Uncertainty Quantification of Large Language Models through Multiple Uncertainty Sources

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains. However, the reliability of responses from LLMs remains a question. Uncertainty quantification (UQ) of LLMs is crucial for ensuring their reliability, especially in areas such as healthcare. Existing UQ methods, often designed around a single resource such as Natural Language Inference (NLI) or graph-based metrics, fail to capture the multifaceted nature of uncertainty in natural language generation. In this work, we propose MS-UQ, a novel Multi-Resource Uncertainty Quantification framework that integrates heterogeneous uncertainty signals into a unified measure. Our approach concatenates matrices from diverse resources and employs tensor decomposition to orthogonally disentangle unique and shared information. To ensure scalability, we construct an adaptive ensemble of outputs from different decomposition methods, enabling the incorporation of new uncertainty sources. Experiments on CoQA, NQ_Open, and HotpotQA demonstrate that MS-UQ consistently outperforms existing methods, offering a comprehensive and scalable solution for uncertainty estimation in black-box LLMs and a more robust framework for enhancing LLM reliability in high-stakes applications. Our code can be accessed at *https://anonymous.4open.science/r/MDUQ-First-202E/README.md*.

1 Introduction

Uncertainty quantification for LLMs (Liu et al., 2025) is essential for ensuring their reliability in high-stakes applications, yet remains challenging due to the open-ended and underspecified nature of natural language generation. Unlike classification tasks, where uncertainty can often be estimated from output probabilities or model variance (Gal and Ghahramani, 2016; Ye et al., 2024), LLM outputs may vary lexically while remaining semantically consistent (Kuhn et al., 2023). Recent methods address this by modeling uncertainty through specific signals, such as semantic agreement via natural language inference (Kuhn et al., 2023), response diversity across generations (Lin et al., 2023), or factual consistency via claim-level graphs (Da et al., 2024). Each of these approaches focuses on a single source of uncertainty.

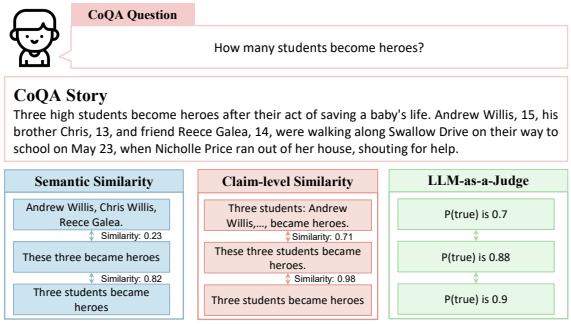

Figure 1: Illustrative example from CoQA: Different uncertainty sources have different information. Our method models uncertainty by integrating such multi-signal information.

However, uncertainty in language model outputs arises from multiple sources, each capturing different forms of variation. For example, a response may be semantically fluent but factually inaccurate. Individual UQ methods capture only one aspect of variability, such as semantic overlap, graph structure, or factual entailment, leading to incomplete assessments. Moreover, some signals reflect local properties (e.g., node degree), while others capture global structure (e.g., matrix spectra). For example, in Figure 1, responses differ in lexical and structural form across different UQ methods but convey the same underlying meaning, underscoring the importance of modeling both shared and source-specific variation when estimating uncertainty. For example, given the question *"How many students became heroes?"*, multiple responses *"Andrew Willis, Chris Willis, Reece Galea"*, *"These three became heroes"*, and *"Three students became*

heroes". These responses differ in surface form and semantic expression, but they convey the same factual content.

The above example highlights the central challenges for developing a comprehensive UQ framework for black-box LLMs: (1) disentangle the unique and shared components of multiple uncertainty signals and (2) scale to accommodate additional signals without redesigning the entire pipeline. Simply concatenating features or averaging scores fails to account for redundancy, may mask important differences, and risks overfitting.

To address these challenges, we propose MS-UQ, a **M**ulti-**S**ource **U**ncertainty **Q**uantification framework designed to integrate and disentangle uncertainty signals from diverse sources. The core intuition is that different signals, such as semantic similarity from NLI models, structural cues from graph statistics, or response diversity, capture complementary aspects of uncertainty, but also exhibit overlapping patterns. Instead of treating these signals in isolation or combining them heuristically, MS-UQ constructs a third-order tensor that stacks the similarity matrices from each source, treating them as separate slices along the source dimension. We then apply tensor decomposition to factor this tensor into low-rank components and compute the reconstruction error, which serves as a unified uncertainty estimate. A low reconstruction error indicates consistent behavior across sources, whereas a high error reflects divergence or ambiguity. This formulation enables us to isolate source-specific variation, avoid redundant signal amplification, and extract uncertainty in a structured way. To ensure scalability as new sources are introduced, we further ensemble reconstruction errors from multiple decomposition methods, allowing the framework to remain robust without redesign or retraining. In summary, our main contributions are:

- We propose a unified framework that integrates diverse uncertainty signals (e.g., NLI similarity, graph statistics) into a shared representation for quantifying model uncertainty.
- We introduce a tensor-based formulation and ensemble framework that models multi-source uncertainty as a third-order similarity tensor and quantifies uncertainty through reconstruction error derived from tensor decomposition. This formulation allows the framework to scale to new sources without redesign.
- Comprehensive experiments on different datasets and models show that MS-UQ consistently outperforms existing state-of-the-art UQ methods. In addition, we show that MS-UQ is more effective than taking the mean of uncertainties from different sources, which may overlook inter-signal dependencies.

2 Related Works

Uncertainty quantification for traditional machine learning problems such as regression has been well studied (Ye et al., 2024; Amini et al., 2020; Sensoy et al., 2018; Ovadia et al., 2019). Most previous works on uncertainty quantification for natural language processing (NLP) consider text classifiers (Jiang et al., 2021; Desai and Durrett, 2020; Kamath et al., 2020) or text regressors (Glushkova et al., 2021; Wang et al., 2022). To transfer NLP tasks into a classification task, previous work may consider using multi-choice question answering datasets or transferring open-ended questions into a multi-choice form (Kamath et al., 2020).

To quantify the uncertainty in open-ended generation tasks, researchers propose semantic entropy (Kuhn et al., 2023), which calculates entropy considering semantic information. However, such an approach still requires the token-related probability values as input. To compute uncertainty for black-box LLMs, previous works (Lin et al., 2023; Chen and Mueller, 2024; Da et al., 2024; Gao et al., 2024; Hou et al., 2024) take a step further compared with semantic entropy and utilize the similarity and consistency between different generated answers from the same query to the LLMs. NLI models are first used to obtain the similarity between answers to get a similarity matrix, upon which eigenvalues are calculated to measure the uncertainty from the graph Laplacian (Lin et al., 2023; Chen and Mueller, 2024; Da et al., 2024; Catak and Kuzlu, 2024). While these works provide useful quantification of uncertainty from different sources, there is a lack of methods to unify them.

3 Preliminaries

UQ aims to estimate the confidence or variability in model $\mathcal{M}$'s predictions. Let an LLM be represented as a probabilistic model $\mathcal{M}$ that generates a response y conditioned on an input x. For the UQ of black-box LLMs, it typically relies on analyzing the variability or consistency across multiple responses sampled from $\mathcal{M}$ given the same x, de-

noted as $\{y_1, y_2, \ldots, y_N\}$, where N responses are sampled from $\mathcal{M}$ for the input x. Formally, it can be defined as:

Problem 1 (Black-box UQ)**.** *Given a black-box LLM $\mathcal{M}$ with input x and a set of responses $\{y_i\}$ sampled n times from $\mathcal{M}$ with the same input x, the goal of UQ task is to compute an uncertainty score U that reflects the variability across $\{y_i\}$.*

To analyze the variability across $\{y_1, y_2, \ldots, y_N\}$, one of the key steps is to compute pairwise similarity scores between responses. The pair-wise similarity matrix $S \in \mathbb{R}^{N \times N}$ can be computed using several complementary approaches capturing different aspects of information from existing work. Common approaches are Jaccard Similarity (Lin et al., 2023) and entailment from NLI (He et al., 2021) that provides semantic similarity with $s_{ij} = \frac{1}{2}\left(P_{\text{entail}}(A^i, A^j) + P_{\text{entail}}(A^j, A^i)\right)$, etc. The uncertainty can then be derived from the matrix S to capture the variability across responses using different methods.

Existing UQ methods provide one source of uncertainty. However, uncertainties in LLMs can come from multiple sources that intertwine. For example, reasoning ambiguities often interact with semantic inconsistencies in complex ways that single-dimensional approaches cannot capture. In this paper, we focus on unifying these sources.

Problem 2 (Multi-source Black-box UQ)**.** *Let $\mathcal{M}$ be a language model and $\{y_1, y_2, \ldots, y_N\}$ be the set of responses generated for an input x. We define a set of D distinct uncertainty signals, each corresponding to a different source (e.g., semantic, factual, internal activation-based). Each source $d \in \{1, \ldots, D\}$ gives rise to a matrix $\mathbf{S}^{(d)} \in \mathbb{R}^{N \times N}$, which quantifies pairwise response similarity or contains the important information under that source. We construct a third-order similarity tensor: $\mathcal{S} \in \mathbb{R}^{N \times N \times D}, \mathcal{S}(:, :, d) = \mathbf{S}^{(d)}$. Our goal is to compute an integrated uncertainty score U that reflects the joint variability encoded across all D sources.*

4 Methodology

We propose MS-UQ, a **Multi-Source Uncertainty Quantification** framework that estimates LLM uncertainty by integrating heterogeneous uncertainty signals through tensor decomposition and reconstruction analysis.

4.1 Overview

LLM uncertainty arises from diverse sources: semantic agreement, factual consistency, and structural variation often overlap but also provide distinct signals. Existing methods typically analyze these sources separately or combine them heuristically, which leads to two problems: (1) **redundancy**, where shared patterns dominate and suppress meaningful variations, and (2) **sensitivity**, where model-specific biases affect the final uncertainty estimate, restricting methods to accommodate additional signals.

To address this, MS-UQ formulates uncertainty estimation as a tensor-based reconstruction problem. We model the structure of multiple uncertainty sources jointly and extract a robust measure of model uncertainty via reconstruction error, which captures how well the model's responses compress under multiple perspectives. As illustrated in fig. 2, our approach integrates these dimensions through three key phases:

1. **Tensor Representation by Uncertainty Sources** (section 4.2): Construct a multi-source similarity tensor $\mathcal{S} \in \mathbb{R}^{N \times N \times D}$ by concatenating similarity matrix $\mathbf{S}$ from all uncertainty sources.

2. **Tensor Decomposition** (section 4.3): Apply orthogonal tensor decomposition to isolate dimension-specific features and suppress redundant information.

3. **Ensemble Scoring** (section 4.4): Combine decomposition residuals across dimensions with ensemble methods to compute the final uncertainty measure U_{final}.

4.2 Multi-Source Uncertainty Representation

Given m uncertainty sources, each providing a pairwise similarity matrix $\mathbf{S}_k \in \mathbb{R}^{N \times N}$ for $k = 1, \ldots, D$, our goal is to jointly model these matrices to estimate uncertainty from all sources simultaneously. A natural approach is to flatten the matrices into a single large $(N \times D) \times N$ matrix via concatenation:

$$S_{\text{concat}} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_D \end{bmatrix} \in \mathbb{R}^{ND \times N}. \quad (1)$$

Uncertainty Source	S_{ij} (Similarity Matrix)	Uncertainty Detail
U_{EigV} $(Agre)$ [25]	$S_{ij} = \mathrm{P}_{\mathrm{entail}}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{\mathrm{entail}}$ is from NLI	$U_{EigV} = \sum_k^m \lambda_k$, λ_k is the k-th eigenvalue
$U_{Ecc}(Agre)$ [25]	$S_{ij} = \mathrm{P}_{\mathrm{entail}}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{\mathrm{entail}}$ is from NLI	$U_{Ecc} = \left\|[V_1^T, \ldots, V_m^T]\right\|_2$, V_k^T is the transform of the eigenvector
$U_{Degree}(Agre)$ [25]	$S_{ij} = \mathrm{P}_{\mathrm{entail}}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{\mathrm{entail}}$ is from NLI	$U_{Degree} = \mathrm{trace}(m\mathrm{I} - \mathrm{D})/\mathrm{m}^2$, D is the degree matrix based S
$U_{EigV}(DisAgre)$ [25]	$S_{ij} = \mathrm{P}_{contra}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{contra}$is from NLI	$U_{EigV} = \sum_k \lambda_k$, λ_k is the k-th eigenvalue
$U_{Ecc}(DisAgre)$ [25]	$S_{ij} = \mathrm{P}_{contra}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{contra}$is from NLI	$U_{Ecc} = \left\|[V_{1,\ldots}^T, V_m^T]\right\|_2$, V_k^T is the transform of the eigenvector
$U_{Degree}(DisAgre)$ [25]	$S_{ij} = \mathrm{P}_{contra}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{contra}$is from NLI	$U_{Degree} = \mathrm{trace}(m\mathrm{I} - \mathrm{D})/\mathrm{m}^2$, D is the degree matrix based S
$p(true)$ [19]	$S_{ii} = P(y_i \; is \; True \mid \theta)$, θ is from LLM.	$U_p = 1 - Mean([p(y_1), \ldots, p(y_N)])$, y_n is the n-th answer.
$D - UE$ [7]	$S_{ij} = \mathrm{P}_{\mathrm{entail}}(\mathrm{y_i}, \mathrm{y_j})$, $\mathrm{P}_{\mathrm{entail}}$ is from NLI	$U_{EigV} = \sum_k \lambda_k$, λ_k is the k-th eigenvalue.

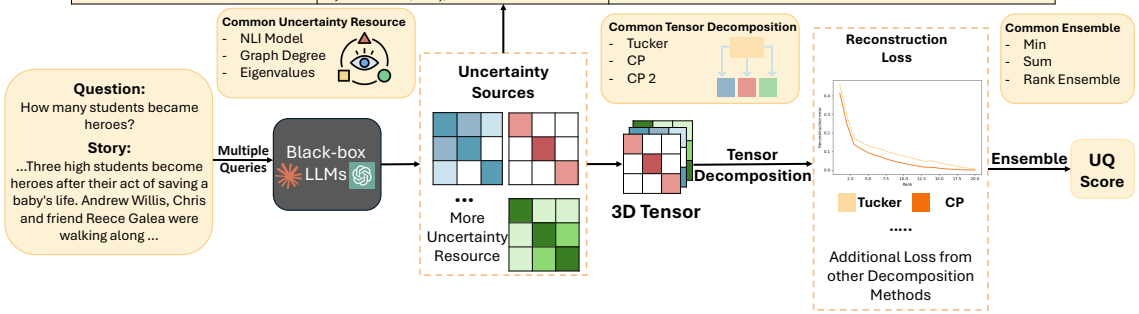

Figure 2: The overall pipeline of MS-UQ. MS-UQ forms the 3D tensor by different uncertainty source matrices and utilizes tensor decomposition methods to ensure the information from all source matrices is effectively incorporated. The tensor decomposition methods allow us to add more source matrices and handle overlapping information.

This strategy allows for the use of standard tools such as singular value decomposition (SVD). However, concatenation introduces a critical problem: **redundancy across sources**. Many uncertainty signals capture overlapping information. For example, NLI-based semantic similarity and graph degree both reflect response agreement. Flattening these signals suppresses source-specific variation.

To quantify this redundancy, we compute the *Spectral Norm Ratio* (SNR) (Wang et al., 2020), which compares the largest eigenvalues (spectral norms) of two matrices: $\mathrm{SNR} = \frac{\|\Lambda_{\mathbf{S}}\|_2}{\|\Lambda_{\mathbf{S}_k}\|_2}$, where $\Lambda_{\mathbf{S}}$ and $\Lambda_{\mathbf{S}_k}$ denote the leading eigenvalues of $\mathbf{S}$ and $\mathbf{S}_k$. An SNR close to 1 indicates that $\mathbf{S}$ and $\mathbf{S}_k$ share dominant structures, implying repeated information. In our experiments on CoQA with Llama2-13b (Touvron et al., 2023), the mean SNR between semantic and claim-level signals is 0.89, indicating substantial shared structure. We observe similar results on HotpotQA and NQ_Open, as reported in section B.

Naïvely concatenating matrices under such redundancy leads to sub-optimal results: shared patterns dominate the decomposition, masking differences that could signal uncertainty, as shown in the experimental part. To mitigate this, we model the collection of matrices as a third-order tensor:

$$\mathcal{S} = [S_1|S_2|...|S_D], \mathcal{S} \in \mathbb{R}^{N \times N \times D}, \qquad (2)$$

where each slice $\mathcal{S}(:,:,d) = \mathbf{S}^{(d)}$ corresponds to one uncertainty source. This formulation preserves the structure of each matrix along dimensions, allowing us to disentangle shared and source-specific variation through tensor decomposition. Unlike concatenation, this approach treats the multi-source signals orthogonally, enabling a more balanced representation of uncertainty across sources.

4.3 Tensor Decomposition for Uncertainty Estimation

Given the constructed tensor $\mathcal{S} \in \mathbb{R}^{N \times N \times D}$, the key question is how to extract an overall uncertainty measure from its structure. Our core intuition is based on **structural compressibility**: if the LLM produces consistent responses across multiple uncertainty sources, the corresponding tensor will exhibit a simple, low-rank structure. In contrast, if the responses diverge semantically, factually, or structurally, the tensor will become more complex and difficult to compress.

We formalize this intuition by applying **tensor decomposition** to approximate $\mathcal{S}$ with lower-rank components and use the reconstruction error as an indicator of uncertainty. High reconstruction error implies that the model's outputs vary significantly across sources, signaling higher uncertainty.

4.3.1 Tucker vs. CP Decomposition.

To extract uncertainty from the tensor $\mathcal{S}$, we apply two complementary tensor decomposition methods: Tucker decomposition and Canonical Polyadic

(CP) decomposition (Kolda and Bader, 2009; Perros et al., 2017). Each provides a distinct trade-off between flexibility, interpretability, and robustness.

• **Tucker decomposition** models $\mathcal{S}$ as a core tensor multiplied by orthogonal factor matrices along each mode (De Lathauwer et al., 2000):

$$\mathcal{S} \approx \mathcal{G} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{3}$$

where $\mathcal{G}$ captures the interactions between modes and $U^{(i)}$ are orthogonal factor matrices that span subspaces of each dimension. Tucker decomposition allows flexible rank selection for each mode, making it well-suited for capturing complex interactions between response pairs and uncertainty.

• **CP decomposition** offers a complementary perspective by representing $\mathcal{S}$ as a sum of rank-one components:

$$\mathcal{S} \approx \sum_{r=1}^{R} \lambda_r \, a_r^{(1)} \circ a_r^{(2)} \circ a_r^{(3)}, \tag{4}$$

where $\circ$ denotes the outer product. CP decomposition is simpler and more robust to noise, but may fail to capture higher-order interactions that Tucker decomposition can model. Both tensor decompositions allow any number from the third mode of the tensor (uncertainty source dimension), which leads to good scalability when considering more uncertainty sources.

4.3.2 Uncertainty Estimation via Reconstruction Error

For each decomposition, we compute the reconstruction error at rank R as:

$$\epsilon_R = \frac{\|\mathcal{S} - \hat{\mathcal{S}}_R\|_F}{\|\mathcal{S}\|_F}, \tag{5}$$

where $\hat{\mathcal{S}}_R$ is the rank-R approximation and $\|\cdot\|_F$ denotes the Frobenius norm. We compute ϵ_R^{cp} and ϵ_R^{tucker} separately. Lower reconstruction error indicates that the tensor structure is simpler, corresponding to more consistent LLM outputs, while higher error suggests greater variability and higher uncertainty. Figure 3 shows the relationship between reconstruction error and LLM output accuracy on CoQA. Samples with higher answer accuracy correspond to lower reconstruction error, confirming that structural compressibility aligns with model confidence.

While Tucker and CP decompositions each have strengths, neither alone fully addresses the challenges of multi-source uncertainty estimation. CP provides a simpler and more robust approximation but may overlook complex interactions between responses and sources. Tucker captures richer structure but can overfit or obscure distinctions between uncertainty sources if all modes are compressed equally. To balance these trade-offs, we adopt a structure-aware decomposition strategy before ensembling them: we leave the third mode (uncertainty source dimension) uncompressed, retaining full rank along this axis while decomposing the response-pair dimensions. This design ensures that the ensemble leverages both CP's robustness and Tucker's expressiveness while preserving source-specific signals. We have the following theorem to show the effectiveness of using reconstruction errors as uncertainties:

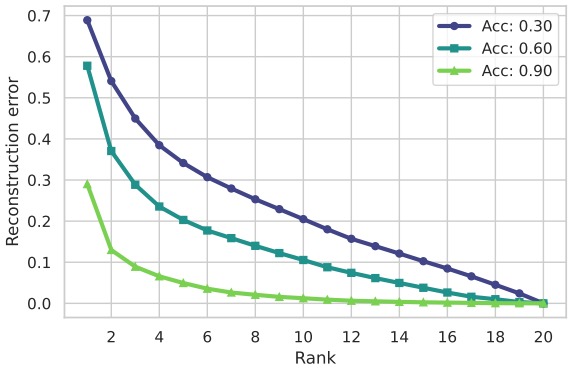

Figure 3: Reconstruction error versus rank for CP decomposition. Higher-accuracy samples produce lower reconstruction error, supporting that reconstruction error as a proxy for uncertainty could separate samples with different accuracies.

Theorem 1. *Let $\mathcal{S}^{(c)} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ be a tensor composed of c mutually orthogonal rank-1 CP components with identical norms. For any $c_1 > c_2$ and rank $R < c_1$, denote the best rank-R CP approximation of $\mathcal{S}^{(c_i)}$ by $\hat{\mathcal{S}}_R^{(c_i)}$. Then*

$$\left\|\mathcal{S}^{(c_1)} - \hat{\mathcal{S}}_R^{(c_1)}\right\|_F > \left\|\mathcal{S}^{(c_2)} - \hat{\mathcal{S}}_R^{(c_2)}\right\|_F.$$

In the MS-UQ pipeline, stronger disagreements among D sources manifest as a larger number of rank-1 blocks in $\mathcal{S}$, and this theorem ensures a larger reconstruction error in this scenario, making the error serve as a good uncertainty value. Complete proof can be found at section C.

4.4 Ensemble Uncertainty

While the reconstruction error from each decomposition provides a proxy for uncertainty, individual estimates may be sensitive to rank choices or

decomposition-specific biases. To address this, we ensemble errors across multiple ranks and methods to obtain a more stable and comprehensive uncertainty measure.

4.4.1 Rank Ensemble

Low-rank reconstruction error is sensitive to the choice of rank R: a small rank may underfit, while a large rank may overfit to noise. To mitigate this, we aggregate reconstruction errors across all ranks $R = 1$ to n for each method. For CP decomposition, the rank-ensemble uncertainty is:

$$U_{\text{CP}} = \sum_{r=1}^{n} \epsilon_r^{\text{CP}}. \tag{6}$$

This strategy ensures that our uncertainty measure reflects the tensor's structural complexity across multiple scales, not just at a fixed rank. We could also do the same ensemble to the Tucker.

4.4.2 Method Ensemble

Since CP and Tucker decompositions capture different structural properties, we further ensemble their rank-aggregated uncertainties. To produce a reliable and scalable uncertainty measure, it is crucial to account for these factors during the ensemble without introducing additional supervision or requiring hyperparameter tuning. We propose two unsupervised combination strategies for the ensemble method:

• **Sum Ensemble:** The final uncertainty is:

$$U = U_{\text{Tucker}} + U_{\text{CP}}. \tag{7}$$

This strategy treats both decomposition methods equally, ensuring that complementary signals from each are combined.

• **Min Ensemble:** Alternatively, we take the minimum of the two uncertainties:

$$U = \min\left(U_{\text{Tucker}}, U_{\text{CP}}\right). \tag{8}$$

If either decomposition method finds a low-complexity explanation for the responses, we interpret the overall uncertainty as low.

5 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of MS-UQ across multiple datasets, LLMs, and uncertainty settings. Our study is designed to answer the following research questions:

- **RQ1:** Does MS-UQ provide more accurate and robust uncertainty quantification compared to existing methods?
- **RQ2:** How do different ensemble strategies and the integration of multiple uncertainty sources impact performance?
- **RQ3:** Is MS-UQ robust across model scales, architectures, and datasets?

Beyond the research questions, we also provide a detailed case study to show why MS-UQ could work in Appendix section E.

5.1 Experimental Setup

5.1.1 Datasets

Following prior work on uncertainty quantification in open-form question answering (Lin et al., 2022), we evaluate MS-UQ on three QA benchmarks with varying levels of reasoning complexity: Coqa (Reddy et al., 2019) is a conversational question-answering dataset that contains dialogues with free-form answers grounded in diverse passages, which is the easiest dataset among all datasets. HotpotQA (Yang et al., 2018) is a multi-hop QA dataset that demands reasoning over multiple Wikipedia paragraphs to derive correct answers. NQ-Open (Kwiatkowski et al., 2019) consists of real-world queries from Google Search, requiring models to retrieve and answer questions without explicit context, which is the hardest dataset.

5.1.2 Models to Evaluate

We test MS-UQ on a diverse set of LLM architectures to evaluate both scalability and generalization. In detail, we use Llama-2-13b and Llama-2-7B (Touvron et al., 2023) to demonstrate the effectiveness of MS-UQ with different model sizes and use Llama-3.1-8B (Dubey et al., 2024) to show that MS-UQ could also work on different versions of Llama. To further demonstrate the generalization ability for other architectures, we also use Phi4 (Abdin et al., 2024) and Deepseek-R1-distill-7B (Guo et al., 2025) in our paper.

5.1.3 Evaluation Metrics

Effective uncertainty measures should correlate with response correctness: higher uncertainty should indicate a higher likelihood of error. Following prior work (Lin et al., 2023), we evaluate uncertainty estimates by using them to predict whether a generated answer is correct. We report Area Under

Table 1: Comparison of our methods with different baselines on various datasets and large language models. The best result is shown in the **bold** and the second best result is shown in the underline. The results show that MS-UQ with sum ensemble consistently outperforms all baseline methods for all models and datasets, which demonstrates that MS-UQ improves the performance of the uncertainty sources it uses.

Methods	Llama2-13B		Llama2-7B		Llama3.1-8B		Phi4		Deepseek-R1-7B	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Dataset: CoQA [Easy]										
Eigv(Dis)	0.7294	0.7775	0.5965	0.9485	0.5762	0.9071	0.6656	0.9120	0.7841	0.9175
Degree(Dis)	0.7369	0.7815	0.5963	0.9473	0.5728	0.9112	0.6677	0.9121	0.7885	0.9189
Eigv(Agre)	0.7541	0.7876	0.5971	0.9507	0.5791	0.9153	0.6399	0.9051	0.7969	0.9234
Degree(Agre)	0.7548	0.7877	0.5908	0.9413	0.5755	0.9097	0.6278	0.8996	0.7930	0.9222
D-UE	0.7566	0.7885	0.5954	0.9481	0.5825	0.9284	0.6503	0.9079	0.7966	0.9228
P(true)	0.7102	0.7088	0.5404	0.9348	0.5816	0.9323	0.6630	0.9087	0.5389	0.8311
MS-UQ-Sum	**0.7802**	**0.7911**	**0.6145**	**0.9521**	**0.6111**	**0.9326**	**0.6851**	**0.9144**	**0.8115**	**0.9332**
Dataset: HotpotQA [Medium]										
Eigv(Dis)	0.6269	0.7770	0.6111	0.7715	0.6099	0.6874	0.5534	0.8614	0.5969	0.5737
Degree(Dis)	0.6336	0.7790	0.6134	0.7714	0.6202	0.7087	0.5666	0.8590	0.5977	0.5756
Eigv(Agre)	0.6235	0.7638	0.6035	0.7648	0.6176	0.7035	0.5328	0.8497	0.6249	0.5878
Degree(Agre)	0.6217	0.7611	0.5973	0.7600	0.6105	0.7016	0.5294	0.8491	0.6278	0.5902
D-UE	0.6252	0.7659	0.6056	0.7669	0.6212	0.7083	0.5335	0.8511	0.6270	0.5893
P(true)	0.6056	0.7591	0.5901	0.7713	0.6362	0.7077	0.5326	0.8513	0.5081	0.4795
MS-UQ-Sum	**0.6461**	**0.7940**	**0.6242**	**0.7809**	**0.6588**	**0.7291**	**0.5874**	**0.8816**	**0.6413**	**0.5956**
Dataset: NQ_Open [Hard]										
Eigv(Dis)	0.6162	0.7300	0.7280	0.6367	0.6742	0.5343	0.7035	0.6035	0.6696	0.2451
Degree(Dis)	0.6130	0.7168	0.7273	0.6318	0.6865	0.5430	0.7090	0.6094	0.6675	0.2463
Eigv(Agre)	0.6258	0.7276	0.7240	0.6327	0.7463	0.5801	0.7515	0.6351	0.7291	0.2758
Degree(Agre)	0.6286	0.7355	0.7290	0.6324	0.7619	0.5885	0.7542	0.6341	0.7353	0.2781
D-UE	0.6281	0.7320	0.7258	0.6342	0.7551	0.5833	0.7539	0.6366	0.7333	0.2801
P(true)	0.6197	0.7289	0.6532	0.5929	0.7061	0.5522	0.7096	0.6049	0.4865	0.1413
MS-UQ-Sum	**0.6435**	**0.7440**	**0.7432**	**0.6455**	**0.7765**	**0.5946**	**0.7596**	**0.6494**	**0.7653**	**0.2921**

Receiver Operating Characteristic (AUROC) and Area Under Accuracy Rejection Curve (AUARC) as evaluation metrics, where a higher AUROC or AUARC demonstrates better uncertainty measures. To compute AUROC and AUARC, the accuracy of each original response is required. To label responses as correct or incorrect, we use a reference LLM, Qwen2.5-32B (Bai et al., 2023), to provide soft correctness scores from 0 to 100. Following previous works (Da et al., 2024; Lin et al., 2023), we treat responses with scores above 70 as correct.

5.2 Compared Methods

We compare MS-UQ against three categories of state-of-the-art baselines, each corresponding to different uncertainty estimation strategies. In our framework, these baselines can also be integrated as distinct uncertainty sources (their details can be found in section A):

• Semantic Similarity (Lin et al., 2023). These methods construct similarity matrices over model outputs and compute graph-based statistics for uncertainty. We consider six variants, which differ by how the similarity matrix is constructed (entailment probability or contradiction logits) and by how uncertainty is induced from the similarity matrix (Eigenvalue, Degree, or Eccentricity). Details can be found at fig. 2.

• Claim-Level Similarity (D-UE) (Da et al., 2024). This method augments LLM outputs with auxiliary claims and evaluates factual consistency to quantify uncertainty.

• LLM-as-a-Judge (p(true)) (Kadavath et al., 2022). This approach directly queries the LLM to estimate the probability that a claim is true, providing a scalar confidence measure.

For MS-UQ, we consider ensemble strategies of Sum from all sources we use and show the difference between Sum and Min later.

5.3 Overall Performance (RQ1)

In this section, we explore whether MS-UQ has better uncertainties compared with state-of-the-art uncertainty quantification methods. In table 1, we compare MS-UQ with baselines across different datasets and models as introduced in section 5.1.

In detail, we have the following observations:

• MS-UQ consistently outperforms all baselines, suggesting that modeling uncertainty from multiple perspectives is more reliable than relying on any single source. Unlike prior methods that model only semantic or factual uncertainty in isolation, MS-UQ combines these heterogeneous signals into a unified representation. The consistent gains across models and datasets imply that this integration captures complementary aspects of uncertainty that single-source methods overlook.

• The most significant gains appear on HotpotQA, indicating that multi-source uncertainty modeling is especially beneficial in reasoning-intensive tasks. HotpotQA requires multi-hop reasoning over multiple documents, where LLM responses can diverge both semantically and factually in complex ways. The results suggest that when model outputs reflect mixed or subtle forms of disagreement on a complex dataset, single-source methods are insufficient. In contrast, MS-UQ disentangles these signals more effectively.

5.4 Ensemble and Ablation Study (RQ2)

In this section, we conduct experiments to examine the importance of multi-source integration and the role of tensor decomposition. Specifically, we compare MS-UQ against (1) variants using only two uncertainty sources, (2) naive averaging of uncertainty scores (U-Mean), (3) SVD applied to a 2D concatenation of all similarity matrices, and (4) different ensemble strategies. Results on HotpotQA with Llama2-13B and Phi4 are shown in Table 2.

The results show that MS-UQ consistently outperforms all ablated variants. The performance gap between MS-UQ and SVD confirms that simple 2D compression fails to address redundancy among sources. Moreover, integrating all three signals achieves better performance than using only two, validating the benefit of combining complementary information. Finally, sum provides slightly more stable results than min strategy and is used by default in our main experiments.

5.5 Sensitivity Analysis (RQ3)

Here, we explore the influence of using different similarity metrics. More Sensitive Analysis about accuracy thresholds can be found at section D.3, and different claim extraction models can be found at section D.2. we present the results that use Jaccard similarity instead of using an NLI model in fig. 4. The results show that using Jaccard similarity will boost the performance for a simple dataset like CoQA but hurt the performance for a difficult dataset like NQ_Open. This is because the answer to a simple question might not have a deeper semantic meaning that requires NLI models. However, MS-UQ can still outperform baseline methods that also use Jaccard similarity.

Table 2: Comparison of different ensemble methods and sources on HotpotQA. The results show that simply using the mean of uncertainty or using SVD cannot fully utilize the information from uncertainty sources as MS-UQ does.

Methods	Llama2-13B		Phi4	
	AUROC	AUARC	AUROC	AUARC
U-Mean (Graph+Claim)	0.6068	0.7715	0.5409	0.8597
SVD (Graph+Claim)	0.6091	0.7732	0.5540	0.8623
MS-UQ (Graph+Claim)	**0.6387**	**0.7855**	**0.5804**	**0.8761**
U-Mean (Claim+Judge)	0.6115	0.7802	0.5588	0.8692
SVD (Claim+Judge)	0.5961	0.7548	0.5256	0.8481
MS-UQ (Claim+Judge)	**0.6312**	**0.7794**	**0.5741**	**0.8724**
U-Mean (All)	0.6089	0.7746	0.5472	0.8631
SVD (All)	0.6016	0.7653	0.5331	0.8532
MS-UQ-Sum (All)	**0.6461**	**0.7940**	**0.5874**	**0.8816**
MS-UQ-Min (All)	0.6413	0.7919	0.5838	0.8785

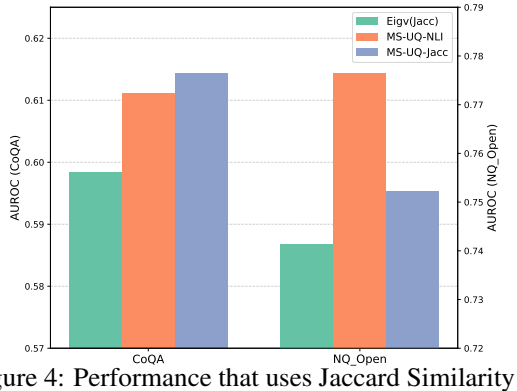

Figure 4: Performance that uses Jaccard Similarity on CoQA and NQ_Open with llama3.1-8b.

6 Conclusion

In conclusion, this study introduces a novel multi-source uncertainty quantification framework, MS-UQ, for large language models, addressing the limitations of conventional uncertainty estimation approaches. By leveraging both semantic similarity and knowledge coherence dimensions, our method disentangles and integrates complementary information to achieve a more robust uncertainty representation. Through the application of tensor decomposition techniques, MS-UQ effectively reduces redundant information and enhances the reliability of uncertainty assessments. Experimental results across multiple datasets and models demonstrate the superiority of our framework in distinguishing uncertain responses, particularly in complex and high-stakes environments.

Limitations

The approach relies on auxiliary large language models for knowledge extraction, which will incur additional costs compared with baselines. Secondly, if the information from the original uncertainty source is completely wrong, it might influence the performance of MS-UQ.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Ameya Agaskar and Yue M Lu. 2013. A spectral graph uncertainty principle. *IEEE Transactions on Information Theory*, 59(7):4338–4356.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ferhat Ozgur Catak and Murat Kuzlu. 2024. Uncertainty quantification in large language models through convex hull analysis. *Discover Artificial Intelligence*, 4(1):90.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200.

Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. 2024. Llm uncertainty quantification through directional entailment graph and claim level response augmentation. *arXiv preprint arXiv:2407.00994*.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. *arXiv preprint arXiv:2109.06352*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *International Conference on Machine Learning*, pages 19023–19042. PMLR.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.

Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. 2017. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 375–384.

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

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dong Wang, Zhike Peng, and Lifeng Xi. 2020. Theoretical and experimental investigations on spectral lp/lq norm ratio and spectral gini index for rotating machine health monitoring. *IEEE Transactions on Automation Science and Engineering*, 18(3):1074–1086.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Kai Ye, Tiejin Chen, Hua Wei, and Liang Zhan. 2024. Uncertainty regularized evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16460–16468.

A Detailed Introduction to Implementation of Each Uncertainty Source

A.1 Claim-level Augmentation Similarity

The claim-level augmentation operates through a structured pipeline that transforms raw responses into factual representations (Da et al., 2024). Given a question Q and its original response A^i, a claim-level representation K^i could be generated through a knowledge mapping process by extracting explicit claims: $K^i = \mathcal{M}_{\text{aux}}(Q, A^i)$ and augmenting the response, where $\mathcal{M}_{\text{aux}}$ denotes for an auxiliary LLM. Specifically, we use an LLM to augment with prompts taking into the question and original response:

Prompt Example for Knowledge Mapping

182: Extract all factual claims from this response $\langle A^i \rangle$, phrased as standalone statements independent of specific wording.
183: Include only information directly relevant to answering the question: $\langle Q \rangle$.

This claim extraction disentangles implicit knowledge from surface semantics and removes stylistic variations while preserving core factual content. Then we could get the similarity matrix as well. To better understand this resource, we compare the differences between the similarity matrices from the original responses and the augmented responses. In detail, we use the NLI model to obtain the similarity matrix.

A.2 Graph Laplacian

By considering each A_i as a node, the similarity between each answer A_i as edges, we could construct a fully connected graph for the question Q. Given the graph, we could use spectral graph principles (Agaskar and Lu, 2013; Lin et al., 2023) to get one resource of uncertainty. In detail, we first build the graph Laplacian that contains more structural information:

$$L = I - D^{-1/2} W D^{-1/2} \tag{9}$$

In graph theory, the eigenvalues of the normalized Laplacian matrix capture key structural properties of a graph, such as connectivity and clustering. Therefore, we could use eigenvalues as one uncertainty resource (Lin et al., 2023):

Dataset	SNR
CoQA	0.8917
Hotpot_QA	0.9127
NQ_Open	0.8379

Table 3: SNR Results for all the datasets on llama2-7b. The results show that the overlapped information is common for all datasets.

$$U_{\text{EigV}} = \sum_{k=1}^{n} \max(0, 1 - \lambda_k). \tag{10}$$

Here, eigenvalues λ_k encode connectivity. Fragmented graphs (low consistency in responses and thus high uncertainty) have smaller eigenvalues. This method is able to capture possible overlapping and continuous semantic relationships, considering the overall structure of the graph.

In the other hand, the degree matrix of the graph encodes the local connectivity of each node structurally. Unlike the eigenvalues that focus on the overall structure, the degree captures more information from individual samples. In detail, we calculate the degree as:

$$d_i = \sum_j S_{ij} \tag{11}$$

Where S_{ij} represents the $(i, j) - th$ value of similarity matrix.

B SNR Results for More Datasets

In table 3, we show the SNR results for three different datasets on llama2-7b. The results show that all datasets contain overlapping information.

C Completed Proof on Theorem 1

Theorem 1 (Monotonic best-rank-R error with equal-norm orthogonal blocks)**.** *For $c \in \mathbb{N}$ let*

$$\mathcal{S}^{(c)} = \sum_{k=1}^{c} \mathcal{U}_k, \qquad \mathcal{U}_k := \lambda_k \, a_k^{(1)} \circ \cdots \circ a_k^{(N)},$$

where the rank-1 blocks satisfy

1. ***Orthogonality:*** $\langle \mathcal{U}_p, \mathcal{U}_q \rangle_F = 0$ *for* $p \neq q$;
2. ***Equal norm:*** $\|\mathcal{U}_k\|_F = \delta > 0$ *for* every k.

Let $R < c$ and denote by

$$\hat{\mathcal{S}}_R^{(c)} = \arg \min_{\substack{Y \\ \text{CP-rank}(Y) \leq R}} \left\| \mathcal{S}^{(c)} - Y \right\|_F$$

the best rank-R CP approximation. Then

$$\left\|\mathcal{S}^{(c)} - \hat{\mathcal{S}}_R^{(c)}\right\|_F \;=\; \sqrt{c-R}\,\delta.$$

Consequently, for any $c_1 > c_2$ and the same $R < c_1$,

$$\left\|\mathcal{S}^{(c_1)} - \hat{\mathcal{S}}_R^{(c_1)}\right\|_F \;>\; \left\|\mathcal{S}^{(c_2)} - \hat{\mathcal{S}}_R^{(c_2)}\right\|_F.$$

Proof. Because the $\mathcal{U}_k$'s are pairwise orthogonal rank-1 tensors, any Y with $\mathrm{CP\text{-}rank}(Y) \le R$ admits an orthogonal expansion

$$Y \;=\; \sum_{k=1}^{c} \alpha_k \mathcal{U}_k + \mathcal{W}, \quad \mathcal{W} \;\perp\; \mathcal{U}_k$$

(The R non-zero coefficients come from expanding each rank-1 component of Y in the orthogonal basis $\{\mathcal{U}_1, \ldots, \mathcal{U}_c\}$.)

Now, let us see the error decomposition. The assumption of orthogonality gives

$$\left\|\mathcal{S}^{(c)} - Y\right\|_F^2 \;=\; \sum_{k=1}^{c} \left\|(1-\alpha_k)\,\mathcal{U}_k\right\|_F^2 \;+\; \|\mathcal{W}\|_F^2.$$

Setting $\mathcal{W} = 0$ clearly reduces the error, so the best approximation lives in $\mathrm{span}\{\mathcal{U}_1, \ldots, \mathcal{U}_c\}$.

Now, let us see how to get the optimal choice of coefficients. For any selected index k we can minimise the summand $\|(1-\alpha_k)\mathcal{U}_k\|_F^2$ by taking $\alpha_k = 1$; for any unselected index we set $\alpha_k = 0$. Because every block has the *same* norm δ, *which* R indices we choose is irrelevant:

$$\left\|\mathcal{S}^{(c)} - \hat{\mathcal{S}}_R^{(c)}\right\|_F^2 = \sum_{k=R+1}^{c} \|\mathcal{U}_k\|_F^2 = (c-R)\,\delta^2.$$

Taking square roots yields

$$\left\|\mathcal{S}^{(c)} - \hat{\mathcal{S}}_R^{(c)}\right\|_F \;=\; \sqrt{c-R}\,\delta. \qquad \text{(a)}$$

, which is the optimal reconstruction loss for any c.

Therefore, if $c_1 > c_2$ and both exceed R, then

$$\sqrt{c_1 - R}\,\delta \;>\; \sqrt{c_2 - R}\,\delta.$$

Using (a) for each tensor gives the strict inequality claimed in the theorem. □

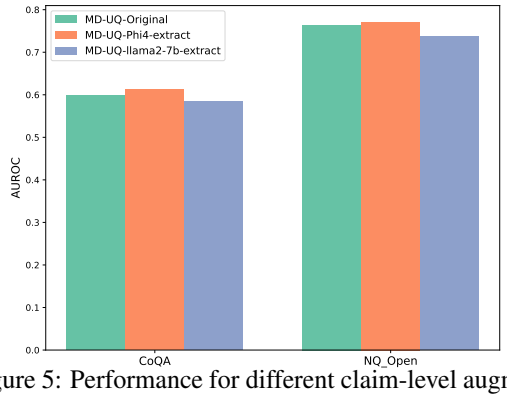

Figure 5: Performance for different claim-level augmentation models on CoQA and NQ_Open with llama3.1.

D More Experimental Results

D.1 Results Compared with Full Uncertainty Sources

In table 4, we compared MS-UQ with all uncertainty sources instead of the sources we are using. The results still show that MS-UQ could exceed all single uncertainty sources.

D.2 Different Models for claim-level augmentation

claim-level augmentation models influence the claim extraction in MS-UQ as stated in section A.1. Therefore, in this section, we test the robustness of MS-UQ on various claim-level augmentation models. We conduct experiments on CoQA and NQ_open using llama2-7b and llama3.1 as the knowledge extracted models. We show the results in fig. 5. From the figure, we can see that MS-UQ is stable to different claim-level agumentataion models.

D.3 Different Accuracy Thresholds

Secondly, we show the influence of different accuracy thresholds. In the previous experiments, we all set the accuracy threshold to 70 as mentioned in section 5.1. To better understand the influence, we choose an extra dataset TriviaQA (Joshi et al., 2017), which is considered the easiest dataset, and NQ_Open, which is the most challenging dataset in our paper, to conduct experiments. We show the results with accuracy thresholds of 70 and 90 in table 5. From the results, we can see that increasing the accuracy threshold decreases the performance of all baselines, while the performance of MS-UQ remains stable or even improves for datasets with higher difficulties, showing the robustness of MS-UQ.

Table 4: Comparison of our methods with different baselines on various datasets and large language models. The best result is shown in the **bold**. The results show that MS-UQ performs better than baselines in general and MS-UQ has a better advantage on more difficult datasets such as NQ_Open.

Methods	Llama2-13B		Llama2-7B		Llama3.1-8B		Phi4		Deepseek-R1-7B	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Dataset: CoQA [Easy]										
Eigv(Dis)	0.7294	0.7775	0.5965	0.9485	0.5762	0.9071	0.6656	0.9120	0.7841	0.9175
Ecc(Dis)	0.6984	0.7553	0.5762	0.9409	0.5802	0.9206	0.6487	0.9066	0.7756	0.9157
Degree(Dis)	0.7369	0.7815	0.5963	0.9473	0.5728	0.9112	0.6677	0.9121	0.7885	0.9189
Eigv(Agre)	0.7541	0.7876	0.5971	0.9507	0.5791	0.9153	0.6399	0.9051	0.7969	0.9234
Ecc(Agre)	0.7593	0.7840	0.5961	0.9480	0.5785	0.9144	0.6335	0.9020	0.7937	0.9224
Degree(Agre)	0.7548	0.7877	0.5908	0.9413	0.5755	0.9097	0.6278	0.8996	0.7930	0.9222
D-UE	0.7566	0.7885	0.5954	0.9481	0.5825	0.9284	0.6503	0.9079	0.7966	0.9228
P(true)	0.7102	0.7088	0.5404	0.9348	0.5816	0.9323	0.6630	0.9087	0.5389	0.8311
MS-UQ-Sum	**0.7802**	**0.7911**	**0.6145**	**0.9521**	**0.6111**	**0.9326**	**0.6851**	**0.9144**	**0.8115**	**0.9332**
Dataset: HotpotQA [Medium]										
Eigv(Dis)	0.6269	0.7770	0.6111	0.7715	0.6099	0.6874	0.5534	0.8614	0.5969	0.5737
Ecc(Dis)	0.6103	0.7774	0.6085	0.7752	0.6044	0.6827	0.5675	0.8691	0.5602	0.5377
Degree(Dis)	0.6336	0.7790	0.6134	0.7714	0.6202	0.7087	0.5666	0.8590	0.5977	0.5756
Eigv(Agre)	0.6235	0.7638	0.6035	0.7648	0.6176	0.7035	0.5328	0.8497	0.6249	0.5878
Ecc(Agre)	0.6233	0.7670	0.6049	0.7666	0.6084	0.6991	0.5469	0.8594	0.6321	0.5907
Degree(Agre)	0.6217	0.7611	0.5973	0.7600	0.6105	0.7016	0.5294	0.8491	0.6278	0.5902
D-UE	0.6252	0.7659	0.6056	0.7669	0.6212	0.7083	0.5335	0.8511	0.6270	0.5893
P(true)	0.6056	0.7591	0.5901	0.7713	0.6362	0.7077	0.5326	0.8513	0.5081	0.4795
MS-UQ-Sum	**0.6461**	**0.7940**	**0.6242**	**0.7809**	**0.6588**	**0.7291**	**0.5874**	**0.8816**	**0.6413**	0.5956
Dataset: NQ_Open [Hard]										
Eigv(Dis)	0.6162	0.7300	0.7280	0.6367	0.6742	0.5343	0.7035	0.6035	0.6696	0.2451
Ecc(Dis)	0.6210	0.7330	0.7167	0.6172	0.6562	0.5007	0.6898	0.5828	0.6607	0.2237
Degree(Dis)	0.6130	0.7168	0.7273	0.6318	0.6865	0.5430	0.7090	0.6094	0.6675	0.2463
Eigv(Agre)	0.6258	0.7276	0.7240	0.6327	0.7463	0.5801	0.7515	0.6351	0.7291	0.2758
Ecc(Agre)	0.6273	0.7311	0.7307	0.6298	0.7612	0.5875	0.7555	0.6370	0.7437	0.2836
Degree(Agre)	0.6286	0.7355	0.7290	0.6324	0.7619	0.5885	0.7542	0.6341	0.7353	0.2781
D-UE	0.6281	0.7320	0.7258	0.6342	0.7551	0.5833	0.7539	0.6366	0.7333	0.2801
P(true)	0.6197	0.7289	0.6532	0.5929	0.7061	0.5522	0.7096	0.6049	0.4865	0.1413
MS-UQ-Sum	**0.6435**	**0.7440**	**0.7432**	**0.6455**	**0.7765**	**0.5946**	**0.7596**	**0.6494**	**0.7653**	**0.2921**

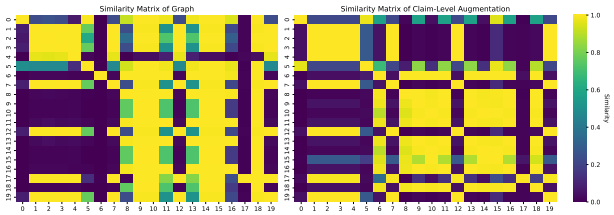

Figure 6: Case1: low accuracy, high overall similarity matrix from Graph-based and a low overall similarity from claim-level augmentation. For the similarity matrix, a color close to yellow indicates a value close to one. A similarity matrix with higher values will lead to a lower U_{EigV}. For the first example, the claim-level augmentation identifies spurious similarities present in the original semantic entropy, which leads to a higher and better uncertainty.

E Case Study

We present two representative cases to compare our method, MS-UQ, with traditional uncertainty quantification methods and highlight the benefits of combining multiple uncertainty sources by MS-UQ. Both cases are selected from the HotpotQA (Yang et al., 2018) dataset using Llama2-13B as the answer generator.

Case 1: *Question: "Is Dappy or Tobias Sammet German?"*

Ground Truth: "Tobias Sammet is German."

This example provides an article that introduces two different persons. The answers generated by Llama2-13B frequently begin with "No", which leads to misleading phrasing. For example, two answers from 20 generations are:

- "No, both artists are from the UK."

Table 5: Comparison of different methods across different accuracy thresholds on TrivialQA and NQ_Open with llama2-13B. The results show that our methods outperform baselines after increasing the accuracy threshold, indicating that our methods have an advantage on more difficult datasets.

Methods	Accuracy Threshold: 0.7		Accuracy Threshold: 0.9	
	AUROC	AUARC	AUROC	AUARC
Dataset: TriviaQA [Easy]				
Eigv(Dis)	0.8261	0.8094	0.8100	0.7604
Degree(Dis)	0.8399	0.8163	0.8259	0.7694
Eigv(Agre)	0.8436	0.8116	0.8351	0.7721
Degree(Agre)	0.8396	**0.8397**	0.8384	0.7739
MS-UQ-Sum	**0.8428**	0.8144	0.8438	0.7749
Dataset: NQ_Open [Hard]				
Eigv(Dis)	0.6162	0.7300	0.5636	0.6017
Degree(Dis)	0.6130	0.7168	0.5662	0.6033
Eigv(Agre)	0.6258	0.7276	0.6146	0.6290
Degree(Agre)	0.6286	0.7355	0.6221	0.6299
MS-UQ-Sum	**0.6435**	**0.7440**	**0.6452**	**0.7447**

- "No, Dappy is English and Tobias Sammet is German."

Although the fact "Tobias Sammet is German" is correct, it contradicts the leading "No" for this yes/no question. As a result, this example has a low sample accuracy of 0.2, meaning only 4 out of 20 answers are correct.

Using the NLI-based semantic similarity, the answers appear highly similar due to the shared misleading prefix 'No', resulting in a similarity matrix with many values close to one, as shown in fig. 6. This causes a lower uncertainty ($U_{EigV} = 0.1577$) for using semantic similarity, which underestimates the true model uncertainty.

In contrast, claim-level augmentation alleviates the influence of such misleading prefixes by cleaning the misleading prefixes and reducing the similarity between contradictory responses. This leads to a higher uncertainty of 0.2198.

Our method, MS-UQ, integrates both sources and captures the underlying disagreement more effectively, resulting in a much higher uncertainty of 0.5108, which is closer to the ideal value.

Case2: *Question: "When the two Coldplay songs U.F.O and Princess of China were written, what factor did they have in common?"*

Ground Truth: "Was written by all four members of the band."

In this example, generated answers are diverse, such as:

- "The truth is that when writing these two songs, it can be said that they were written in the same structure."

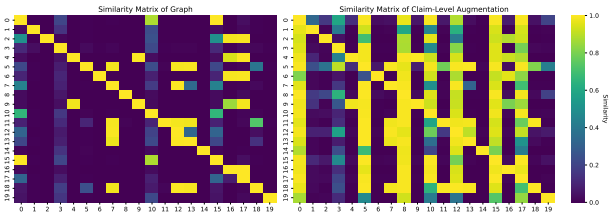

Figure 7: Case 2: low accuracy, low overall similarity from Graph-based and a high overall similarity from claim-level augmentation. For the second example, the uncertainty from semantic similarity is more accurate. For both examples, one uncertainty source works badly, while MS-UQ works better than any single uncertainty source by using the information from all sources dynamically.

- "The first song of the album."

This diversity results in a low accuracy of 0.2. Semantic similarity reflects this variability and gives an uncertainty of 0.6033.

However, claim-level augmentation introduces repetitive and generic statements, e.g.,

- "Coldplay's U.F.O. and Princess of China were written in the same structure... similar pattern."

- "Both U.F.O. and Princess of China were written by Coldplay."

These augmented answers become more similar, which causes the claim-level similarity matrix to contain larger values and reduces the resulting uncertainty to 0.1231, which underestimates the true variability.

MS-UQ overcomes this limitation by dynamically integrating all sources. It achieves an uncertainty of 0.6202, which better reflects the actual answer variability.

In our case analysis, both examples have a low accuracy and using one source only will lead to a suboptimal uncertainty, indicating the necessity of multiple sources of uncertainty. Besides, if we are using a simple way to utilize information from all uncertainty sources, the uncertainty results are still suboptimal, with an uncertainty of 0.1887 for the first example, while an uncertainty of 0.3632 for the second example. However, MS-UQ utilizes all the sources dynamically, deals with the overlapped information and obtains a much better uncertainty, showing the effectiveness.