International Research Journal of Multidisciplinary Sciences **ISSN:3107-930X(ONLINE)**
VOL-2 ISSUE-1 January 2026 PP:23-29

Quantifying Uncertainty in Large Language Models: Sources, Methods, and Applications

[1]Yazdani Hasan(School of Science),[2]Kavita Kumari(PhD Scholar),[3]Anjali Jagtiani(PhD Scholar)
[1] Noida International University,[2] Noida International University, [3] Noida International University

Abstract —Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet their tendency to generate plausible but incorrect outputs—often termed hallucinations—poses significant challenges for real world deployment. This paper provides a comprehensive overview of uncertainty estimation techniques for LLMs, with a focus on understanding the fundamental sources of uncertainty and methods for quantifying them. We present a taxonomy distinguishing between data uncertainty (aleatoric), model uncertainty (epistemic), and emerging user induced uncertainty arising from prompt sensitivity. We then survey estimation methods including logit based approaches, consistency measures, and ensemble techniques, along with evaluation metrics for assessing uncertainty quality. Finally, we examine applications in high stakes domains and identify open challenges including computational efficiency and calibration across diverse tasks. This review aims to provide researchers with a structured foundation for developing more reliable and trustworthy LLM systems.

Keywords —Large Language Models, Uncertainty Estimation, Trustworthy AI, Hallucination Detection, Model Calibration

1. Introduction

The rapid advancement of Large Language Models (LLMs) has transformed artificial intelligence, enabling breakthroughs in natural language understanding, generation, and reasoning. Models such as GPT 4, LLaMA, and PaLM have demonstrated impressive performance across tasks ranging from code generation to scientific reasoning . However, as these models are deployed in high stakes domains including healthcare, finance, and legal analysis, a critical question emerges: to what extent can we trust their outputs?

LLMs are fundamentally probabilistic systems that generate responses by sampling from learned distributions over tokens. This stochastic nature, combined with limitations in training data coverage and inherent ambiguities in language, means that model outputs carry varying degrees of uncertainty . An erroneous prediction in medical diagnosis or financial advice could have severe consequences, making uncertainty estimation not merely beneficial but essential for responsible deployment.

Despite growing research interest, the sources of uncertainty in LLMs remain insufficiently understood. Traditional machine learning frameworks distinguish between aleatoric uncertainty (inherent data noise) and epistemic uncertainty (model knowledge limitations) . However, LLMs introduce unique challenges: their massive scale, opaque training data, and sensitivity to prompt phrasing create novel uncertainty sources that require specialized treatment.

This paper provides a comprehensive overview of uncertainty estimation for LLMs, organized around three research questions: (1) What are the fundamental sources of uncertainty in LLM predictions? (2) What methods exist for quantifying these uncertainties? (3) How can uncertainty estimates be evaluated and applied in practice? By synthesizing recent research, we aim to provide researchers and practitioners with a structured foundation for building more trustworthy LLM systems.

The remainder of this paper is organized as follows. Section 2 presents a taxonomy of uncertainty sources specific to LLMs. Section 3 surveys estimation methods ranging from logit based approaches to consistency measures. Section 4 discusses evaluation metrics and benchmarks. Section 5 examines applications in high stakes domains. Section 6 identifies open challenges and future directions, and Section 7 concludes.

2. Sources of Uncertainty in Large Language Models

Understanding the origins of uncertainty is fundamental to developing effective estimation methods. While traditional machine learning distinguishes between aleatoric and epistemic uncertainty, LLMs introduce additional complexity due to their unique characteristics. This section presents a three part taxonomy adapted from recent surveys .

2.1 Data Uncertainty (Aleatoric)

Data uncertainty arises from inherent variability and ambiguity in the text data itself. Unlike images or structured data, text carries semantic ambiguity at multiple levels: word sense, syntactic structure, and pragmatic interpretation. A single prompt may admit multiple valid responses depending on context or user intent. For example, the question "What is the capital of France?" has a single correct answer, but "Explain the concept of justice" invites diverse valid interpretations.

This uncertainty is irreducible—even perfect knowledge would not eliminate it because language is fundamentally ambiguous. Tokenization choices further compound data uncertainty, as different subword segmentations can alter model predictions . Research on multilingual LLMs has shown that uncertainty patterns vary significantly across languages, with lower resource languages exhibiting higher uncertainty due to limited training data coverage .

2.2 Model Uncertainty (Epistemic)

Model uncertainty stems from limitations in the model's knowledge and approximation capabilities. This includes uncertainty about parameters, insufficient training data for certain domains, and architectural constraints that limit representational capacity . Unlike data uncertainty, model uncertainty is potentially reducible through better training, more data, or improved architectures.

For LLMs, model uncertainty manifests in several ways. First, the opacity of training data means users cannot know whether specific knowledge was present during training. Second, the massive parameter space creates uncertainty about which parameters capture which knowledge. Third, distribution shift between training and deployment contexts introduces uncertainty about model behavior on out of domain inputs .

Recent work on reasoning failures in LLMs has shown that model uncertainty is particularly high for tasks requiring multi step logical inference, where errors compound across reasoning steps . This suggests that uncertainty estimation must account for task complexity, not just input characteristics.

2.3 User Induced Uncertainty

A novel source of uncertainty emerging in LLM research relates to human interaction. LLMs are exceptionally sensitive to prompt phrasing, with minor variations potentially producing dramatically different outputs . This creates

what we term user induced uncertainty—uncertainty arising from the under specification of user intent and the model's interpretation of ambiguous prompts.

Research on prompt engineering has demonstrated that chain of thought prompting, few shot examples, and instruction phrasing significantly influence output distributions . From an uncertainty perspective, this means that the same underlying query, expressed differently, may yield responses with widely varying confidence. This sensitivity challenges traditional notions of uncertainty, which typically assume a fixed input distribution.

Multi agent LLM systems introduce additional complexity, as uncertainty propagates and potentially amplifies through agent interactions . When multiple LLM agents collaborate on tasks, each agent's uncertainty contributes to collective uncertainty in ways not yet well understood.

3. Uncertainty Estimation Methods

A variety of methods have been developed to estimate uncertainty in LLM predictions. These range from simple logit based approaches to sophisticated consistency measures and ensemble techniques .

3.1 Logit Based Methods

The most direct uncertainty estimates derive from the model's output probabilities. For classification tasks, the softmax distribution over tokens provides a natural confidence measure. However, LLMs generate sequences, requiring aggregation across tokens. Common approaches include:

Maximum token probability : The probability of the highest likelihood token at each position, averaged or multiplied across the sequence.
Perplexity : The exponentiated negative log likelihood of the generated sequence, reflecting overall model confidence.
Entropy : Shannon entropy of the token distribution, capturing dispersion in model predictions.

Logit based methods are computationally efficient but have significant limitations. Modern LLMs are often poorly calibrated, with overconfident predictions even when incorrect . Furthermore, these methods capture only the model's internal confidence, not whether that confidence is justified.

3.2 Consistency Based Methods

Consistency methods estimate uncertainty by examining variability across multiple model outputs. The intuition is that if the model produces diverse responses to the same prompt, uncertainty is high. Key approaches include:

Self consistency : Generating multiple responses via sampling and measuring agreement . Metrics include lexical similarity (e.g., BLEU, ROUGE) and semantic similarity using embedding models.
Verbalized uncertainty : Prompting the model to express its confidence in natural language (e.g., "On a scale of 1 10, how confident are you in this answer?"). Research shows verbalized confidence correlates with accuracy but requires careful prompt design .
Metacognitive prompting : Techniques like "Let's think step by step" that encourage the model to reason about its own uncertainty .

Consistency methods capture epistemic uncertainty better than logit based approaches but require multiple model calls, increasing computational cost.

3.3 Ensemble and Sampling Methods

Ensemble methods combine predictions from multiple models or multiple forward passes to estimate uncertainty. For LLMs, practical approaches include:

Monte Carlo dropout : Applying dropout at inference time to generate multiple stochastic forward passes .
Deep ensembles : Training multiple independent models, though computationally prohibitive for LLMs.
Temperature scaling : Adjusting sampling temperature to explore the prediction distribution more broadly.

Recent work has explored efficient alternatives including last layer ensembles and parameter efficient fine tuning to create ensemble diversity at lower cost .

3.4 Retrieval Augmented Uncertainty

Retrieval augmented generation (RAG) offers a novel approach to uncertainty reduction by grounding LLM outputs in external knowledge . By retrieving relevant documents and conditioning generation on them, RAG systems can reduce uncertainty for knowledge intensive tasks. Uncertainty estimation in RAG systems must account for both retrieval quality and generation confidence, creating a two stage estimation problem.

Research on biomedical applications shows that RAG with structured knowledge graphs can significantly reduce hallucinations and improve traceability . However, uncertainty estimation in RAG remains underdeveloped, with most work focusing on performance rather than reliability metrics.

4. Evaluation Metrics and Benchmarks

Evaluating uncertainty estimates requires metrics that assess both accuracy and calibration quality. This section reviews key evaluation approaches .

4.1 Calibration Metrics

Calibration measures alignment between predicted confidence and actual accuracy. A well calibrated model should be correct 80% of the time when it assigns 80% confidence. Common metrics include:

Expected Calibration Error (ECE) : The weighted average of the absolute difference between accuracy and confidence across bins.
Brier Score : The mean squared difference between predicted probabilities and actual outcomes, combining calibration and sharpness.
Reliability diagrams : Visual plots of accuracy versus confidence that reveal systematic miscalibration patterns.

Research on LLM calibration has found significant variation across tasks and model families. Larger models tend to be

better calibrated, but all models show degradation on out of distribution inputs .

4.2 Hallucination Detection

Uncertainty estimation is closely linked to hallucination detection—identifying when model outputs are factually incorrect. Benchmarks for hallucination detection evaluate whether uncertainty scores correlate with factual accuracy. Recent surveys categorize hallucinations by source (data, training, inference) and modality .

Key challenges include distinguishing between factual errors and creative generation (where uncertainty may be appropriate) and evaluating open ended generation where ground truth is ambiguous .

4.3 Task Specific Benchmarks

Different tasks require different uncertainty properties. For multiple choice questions, token level probabilities may suffice. For open ended generation, semantic consistency across samples is more relevant. Recent benchmarks including TruthfulQA and HaluEval provide task specific evaluation frameworks .

5. Applications in High Stakes Domains

Uncertainty estimation enables trustworthy LLM deployment in domains where errors carry significant consequences.

5.1 Healthcare and Biomedicine

LLMs are increasingly applied to clinical decision support, literature mining, and patient data analysis . In these contexts, uncertainty estimates can flag low confidence predictions for human review, preventing automated decisions based on unreliable outputs. Research on biomedical LLMs emphasizes the need for traceability and evidence grounding, with uncertainty estimation serving as a quality control mechanism .

5.2 Finance and Legal Analysis

Financial and legal applications require high reliability due to regulatory requirements and potential liability. Uncertainty estimates enable risk based decision making, where low confidence predictions trigger additional verification. Recent work on LLM agents for financial analysis incorporates uncertainty into trading decisions .

5.3 Education and Research

In educational applications, uncertainty estimates help identify topics where models lack knowledge, enabling appropriate scaffolding or referral to human experts . For research assistance, uncertainty quantification supports scientific rigor by indicating confidence in generated hypotheses or literature summaries .

6. Challenges and Future Directions

Despite progress, significant challenges remain in uncertainty estimation for LLMs.

6.1 Computational Efficiency

Most uncertainty methods require multiple model calls or ensembles, creating computational overhead that may be prohibitive for real time applications. Developing efficient uncertainty estimators that leverage internal model representations rather than multiple forward passes is an important research direction .

6.2 Calibration Across Tasks

LLMs are deployed across diverse tasks, but uncertainty methods optimized for one task may fail on others. Task adaptive uncertainty estimation that adjusts based on input characteristics and task demands remains an open problem .

6.3 Uncertainty in Multi Agent Systems

As LLM based multi agent systems become more common, understanding and estimating collective uncertainty becomes critical. Uncertainty may propagate or amplify through agent interactions, requiring new estimation frameworks .

6.4 Privacy and Memorization

Uncertainty estimation must account for the risk that low uncertainty predictions may reflect memorization of training data, raising privacy concerns . Balancing uncertainty estimation with privacy preservation is an emerging research area.

7. Conclusion

Uncertainty estimation is essential for deploying LLMs in high stakes applications where reliability is paramount. This paper has provided a structured overview of uncertainty sources, estimation methods, evaluation metrics, and applications. We have highlighted the unique characteristics of LLMs—including prompt sensitivity, opaque training data, and massive scale—that distinguish uncertainty estimation in this context from traditional machine learning approaches.

Key takeaways include: (1) Uncertainty in LLMs arises from data, model, and user interaction sources requiring different estimation strategies; (2) No single estimation method suffices for all tasks; (3) Evaluation must consider both calibration and task specific requirements; (4) Applications in healthcare, finance, and education demonstrate practical value but reveal open challenges.

Future research should focus on efficient uncertainty estimation, task adaptive methods, and frameworks for multi agent systems. As LLMs continue to evolve, uncertainty estimation will remain critical for ensuring these powerful models are deployed responsibly and reliably.

References

[1] P. Song et al., "Large Language Model Reasoning Failures," arXiv preprint arXiv:2602.06176 , 2026.

[2] J. Haase and S. Pokutta, "The Hidden Cost of Tokenization: Why (most) Non English Speakers Pay More for Less," arXiv preprint , 2026.

[3] J. Haase et al., "Building Socially Grounded Multi Agent LLM Systems," arXiv preprint , 2026.

[4] J. Haase and S. Pokutta, "Beyond Static Responses: Multi Agent LLM Systems as a New Paradigm for Social Science Research," arXiv preprint , 2025.

[5] M. Zimmer et al., "PERP: Rethinking the Prune Retrain Paradigm in the Era of LLMs," arXiv preprint , 2023.

[6] Y. Mou et al., "Comparative analyses of hybrid LLMs with Knowledge base integration and RAGs in biomedical domain," RWTH Aachen University , 2026.

[7] S. Okabe et al., "Machine Translation for (very) low resource languages," TUM Department of Computer Science , 2026.

[8] M. Di Marco, "In context learning for translating low resourced languages," TUM Department of Computer Science , 2026.

[9] Z. Liu et al., "Survey of uncertainty estimation in LLMs Sources, methods, applications, and challenges," Information Fusion , vol. 130, 104057, 2026.

[10] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," Artificial Intelligence Review , 2023.

[11] Y. Huang et al., "Trustworthiness in LLMs: a survey," arXiv preprint , 2024.

[12] A. Satvaty et al., "Undesirable Memorization in Large Language Models: A Survey," arXiv preprint arXiv:2410.02650 , 2026.

[13] T. Brown et al., "Language Models are Few Shot Learners," NeurIPS , 2020.

[14] J. Wei et al., "Chain of Thought Prompting Elicits Reasoning in Large Language Models," NeurIPS , 2022.

[15] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," ICLR , 2023.

[16] N. Shinn et al., "Reflexion: Language Agents with Verbal Reinforcement Learning," NeurIPS , 2023.

[17] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," arXiv preprint , 2023.

[18] S. Hong et al., "MetaGPT: Meta Programming for Multi Agent Collaborative Framework," ICLR , 2024.

[19] G. Li et al., "CAMEL: Communicative Agents for 'Mind' Exploration of Large Scale Language Model Society," NeurIPS , 2023.

[20] J. Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," UIST , 2023.