peaks around $k = 5$–7 for all variants, with the combined method consistently on top. Performance saturates beyond $k \approx 7$, indicating diminishing returns and that a modest budget ($N = 5$–7) is sufficient for stable uncertainty estimates. Detailed AUROC values for each

sampling size are reported in Appendix E.

Sampling Temperature. Fig. 5 analyzes how decoding temperature affects both answer accuracy and uncertainty discrimination. As temperature increases, sampling becomes more diverse, shifting the distribution of reasoning paths and thus impacting UQ performance. Overall, CoSu-UQ remains robust across a wide temperature range and consistently achieves strong AUROC, indicating that jointly modeling confidence and cross-sample semantic support is stable under different sampling regimes. We also observe that moderate temperatures strike a favorable balance between answer correctness and semantic diversity, while overly low or high temperatures reduce sample informativeness or reliability.

V. Conclusion

We propose a unified uncertainty quantification approach for reasoning-centric language models that fuses generation confidence with cross-sample semantic support. Answer-level bidirectional entailment clustering captures both internal uncertainty and cross-response disagreement. Extensive experiments across diverse reasoning tasks show this joint modeling yields more stable and reliable uncertainty estimates than probability or semantics-based baselines.

VI. Limitations

While the proposed method performs stably across models and tasks, it relies on multi-sample generation and extra semantic modeling, increasing overhead versus single-pass probability-based methods. Future work may explore more efficient implementations to reduce inference cost.

Acknowledgment

This work was supported in part by the Research Project of Quan Cheng Laboratory, China (Grant No. QCL20250203), and in part by the National Natural Science Foundation of China under Grant 62172053 and Grant 62302059.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[3] T. OpenAI, "Learning to reason with llms," *OpenAI Blog*, 2024.

[4] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[6] J. Long, "Large language model guided tree-of-thought," *arXiv preprint arXiv:2305.08291*, 2023.

[7] A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," *arXiv preprint arXiv:2205.09712*, 2022.

[8] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

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

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. Ghaffari Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen *et al.*, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, p. 141, 2023.

[12] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, and H. Wei, "Uncertainty quantification and confidence calibration in large language models: A survey," *arXiv preprint arXiv:2503.15850*, 2025.

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

[14] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," *arXiv preprint arXiv:2002.07650*, 2020.

[15] L. Kuhn, Y. Gal, S. Farquhar, and J. Kossen, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," *arXiv preprint arXiv:2302.09664*, 2023.

[16] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[17] J. Duan, H. Cheng, S. Wang, A. Zavalny, C. Wang, R. Xu, B. Kailkhura, and K. Xu, "Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5050–5063. [Online]. Available: https://aclanthology.org/2024.acl-long.276/

[18] C. Zhang, F. Liu, M. Basaldella, and N. Collier, "LUQ: Long-text uncertainty quantification for LLMs," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5244–5262. [Online]. Available: https://aclanthology.org/2024.emnlp-main.299/

[19] T. Fu, J. Conde, G. Martínez, M. Grandury, and P. Reviriego, "Multiple choice questions: Reasoning makes large language models (llms) more self-confident even when they are wrong," *arXiv preprint arXiv:2501.09775*, 2025.

[20] S. J. Mielke, A. Szlam, E. Dinan, and Y.-L. Boureau, "Reducing conversational agents' overconfidence through linguistic calibration," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 857–872, 2022.

[21] S. Lin, J. Hilton, and O. Evans, "Teaching models to express their uncertainty in words," *arXiv preprint arXiv:2205.14334*, 2022.

[22] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu, "A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation," *arXiv preprint arXiv:2307.03987*, 2023.

[23] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[24] B. Zhang and R. Zhang, "CoT-UQ: Improving response-wise uncertainty quantification in LLMs with chain-of-thought," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 26 114–26 133. [Online]. Available: https://aclanthology.org/2025.findings-acl.1339/

[25] Z. Li, S. Shen, W. Yang, R. Jin, H. Chen, J. Ren *et al.*, "Enhancing uncertainty quantification in large language models through semantic graph density," in *The 41st Conference on Uncertainty in Artificial Intelligence*.

[26] Y. Li, S. Wang, L. Huang, and L.-P. Liu, "Graph-based confidence calibration for large language models," *arXiv preprint arXiv:2411.02454*, 2024.

[27] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spacy: Industrial-strength natural language processing in python," 2020.

[28] Y. Huang, J. Song, Z. Wang, S. Zhao, H. Chen, F. Juefei-Xu, and L. Ma, "Look before you leap: An exploratory study of uncertainty analysis for large language models," *IEEE Transactions on Software Engineering*, vol. 51, no. 2, pp. 413–429, 2025.
[29] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD
[30] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388
[31] M. AI, "The LLaMA 3 herd of models," *arXiv preprint arXiv:2404.14219*, 2024.
[32] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.
[33] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Guo, D. Liang, S. Song, and J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," in *Advances in Neural Information Processing Systems*, 2021.
[34] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
[35] N. Ho, J. Nguyen, N. Tran, N. Tuan, and T. Nguyen, "Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.
[36] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/14/6421
[37] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.

APPENDIX

A. Dataset Processing Details

We benchmark on GSM8K [32], MATH [33], 2WikiMultiHopQA [35], HotpotQA [34], and MedQA [36]. For MATH, we clean and normalize answers by retaining only parsable numeric values or simple expressions. For multi-hop QA datasets, we do not use supporting documents and rely only on the question text, treating them as generative reasoning tasks.

B. One-shot Chain-of-Thought Prompting.

We use a one-shot Chain-of-Thought (CoT) prompting strategy for all experiments. We prepend a single demonstration example with step-by-step reasoning, encouraging the model to produce structured reasoning steps and a final answer in a consistent format:

```
System: Please reason the following
question step by step. Label each reasoning
step as ``Step i'', and provide the final
answer on a separate line starting with
``Final Answer:''.
User: Question: [One-shot example question]
Assistant: Step 1: ...
Final Answer: [Example final answer]
User: Question: [Target question]
```

For MedQA, we use the same multi-turn one-shot construction as GSM8K, with an explicit multiple-choice answer format constraint:

```
System: Please reason the following
question step by step. Label each reasoning
step as ``Step i'', and provide the final
answer on a separate line starting with
``Final Answer:''. Extremely Important: The
Final Answer must follow the format ``Final
Answer: [Option Letter] ([Option Text])''.
User: Question: [One-shot example question]
Assistant: Step 1: ...
Step 2: ...
Final Answer: [Option Letter] ([Option
Text])
User: Question: [Target question]
```

In implementation, prompts are instantiated from structured templates by replacing placeholders for the demonstration question/answer and the target question. We use the same sampling configuration as in Sec. IV-A: each query is sampled $N = 5$ times with temperature 1.0, and all models follow identical inference settings for fair comparison. For MedQA, we additionally enforce answer-format validity during parsing: only outputs matching "Final Answer: [Option Letter] ([Option Text])" are accepted as valid final answers for downstream uncertainty computation. This constraint reduces formatting noise and improves consistency of answer extraction across samples.

C. Human Annotation Study

To validate the reliability of LLM-as-judge labeling, we conduct a human annotation study on two model sizes (Qwen3-4B and Qwen3-32B) across three datasets (MATH, HotpotQA, and MedQA), as summarized in Table III. For each model–dataset pair, we sample 100 instances and annotate one response per instance (600 samples total), assigning binary labels based only on final-answer correctness. Automatic labels are obtained via majority voting among three judge models (GPT-4.1-mini[1], DeepSeek-V3.2-Fast[2], and Grok-4-1-Fast-Reasoning[3]) to reduce individual bias.

TABLE III: Human annotation subsets by model size and dataset.

Model	Dataset	Consistency
Qwen3-4B	GSM8K	98%
	HotpotQA	98%
	MedQA	99%
Qwen3-32B	GSM8K	99%
	HotpotQA	98%
	MedQA	99%

Overall, agreement between LLM-as-judge labels and human annotations is consistently high (98%–99%) across all evaluated subsets, supporting the reliability of our automatic labeling protocol for final-answer correctness.

[1]https://openai.com/index/gpt-4-1/
[2]https://chat.deepseek.com/
[3]https://x.ai/grok

D. Clustering and Keyword Filtering Ablations.

Table IV reports additional AUROC results for answer-level clustering versus full-reasoning clustering, as well as keyword-based confidence aggregation versus direct aggregation.

TABLE IV: Ablations on clustering granularity and keyword filtering (AUROC). Ours uses final-answer clustering + keyword aggregation. w/o FinalAns replaces final-answer clustering with full-reasoning clustering. w/o Keywords replaces keyword aggregation with direct aggregation.

Model	Dataset	Ours	w/o FinalAns	w/o Keywords
Llama-3.1-8B	GSM8K	0.8261	0.8159	0.8206
Llama-3.1-8B	MATH	0.9066	0.8948	0.9073
Llama-3.1-8B	HotpotQA	0.8408	0.7950	0.8389
Llama-3.1-8B	2WikiMultiHopQA	0.8004	0.6078	0.7950
Qwen3-4B	GSM8K	0.8612	0.8593	0.8494
Qwen3-4B	MATH	0.9580	0.7592	0.9631
Qwen3-4B	HotpotQA	0.8653	0.7865	0.8738
Qwen3-4B	2WikiMultiHopQA	0.8335	0.6651	0.8254
Qwen3-14B	GSM8K	0.8921	0.8874	0.8874
Qwen3-14B	MATH	0.9865	0.9255	0.9837
Qwen3-14B	HotpotQA	0.8177	0.7602	0.8223
Qwen3-14B	2WikiMultiHopQA	0.8247	0.7641	0.8217
Qwen3-32B	GSM8K	0.9465	0.9188	0.9409
Qwen3-32B	MATH	0.9732	0.9212	0.9714
Qwen3-32B	HotpotQA	0.8736	0.7738	0.8664
Qwen3-32B	2WikiMultiHopQA	0.8452	0.7850	0.8389

E. Sampling Size Ablation Table

Table V reports AUROC for confidence-only, support-only, and combined variants across different sampling sizes.

TABLE V: Sampling size ablation with different uncertainty signals (AUROC). Bold indicates the best result and † indicates the second best within each row.

Samples (k)	Confidence-Only	Support-Only	Combined
2	**0.8362**	0.7427	0.8005†
3	**0.8205**	0.7425	0.7838†
4	0.8124†	0.7760	**0.8302**
5	0.8241†	0.8220	**0.8282**
6	0.8332†	0.8144	**0.8484**
7	0.8280†	0.8232	**0.8325**
9	0.8093†	0.7968	**0.8096**
12	0.8169	0.8197†	**0.8261**
15	0.7796	0.7862†	**0.7878**
20	0.7865†	0.7839	**0.7911**