Uncertainty-Aware Robustness Enhancement of Large Language Model Agents for High-Stakes Medical Diagnosis and Treatment Recommendation

Nils Bowers
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
contactnils@buffalo.edu

Siddharth J. Prasad
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
siddharthwork@binghamton.edu

Fedfrey Waber
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
hellojeffrey@ucf.edu

Enzo Mills
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
enzomail@oregonstate.edu

Abstract

Large language model (LLM) agents are increasingly being proposed for clinical decision support, yet their deployment in high-stakes medical diagnosis and treatment recommendation remains fraught with unresolved uncertainty challenges. This paper presents a systems-level analysis of uncertainty-aware robustness enhancement for LLM agents operating in clinical environments. We examine the interplay between epistemic and aleatoric uncertainty sources, adversarial vulnerabilities, and the architectural trade-offs inherent in designing agents that can detect, quantify, and appropriately communicate uncertainty. A central argument advanced is that robustness cannot be attained through post-hoc calibration or prompting alone; instead, it requires deeply integrated architectural components such as Bayesian reasoning modules, retrieval-augmented evidence fusion, reinforcement learning with explicit safety constraints, and federated learning frameworks that preserve patient privacy while enabling continuous model improvement. The discussion extends to infrastructure requirements, including latency-accuracy trade-offs in real-time clinical settings, data governance, and the alignment of LLM agent behavior with evolving regulatory standards. We further address fairness considerations, highlighting how uncertainty-aware mechanisms can mitigate disparate performance across demographic subgroups by flagging low-confidence decisions for human review. Long-term sustainability and maintenance of these systems in hospital workflows are examined through the lens of model drift, concept shift, and the need for institutional oversight structures. By synthesizing insights from machine learning, medical informatics, and socio-technical systems theory, the paper offers a roadmap for building trustworthy LLM agents that do not merely generate plausible text but actively manage the limits of their own knowledge in life-critical settings.

Keywords

large language model agents, medical decision-making, uncertainty quantification, adversarial robustness, clinical decision support, system safety, fairness.

1. Introduction

The rapid evolution of large language models has generated considerable interest in their application as intelligent agents for medical diagnosis and treatment recommendation. These models, when augmented with tool-use capabilities, memory, and long-term planning, can potentially integrate disparate clinical data sources, interpret imaging reports, summarize patient histories, and propose therapeutic strategies with a fluency that mimics expert consultation. Yet the very feature that makes them compelling—their capacity to generate coherent, contextually appropriate language—also masks a fundamental brittleness. In high-stakes medical contexts, a confident but erroneous recommendation can result in irreversible harm, making uncertainty awareness and robustness not merely desirable properties but absolute prerequisites for clinical deployment. As LLM agents move from laboratory demonstrations into pilot clinical workflows, the systems community must confront a constellation of interdependent challenges that span model architecture, infrastructure design, governance frameworks, and long-term sustainability.

Existing approaches often treat uncertainty quantification as a separate post-processing step, applying calibration techniques after the model has already generated an output. This layer of retrofitted caution is insufficient when the agent's internal representations fail to capture the multimodal, temporally evolving, and often contradictory nature of clinical data. A more profound integration is necessary, one in which uncertainty estimation is woven into the agent's reasoning pipeline, influencing not only the final recommendation but also the decision to defer to human clinicians, to request additional diagnostic information, or to adjust the confidence level conveyed in its communication. Equally important is robustness against adversarial perturbations, which in a medical setting may arise from unintentional data corruption, systematic biases in electronic health records, or deliberate manipulation of input text designed to mislead the model into dangerous conclusions. The consequences of such failures extend beyond individual patient safety to institutional liability, erosion of clinician trust, and the potential exacerbation of healthcare disparities.

This paper undertakes a comprehensive examination of the systems-level requirements for developing uncertainty-aware, robust LLM agents in medical diagnosis and treatment recommendation. We ground the analysis in a broad body of prior work spanning uncertainty estimation in deep learning, adversarial robustness in natural language processing, and the socio-technical realities of clinical practice. By synthesizing these perspectives, we articulate a set of architectural principles and infrastructure design choices that can collectively enhance the trustworthiness of LLM agents. The discussion emphasizes structural trade-offs, such as the tension between computational overhead and real-time responsiveness, between centralized learning and privacy-preserving decentralization, and between model autonomy and human oversight. Throughout, we maintain a forward-looking stance, identifying open problems and policy implications that will shape the next generation of medical AI systems.

2. Foundational Challenges in Deploying LLM Agents for Clinical Decision-Making

2.1. Sources of Epistemic and Aleatoric Uncertainty

A clinical LLM agent faces two broad categories of uncertainty. Epistemic uncertainty, or model uncertainty, arises from limitations in the training data, architectural capacity, or the agent's incomplete understanding of rare disease presentations. Aleatoric uncertainty stems

from inherent noise in the data itself—ambiguous symptom descriptions, inconsistent laboratory measurements, and the stochastic nature of disease progression. In standard classification tasks, these uncertainties can be partially captured through predictive entropy or ensemble disagreement, but an LLM agent that generates free-text recommendations introduces additional layers of complexity. The linguistic surface of a recommendation can convey certainty through phrasing even when the underlying probability distribution is flat, and explicitly requesting the model to verbalize its uncertainty often produces miscalibrated self-assessments.

Early work on Bayesian neural networks demonstrated that placing distributions over model parameters can provide principled epistemic uncertainty estimates, but scaling such methods to the size of modern LLMs remains computationally prohibitive. Approximate methods such as Monte Carlo dropout, deep ensembles, and Laplace approximations have been adapted for transformer architectures, yet their integration into agentic pipelines requires careful coordination with the planning and tool-use modules that characterize LLM agents. When an agent retrieves relevant studies or lab values, the uncertainty associated with retrieval relevance and source reliability must be propagated alongside the model's internal uncertainty, a challenge that calls for holistic uncertainty propagation frameworks rather than isolated point estimates.

2.2. Robustness Vulnerabilities and Adversarial Threats

Medical LLM agents are susceptible to both inadvertent and adversarial input perturbations. A minor typographical error in a patient's medication list or a seemingly innocuous rephrasing of a symptom can shift the model's recommendation from a conservative management plan to an aggressive surgical intervention. More concerning are targeted adversarial attacks crafted to exploit the model's linguistic sensitivities, a threat that becomes increasingly plausible as these agents are integrated into public-facing triage systems or automated prescription platforms. Research on adversarial examples in natural language processing has revealed that even state-of-the-art models can be misled by semantically equivalent paraphrases, synonym substitutions, or carefully placed distracting sentences. In the medical domain, such vulnerabilities directly endanger patient safety and undermine the legal defensibility of AI-assisted decisions.

Security-oriented investigations have begun to address the specific challenges of adversarial attacks against LLM-based medical decision agents. Recent work has proposed security enhancement methods that strengthen the model's resilience through adversarial training, input sanitization, and anomaly detection modules designed to identify manipulative patterns before they reach the generation stage [9]. However, a purely security-focused lens must be broadened to encompass the socio-technical dimensions of robustness: an agent may be algorithmically robust against known attack vectors yet still fail in practice due to distribution shifts, data drift, or the sheer diversity of real-world clinical language. A comprehensive robustness strategy must therefore combine defensive training with continuous monitoring, feedback loops that capture near-miss events, and institutional processes for rapid model patching.

3. Architectural Paradigms for Uncertainty-Aware LLM Agent Systems

3.1. Bayesian Inference and Ensemble-Based Calibration

Integrating uncertainty quantification directly into the agent architecture requires a departure from the standard autoregressive generation paradigm. One promising direction is to equip the

LLM core with a Bayesian inference layer that maintains a distribution over possible interpretations of the clinical context. This can be implemented through lightweight adapter modules trained to output both a point estimate and a variance measure for each generated token, allowing the agent to internally flag segments of a recommendation that are highly uncertain. Ensemble methods, in which multiple model instances or checkpoints are queried in parallel, offer another practical route. Disagreement among ensemble members can serve as a signal to initiate a clarification dialogue with the clinician or to append a confidence disclaimer to the output. The computational cost of ensembles must be weighed against the safety benefits, and hybrid architectures that activate ensembling only for high-risk decision categories—such as chemotherapy regimen selection or surgical candidacy assessments—can strike a balance between latency and reliability.

Beyond token-level uncertainty, there is a need for decision-level calibration that considers the downstream clinical workflow. An agent that accurately quantifies its uncertainty but fails to translate it into actionable caution is of limited value. Architectural designs that link uncertainty thresholds to predefined escalation protocols, such as automatic referral to a specialist or suppression of a recommendation entirely, transform uncertainty awareness from a passive diagnostic feature into an active safety mechanism. Such designs require close collaboration between system engineers and clinical domain experts to determine appropriate thresholds, a process that must be periodically revisited as the model is updated and as clinical guidelines evolve.

3.2. Reinforcement Learning with Human Feedback and Safety Constraints

Reinforcement learning from human feedback (RLHF) has become a standard technique for aligning LLM behavior with human preferences, yet its application in medical contexts demands an extension of the preference model to explicitly capture uncertainty and risk tolerance. Clinicians do not simply prefer accurate answers; they prefer answers that are appropriately qualified, that acknowledge ambiguous evidence, and that err on the side of caution when data are insufficient. Training an agent to internalize these nuanced preferences requires a reward model that penalizes overconfidence and rewards honest expressions of uncertainty. Furthermore, hard safety constraints can be encoded through constrained policy optimization, ensuring that the agent never recommends contraindicated drug combinations or exceeds safe dosage limits regardless of its confidence level.

The integration of safety constraints with uncertainty awareness opens up a design space in which the agent's policy can dynamically adjust its behavior: when uncertainty is low and constraints are satisfied, the agent may offer a definitive recommendation; when uncertainty is high, it may pivot to a differential diagnosis format that enumerates possibilities along with their estimated probabilities. This dynamic policy regime mirrors the cognitive strategies of experienced clinicians who modulate their decisiveness based on the clarity of the clinical picture, and embedding such strategies into the agent architecture promotes a more natural and trustworthy human-agent interaction.

3.3. Retrieval-Augmented Generation and Evidence Fusion

Retrieval-augmented generation (RAG) architectures, which allow LLMs to ground their outputs in external knowledge bases, offer a powerful mechanism for reducing epistemic uncertainty by anchoring recommendations in up-to-date medical literature, clinical practice guidelines, and institutional protocols. However, the retrieval step introduces its own uncertainty: the relevance ranking of retrieved documents may be noisy, the documents

themselves may contain conflicting evidence, and the provenance of information must be traceable for medico-legal accountability. An uncertainty-aware RAG agent must therefore maintain a meta-level representation of retrieval confidence and fuse evidence from multiple sources in a manner that reflects the strength and consistency of the underlying data.

Such evidence fusion can be approached through hierarchical attention mechanisms that weight retrieved passages not only by relevance but also by the credibility of the source—distinguishing, for example, between a randomized controlled trial and an anecdotal case report. When evidence is contradictory, the agent should explicitly surface this disagreement rather than selectively summarizing the majority view. Architectural components that perform structured evidence synthesis, perhaps generating a brief literature review alongside a recommendation, can enable supervising clinicians to independently assess the quality of the agent's reasoning. This transparency transforms the agent from an opaque oracle into a collaborative decision support tool whose outputs are accompanied by a visible audit trail.

4. System-Level Design and Infrastructure Considerations

4.1. Data Governance, Privacy, and Federated Learning

Deploying uncertainty-aware LLM agents in clinical environments demands a data governance framework that reconciles the need for continuous model improvement with stringent patient privacy protections. Centralized aggregation of clinical data for retraining is often infeasible due to regulatory constraints and institutional data silos. Federated learning provides a compelling alternative, allowing individual hospitals to collaboratively train shared models without exchanging raw patient records. Differential privacy guarantees can be layered onto federated updates to bound the information leakage from model gradients. However, the heterogeneity of clinical data across sites—differences in patient demographics, documentation practices, and disease prevalence—introduces additional uncertainty that must be communicated to end users. An agent trained via federated learning should ideally report not only its own predictive uncertainty but also the distributional characteristics of the training data, enabling clinicians to gauge whether the model's knowledge base adequately reflects their local patient population.

The infrastructure must also support fine-grained access control and audit logging, recording every query, recommendation, uncertainty score, and clinician override. Such logs serve dual purposes: they enable retrospective analysis of model failures and near misses, and they provide the evidentiary foundation for regulatory inspections. The design of these logging systems must balance completeness against storage costs and must be resilient to tampering, requirements that align with established practices in high-integrity healthcare IT systems.

4.2. Scalable Deployment and Latency-Accuracy Trade-offs

Real-time clinical workflows impose strict latency requirements. A triage agent that takes several seconds to generate a recommendation with well-calibrated uncertainty may be viable in a primary care setting, but an intraoperative decision support system must deliver guidance within sub-second time scales. The computational intensity of ensemble methods, Bayesian inference layers, and retrieval-augmented generation can conflict with these latency budgets. System designers must therefore engineer tiered inference pipelines that trade off depth of uncertainty analysis against response time. A fast, approximate uncertainty estimate can be computed first, with a more thorough analysis triggered only when the initial estimate exceeds a critical threshold or when the clinical scenario is flagged as high-risk by a lightweight rule-based pre-filter.

Model compression techniques such as quantization, distillation, and speculative decoding can reduce the runtime footprint of uncertainty-aware modules, but these optimizations must be carefully validated to ensure that they do not inadvertently degrade the quality of uncertainty estimates. Edge deployment on hospital-local servers, rather than reliance on cloud APIs, can further reduce latency and address data sovereignty concerns, though it shifts the burden of hardware maintenance onto healthcare institutions. This shift necessitates standardized containerized deployment models and continuous integration pipelines that deliver model updates without disrupting clinical operations, an area where the broader DevOps community's practices can be adapted to the unique safety requirements of medical AI.

5. Fairness, Accountability, and Regulatory Alignment

Uncertainty awareness has a critical role to play in promoting fairness across demographic subgroups. LLMs trained on historically biased clinical data may exhibit lower accuracy for underrepresented populations; a conventional agent might produce incorrect recommendations with high confidence, thereby perpetuating health disparities. An uncertainty-aware agent, by contrast, can detect regions of the input space where its training data are sparse or where performance on a particular subgroup is known to degrade, and it can respond by lowering its confidence, requesting additional data, or deferring to a human clinician. This dynamic recalibration of decisional authority based on uncertainty estimates operationalizes a form of procedural fairness that does not require explicit demographic parity constraints on every recommendation. However, the design of the uncertainty estimation modules themselves must be scrutinized for bias, as miscalibrated uncertainty can systematically disadvantage certain groups if the model is overconfident for the majority population and underconfident for minorities.

Accountability frameworks for LLM agents must delineate the responsibilities of model developers, deploying institutions, and supervising clinicians. When an adverse event occurs following an agent's recommendation, the ability to trace the decision pathway—including the uncertainty estimates that were presented to the clinician and the clinician's response—is essential for fair assignment of liability. This traceability requirement has implications for system architecture, as it demands that uncertainty scores and their provenance be embedded in the permanent medical record in a tamper-evident format. Regulatory bodies such as the U.S. Food and Drug Administration are beginning to articulate expectations for adaptive AI systems, and uncertainty awareness is likely to feature prominently in future guidance, particularly as it relates to the requirement for predetermined change control plans that describe how models will be monitored and updated post-deployment.

6. Sustainability and Long-Term Maintenance

The long-term sustainability of medical LLM agents depends on institutional capacity for ongoing monitoring, retraining, and governance. Clinical knowledge evolves continuously, with new guidelines, drug approvals, and emerging disease patterns rendering static models obsolete. Uncertainty-aware agents possess an advantage in this regard, as they can detect when their knowledge is becoming stale by monitoring shifts in the distribution of their own uncertainty signals. A gradual increase in epistemic uncertainty across a class of queries may indicate that the underlying clinical evidence base has changed, triggering a review and potential retraining cycle. Designing monitoring dashboards that track population-level uncertainty metrics, rather than individual query logs, can provide operational intelligence to

hospital IT departments and clinical governance committees without overwhelming them with data.

Sustainability also encompasses the environmental and financial costs of retraining large models. Incremental learning strategies that update agent components without full retraining, combined with modular architectures where evidence retrieval modules can be updated independently of the language model core, offer a path toward more sustainable maintenance. Furthermore, the development of shared uncertainty benchmarks and validation suites across institutions can distribute the cost of safety assurance while fostering interoperability. The eventual establishment of multi-institutional consortia for continuous safety monitoring of medical AI agents, analogous to pharmacovigilance networks for drugs, would represent a mature socio-technical infrastructure for long-term governance.

7. Conclusion

The integration of large language model agents into high-stakes medical diagnosis and treatment recommendation demands a fundamental reorientation from performance-centric evaluation to safety-centric system design. Uncertainty awareness, when implemented as a core architectural property rather than an afterthought, can serve as the connective tissue linking robustness, fairness, accountability, and sustainability. This paper has argued that such awareness must span the entire decision pipeline, from Bayesian and ensemble-based uncertainty quantification at the model level to evidence fusion in retrieval-augmented architectures, from safety-constrained policy optimization to federated learning frameworks that respect privacy while enabling continuous improvement. The structural trade-offs analyzed—between latency and depth of uncertainty analysis, between centralized and decentralized learning, between model autonomy and human oversight—highlight that no single architectural choice can satisfy all clinical contexts, and that a portfolio of configurable uncertainty management strategies is required.

Infrastructure design must evolve in parallel, with institutional governance structures that leverage uncertainty signals for dynamic risk management, equitable care delivery, and regulatory compliance. Looking forward, the maturation of LLM agents in medicine will depend not only on technical innovation but also on the cultivation of interdisciplinary standards, shared validation resources, and a culture of safety that treats uncertainty not as an embarrassing limitation but as clinically actionable information. By embracing this perspective, the research community can guide the development of intelligent agents that genuinely enhance, rather than merely simulate, expert medical judgment.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics.

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).

3. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2022). Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

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

5. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

6. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

7. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (pp. 39–57). IEEE.

8. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine, 15(11), e1002686.

9. 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.

10. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.

11. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

12. 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.

13. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

14. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474).

16. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

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

18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

19. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231–237.

20. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265.

21. Schulam, P., & Saria, S. (2019). Can you trust this prediction? Auditing pointwise reliability after learning. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (pp. 1022–1031). PMLR.

22. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning (pp. 1885–1894). PMLR.