Understanding Uncertainty In LLMs

Chandan Kumar Sah
School of Computer Science and Engineering, Beihang University
Haidian, Beijing, China
sahchandan98@buaa.edu.cn

Abstract

Large Language Models (LLMs) have revolutionized AI, yet their inherent uncertainties pose significant challenges to reliable deployment. This paper presents a comprehensive systematic review of uncertainty in LLMs, bridging theoretical foundations and cutting-edge methodologies. We analyze over 45 papers from top venues—including ASE, NeurIPS, ICML, and Nature—to trace the evolution of uncertainty quantification (UQ). We categorize uncertainty into aleatoric and epistemic types, detailing probabilistic modeling, confidence estimation, and calibration techniques. Through illustrative case studies in high-stakes domains such as medical diagnosis and code generation, we demonstrate UQ's pivotal role in enhancing reliability. We further discuss limitations, ethical considerations, and future directions, emphasizing the need for granular interpretability and human-AI collaboration. This work advances the understanding of LLM uncertainty to enable safer, trustworthy, and responsible real-world integration.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Information systems** → Uncertainty quantification.

Keywords

Large Language Models, Uncertainty, Natural Language Processing

ACM Reference Format:
Chandan Kumar Sah. 2018. Understanding Uncertainty In LLMs. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 RESEARCH PROBLEM AND MOTIVATION

Modern AI systems increasingly make high-stakes decisions, yet they often lack calibrated confidence. In safety-critical domains (healthcare, autonomous driving, law enforcement), unquantified uncertainty in LLM-driven tools can lead to catastrophic outcomes [8, 16]. For example, biased facial-recognition systems have produced wrongful arrests, and LLM chatbots routinely "hallucinate" confidently with false information [11, 16]. In healthcare, AI diagnostic models frequently give fixed predictions with no "I don't know" option, causing unchecked medical errors [1]. Likewise, autonomous vehicles must recognize rare outliers (e.g. unexpected pedestrians) to avoid crashes, but deep models struggle with long-tail events [21, 38]. A better understanding of uncertainty and how people deal with uncertainty [28]. A key challenge is that modern neural models are notoriously overconfident and poorly calibrated [35, 38], and they lack robust detection of out-of-distribution inputs [3, 6, 8]. Without selective prediction or reliable confidence estimates, AI outputs cannot be trusted. Thus, rigorous uncertainty quantification is essential for building safe, trustworthy AI systems.

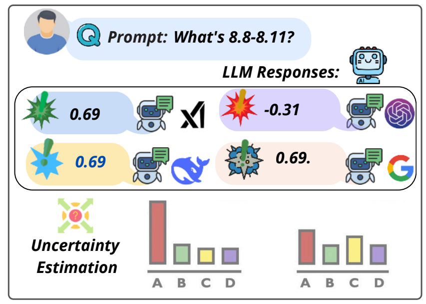


Figure 1: Uncertainty Estimation in Question Answering: Comparing Accurate and Risky Predictions Across LLMs.

2 RELATED WORKS

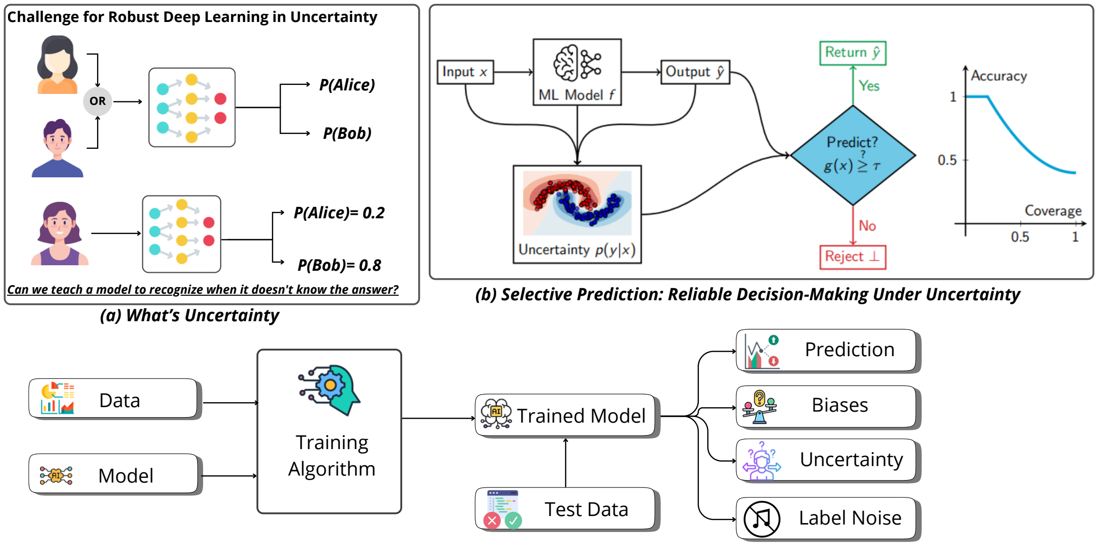
2.1 Uncertainty of LLMs

Large language models (LLMs) are increasingly vital across domains, necessitating robust uncertainty estimation to assess prediction confidence, especially in high-stakes fields like medical diagnosis where errors can be critical [9, 30, 34]. This estimation also helps mitigate LLM hallucinations by identifying knowledge boundary issues [23], enhancing trust in transformer-based outputs [17]. Uncertainty reflects output distribution variability, distinct from confidence in prediction accuracy. Research in [36] explores LLM confidence in code token accuracy, finding a strong correlation between entropy-based uncertainty [23] and token correctness in code completion tasks [47]. High uncertainty often signals potential errors, which highlights its role in improving the reliability of code generation.

2.2 Optimizing LLM Code Generation

The rapid advancement of LLMs like GPT-4 [25], GPT-5 [24] and Grok-4 [42] has revolutionized code generation, leading to specialized Code LLMs such as CodeLlama [29], Deepseek-coder [10], and Qwen-coder [2]. These models excel in multi-language programming, code completion, debugging, and refactoring [4, 20, 45], trained on vast codebases to grasp complex logic and intents [11, 18]. Enhancement techniques include prompting with domain knowledge [15, 19, 33], fine-tuning on specific datasets [41], and decoding

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX



(a) What's Uncertainty

(b) Selective Prediction: Reliable Decision-Making Under Uncertainty

(c) Model-Agnostic Framework for Risk Estimation in Decision Making

Figure 2: Illustration of Uncertainty Challenges in Deep Learning: (a) Defining Uncertainty with Recognition Examples; (b) Selective Prediction for Reliable Decisions; (c) Model-Agnostic Framework for Risk Estimation in Robust AI Systems.

strategies using test cases and feedback [48]. Chain-of-Thought (CoT) prompting [40] addresses reasoning bottlenecks by generating intermediate steps, validated in zero-shot and simple instructions. Derivatives like self-consistency and integrations in models like OpenAI o1 [37] and Deepseek-R1 boost performance. In code generation, CoT variants such as Self-planning, SCoT and LVLMs [14] incorporate planning, structures, and reflection to simplify problem-solving.

3 APPROACH

We systematically categorize uncertainty (e.g. epistemic vs. aleatoric) and survey state-of-the-art UQ methods (Bayesian models, deep ensembles, calibration, etc.) for LLMs. Illustrative case studies in medical diagnosis and code generation ground our discussion. Example of Question Answering with correct response and risky response Figure 1 exemplifies our analysis: it compares three LLMs answering a numerical question, showing one model's confident (0.69) correct prediction versus another's low-confidence (−0.31) error. Meanwhile, OpenAI claims GPT-5 model boosts ChatGPT to **PhD level** [13]. Figure 2 presents a conceptual illustration of Uncertainty Challenges, Selective Prediction, and Model-Agnostic Risk Estimation in Deep Learning. These figures clarify how LLM output distributions reflect different uncertainty sources, highlighting how quantifying uncertainty enables models to defer or warn when predictions may be unreliable.

4 RESULTS AND CONTRIBUTIONS

Across LLM tasks (code generation, QA, summarization, MT), enhanced uncertainty measurement via perturbation strategies proves valuable yet insufficient for full risk assessment (as conceptualized in Figure 2). This necessitates optimized prompting for researchers and an "ask more, get more" interactive strategy for developers, alongside future research priorities (outlined in Table 1) for trustworthy deployment.

Contributions. Our work introduces a unified framework for categorizing and quantifying uncertainty in LLMs (aleatoric/epistemic) through Bayesian, calibration, and ensemble methods. We comprehensively review 45+ papers, supported by case studies in medical diagnosis and code generation. Finally, we outline future directions—granular uncertainty, trustworthy AI, and scalable UQ—to guide research and deployment as show in the Table 1.

Table 1: Future Directions in Uncertainty Research

Area	Future Directions	src.
Uncertainty in Modern Models, Suitability and Meta Learning	Scalability, over-parameterization, predictive distributions, data shift, label-free detection, agentic inference, meta learning, compositional generalization, causal inference, synthetic data, TL techniques.	[7, 22, 27, 39, 43, 46]
UnCert-CoT	Hyperparameter robustness.	[47]
Uncertainty Quant.	Knowledge redundancy assessment, reasoning structure insights.	[5, 18]
Trustworthy AI	Diagnosis uncertainty, bias mitigation, system improvement.	[8, 26, 39]
Industry Use	Trustworthy LLMs for industry.	[7, 12]
Data & Bench.	Datasets for UQ, challenges, benchmarking.	[31, 32, 44]

5 Acknowledgments

We are grateful to Prof. Dr. Lian Xiaoli and Prof. Dr. Zhang Li for enlightening discussions and comments. The Chinese Government Scholarship and Beihang University supported this work. We thank Stephan Rabanser, Qinghua Xu, and Sadhana Lolla for their support.

References

[1] Zahra Atf, Seyed Amir Ahmad Safavi-Naini, Peter R Lewis, Aref Mahjoubfar, Nariman Naderi, Thomas R Savage, and Ali Soroush. 2025. The challenge of uncertainty quantification of large language models in medicine. *arXiv preprint arXiv:2504.05278* (2025).

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[3] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. 2020. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems* 33 (2020), 16085–16095.

[4] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[5] Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. [n. d.]. Understanding the Uncertainty of LLM Explanations from Reasoning Topology. ([n. d.]).

[6] Antoine de Mathelin, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis. 2025. Deep Out-of-Distribution Uncertainty Quantification via Weight Entropy Maximization. *Journal of Machine Learning Research* 26, 4 (2025), 1–68.

[7] Gianluca Detommaso, Alberto Gasparin, Michele Donini, Matthias Seeger, Andrew Gordon Wilson, and Cedric Archambeau. 2024. Fortuna: A library for uncertainty quantification in deep learning. *Journal of Machine Learning Research* 25, 238 (2024), 1–7.

[8] Jessica Deuschel, Andreas Foltyn, Karsten Roscher, and Stephan Scheele. 2024. The role of uncertainty quantification for trustworthy AI. In *Unlocking Artificial Intelligence: From Theory to Applications*. Springer, 95–115.

[9] Renee C Fox. 1980. The evolution of medical uncertainty. *The Milbank Memorial Fund Quarterly. Health and Society* (1980), 1–49.

[10] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming–The Rise of Code Intelligence. *arXiv preprint arXiv:2401.14196* (2024).

[11] Joe B Hakim, Jeffery L Painter, Darmendra Ramcharran, Vijay Kara, Greg Powell, Paulina Sobczak, Chiho Sato, Andrew Bate, and Andrew Beam. 2025. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. *Scientific Reports* 15, 1 (2025), 27886.

[12] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering* (2025).

[13] Lily Jamali and Liv McMahon. 2025. OpenAI claims GPT-5 model boosts ChatGPT to 'PhD level'. *BBC News* (8 aug 2025). https://www.bbc.com/news/articles/cy5prvgw0r1o North America Technology correspondent (Lily Jamali); Technology reporter (Liv McMahon).

[14] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 525–534.

[15] Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. 2024. Self-planning code generation with large language models. *ACM Transactions on Software Engineering and Methodology* 33, 7 (2024), 1–30.

[16] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 4.

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

[18] Wonbin Kweon, Sanghwan Jang, SeongKu Kang, and Hwanjo Yu. 2025. Uncertainty Quantification and Decomposition for LLM-based Recommendation. In *Proceedings of the ACM on Web Conference 2025*. 4889–4901.

[19] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology* 34, 2 (2025), 1–23.

[20] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).

[21] Henry X Liu and Shuo Feng. 2024. Curse of rarity for autonomous vehicles. *nature communications* 15, 1 (2024), 4808.

[22] Tao Liu, Jiahao Liu, Dong Li, and Shan Tan. 2025. Bayesian deep-learning structured illumination microscopy enables reliable super-resolution imaging with uncertainty quantification. *Nature Communications* 16, 1 (2025), 5027.

[23] Joan M. Morrissey. 1990. Imprecise information and uncertainty in information systems. *ACM Transactions on Information Systems (TOIS)* 8, 2 (1990), 159–180.

[24] OpenAI. 2025. GPT-5 System Card. (August 7 2025). https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf

[25] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2, 5 (2023), 1.

[26] W Pisciotta, P Arina, D Hofmaenner, and M Singer. 2023. Difficult diagnosis in the ICU: making the right call but beware uncertainty and bias. *Anaesthesia* 78, 4 (2023), 501–509.

[27] Stephan Rabanser. 2025. Uncertainty-Driven Reliability: Selective Prediction and Trustworthy Deployment in Modern Machine Learning. *Department of Computer Science, University of Toronto* (2025).

[28] William D Rowe. 1994. Understanding uncertainty. *Risk analysis* 14, 5 (1994), 743–750.

[29] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).

[30] Michele Salvagno, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. Artificial intelligence hallucinations. *Critical Care* 27, 1 (2023), 180.

[31] Simon Schmitt, John Shawe-Taylor, and Hado van Hasselt. 2025. General Uncertainty Estimation with Delta Variances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 20318–20328.

[32] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Comput. Surveys* (2025).

[33] Disha Shrivastava, Hugo Larochelle, and Daniel Tarlow. 2023. Repository-level prompt generation for large language models of code. In *International Conference on Machine Learning*. PMLR, 31693–31715.

[34] Arabella Simpkin and Richard Schwartzstein. 2016. Tolerating uncertainty—the next medical revolution? *New England Journal of Medicine* 375, 18 (2016).

[35] Christian Soize. 2017. *Uncertainty quantification*. Vol. 23. Springer.

[36] Claudio Spiess, David Gros, Kunal Suresh Pai, Michael Pradel, Md Rafiqul Islam Rabin, Amin Alipour, Susmit Jha, Prem Devanbu, and Toufique Ahmed. 2024. Calibration and correctness of language models for code. *arXiv preprint arXiv:2402.02047* (2024).

[37] Mohamad-Hani Temsah, Amr Jamal, Khalid Alhasan, Abdulkarim A Temsah, and Khalid H Malki. 2024. OpenAI o1-preview vs. ChatGPT in healthcare: a new frontier in medical AI reasoning. *Cureus* 16, 10 (2024).

[38] Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. 2021. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems* 34 (2021), 11809–11820.

[39] Meng Wang, Tian Lin, Lianyu Wang, Aidi Lin, Ke Zou, Xinxing Xu, Yi Zhou, Yuanyuan Peng, Qingquan Meng, Yiming Qian, et al. 2023. Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications* 14, 1 (2023), 6757.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[41] Martin Weyssow, Xin Zhou, Kisub Kim, and David Lo. 2023. Exploring parameter-efficient fine-tuning techniques for code generation with large language models. *ACM Transactions on Software Engineering and Methodology* (2023).

[42] xAI. 2025. Grok-4. *x.ai News* (jul 2025). https://x.ai/news/grok-4 Accessed: 2025-08-12.

[43] Qinghua Xu, Shaukat Ali, Tao Yue, and Maite Arratibel. 2022. Uncertainty-aware transfer learning to evolve digital twins for industrial elevators. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1257–1268.

[44] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems* 37 (2024), 15356–15385.

[45] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Shen, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5673–5684.

[46] Jianhan Zhu, Jun Wang, Ingemar J Cox, and Michael J Taylor. 2009. Risky business: modeling and exploiting uncertainty in information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 99–106.

[47] Yuqi Zhu, Ge Li, Xue Jiang, Jia Li, Hong Mei, Zhi Jin, and Yihong Dong. 2025. Uncertainty-guided chain-of-thought for code generation with llms. *arXiv preprint arXiv:2503.15341* (2025).

[48] Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. 2024. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 437–445.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009