GENUINE: Graph Enhanced Multi-level Uncertainty Estimation for Large Language Models

Anonymous ACL submission

Abstract

Uncertainty estimation is essential for enhancing the reliability of Large Language Models (LLMs), particularly in high-stakes applications. Existing methods often overlook semantic dependencies, relying on token-level probability measures that fail to capture structural relationships within the generated text. We propose GENUINE: **G**raph **EN**hanced m**U**lti-level uncerta**IN**ty **E**stimation for Large Language Models, a structure-aware framework that leverages dependency parse trees and hierarchical graph pooling to refine uncertainty quantification. By incorporating supervised learning, GENUINE effectively models semantic and structural relationships, improving confidence assessments. Extensive experiments across NLP tasks show that GENUINE achieves up to 29% higher AUROC than semantic entropy-based approaches and reduces calibration errors by over 15%, demonstrating the effectiveness of graph-based uncertainty modeling. The code is available at `https://anonymous.4open.science/r/GUQ-39E7`.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in conversation (Wu et al., 2023), logical reasoning (Wang et al., 2023), and scientific discovery (Shojaee et al., 2024). Models such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), and DeepSeek (Liu et al., 2024a), trained on vast corpora and aligned to human preferences, have significantly expanded the potential of artificial intelligence. However, despite these advancements, LLMs are prone to well-documented reliability issues, including hallucinations and factual inaccuracies (Huang et al., 2025; Liu et al., 2024c). These issues pose serious risks, particularly in high-stakes applications such as medical diagnosis (Panagoulias et al., 2024), financial decision-making (de Zarzà et al., 2023), and legal advisory systems (Cheong et al., 2024), where users must rely on the model's outputs with confidence. Therefore, uncertainty quantification, which assesses the trustworthiness of an LLM response, is essential for safe and effective human and artificial intelligence interaction.

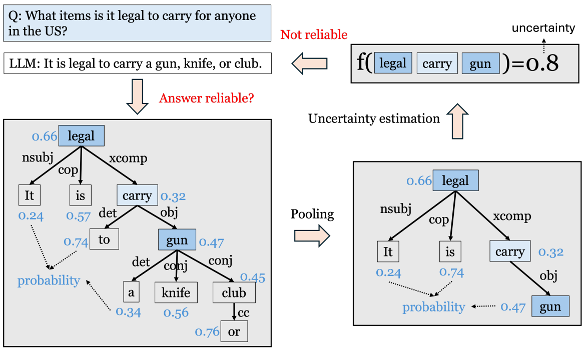


Figure 1: An example highlighting the role of token significance in uncertainty estimation. The model misinterprets "legal", generating an incorrect response. The dependency parse tree helps identify critical tokens, improving uncertainty quantification by reducing the influence of less relevant tokens.

Quantifying uncertainty in LLM-generated outputs presents several challenges. First, LLMs often produce long-form textual responses, making attributing uncertainty to specific components difficult. Second, hallucinations and uncertainties may affect only a few critical tokens within an otherwise coherent response, undermining the reliability of the entire output. Third, identifying and aggregating uncertainty across multiple tokens in lengthy outputs is non-trivial, requiring distinguishing semantically pivotal tokens from those not pivotal.

Previous studies have explored various approaches to quantify uncertainty in LLM outputs. Some methods rely on self-evaluation through modified prompts (Tian et al., 2023b), though they often inherit the model's biases. Others use token-level uncertainty measures based on logits, entropy, or probability distributions (Kuhn et al., 2023; Malinin and Gales, 2020, 2021). Recent advancements, such as semantic entropy, cluster semantically equivalent generations and measure entropy as an uncertainty indicator (Kuhn et al., 2023). However, most existing methods treat all tokens equally, overlooking findings that certain tokens carry more semantic weight in determining output validity (Liu et al., 2024b; Duan et al., 2024; Cheng and Vlachos, 2024). Additionally, some approaches (Duan et al., 2024) depend on external smaller models to estimate token importance, but these models often operate independently of the LLM's internal representations. As a result, they may introduce inconsistencies, misinterpret token dependencies, or fail to capture the structural relationships within the generated text, leading to inaccurate uncertainty estimates.

To illustrate this issue, consider the example in Fig. 1. A user inquires about legal items to carry in the United States, but the model responds with a list of illegal items, such as a gun, knife, and club. The misunderstanding stems from the token "legal," which is central to the query's meaning. A minor modification, replacing the word legal with illegal, would render the response appropriate. This example underscores two insights: certain tokens are disproportionately influential in determining output validity. Dependency parse trees effectively capture the hierarchical structure of sentence meaning by identifying core decision points. Building on these insights, we propose leveraging dependency parse trees and graph pooling techniques to infer LLM prediction uncertainty in a structured and interpretable manner.

Modeling uncertainty estimation as a graph-based problem offers several advantages. Graphs inherently capture dependencies between generated tokens, reflecting the autoregressive nature of LLMs, where each token influences subsequent ones. By representing an LLM response as a structured graph, we can propagate and aggregate critical information across tokens, ensuring that semantically significant tokens contribute more substantially to the overall uncertainty estimate. However, this approach introduces several challenges. Determining the optimal graph structure that accurately represents token dependencies remains an open question. Selecting appropriate graph pooling techniques that summarize uncertainty information effectively without losing essential context is difficult. Addressing these challenges is essential to realize the potential of graph-based uncertainty estimation fully.

Our approach integrates multiple uncertainty features to enhance robustness. Specifically, we utilize probability distributions, entropy-based measures, and LLM embeddings to model uncertainty. We introduce a hierarchical strategy to address the challenge of aggregating uncertainty over long-form text. We construct a dependency parse tree for each sentence to extract structural and semantic relationships. We merge sentence-level trees into a document-level graph by connecting their root nodes. We apply graph pooling techniques to model uncertainty across the entire paragraph efficiently. GENUINE involves learning pooling functions that adaptively fuse different features, capturing both local and global dependencies within the text. Experimental results demonstrate that this approach outperforms existing methods, highlighting the critical role of structural relationships in uncertainty estimation. Furthermore, we compare the effectiveness of probability-based and embedding-based features across various datasets and LLMs, offering insights into their respective utilities. Given that commercial LLMs typically provide only probability and entropy features, our findings suggest an intriguing direction for future research. Exploring whether open-source LLMs, which offer both probability and embedding features, can facilitate superior uncertainty quantification compared to their commercial counterparts.

The following are our main contributions:

- We highlight the role of semantically significant tokens in uncertainty estimation, demonstrating how structural relationships can enhance model uncertainty assessment.
- We propose a graph-based framework for LLM uncertainty quantification, integrating dependency parse trees and graph pooling to capture structural and semantic relationships in the generated text.
- We develop an adaptive graph pooling mechanism that effectively propagates and aggregates uncertainty information by learning to fuse multiple uncertainty features.
- We conduct extensive experiments on real-world datasets, evaluating different uncertainty features and demonstrating that GENUINE outperforms existing uncertainty quantification methods in assessing the trustworthiness of LLM-generated responses.

2 Related Works

This section reviews prior approaches in uncertainty quantification and graph pooling techniques, highlighting their limitations and the need for a structured, context-aware framework like GENUINE.

Uncertainty Quantification in LLMs. Uncertainty quantification has been extensively studied in traditional machine learning (Chen et al., 2019; Zhao et al., 2020), but it remains a developing challenge for LLMs. Unlike conventional models with well-defined solution spaces, LLMs generate open-ended responses where multiple outputs may be valid as long as they align with the semantic meaning of the input. This flexibility complicates uncertainty estimation, requiring approaches beyond standard predictive confidence measures. Existing methods for uncertainty quantification in LLMs can be categorized into two primary approaches. The first involves self-assessment, where the model is prompted to estimate its own uncertainty (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023a). While intuitive, these methods often inherit the model's biases and inconsistencies. The second category relies on external uncertainty measures, such as analyzing the consistency of multiple generations (Manakul et al., 2023) or computing entropy over predictive distributions (Malinin and Gales, 2020). However, these techniques typically treat all tokens as equally important, overlooking the fact that certain tokens contribute more to the overall reliability of a response. Recent work addresses this limitation by incorporating semantic-aware uncertainty estimation. Semantic entropy (SE) (Kuhn et al., 2023) measures uncertainty at the semantic level, grouping semantically equivalent outputs to reduce redundancy. Other approaches argue that not all tokens are equally important and propose methods to weight different tokens accordingly (Duan et al., 2024). Another direction of studies shows that hidden layer activations offer valuable uncertainty signals by capturing internal model representations (Liu et al., 2024b). Building on this, we integrate dependency parse trees to identify key tokens shaping response meaning, while hidden activations provide semantic context. This combination enables a structured and context-aware approach to uncertainty estimation in LLMs.

Graph Pooling Approaches. The graph pooling mechanism is essential in condensing the input graph into a smaller-sized graph while preserving essential structural and semantic information. Graph pooling methods generally fall into two categories: flat pooling, which applies simple aggregation functions like averaging or summation (Xu et al., 2019; Duvenaud et al., 2015), and hierarchical pooling, which progressively coarsens the graph to capture multi-level relationships (Ying et al., 2018). Among hierarchical methods, DiffPool (Ying et al., 2018) uses Graph Neural Networks (GNNs) to learn adaptive pooling assignments, while StructPool (Yuan and Ji, 2020) extends this by incorporating high-order structural dependencies. Additional strategies include memory-based pooling (Khasahmadi et al., 2020), spectral filtering (Defferrard et al., 2016), and expressive pooling architectures (Bianchi and Lachi, 2023). Unsupervised pooling techniques, such as mutual information maximization (Liu et al., 2022), further enable structure-preserving compression without requiring labeled data. This work proposes a hierarchical graph pooling strategy that leverages dependency tree structures to refine uncertainty estimation. By representing LLM outputs as dependency graphs, GENUINE captures both semantic and structural relationships, enabling a more context-aware evaluation of uncertainty. The proposed framework effectively prioritizes key tokens influencing response reliability, leading to a more precise and interpretable confidence assessment.

3 Background

This section defines the problem statement and provides the necessary background on dependency parsing trees and features helpful for uncertainty estimation in LLMs, laying the foundation for our proposed approach.

3.1 Problem Setup

Uncertainty quantification in LLMs involves assessing confidence in LLM-generated responses based on input prompts. Given a prompt $\mathbf{x} = \{x_1, x_2, ..., x_k\}$, an LLM generates an output sequence $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, where each token y_j is sampled from a probability distribution conditioned on the prompt and prior tokens:

$$y_j \sim p_\theta(\cdot|\mathbf{x}, y_1, y_2, ..., y_{j-1}), \tag{1}$$

where p_θ represents the model's learned parameters. This next-token probability reflects how likely the model is to generate a particular token given the preceding context. Following (Liu et al., 2024b), uncertainty estimation is framed as a function $g(\mathbf{x}, \mathbf{y})$ that predicts the expected correctness of a response:

$$g(\mathbf{x}, \mathbf{y}) \approx \mathbb{E}\left[s(\mathbf{y}, \mathbf{y}_{\text{true}})|\mathbf{x}, \mathbf{y}\right]. \tag{2}$$

Here, $s(\mathbf{y}, \mathbf{y}_{\text{true}})$ denotes an evaluation metric comparing the generated response $\mathbf{y}$ with a ground-truth reference $\mathbf{y}_{\text{true}}$. The expectation is taken considering the semantic flexibility of natural language. The uncertainty arises from the input prompt $\mathbf{x}$ and the LLM itself rather than from a single absolute reference answer.

3.2 Dependency Parse Trees in NLP

Dependency parse trees provide a structured representation of syntactic relationships, defining hierarchical dependencies such as *subjects*, *objects*, and *modifiers* within a sentence. These structures have been widely applied in various NLP tasks, including relation extraction (RE) (Fundel et al., 2006; Björne et al., 2009), named entity recognition (NER) (Jie et al., 2017), and semantic role labeling (SRL) (Marcheggiani and Titov, 2017). They also enhance summarization by prioritizing salient information while filtering redundant content (Li et al., 2014; Xu and Durrett, 2019).

This work uses dependency parse trees to model structural relationships in LLM-generated text. These trees serve two key purposes: (1) They provide a hierarchical organization of tokens, helping distinguish pivotal words that shape response meaning, (2) They offer a consistent structure across different sentence formations, making them adaptable for modeling uncertainty in diverse LLM outputs.

3.3 Features for Uncertainty Estimation

Uncertainty estimation in LLMs relies on extracting meaningful features from the generated text. Prior studies (Xiao et al., 2022; Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023a; Kuhn et al., 2023; Liu et al., 2024b) have demonstrated the effectiveness of token-level probability metrics, such as entropy, in uncertainty estimation. We categorize these features based on their sources (Liu et al., 2024b):

White-box features: These features are derived from hidden-layer activations, capturing the internal representation of tokens and providing insights into model confidence. These features are available only in open-source LLMs.

Grey-box features: These include *token probabilities* and transformations such as entropy, offering uncertainty signals applicable to both open-source and commercial LLMs. The entropy of a discrete distribution p over the vocabulary $\mathcal{V}$ is defined as $H(p) = -\sum_{v \in \mathcal{V}} p(v) \log (p(v))$. Given a prompt-response pair $(\mathbf{x}, \mathbf{y}) = (x_1, ..., x_k, y_1, ..., y_n)$, the entropy features for the j-th output token are given by $H(q_\theta(y_j | \mathbf{x}, y_1, ..., y_{j-1}))$, where q_θ denotes the LLM. The detailed mathematical definition of the features is provided in Appendix A.1.

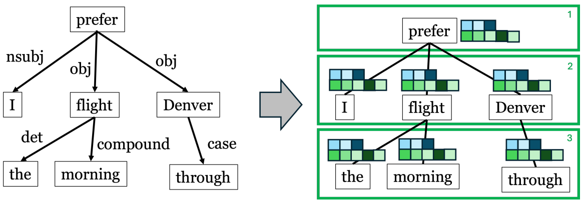


Figure 2: Dependency parse tree example

4 Approach

This section details our approach, including graph formulation, hierarchical learning, and joint optimization, enabling a more structured and context-aware uncertainty estimation for LLMs.

4.1 Graph Formulation

We transform dependency parse trees into graphs to structure LLM-generated text for uncertainty estimation. We first obtain the dependency tree using the Stanford NLTK parser, where each word serves as a node, and directed edges represent dependency relations. As shown in Fig. 2, the root word, such as "prefer," has dependent words like "I" and "flight," forming a tree-like structure.

To extend this formulation beyond individual sentences, we construct a paragraph-level graph by linking the root nodes of multiple sentence-level dependency trees. Prior work (Duan et al., 2024) estimates uncertainty at the sentence level using a separate model to compute similarity, but such approaches may overlook deeper semantic relationships between sentences. Instead, GENUINE learns inter-sentence relations directly, ensuring a more cohesive uncertainty estimation. Connecting root nodes across sentences enables cross-sentence token interactions, allowing uncertainty information to propagate effectively across the entire output. This formulation ensures that pivotal words influence the overall confidence estimation. The resulting global dependency graph provides a structured representation of LLM output, enhancing the ability of the proposed approach to assess uncertainty in LLM-generated text.

4.2 Hierarchical Learning

Transforming dependency parse trees into graphs enables us to frame uncertainty estimation as a graph aggregation problem, where each LLM-generated output is represented as a graph with nodes corresponding to words and edges capturing dependency relations. Each node has token-level features, such as next-token probability, entropy, and hidden state embeddings. We propose a hierarchical graph pooling approach inspired by semantic parsing trees (Song and King, 2022) to aggregate this information efficiently.

In a dependency graph (Fig. 2), words appear at different levels based on their distance from the root token, which often signifies their semantic importance. Higher-level words generally play a more critical role in defining the sentence's meaning and, consequently, have a greater impact on uncertainty. To capture this, we introduce graph pooling, which groups tokens at different hierarchical levels, mitigating the effect of noisy words while assigning appropriate contributions to each token's uncertainty estimate.

Formally, given a dependency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents words and $\mathcal{E}$ defines their syntactic relations, we define an adjacency matrix $\mathcal{A} \in \mathbb{R}^{n \times n}$ and a feature matrix $\mathcal{X} \in \mathbb{R}^{n \times d}$. Inspired by hierarchical graph pooling methods (Ying et al., 2018), we define the node clustering process using a learned soft assignment matrix:

$$\mathcal{S}^l = \text{Softmax}(f(\mathcal{A}^l, \mathcal{X}^l, \theta_s)), \tag{3}$$

where $\mathcal{A}^l$ and $\mathcal{X}^l$ represent the adjacency and feature matrices at pooling layer l, and f in a GNN with learnable parameters θ_s.

Before pooling, information propagates across the graph to model connectivity between clusters:

$$\mathcal{Z}^l = f(\mathcal{A}^l, \mathcal{X}^l, \theta_z), \tag{4}$$

where θ_z are the parameters of the GNN responsible for feature transformation. Using the learned assignment matrix $\mathcal{S}^l$, the graph is iteratively coarsened to generate a more compact representation:

$$\begin{aligned} \mathcal{X}^{l+1} &= \mathcal{S}^l \mathcal{Z}^l \in \mathbb{R}^{n_{l+1} \times d}, \\ \mathcal{A}^{l+1} &= \mathcal{S}^l \mathcal{A}^l \mathcal{S}^{l^T} \in \mathbb{R}^{n_{l+1} \times n_{l+1}}. \end{aligned} \tag{5}$$

Here, $\mathcal{X}^l$ and $\mathcal{A}^l$ are iteratively refined representations at each pooling level, ensuring that semantically important tokens retain greater influence in uncertainty estimation. By hierarchically aggregating token-level uncertainty, GENUINE enhances interpretability and robustness, providing a structured estimation of confidence in LLM-generated responses.

4.3 Joint Optimization

Uncertainty estimation in LLMs relies on multiple feature types as discussed in Section 3.3, including hidden states (white-box features), probability, and entropy (grey-box features), each contributing differently to uncertainty estimation. Prior work (Liu et al., 2024b) highlights that hidden states encode valuable uncertainty information due to the misalignment between model pretraining and uncertainty estimation. Additionally, hidden states capture semantic relationships among tokens, making them crucial for confidence evaluation.

We propose a joint optimization framework to integrate these features effectively. The framework of GENUINE shown in Fig. 3 consists of a semantic pooling module, which leverages hidden state embeddings, and a structural pooling module, which utilizes probability and entropy features. Both modules share the same dependency parse tree, ensuring a unified structural representation. The outputs from these modules are then fed into a fusion module, which learns a joint graph pooling matrix to combine information from both the semantic and structural components. This fused graph pooling matrix is further optimized to balance structural and semantic uncertainty signals, refining the final uncertainty estimation.

Rather than merging features at the node level, we fuse them at the assignment matrix level to ensure a balanced integration of structural and semantic information. There are three key reasons for this choice. First, direct feature fusion would overemphasize embeddings, as their dimensionality is significantly larger than probability and entropy features. Second, embeddings contain rich semantic context but provide limited insights into generation-specific uncertainty, whereas probability and entropy features offer more precise confidence indicators. Third, the assignment matrix inherently captures token importance and relationships, making it a more suitable fusion point for diverse feature types.

To achieve this, we introduce an end-to-end learnable fusion module, where the fused assignment matrix is computed as:

$$\mathcal{S}_*^l = \text{Softmax}(g(\mathcal{S}_{\text{grey}}^l, \mathcal{S}_{\text{white}}^l, \theta_{s*})), \tag{6}$$

where $\mathcal{S}_{\text{grey}}^l$ and $\mathcal{S}_{\text{white}}^l$ are the assignment matrices at pooling layer l from the structural and semantic modules, respectively, and θ_{s*} denotes the learnable parameters of the fusion function g.

Following this, a GNN propagates information across the graph, refining node representations through:

$$\mathcal{Z}_*^l = f(\mathcal{A}_*^l, \mathcal{X}_*^l, \theta_{z*}), \tag{7}$$

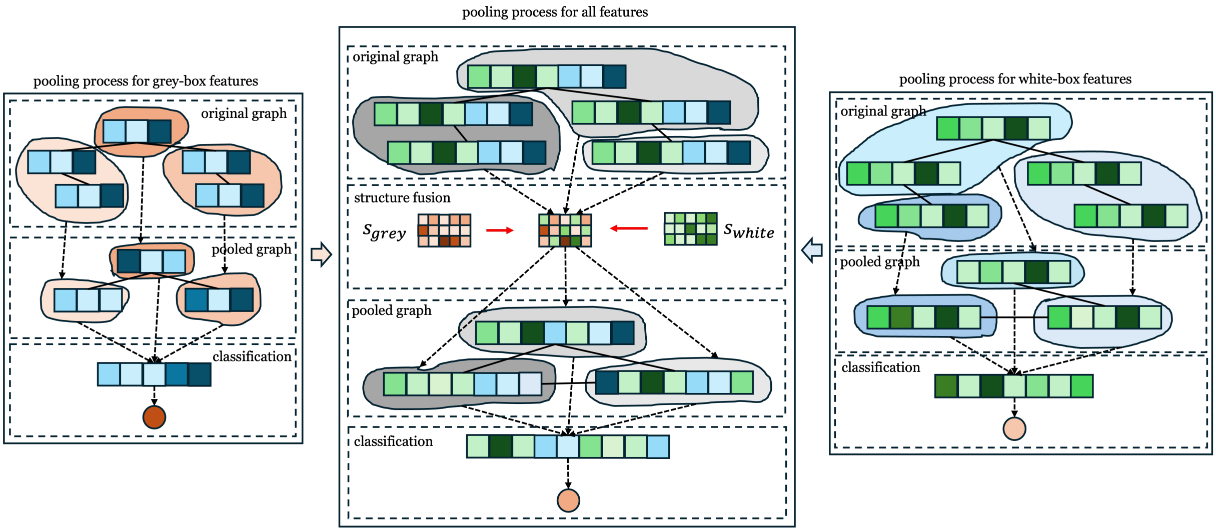


Figure 3: The Overview of GENUINE, composed of three modules: (1) pooling based on grey-box features, (2) pooling based on white-box features, and (3) a learnable fusion process integrating both modules.

where f is a GNN with learnable parameters θ_{z*}. These updated assignment and node embedding matrices are used to refine the graph iteratively:

$$\mathcal{X}_*^{l+1} = \mathcal{S}_*^l \mathcal{Z}_*^l, \\ \mathcal{A}_*^{l+1} = \mathcal{S}_*^l \mathcal{A}_*^l \mathcal{S}_*^{l^T}. \tag{8}$$

Here, $\mathcal{X}_*$ encodes probability and entropy features, while embeddings enhance the model's semantic understanding. The independent assignment matrices $\mathcal{S}_{\text{grey}}^l$ and $\mathcal{S}_{\text{white}}^l$ are jointly optimized to capture both structural and contextual uncertainty, improving the robustness of LLM confidence evaluation.

5 Experiments

This section evaluates GENUINE across multiple dimensions: (1) its effectiveness in assessing uncertainty (Section 5.2), (2) an ablation study to analyze the role of the fused modules (Section 5.3), (3) a scalability test to assess computational efficiency (Section 5.4), (4) the impact of dependency parse trees on uncertainty estimation (Appendix B.2), (5) a parameter analysis to determine the sensitivity of GENUINE to hyperparameter tuning (Appendix B.3), and (6) the impact of LLM parameters on GENUINE's uncertainty estimation performance (Appendix B.4). Due to space constraints, the results for experiments on dimensions 4, 5, and 6 are presented in Appendix B.

5.1 Experimental Setup

We evaluate GENUINE using different LLM architectures, multiple datasets spanning various NLP tasks, and state-of-the-art baselines. All experiments are conducted on a Linux server with 64 AMD EPYC 7313 CPUs and an Nvidia Tesla A100 SXM4 GPU with 80 GB memory.

LLMs. We consider open-source LLMs, including Llama2-7B, Llama2-13B, Llama3-8B (Touvron et al., 2023), as well as Gemma-7B and Gemma2-9B (Gemma Team et al., 2024). The respective tokenizers provided by Hugging Face are used, and model parameters remain unchanged.

Datasets. We evaluate uncertainty estimation on three NLP tasks: question answering, machine translation, and summarization. Each dataset is split into training (60%), validation (10%), and test (30%) sets, with five runs performed to mitigate the effects of randomness in parameter optimization. Few-shot prompting is adopted, with templates detailed in Appendix A.2.

Question Answering. We use the CoQA (Reddy et al., 2019) and TriviaQA (Joshi et al., 2017) datasets to assess LLMs' ability to generate responses based on contextual understanding and pre-trained knowledge. Additionally, we include the Finance QA dataset (Taori et al., 2023), which evaluates domain-specific knowledge in financial contexts. Rouge-1 (Lin and Och, 2004) is used as the scoring function, labeling a response $\mathbf{y}_i$ as correct if $s(\mathbf{y}_i, \mathbf{y}_{i,\text{true}}) \geq 0.3$.

Machine Translation. We evaluate translation quality using the WMT 2014 dataset (Bojar et al., 2014), with BLEU score (Papineni et al., 2002) as the metric. A response $\mathbf{y}_i$ is considered correct if $s(\mathbf{y}_i, \mathbf{y}_{i,\text{true}}) \geq 0.3$.

Summarization. The CNN (Hermann et al., 2015) dataset is used for summarization task, where generated outputs are labeled as correct if they achieve a Rouge-L score of at least 0.35, following (Quach

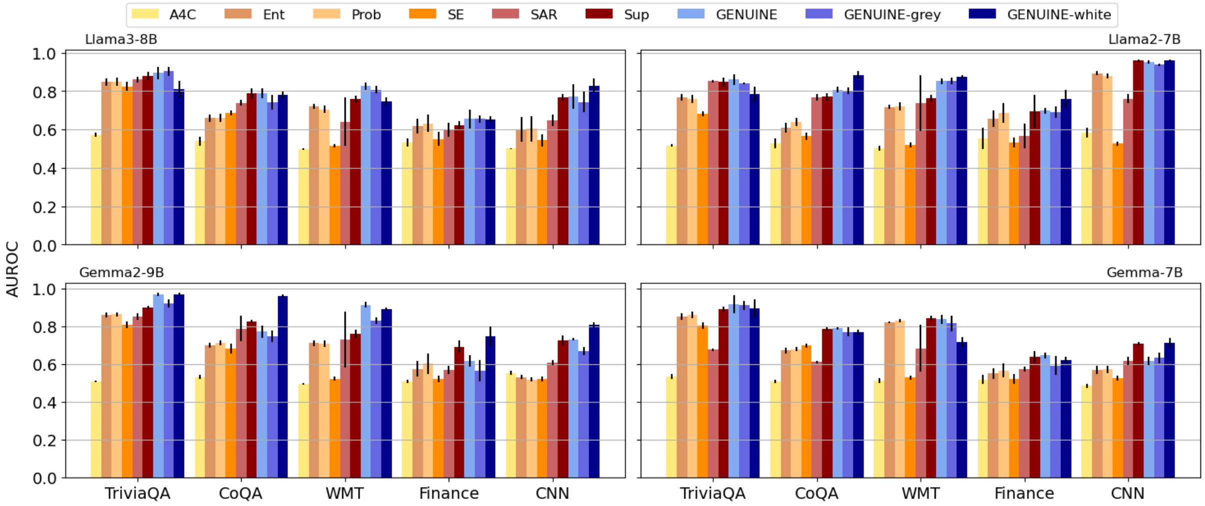


Figure 4: Comparison of AUROC on five datasets, four LLMs, and six baselines. Error bars denote variance over five runs. GENUINE and its transformations outperform baselines for all datasets and LLMs.

et al., 2024).

Baselines. We compare GENUINE against four categories of state-of-the-art baselines: (1) A4C (Tian et al., 2023b), which directly queries the LLM for its self-assessed uncertainty, (2) Entropy and probability-based methods, including Avg Probability (Prob) and Avg Entropy (Ent), as defined in Table 2 in the Appendix A.1, (3) Semantic-aware methods, such as Semantic Entropy (SE) (Kuhn et al., 2023) and SAR (Duan et al., 2024), and (4) A supervised uncertainty estimation approach (Sup) (Liu et al., 2024b). Details of the prompt templates for each dataset are provided in the Appendix A.2.

Evaluation Metrics. We adopt the methodology from (Liu et al., 2024b; Kuhn et al., 2023) to assess how well GENUINE distinguishes between correct and incorrect responses using uncertainty scores. The primary metric is the area under the receiver operating characteristic curve (AUROC), which measures the ability to rank correct responses higher than incorrect ones based on uncertainty. In addition to AUROC, we evaluate calibration performance using Expected Calibration Error (ECE) (Naeini et al., 2015), quantifying the deviation between estimated probabilities and actual correctness. Furthermore, we report the Brier score (Hernández-Orallo et al., 2011) and negative log-likelihood (NLL) (Hastie et al., 2001) to assess how well GENUINE aligns uncertainty estimates with true confidence levels. We present the AUROC performance in the main paper, while additional results for ECE, Brier score, and NLL are provided in Appendix B.1.

5.2 Performance of Uncertainty Estimation

We evaluate GENUINE on the AUROC metric, comparing its performance against state-of-the-art baselines. The results in Fig. 4 demonstrate that GENUINE consistently outperforms existing methods, particularly in tasks involving long-form text generation, such as WMT, Finance, and CNN. Integrating dependency-based structural modeling enhances uncertainty estimation by mitigating error propagation over extended sequences. This is especially valuable for applications requiring contextually accurate long-form responses, such as legal, medical, and financial AI assistants. Additionally, GENUINE exhibits superior calibration performance, as confirmed by ECE, NLL, and Brier score results in Appendix B.1, reducing uncertainty misalignment in downstream tasks.

The results further highlight that response length significantly impacts uncertainty estimation. Table 4 in the Appendix presents the dataset and LLM output lengths. In shorter responses, such as those in TriviaQA and CoQA, GENUINE achieves modest gains over SAR, improving AUROC by 2% for Llama3-8B and 1% for Llama2-7B. However, the advantage becomes more pronounced in longer responses, such as those in WMT, Finance, and CNN, with 29% and 15% improvements in WMT for Llama3-8B and Llama2-7B, respectively. Traditional token-wise uncertainty models struggle with cumulative errors over long sequences, making structured uncertainty estimation essential for trustworthy AI applications, including conversational agents and document summarization.

Feature selection also plays a crucial role in uncertainty estimation. While combining multi-

Table 1: Ablation study of fusion process on TriviaQA

Methods	Llama3-8B	Gemma2-9B
	AUROC ↑	AUROC ↑
Simple concat	0.809±0.096	0.963±0.015
GENUINE	0.894±0.032	0.969±0.009

ple features generally improves performance (TriviaQA, CoQA, WMT), hidden-layer embeddings alone (GENUINE-white) perform best on Finance and CNN datasets. Longer sequences accumulate token-wise uncertainty errors, affecting entropy-based methods, whereas hidden-layer features provide a stable representation, remaining robust regardless of sampling strategy. These findings suggest that open-source LLMs with access to internal representations offer a significant advantage in uncertainty modeling. Future research should explore dynamic feature selection mechanisms, optimizing uncertainty estimation based on sequence length and task-specific requirements.

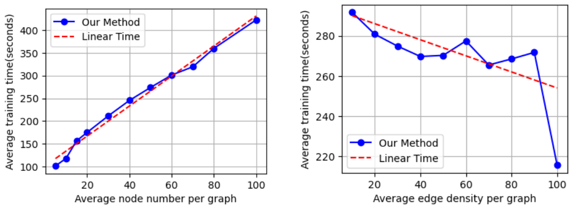
These results underscore the importance of structured uncertainty estimation and suggest that incorporating hierarchical representations, adaptive modeling, and task-aware feature selection can further enhance uncertainty estimation in real-world NLP applications. By leveraging dependency parsing, GENUINE provides a robust framework for improving confidence calibration, making LLMs more reliable for high-stakes decision-making.

5.3 Ablation Study

To assess the impact of the fused assignment matrix, we conduct ablation experiments on the TriviaQA dataset using Llama3-8B and Gemma2-9B. As shown in Table 1, the fusion process (Fig. 3) improves AUROC by 10.5% for Llama3-8B and 0.6% for Gemma2-9B compared to simple concatenation. These results demonstrate that the fusion strategy effectively integrates structural and semantic uncertainty signals, enabling more robust uncertainty propagation across tokens. In contrast, simple concatenation fails to capture meaningful relationships between uncertainty features, leading to suboptimal performance. The consistent improvement across models highlights the importance of structured feature fusion in uncertainty estimation. By jointly optimizing structural and semantic representations, GENUINE enhances both robustness and interpretability, making it well-suited for uncertainty-aware applications.

5.4 Scalability

We evaluate the scalability of GENUINE by analyzing its computational efficiency as the number of nodes and graph density increase. The results in Fig. 5a and Fig. 5b highlight GENUINE's ability to handle increasing structural complexity efficiently.



(a) Scalability test on number of nodes per graph

(b) Scalability test on graph density

Figure 5: Scalability test on the node number and edge density

As shown in Fig. 5a, training time scales near-linearly with the number of nodes, demonstrating that GENUINE remains computationally feasible even for larger graphs. This suggests that the model can efficiently process uncertainty in large-scale LLM outputs without excessive overhead. In Fig. 5b, computational cost decreases as graph density increases, indicating that denser graphs facilitate more efficient uncertainty aggregation. Sparse graphs (e.g., 10% density) require 1.5 times more processing time than fully connected graphs (100% density), emphasizing the trade-off between structure complexity and efficiency.

These findings confirm that GENUINE scales effectively with increasing graph complexity, making it well-suited for high-dimensional NLP tasks such as document summarization, multi-turn dialogue, and knowledge-intensive reasoning. Its ability to maintain efficiency while capturing semantic and structural relationships ensures its adaptability to real-world LLM evaluation scenarios.

6 Conclusion

This paper introduces dependency-based semantic structures for uncertainty estimation in LLMs. Our findings prove that incorporating structural information enhances uncertainty modeling, leading to more accurate and calibrated estimates. GENUINE outperforms existing uncertainty estimation methods (AUROC), particularly in long-form text generation, while also improving calibration metrics (ECE, NLL, Brier). Our results show that semantic graphs derived from dependency parse trees enhance uncertainty modeling, making them valuable for evaluating LLMs' outputs and guiding future improvements in adaptive uncertainty estimation in dynamic, real-world settings.

7 Ethical Consideration

GENUINE enhances the credibility and reliability of LLMs by improving uncertainty estimation, helping to mitigate the risks of misinformation. By refining confidence assessment, GENUINE reduces misinformation and promotes more trustworthy AI-generated content.

However, several ethical limitations must be considered. Uncertainty estimation does not prevent misinformation but provides a measure of confidence, which still requires human interpretation. Over-reliance on uncertainty scores could lead to misjudgments, either overestimating or underestimating the reliability of LLM outputs. Additionally, GENUINE's effectiveness depends on dependency parsing and feature selection, which may introduce biases if trained on imbalanced datasets. Furthermore, while GENUINE improves model calibration, uncertainty quantification remains imperfect, and its reliability may vary across domains, particularly in high-stakes applications such as healthcare, finance, and law. Addressing these challenges requires ongoing evaluation, transparency, and responsible deployment to ensure ethical and fair AI use.

8 Limitations

GENUINE introduces a graph-based approach for confidence evaluation in LLMs, but certain limitations remain. GENUINE relies on token logits and embeddings, which, though widely available in open-source and commercial LLMs, may limit its applicability in black-box scenarios where such information is restricted. Additionally, its performance is influenced by generation length and labeled data availability, making it sensitive to dataset variability. Finally, this study focuses on NLP tasks and datasets, leaving open the exploration of its effectiveness in multimodal and cross-domain applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Filippo Maria Bianchi and Veronica Lachi. 2023. The expressive power of pooling in graph neural networks. *Preprint*, arXiv:2304.01575.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. 2019. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3363–3370.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469.

I de Zarzà, J de Curtò, Gemma Roig, and Carlos T Calafate. 2023. Optimized financial planning: integrating individual and cooperative budgeting models with llm recommendations. *AI*, 5(1):91–114.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. 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)*, pages 5050–5063.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2224–2232, Cambridge, MA, USA. MIT Press.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2006. RelEx – Relation extraction using dependency parse trees. *Bioinformatics*.

Thomas Mesnard Gemma Team, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. 2024. Gemma.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

José Hernández-Orallo, Peter A Flach, and Cèsar Ferri Ramirez. 2011. Brier curves: a new cost-based visualisation of classifier performance. In *Icml*, pages 585–592.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Zhanming Jie, Aldrian Obaja Muis, and Wei Lu. 2017. Efficient dependency-guided named entity recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Amir Hosein Khasahmadi, Kaveh Hassani, Parsa Moradi, Leo Lee, and Quaid Morris. 2020. Memory-based graph networks. *Preprint*, arXiv:2002.09518.

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

Chen Li, Yang Liu, Fei Liu, Lin Zhao, and Fuliang Weng. 2014. Improving multi-documents summarization by sentence compression based on expanded constituent parse trees. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 691–701, Doha, Qatar. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024b. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Ning Liu, Songlei Jian, Dongsheng Li, and Hongzuo Xu. 2022. Unsupervised hierarchical graph pooling via substructure-sensitive mutual information maximization. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1299–1308, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024c. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *Preprint*, arXiv:2308.05374.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. *Preprint*, arXiv:2002.07650.

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

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. 2024. Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. *Preprint*, arXiv:2306.10193.

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.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2024. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.

Zixing Song and Irwin King. 2022. Hierarchical heterogeneous graph attention network for syntax-aware summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11340–11348.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023a. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *Preprint*, arXiv:2305.13160.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3292–3303, Hong Kong, China. Association for Computational Linguistics.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? *Preprint*, arXiv:1810.00826.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

Hao Yuan and Shuiwang Ji. 2020. Structpool: Structured graph pooling via conditional random fields. In *Proceedings of the 8th international conference on learning representations*.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33:12827–12836.

Appendix

A Implementation Details

This section provides an overview of the implementation details of GENUINE.

A.1 Details of Features

This section provides the mathematical definitions of the features used in our uncertainty estimation framework. A detailed breakdown is presented in Table 2.

Table 2: Features used for the supervised task of uncertainty estimation for LLMs.

Name	Definition
Ent	$H(p_\theta(\cdot\|\mathbf{x}, y_1, \ldots, y_{j-1}))$
Max Ent	$\max_{j\in\{1,\ldots,n\}}\ H(p_\theta(\cdot\|\mathbf{x}, y_1, \ldots, y_{j-1}))$
Min Ent	$\min_{j\in\{1,\ldots,n\}}\ H(p_\theta(\cdot\|\mathbf{x}, y_1, \ldots, y_{j-1}))$
Avg Ent	$\frac{1}{n}\sum_{j=1}^{n} H(p_\theta(\cdot\|\mathbf{x}, y_1, \ldots, y_{j-1}))$
Std Ent	$\sqrt{\frac{\sum_{j=1}^{n}(H(p_\theta(\cdot\|\mathbf{x},y_1,\ldots,y_{j-1}))-\text{Avg Ent})^2}{n-1}}$
Prob	$p_\theta(y_j\|\mathbf{x}, y_1, \ldots, y_{j-1})$
Max Prob	$\max_{j\in\{1,\ldots,n\}}\ p_\theta(y_j\|\mathbf{x}, y_1, \ldots, y_{j-1})$
Min Prob	$\min_{j\in\{1,\ldots,n\}}\ p_\theta(y_j\|\mathbf{x}, y_1, \ldots, y_{j-1})$
Avg Prob	$\frac{1}{n}\sum_{j=1}^{n} p_\theta(y_j\|\mathbf{x}, y_1, \ldots, y_{j-1})$
Std Prob	$\sqrt{\frac{\sum_{j=1}^{n}(p_\theta(y_j\|\mathbf{x},y_1,\ldots,y_{j-1})-\text{Avg Prob})^2}{n-1}}$

A.2 Prompt Template

We adopt a few-shot prompting strategy, following the approach of (Liu et al., 2024b). Each prompt comprises four components: introduction, examples, question, and answer. The examples are user-defined question-answer pairs structured identically to the target task, ensuring consistency in format. The model receives the formatted template along with the reference question and is prompted to generate an appropriate response. This structured approach helps standardize uncertainty estimation across different tasks.

TriviaQA

```
Answer the question as following examples. Examples: Q: What star sign is Michael Caine? A: Pisces. Q: Which George invented the Kodak roll-film camera? A: Eastman. Q: ... A: ...
Q: In which decade was Arnold Schwarzenegger born? A: 1950s
```

CoQA

```
Reading the passage and answer given questions accordingly. Passage: The Vatican Apostolic Library, more commonly called the Vatican Library or simply the Vat, is the library of the Holy See, located in Vatican City. ... Examples: Q: When was the Vat formally opened? A: It was formally established in 1475. Q: ... A: ...
Q: what was started in 2014? A: a project.
```

WMT

```
What is the English translation of the following sentence? Q: Spectaculaire saut en ẅingsuitäu-dessus de Bogota. A: Spectacular Wingsuit Jump Over Bogota. Q: ... A: ...
Q: Une boîte noire dans votre voiture ? A: A black box in your car?
```

Finance

```
Answer the question as following examples. Examples: Q: For a car, what scams can be plotted with 0% financing vs rebate? A: he car deal makes money 3 ways. If you pay in one lump payment. ... Q: ... A: ...
Q: Where should I be investing my money? A: Pay off your debt. As you witnessed, no "investment" % is guaranteed. ...
```

Finance

```
What are the highlights in this paragraph? Examples: Q: LONDON, England (Reuters) – Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune ... A: Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . ... Q: ... A: ...
Q: Editor's note: In our Behind the Scenes series, CNN correspondents share ... A: Mentally ill inmates in Miami are housed on the "forgotten floor" ...
```

B Additional Experiments

In this section, we first assess model calibration performance through ECE, NLL, and Brier score metrics, shown in Figures 6, 7, and 8, respectively, comparing GENUINE's reliability against baselines. Then, we present additional experimental results evaluating GENUINE across three key dimensions: (1) the impact of dependency parse trees on uncertainty estimation (Section B.2), (2) a parameter analysis to determine the sensitivity of GENUINE to hyperparameter tuning (Section B.3), and (3) the impact of LLM parameters on GENUINE's uncertainty estimation performance (Section B.4).

B.1 Calibration Performance of GENUINE

Calibration ensures that model confidence aligns with actual correctness, making uncertainty estimation more reliable and interpretable. We assess GENUINE and baseline methods using Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier score, as shown in Figures 6, 7, and 8.

The ECE results (Fig. 6) reveal that while GENUINE outperforms baselines in WMT, Finance, and CNN datasets, it does not consistently achieve the lowest calibration error in TriviaQA and CoQA. This suggests that token-level methods such as SAR and entropy-based approaches remain competitive in capturing uncertainty effectively for shorter responses. However, in longer text generation tasks, where error accumulation can distort confidence estimates, GENUINE demonstrates superior calibration by leveraging dependency structures to refine uncertainty aggregation.

The NLL results (Fig. 7) further reinforce these trends. GENUINE consistently achieves lower NLL across all datasets, indicating that it assigns more accurate probability distributions to correct and incorrect responses compared to baselines. The advantage is particularly pronounced in WMT, Finance, and CNN datasets, where long-form responses make token-level uncertainty estimation less effective. Baselines like A4C and SE, which rely on self-evaluation or direct entropy measures, exhibit significantly higher NLL, suggesting that they struggle to generalize confidence estimates across diverse text lengths and response structures.

The Brier score results (Fig. 8) show that GENUINE achieves competitive performance across all datasets, with particularly strong improvements in WMT, Finance, and CNN datasets, aligning with its NLL performance. The gap between GENUINE and its grey-box and white-box variants indicates that hidden layer representations significantly improve calibration, especially for longer outputs. However, the higher ECE in TriviaQA and CoQA suggests that while structural modeling improves overall uncertainty estimation, it may not always provide the best confidence calibration for shorter text generations, where simpler token-wise approaches remain effective.

These results highlight that GENUINE excels in modeling uncertainty for long-form text but is less dominant in short-response tasks, where entropy-based methods can still provide competitive calibration. The findings reinforce the need for task-specific uncertainty estimation strategies, where dependency-aware modeling is particularly beneficial for applications involving complex text structures and extended reasoning.

B.2 Graph Structure and Uncertainty Estimation

Understanding the impact of graph structure on uncertainty estimation is essential for refining confidence assessment in LLM-generated responses. This section evaluates the effectiveness of dependency parse trees and analyzes graph structure variations across datasets and LLMs, using results from Table 3 and Table 4.

Dependency Parse Trees vs. Next-Token Graphs. To assess the impact of different graph structures, we compare the dependency parse tree (DPT) against the next-token graph (NTG), where edges only connect adjacent words in a sentence. The results in Table 3 clearly demonstrate that DPT-based graphs consistently outperform NTG-based graphs across all evaluation metrics, reinforcing the importance of semantic structure in uncertainty estimation.

For Llama3-8B, DPT achieves an AUROC of 0.894, improving over NTG (0.885), while also achieving lower ECE (0.246 vs. 0.264), NLL (0.362 vs. 0.437), and Brier score (0.094 vs. 0.130). Similar trends hold for Gemma2-9B, where DPT significantly outperforms NTG with an AUROC improvement of nearly 6% (0.905 vs. 0.846) and lower calibration errors. These results confirm that structural relationships encoded in dependency graphs improve uncertainty estimation, providing richer contextual information than simple word adjacency models.

When comparing grey-box vs. white-box features, we observe that DPT consistently performs better than NTG in both settings. For instance, DPT w/ grey achieves an AUROC of 0.903 for Llama3-8B, outperforming NTG w/ grey (0.897) while maintaining better calibration across ECE, NLL, and Brier scores. The trend holds for white-box features, where DPT w/ white achieves 0.809 AUROC vs. 0.795 for NTG w/ white, showing

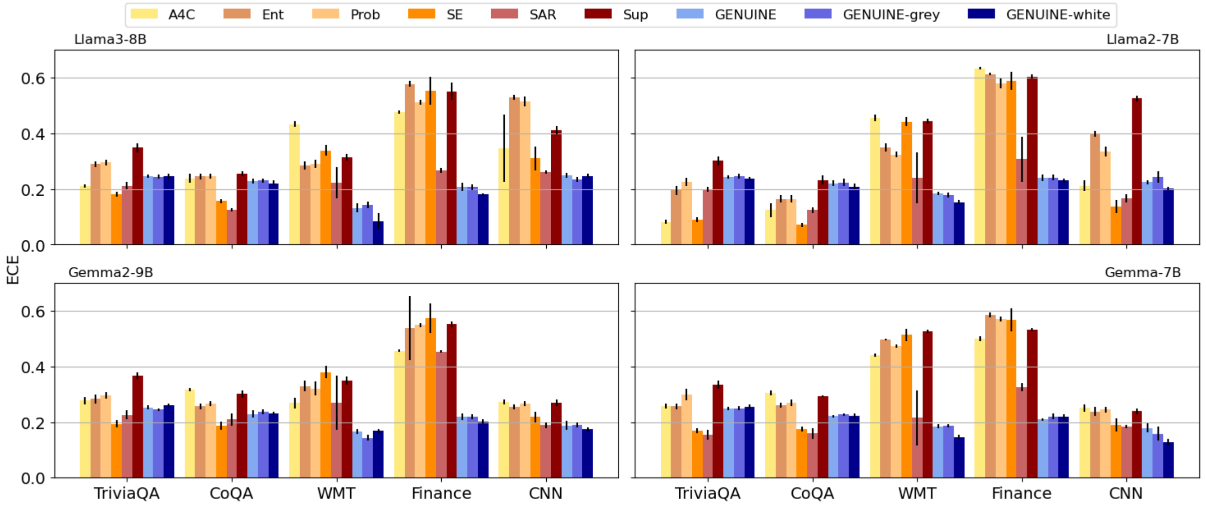


Figure 6: Comparison of ECE on five datasets, four LLMs, and six baselines. Error bars denote variance over five runs.

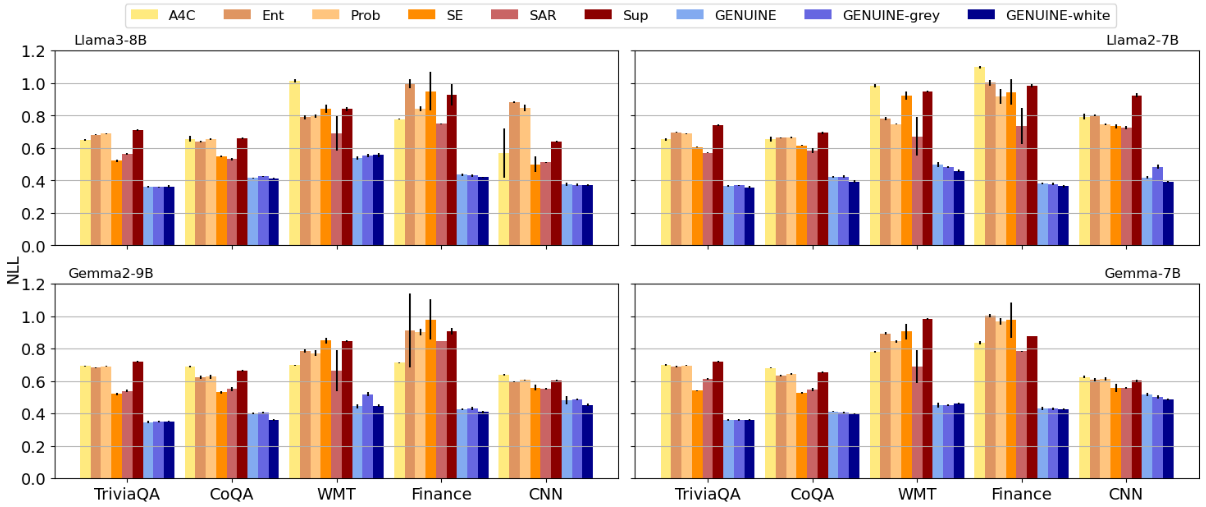


Figure 7: Comparison of NLL on five datasets, four LLMs, and six baselines. Error bars denote variance over five runs.

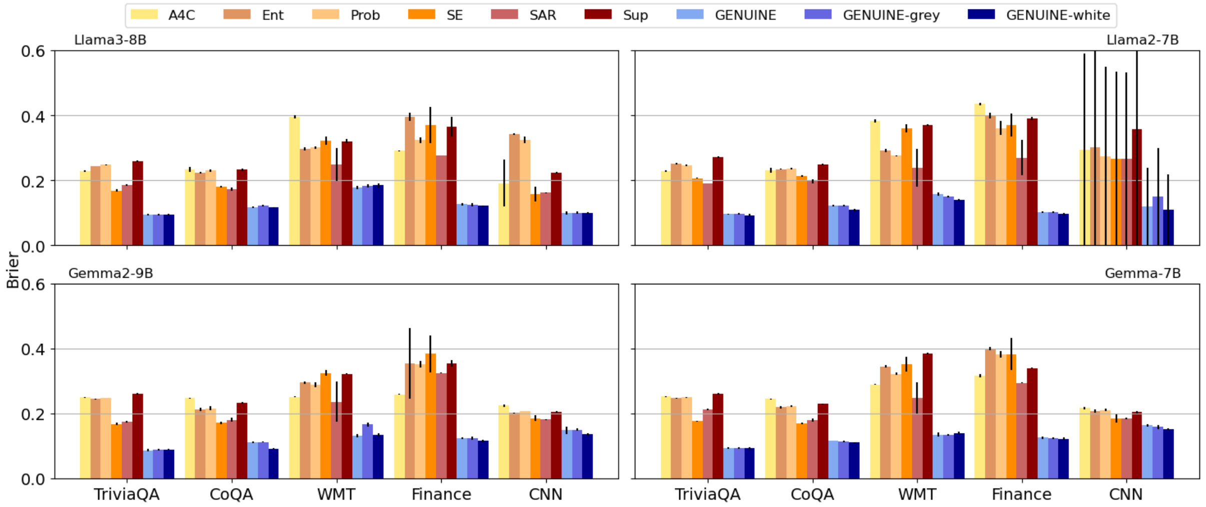


Figure 8: Comparison of Brier scores on five datasets, four LLMs, and six baselines. Error bars denote variance over five runs.

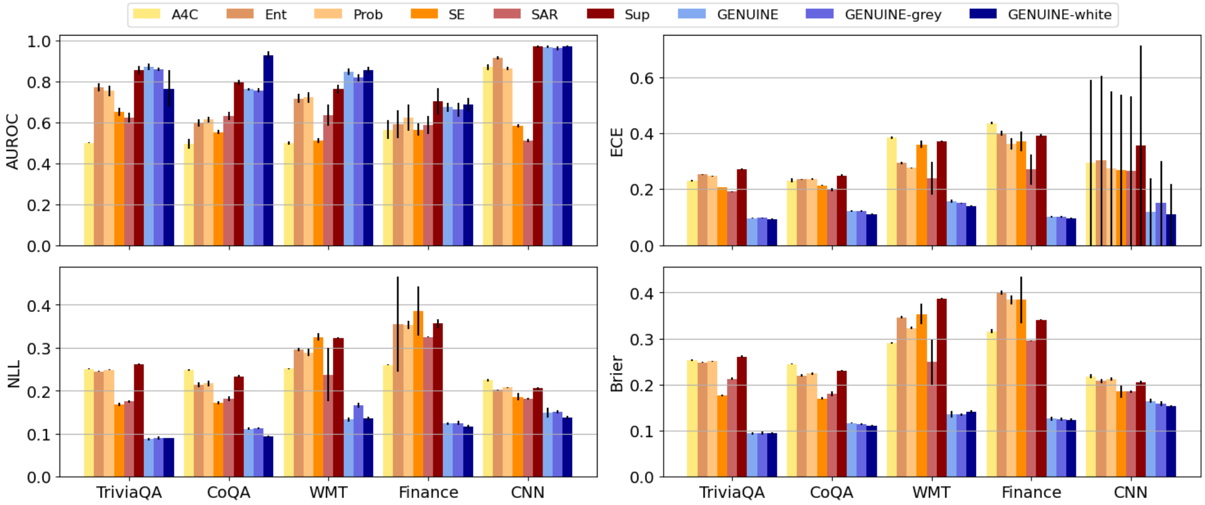


Figure 9: Experimental results on five datasets and six baseline models on Llama2-13B model. Error bars denote variance over five runs.

Table 3: Comparison of different graph structures for uncertainty estimation on TriviaQA. NTG refers to the next-token graph utilizing both white-box and grey-box features, while DPT represents the dependency parse tree graph with the same feature set. NTG w/ grey and DPT w/ grey denote the respective graphs using only grey-box features, whereas NTG w/ white and DPT w/ white correspond to configurations using only white-box features.

Graphs	Llama3-8B				Gemma2-9B			
	AUROC ↑	ECE ↓	NLL ↓	Brier ↓	AUROC ↑	ECE ↓	NLL ↓	Brier ↓
NTG	0.885±0.048	0.264±0.040	0.437±0.133	0.130±0.062	0.846±0.088	0.312±0.082	0.442±0.122	0.131±0.056
DPT	0.894±0.032	0.246±0.007	0.362±0.005	0.094±0.002	0.905±0.041	0.248±0.009	0.356±0.004	0.092±0.002
NTG w/ grey	0.897±0.039	0.245±0.007	0.363±0.007	0.095±0.003	0.914±0.041	0.251±0.006	0.354±0.006	0.091±0.003
DPT w/ grey	0.903±0.025	0.244±0.008	0.360±0.003	0.094±0.002	0.922±0.021	0.245±0.005	0.352±0.006	0.090±0.003
NTG w/ white	0.795±0.049	0.249±0.010	0.364±0.007	0.095±0.003	0.960±0.019	0.261±0.009	0.357±0.006	0.092±0.003
DPT w/ white	0.809±0.044	0.246±0.009	0.362±0.007	0.094±0.003	0.970±0.010	0.261±0.006	0.353±0.003	0.090±0.001

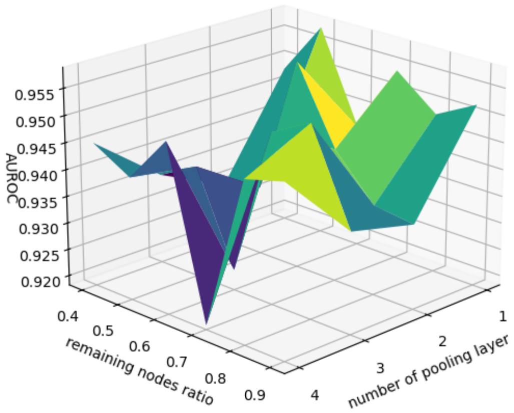


Figure 10: Parameter analysis test on number of pooling layers and remaining nodes ratio for each pooling layer

Table 4: Graph Statistics. Here # Node denotes the average node number and Density denotes the average edge density.

Datasets	Llama3-8B		Llama2-7B	
	# Node	Density	# Node	Density
TriviaQA	3.86	0.56	3.77	0.58
CoQA	5.60	0.47	5.59	0.50
WMT	24.01	0.11	21.75	0.13
Finance	46.46	0.05	21.70	0.15
CNN	61.21	0.04	87.98	0.11
Datasets	Gemma2-9B		Gemma-7B	
	# Node	Density	# Node	Density
TriviaQA	3.83	0.56	3.84	0.56
CoQA	5.28	0.47	5.19	0.48
WMT	23.65	0.12	27.14	0.10
Finance	43.61	0.05	42.92	0.06
CNN	175.33	0.01	162.96	0.01

that dependency parsing enhances uncertainty modeling even when using only hidden-layer embeddings.

These findings suggest that semantic-aware uncertainty estimation is essential, especially for longer text sequences where sequential token dependencies alone fail to capture structural nuances. By modeling hierarchical relations, DPT-based uncertainty estimation improves both reliability and calibration, making it particularly useful for structured prediction tasks.

Graph Variations Across Datasets and LLMs. Beyond structural differences, graph complexity varies significantly across datasets and LLM architectures, as shown in Table 4. We observe several key trends.

First, dataset complexity impacts graph structure. TriviaQA produces the shortest outputs, leading to small graphs with an average of 3.8 nodes, while CNN generates significantly longer responses, resulting in much larger graphs (61.2 nodes for Llama3-8B, 175.3 for Gemma2-9B). This confirms that longer text generations create more intricate dependency structures, further reinforcing why graph-based uncertainty estimation is particularly beneficial for longer responses.

Second, LLM architectures influence graph statistics. While Llama models tend to produce slightly longer responses than Gemma models in shorter datasets like TriviaQA and CoQA, this trend reverses in long-form datasets such as CNN, where Gemma models generate significantly longer outputs than Llama models (e.g., 175.3 nodes vs. 61.2 nodes in CNN for Gemma2-9B and Llama3-8B, respectively). This suggests that some LLM families prioritize brevity while others favor more detailed responses, impacting uncertainty estimation requirements.

Lastly, graph density plays a role in structural complexity. Datasets with shorter outputs (TriviaQA, CoQA) tend to have higher edge density, while longer outputs (CNN, Finance) exhibit lower density, indicating that dependency structures become more sparse as response length increases. This suggests that uncertainty estimation models should be designed to handle both dense, local dependencies and sparse, long-range relationships effectively.

Impact on Uncertainty Estimation Performance: The trends in graph statistics correlate directly with AUROC improvements in Figure 4, showing that graph-based uncertainty estimation is particularly beneficial for longer text. The WMT dataset, for example, shows substantial AUROC gains when using graph structures, emphasizing that graph-based methods provide the most value in tasks requiring extended reasoning and structured generation.

Overall, these findings confirm that dependency parsing enhances uncertainty estimation by providing hierarchical token relationships, making it particularly valuable for long-form generation, structured prediction, and document-level tasks. The graph structure directly influences uncertainty estimation effectiveness, reinforcing the need for adaptive modeling strategies based on dataset and model characteristics.

B.3 Parameter Sensitivity

Understanding the impact of hyperparameters on GENUINE's performance is essential for optimizing uncertainty estimation while ensuring efficiency. We evaluate two key parameters: the number of pooling layers (ranging from 1 to 4) and the remaining node ratio at each pooling step. The results, shown in Fig. 10, reveal important trends that highlight GENUINE's robustness and adaptability.

The results indicate that AUROC remains high with fewer pooling layers, suggesting that a deep hierarchy is not necessary for effective uncertainty estimation. As the number of pooling layers increases, performance fluctuates, indicating that excessive pooling may lead to loss of critical structural information, reducing the model's ability to capture meaningful uncertainty signals. This trend suggests that GENUINE achieves optimal results with a moderate number of pooling layers, avoiding unnecessary complexity while maintaining strong predictive performance.

Additionally, the remaining node ratio plays a crucial role in uncertainty estimation. The model may struggle with redundant information when too many nodes are retained, leading to slightly lower AUROC. However, when the number of retained nodes is optimized, performance improves, reinforcing the idea that removing less informative nodes enhances uncertainty representation. Interestingly, when the remaining ratio is lower, but the number of pooling layers is set appropriately, AUROC reaches peak performance, highlighting the benefits of structured feature reduction in refining uncertainty quantification.

Overall, these findings demonstrate that GENUINE is robust to hyperparameter choices, requiring minimal tuning to achieve strong performance. The ability to maintain high AUROC across a range of configurations suggests that GENUINE can be easily applied to various tasks and LLMs without extensive parameter optimization, making it highly adaptable for real-world deployment.

B.4 Impact of LLM Parameters

Understanding how LLM architecture and scale affect uncertainty estimation is crucial for assessing the generalizability of GENUINE. We compare the performance of Llama2-13B (Fig. 9) against Llama3-8B and Llama2-7B, analyzing its effectiveness across AUROC, calibration metrics (ECE, NLL, and Brier scores), and overall robustness.

Uncertainty Estimation Across LLM Variants. Llama2-13B achieves strong AUROC performance across all datasets, often matching or surpassing Llama3-8B and Llama2-7B. The improvements are particularly evident in WMT, Finance, and CNN datasets, where Llama2-13B consistently outperforms its smaller counterparts. This suggests that larger models benefit from enhanced representation learning, leading to more stable and accurate uncertainty estimation in complex, long-form text generation tasks. However, in TriviaQA and CoQA, the AUROC gains are marginal, indicating that the advantages of increased model size are less pronounced for shorter responses.

Calibration Trends: ECE, NLL, and Brier Score Analysis. One notable observation is that GENUINE outperforms baselines in ECE for TriviaQA and CoQA on Llama2-13B, whereas this trend is not observed in Llama3-8B and Llama2-7B. This suggests that larger models may allow GENUINE to better align confidence scores with correctness probabilities in short-response tasks, where previous versions struggled to outperform entropy-based baselines. The ECE results (Fig. 6) further confirm that in WMT, Finance, and CNN, Llama2-13B achieves lower calibration errors, highlighting its ability to generate better-aligned confidence estimates for longer outputs.

The NLL and Brier score results (Figures 7 and 8) reinforce these findings. Llama2-13B consistently achieves lower NLL and Brier scores across datasets, particularly in WMT, Finance, and CNN, where uncertainty estimation benefits from structured confidence propagation. This suggests that larger models improve AUROC and provide better-calibrated uncertainty estimates, making them well-suited for tasks requiring complex reasoning and structured text.

The results indicate that larger models significantly enhance both uncertainty estimation and confidence calibration, particularly in short-response tasks like TriviaQA and CoQA, where GENUINE surpasses entropy-based baselines in ECE for the first time. This suggests that model size can influence calibration effectiveness differently across datasets, with larger architectures improving both long-form uncertainty quantification and short-text confidence alignment. Future research should explore adaptive calibration strategies tailored to different response lengths, ensuring that LLMs remain reliable across diverse NLP applications.

Overall, these findings reinforce that GENUINE scales effectively across different LLM architectures, maintaining robust uncertainty estimation and calibration performance while highlighting areas where model size influences uncertainty quantification.