Uncertainty-Aware LLM Agents for Safe Medical Decision-Making in Noisy Clinical Environments

Abhishek Banerjee
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
abhishekwork@oregonstate.edu

Abstract

Large language model (LLM) agents are increasingly being considered for clinical decision support, yet their deployment in noisy hospital environments raises fundamental concerns about safety, reliability, and governance. This paper proposes a system-level framework for uncertainty-aware LLM agents that can operate robustly under the high-variability conditions typical of real-world clinical settings. We argue that current LLM architectures lack formal mechanisms to quantify and communicate epistemic and aleatoric uncertainty, leading to overconfident recommendations that may endanger patients. Drawing on principles from probabilistic machine learning, human-in-the-loop design, and socio-technical systems theory, we present a multi-layered architecture that integrates uncertainty estimation, deferral protocols, and continuous monitoring. We examine structural trade-offs between autonomy and oversight, the role of regulatory infrastructure, and the challenges of fairness across diverse patient populations. By comparing uncertainty-aware approaches in autonomous driving and financial risk assessment, we derive lessons for clinical deployment. The paper further addresses sustainability implications of running large models in resource-constrained healthcare environments and discusses policy frameworks for certification and liability. We conclude that uncertainty-aware LLM agents, while not a panacea, represent a necessary evolution toward trustworthy AI in medicine, provided they are embedded within robust institutional governance structures.

Keywords

uncertainty quantification, large language models, clinical decision support, safe AI, socio-technical systems, healthcare infrastructure.

1. Introduction

The integration of large language models into clinical workflows promises to augment diagnostic accuracy, streamline documentation, and support treatment planning [1]. However, the deployment of these models in noisy clinical environments—characterized by incomplete records, sensor errors, varying staff expertise, and unpredictable patient presentations—introduces profound safety risks [2]. Traditional AI systems in medicine, such as those for image analysis, have been validated under controlled conditions, but LLMs operate on natural language inputs that are inherently ambiguous and context-dependent [3]. The absence of explicit uncertainty quantification in most LLM agents leads to a dangerous illusion of certainty: when a model outputs a recommendation without indicating its confidence, clinicians may over-rely on that output, especially under time pressure [4]. This paper argues that uncertainty-aware LLM agents, which can estimate and communicate their own predictive uncertainty, are a critical requirement for safe medical decision-making in noisy settings. We adopt a system-level perspective that encompasses not only algorithmic design

but also governance, infrastructure, deployment, and sustainability. By examining the structural trade-offs between agent autonomy and human oversight, we propose an architecture that balances efficiency with safety. We also consider the implications for fairness, as uncertainty estimates may vary across demographic groups and data distributions [5]. The paper is structured as follows. Section 2 reviews related work in uncertainty quantification for neural networks and clinical AI. Section 3 describes the proposed system architecture. Section 4 discusses governance and policy implications. Section 5 examines robustness and fairness challenges. Section 6 addresses deployment and sustainability. Section 7 presents case illustrations and cross-domain comparisons. Section 8 concludes with future directions.

2. Background and Related Work

Uncertainty quantification has long been studied in machine learning, with early work on Bayesian neural networks providing theoretical foundations for estimating epistemic and aleatoric uncertainty [6]. In clinical settings, uncertainty-aware models have been applied to medical imaging and time-series analysis, where probabilistic outputs have been shown to improve calibration and safe decision-making [7]. The advent of large language models has introduced new challenges: the massive parameter counts and autoregressive generation make full Bayesian inference intractable, and most deployment pipelines rely on deterministic sampling or simple temperature scaling [8]. Recent research has explored approaches such as conformal prediction, Monte Carlo dropout, and ensemble methods to obtain uncertainty estimates from LLMs [9]. However, these methods are often evaluated on benchmark tasks rather than real clinical workflows, and their computational overhead can be prohibitive in resource-constrained environments [10]. Furthermore, the notion of uncertainty in language generation is not simply about probability but about semantic coherence and factual accuracy [11]. In medical decision-making, uncertainty must be communicated to clinicians in an interpretable manner, ideally with actionable deferral suggestions [12]. The work by Hu (2026) highlights security enhancement methods for adversarial robustness in LLM agents for medical tasks, underscoring the need to protect uncertainty mechanisms from manipulation [8]. While the present paper focuses on uncertainty awareness rather than adversarial robustness, we recognize that safe deployment requires both. Beyond technical approaches, the socio-technical literature emphasizes that AI systems in healthcare must be embedded within organizational routines, error reporting systems, and regulatory oversight [13]. The "noisy clinical environment" refers not only to data noise but to the chaotic interplay of human factors, interruptions, and resource limitations [14]. Therefore, our framework integrates technical uncertainty quantification with governance structures that support appropriate reliance and accountability.

3. System Architecture for Uncertainty-Aware LLM Agents

We propose a multi-layered architecture that enables an LLM agent to estimate, represent, and act upon uncertainty in real time. The first layer is the perception and input processing module, which handles noisy and incomplete clinical data. This module standardizes free-text notes, lab results, and sensor feeds, and flags missing or contradictory information. The second layer is the uncertainty estimation engine, which computes both epistemic uncertainty (model uncertainty due to limited training data) and aleatoric uncertainty (inherent randomness in the data). For epistemic uncertainty, we recommend an ensemble of LLMs with different random initializations or fine-tuned on different subsets of clinical corpora, combined with a Bayesian aggregation method. For aleatoric uncertainty, we use a calibration

module that estimates the entropy of the output distribution adjusted for input ambiguity. The third layer is the decision policy layer, which uses the computed uncertainty estimates to decide whether to generate a recommendation, defer to a human clinician, or request additional information. This policy is informed by a risk threshold that can be adjusted based on the criticality of the clinical scenario (e.g., higher thresholds for emergency interventions). The fourth layer is the communication and interface layer, which presents the recommendation along with a confidence interval, a qualitative description of uncertainty, and a deferral option. This layer also logs the interaction for post-hoc analysis. The architecture is designed to be modular so that components can be updated independently—for instance, the uncertainty estimation engine can be replaced with a conformal prediction module as methods mature [15]. A key structural trade-off is between the computational cost of running multiple models and the latency requirements of clinical decision support. In noisy environments, we argue that the cost of overconfidence is far greater than the cost of deferral; thus, uncertainty-aware systems should be tuned to be conservative. Additionally, the system must handle longitudinal interactions: as the same patient is monitored over time, uncertainty estimates should incorporate previous feedback cycles to reduce epistemic uncertainty gradually [16]. This architecture aligns with the principle of "human-on-the-loop" rather than "human-in-the-loop," where the agent operates autonomously only within predefined uncertainty bounds, and escalates when those bounds are exceeded.

4. Governance and Policy Implications

The deployment of uncertainty-aware LLM agents in clinical settings raises profound governance questions. First, who is liable when an agent's recommendation leads to patient harm? If the agent expresses high uncertainty and the clinician overrides or ignores it, liability may rest with the clinician. But if the agent expresses low uncertainty and is later found to be wrong, the system designer and the deploying institution share responsibility [17]. Regulatory bodies such as the FDA have begun to develop frameworks for AI/ML-based medical devices, but these frameworks largely assume deterministic outputs with fixed performance metrics [18]. Uncertainty-aware systems challenge such approaches because performance can vary with input noise and calibration. We propose a governance model that requires continuous post-market surveillance of uncertainty calibration across different clinical sites. Institutions should maintain an audit trail that records every agent recommendation along with its uncertainty estimate, the clinician's response, and the patient outcome. This data can be used to recalibrate the system and to identify systematic biases. Furthermore, policy must address fairness: if an LLM agent's uncertainty estimates are systematically higher for underrepresented populations due to training data imbalances, then the agent may defer more often for those patients, leading to disparate treatment [19]. Regulatory guidelines should mandate that uncertainty calibration be evaluated across demographic subgroups and that thresholds be adjusted to ensure equitable deferral rates. Another governance dimension is data privacy: the uncertainty estimation engine may require access to raw patient data, raising concerns under HIPAA and GDPR. A possible solution is to use on-device deployment or federated uncertainty quantification, where the uncertainty computation is performed locally without transmitting sensitive data [20]. Finally, the introduction of uncertainty-aware agents must be accompanied by training programs for clinicians to interpret uncertainty estimates correctly—a challenge that has been studied in the context of AI-assisted diagnosis [21].

5. Robustness and Fairness in Noisy Clinical Environments

Noisy clinical environments degrade model performance in multiple ways. Input noise can arise from transcription errors, missing values, or ambiguous language (e.g., “chest pain” may refer to cardiac, musculoskeletal, or gastrointestinal causes). An uncertainty-aware LLM agent should be robust to such perturbations: the uncertainty estimation engine should detect when input noise increases aleatoric uncertainty and respond by raising deferral rates. However, traditional robustness metrics like adversarial accuracy do not directly translate to clinical safety. Instead, we propose a “safe region” concept: a set of input conditions under which the agent’s uncertainty is guaranteed to be calibrated. Outside this region, the agent must defer. This safe region can be learned from historical data and updated as the environment changes. Fairness implications emerge because noise is not uniformly distributed: patients with rare diseases, those in under-resourced clinics, or those using non-standard terminology may experience higher input noise, leading to systematically higher uncertainty estimates and thus more deferrals. This could paradoxically reduce access to AI-assisted care for the very populations that might benefit most. To mitigate this, the architecture should include a fairness-aware calibration step that adjusts uncertainty thresholds for known biases in training data. Additionally, the system should be resilient to adversarial manipulation: an attacker could craft inputs to artificially reduce the model’s uncertainty estimate, causing the agent to give a confident but wrong recommendation. The work by Hu (2026) on adversarial robustness for LLM medical agents directly addresses this vulnerability, proposing security enhancement methods that protect the uncertainty estimation pipeline from manipulation [8]. Integrating such adversarial defenses is crucial, as the clinical environment may include malicious actors (e.g., hacking of electronic health records) or inadvertent inputs from stressed staff. Robustness also includes temporal stability: the agent’s uncertainty estimates should not fluctuate erratically with minor changes in input, as that would erode clinician trust. We recommend using a smoothing filter over recent uncertainty estimates in longitudinal monitoring.

6. Deployment and Sustainability Considerations

Deploying large language models in healthcare settings requires substantial computational infrastructure, which may be unavailable in low-resource hospitals. Uncertainty-aware agents, which often use ensembles or Monte Carlo methods, exacerbate computational demands. A sustainable deployment strategy involves tiered model sizes: a small, fast model for routine cases with low uncertainty, and a larger, more accurate model for complex cases or when uncertainty is high. This tiered approach reduces energy consumption and latency. Moreover, healthcare institutions can adopt federated deployment where uncertainty computations are partially done on edge devices, with cloud fallback for high-uncertainty situations. Sustainability also encompasses the cost of continuous monitoring and recalibration. We argue that uncertainty-aware systems can actually reduce overall computational cost by minimizing unnecessary full model runs: if the system detects high uncertainty early, it can defer to a human without generating a full recommendation. Another sustainability dimension is the carbon footprint of training and fine-tuning LLMs. While this paper focuses on inference, we note that training uncertainty-aware models requires not just base LLM training but also training the uncertainty estimation module. Transfer learning and parameter-efficient fine-tuning can mitigate this. From a policy perspective, payers (insurance companies, government health programs) need to decide whether to reimburse for AI-assisted decisions that incorporate uncertainty. One model is to reimburse only when the agent provides a recommendation within its safe region, with lower reimbursement for deferred cases. This creates economic incentives for uncertainty-aware design. Finally, clinician acceptance is a

major sustainability factor: if uncertainty estimates are perceived as unhelpful or frequently wrong, clinicians will ignore them. User studies should inform the design of uncertainty communication—for example, using color-coded confidence bars instead of raw probabilities [22].

7. Case Illustrations and Cross-Domain Comparisons

To ground the architecture, consider a case of sepsis prediction in an intensive care unit. An LLM agent processes nursing notes, vital signs, and lab trends. A traditional agent might output "sepsis likely" with high confidence even if the patient's history is atypical. An uncertainty-aware agent would output the same prediction but also report high epistemic uncertainty because the training data contained few similar cases. The agent would then suggest a consult with a senior intensivist. This deferral prevents a false positive that could lead to unnecessary antibiotics. Over time, as similar cases are documented and feedback is given, the epistemic uncertainty decreases for that profile. Another case: radiology report generation. The LLM generates a descriptive report from an image. If the image quality is poor (high aleatoric uncertainty), the agent flags regions of the report with low confidence and prompts the radiologist to review those specific findings. Cross-domain comparisons are illuminating. In autonomous driving, uncertainty-aware systems have been mandatory for safety: vehicles must estimate the probability of collision and trigger emergency braking. The automotive industry has developed formal verification methods to guarantee that uncertainty estimates are conservative under worst-case scenarios [23]. Clinical applications can adopt similar formal guarantees, but the complexity of medical reasoning makes full verification infeasible. In financial risk assessment, uncertainty quantification is used to set capital reserves; similarly, clinical systems could set "safety reserves" such as mandatory double-checks when uncertainty exceeds a threshold. However, financial models operate on structured data with clear loss functions, whereas clinical outcomes are multifaceted. Another domain is legal AI, where uncertainty about case law is displayed to judges. Studies show that presenting uncertainty can improve decision accuracy when the uncertainty is well-calibrated [24]. These comparisons reinforce the need for interdisciplinary design that borrows safety engineering principles while respecting the uniqueness of medicine.

8. Conclusion

Uncertainty-aware LLM agents represent a critical step toward safe and trustworthy AI in clinical environments. This paper has presented a system-level architecture that integrates uncertainty estimation, deferral policies, and robust governance. We have argued that deploying LLMs without explicit uncertainty awareness is risky in noisy clinical settings, where data quality and model applicability are variable. The proposed architecture balances autonomy with oversight through a tiered decision policy that escalates when uncertainty exceeds thresholds. Governance measures including continuous monitoring, demographic fairness evaluation, and liability frameworks are essential for responsible adoption. Sustainability concerns, particularly computational cost, can be mitigated through tiered model deployment and energy-efficient uncertainty estimation. Cross-domain lessons from autonomous driving and finance highlight the feasibility of uncertainty-aware systems but also the need for clinical-specific adaptations. Future research should focus on developing interpretable uncertainty communication methods, improving adversarial robustness as highlighted by Hu (2026) [8], and conducting large-scale clinical trials to validate safety and utility. Ultimately, the safe deployment of LLM agents in medicine depends not only on

algorithmic advances but on the socio-technical infrastructure that embeds them within human decision-making processes.

References

1. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

2. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

3. Liu, Y., Shi, Z., Wei, Y., & Jiang, H. (2024). Large language models in healthcare: A systematic review. Journal of Biomedical Informatics, 149, 104578.

4. Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. Science, 373(6554), 284–286.

5. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

6. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1050–1059.

7. Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports, 7(1), 17816.

8. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

9. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

10. Mena, J., Pujol, O., & Vitrià, J. (2021). A survey on uncertainty estimation in deep learning. Artificial Intelligence Review, 54, 5935–6002.

11. Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. Proceedings of the 11th International Conference on Learning Representations.

12. Raghu, M., Blumer, K., Corrado, G., & Kleinberg, J. (2019). The algorithmic automation problem: Prediction, triage, and human effort. arXiv preprint arXiv:1903.12220.

13. Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Quality and Safety in Health Care, 19(Suppl 3), i68–i74.

14. Carayon, P., & Wood, K. E. (2010). Patient safety: The role of human factors and systems engineering. Studies in Health Technology and Informatics, 153, 23–46.

15. Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. Advances in Neural Information Processing Systems, 32.

16. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452–459.

17. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

18. U.S. Food and Drug Administration. (2021). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). FDA.

19. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91.

20. Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

21. Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 33, 7089–7096.

22. Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305.

23. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374.

24. Varshney, K. R. (2019). Trustworthy machine learning and artificial intelligence. XRDS: Crossroads, 25(3), 26–29.

25. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.