M-∂KG: A Theoretical Framework for Uncertainty Quantification in Large Language Model

Md Tauhidul Islam[1] TAUHIDSUMAN10@GMAIL.COM
Md Wahiduzzaman Suva[2] 22-47171-1@STUDENT.AIUB.EDU
Md Tanzib Hosain[2] 20-42737-1@STUDENT.AIUB.EDU
Nafiz Fahad[3] NAFIZ.FAHAD@STUDENT.MMU.EDU.MY
Md Kishor Morol[4] MMOROL@CORNELL.EDU
Dip Nandi[2] DIP.NANDI@AIUB.EDU
Mashiour Rahman[2] MASHIOUR@AIUB.EDU
Mohammad Ali Moni[*5] M.MONI@UQ.EDU.AU

[1] *NI Lobachevsky National Research Nizhny Novgorod State University, Nizhny Novgorod, Russia*
[2] *American International University-Bangladesh, Dhaka, Bangladesh*
[3] *Multimedia University, Melaka, Malaysia*
[4] *Cornell University, New York, United States of America*
[5] *The University of Queensland, Queensland, Australia*

Editors: Under Review for MIDL 2026

Abstract

Large Language Models (LLMs) have shown strong potential in healthcare, but their medical deployment remains limited by challenges in uncertainty quantification and interpretability. Although probabilistic uncertainty estimation has been widely studied, formal graph-theoretical frameworks for quantifying uncertainty in LLM reasoning are still lacking. This paper proposes a novel mathematical framework that models LLM reasoning as traversals over knowledge graphs and defines uncertainty using graph-theoretical properties. The framework decomposes uncertainty into epistemic (model) and aleatoric (data) components, derives theoretical bounds on uncertainty estimates, and enables formal comparison with existing probabilistic approaches. We further demonstrate its use in medical reasoning tasks, including diagnosis, treatment planning, and prognosis, showing that graph-based properties provide more interpretable and theoretically grounded uncertainty estimates. Finally, through mathematical analysis and simulation, we empirically validate the theoretical bounds and relationships established by the framework.

Keywords: Uncertainty quantification, medical knowledge graph, epistemic uncertainty, clinical decision support.

1. Introduction

Large Language Models (LLMs) have demonstrated strong potential in healthcare, including medical question answering, diagnosis support, and treatment planning Chen et al. (2025); Singhal et al. (2023), yet their safe medical deployment remains limited by challenges in uncertainty quantification and interpretability. In healthcare, reliable confidence estimation

* Corresponding author

is not only a technical necessity but also an ethical requirement Hosseini et al. (2023). Recent medical LLMs such as Med-PaLM 2 Singhal et al. (2023) and Med-R1 8B Lai et al. (2025) have shown near-expert performance on medical reasoning tasks, but they may still produce overly confident incorrect answers, which is problematic for medical decision support Chen et al. (2025). Although uncertainty quantification has been widely explored in deep learning through Bayesian, ensemble, and dropout-based methods Abdar et al. (2021), applying these techniques to LLMs in healthcare remains challenging. Prior work has begun addressing this issue by studying epistemic and aleatoric uncertainty in medical LLMs Chen et al. (2025), and knowledge graphs have also been incorporated into healthcare LLM systems Pan et al. (2024).

However, a formal graph-theoretical framework for quantifying uncertainty in LLM reasoning is still missing. To address this gap, this paper proposes a mathematical framework that models LLM reasoning as paths in knowledge graphs, defines graph-based uncertainty measures, distinguishes epistemic and aleatoric uncertainty through graph structure, and establishes theoretical uncertainty bounds grounded in graph properties.

2. M-∂KG Framework

M-∂KG models medical knowledge as a directed, weighted, labeled graph $G = (V, E, \mathcal{L}_V, \mathcal{L}_E, w)$, where vertices represent medical concepts, edges represent relations, and $w : E \to [0, 1]$ assigns confidence scores. For a medical query q, the relevant subgraph is the reasoning graph G_q, and an LLM's reasoning is represented by one or more paths $P = (v_1, e_1, \ldots, e_{n-1}, v_n)$. Path-level uncertainty is captured through path confidence $C(P) = \prod_{i=1}^{n-1} w(e_i)$, path entropy $H(P) = -\sum_{i=1}^{n-1} w(e_i) \log w(e_i)$, and path-length complexity $L(P) = 1 - e^{-\lambda|P|}$. Graph-level epistemic uncertainty is measured using density uncertainty $U_D(G_q) = 1 - \frac{|E_q|}{|V_q|(|V_q|-1)}$, connectivity uncertainty $U_C(G_q) = 1 - \frac{\lambda_2(G_q)}{\lambda_1(G_q)}$, and centrality divergence $U_{CD}(G_q) = \frac{1}{|V_q|}\sum_{v \in V_q} |BC(v) - CC(v)|$. For multi-path reasoning, agreement between two paths is $A(P_i, P_j) = \frac{|V(P_i) \cap V(P_j)|}{|V(P_i) \cup V(P_j)|}$, diversity is $D(\mathcal{P}) = 1 - \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} A(P_i, P_j)$, and multi-path uncertainty is $U_{MP}(\mathcal{P}) = D(\mathcal{P}) \left(1 - \frac{1}{k} \sum_{i=1}^{k} C(P_i)\right)$. M-∂KG decomposes total uncertainty into epistemic and aleatoric parts: $U_E(G_q) = \alpha_1 U_D(G_q) + \alpha_2 U_C(G_q) + \alpha_3 U_{CD}(G_q)$, $U_A(G_q, \mathcal{P}) = \beta_1 D(\mathcal{P}) + \beta_2 \left(\frac{1}{k} \sum_{i=1}^{k} H(P_i)\right) + \beta_3 \left(\frac{1}{k} \sum_{i=1}^{k} L(P_i)\right)$, and combines

Algorithm 1 M-∂KG Computing Protocol

Require: Medical query q, medical knowledge graph G, number of reasoning paths k
Ensure: Total uncertainty $\phi(U_T)$

1: Construct reasoning graph $G_q = (V_q, E_q)$ from G based on q
2: Generate k paths $\mathcal{P} = \{P_1, ..., P_k\}$ in G_q using LLM
3: Compute $C(P_i)$, $H(P_i)$, $L(P_i)$ for each $P_i \in \mathcal{P}$
4: Compute $U_D(G_q)$, $U_C(G_q)$, $U_{CD}(G_q)$
5: Compute $D(\mathcal{P})$, $U_{MP}(\mathcal{P})$
6: Compute $U_E(G_q) = \alpha_1 U_D(G_q) + \alpha_2 U_C(G_q) + \alpha_3 U_{CD}(G_q)$
7: Compute $U_A(G_q, \mathcal{P}) = \beta_1 D(\mathcal{P}) + \beta_2 \left(\frac{1}{k} \sum_{i=1}^{k} H(P_i)\right) + \beta_3 \left(\frac{1}{k} \sum_{i=1}^{k} L(P_i)\right)$
8: Compute $U_T(G_q, \mathcal{P}) = \gamma_1 U_E(G_q) + \gamma_2 U_A(G_q, \mathcal{P})$
9: Calibrate: $\phi(U_T) = \frac{1}{1+e^{-\sigma(U_T - \mu)}}$
10: **return** $\phi(U_T)$

them as $U_T(G_q, \mathcal{P}) = \gamma_1 U_E(G_q) + \gamma_2 U_A(G_q, \mathcal{P})$. Finally, the uncertainty is calibrated into an interpretable probability using $\phi(U_T) = \frac{1}{1+e^{-\sigma(U_T-\mu)}}$. Refer to algorithm 1.

3. Empirical Validation

This section empirically validates the proposed graph-theoretical uncertainty framework through simulation. We implemented a simplified Python framework using `NetworkX`, generated random medical knowledge graphs with controlled numbers of vertices and edge probabilities, simulated reasoning paths from symptoms to diagnoses, and computed the uncertainty measures. The results confirm the theoretical upper bound on epistemic uncertainty, with all measured values satisfying $U_E \leq 1 - \frac{1}{|V_q|}$, as illustrated in Figure 1. In addition, the knowledge graph representation for the MedQA setting, the simulation used graphs with $|V| = 110$, $|E| = 341$ for diagnosis and treatment (258 for prognosis), $k = 5$ reasoning paths, and 29 connected components. The corresponding path-based metrics were average path confidence $C(P) = 0.950$, path entropy $H(P) = 0.070$, path complexity $L(P) = 0.095$, path diversity $D(\mathcal{P}) \in \{0.767, 0.933, 1.000\}$, and multi-path uncertainty $U_{MP} \in \{0.038, 0.047, 0.050\}$. The graph-based metrics were $U_D = 0.979$, $U_C = 0.016$, and $U_{CD} = 0.040$, which yielded epistemic uncertainty $U_E = 0.345$, aleatoric uncertainty $U_A \in \{0.311, 0.366, 0.389\}$, total uncertainty $U_T \in \{0.328, 0.356, 0.367\}$, and calibrated uncertainty $\phi(U_T) \in \{0.457, 0.464, 0.467\}$. These results show that the framework can consistently quantify both epistemic and aleatoric uncertainty in LLM reasoning while empirically satisfying the theoretical bounds.

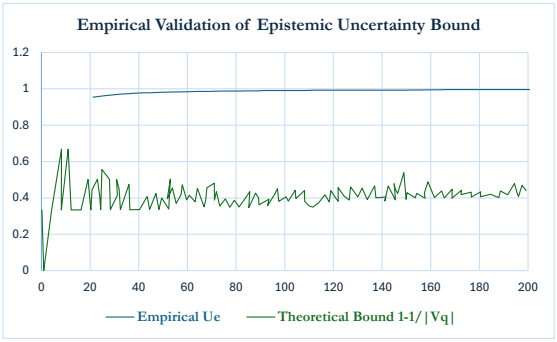

Figure 1: Empirical validation of the theoretical upper bound on epistemic uncertainty on MedQA. The scatter points represent epistemic uncertainty values computed for reasoning graphs of varying sizes, while the dashed line represents the theoretical upper bound $1 - \frac{1}{|V_q|}$ established.

4. Conclusion

This paper presents a novel graph-theoretical framework for uncertainty quantification in LLM reasoning for healthcare. By formalizing LLM reasoning as traversals through knowledge graphs, the framework defines uncertainty measures grounded in graph-theoretical properties and provides a more interpretable and theoretically rigorous alternative to conventional probabilistic approaches. We establish theoretical bounds on uncertainty estimates, compare the framework formally with existing methods, demonstrate its applicability to medical reasoning tasks such as diagnosis, treatment planning, and prognosis, and validate its behavior through simulation. Overall, the proposed framework strengthens the foundation for trustworthy AI in healthcare by enabling more reliable uncertainty estimates for high-stakes medical decision-making. Future work may extend this framework to real-world medical datasets, dynamic and multimodal knowledge graphs, and hybrid models that integrate graph-based and linguistic uncertainty measures.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 5 2021. doi: 10.1016/j.inffus.2021.05.008. URL https://doi.org/10.1016/j.inffus.2021.05.008.

Zizhang Chen, Peizhao Li, Xiaomeng Dong, and Pengyu Hong. Uncertainty Quantification for Clinical Outcome Predictions with (Large) Language Models. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 7512–7523, 1 2025. doi: 10.18653/v1/2025.findings-naacl.419. URL https://doi.org/10.18653/v1/2025.findings-naacl.419.

Mohammad Hosseini, Catherine A. Gao, David M. Liebovitz, Alexandre M. Carvalho, Faraz S. Ahmad, Yuan Luo, Ngan MacDonald, Kristi L. Holmes, and Abel Kho. An exploratory survey about using ChatGPT in education, healthcare, and research. *PLoS ONE*, 18(10):e0292216, 10 2023. doi: 10.1371/journal.pone.0292216. URL https://doi.org/10.1371/journal.pone.0292216.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2503.13939.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 1 2024. doi: 10.1109/tkde.2024.3352100. URL https://doi.org/10.1109/tkde.2024.3352100.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 7 2023. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.