Blockchain-Enabled Secure Collaboration of Multi-Agent Large Language Model Systems for Clinical Decision Support

Maxime Fdwards
Department of Computer Science, University of North Texas, Denton, TX, USA.
maximemail@unt.edu

Tobias Simpson
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
tobiass@uab.edu

Abstract

The integration of large language models into clinical decision support introduces transformative potential for diagnostic reasoning, treatment planning, and patient communication. However, real-world deployment of such models within high-stakes medical environments demands robust mechanisms for inter-model coordination, auditability, privacy preservation, and adversarial resilience. This paper proposes a system-level architecture in which multiple specialized large language model agents, each acting as a distinct clinical reasoning entity, collaborate through a blockchain-mediated coordination protocol. The blockchain fabric serves not merely as a tamper-evident log but as a decentralized governance layer that enforces access policies, tracks provenance of medical inferences, and enables cryptographically assured consensus among autonomous agents. We examine the structural trade-offs inherent in coupling multi-agent language intelligence with distributed ledger technology, including latency, throughput, and the tension between transparent audit trails and patient data confidentiality. A central analytical focus concerns the adversarial robustness of language model agents operating in clinical contexts, where manipulated inputs or coordinated poisoning attacks could subvert collective decision processes. Through a cross-domain lens that integrates insights from federated learning, decentralized identity management, and distributed systems governance, the paper discusses architectural principles for sustainable, fair, and regulation-compliant multi-agent clinical intelligence. The analysis extends to policy implications, infrastructure requirements, and the long-term sustainability of such hybrid computational ecosystems. Ultimately, the paper argues that blockchain-enabled secure collaboration offers a pathway toward verifiable, resilient, and ethically governed clinical AI, while also delineating the substantial engineering and institutional challenges that must be overcome.

Keywords

clinical decision support, large language models, multi-agent systems, blockchain, secure collaboration, adversarial robustness, governance, decentralized infrastructure.

1. Introduction

The rapid advancement of large language models (LLMs) has begun to reshape the landscape of clinical decision support, offering capabilities that span automated differential diagnosis

generation, literature synthesis, and patient-specific treatment recommendations. Architectures such as the transformer model [1] and subsequent instantiations exemplified by GPT-3 [2] have demonstrated emergent reasoning capabilities that, when carefully directed, can augment clinical cognition. Yet the translation of these models into operational healthcare environments necessitates a departure from monolithic, single-model deployments. Real-world clinical reasoning is a distributed cognitive process involving multiple specialties, institutional perspectives, and evidentiary standards. A multi-agent architecture composed of domain-specific LLM agents, each acting with bounded expertise and agency, more naturally mirrors the collaborative nature of medical decision-making. At the same time, the sensitivity of patient data, the requirement for explanation and auditability, and the threat landscape that includes adversarial manipulation demand an infrastructure that goes beyond traditional client-server or centralized orchestration models. The present work examines the architectural intersection of multi-agent LLM systems and blockchain technologies, proposing a secure collaboration framework in which a consortium blockchain provides the coordination substrate for clinical reasoning agents. Throughout this analysis, the focus remains on system-level properties: structural trade-offs, governance, sustainability, fairness, and the policy dimensions that arise when autonomous language agents are entrusted with clinical influence.

The central motivation for integrating blockchain into multi-agent clinical reasoning is not technological novelty but the resolution of deep-seated trust and coordination failures. In a hospital network or a regional health information exchange, different institutions and care teams may operate their own fine-tuned LLM agents that reflect local population demographics, clinical guidelines, and privacy constraints. Federating these agents through a centralized hub often reintroduces a single point of control and a potential bottleneck for both performance and policy enforcement. A blockchain-based coordination layer, by contrast, distributes the verification of agent outputs, the tracing of decision provenance, and the enforcement of access control policies across nodes that represent legally and clinically independent stakeholders. This design aligns with the broader movement toward decentralized health data governance exemplified by MedRec [6] and related blockchain-based medical record frameworks. However, the addition of LLM agents introduces novel dimensions: the models themselves are probabilistic black boxes that can be co-opted, and their collaborative interactions create an emergent decision surface that must be defended against both data-level and model-level adversarial threats. Consequently, the paper dedicates considerable attention to the intersection of adversarial robustness and multi-agent coordination, drawing on recent investigations into security enhancement methods for adversarial robust LLM agents in medical decision-making tasks [11]. The aim is to provide a synthetic, forward-looking analysis that bridges system architecture, security engineering, and health policy.

2. Architectural Foundations of Multi-Agent LLM Systems for Clinical Contexts

The architecture of a multi-agent clinical reasoning system extends the widely adopted paradigm of single-model LLM inference by introducing a population of specialized agents, each with a defined role, access scope, and reasoning protocol. A typical configuration includes a generalist agent that synthesizes patient history, a specialist agent that focuses on a particular disease domain, a pharmacology agent that evaluates drug interactions, and an evidence-based medicine agent that grounds recommendations in current clinical guidelines. These agents communicate through structured message formats that encapsulate partial clinical assessments, confidence estimates, and supporting evidence references. The design

challenge is not merely in training or fine-tuning such agents—techniques for domain adaptation of foundation models are well established [3, 4]—but in architecting the coordination layer so that the emergent collective behavior is clinically safe, auditable, and resilient to both random errors and deliberate subversion.

A fundamental structural trade-off in multi-agent LLM systems lies between deliberation richness and decision latency. Clinical workflows demand timely decisions, yet high-quality collaborative reasoning often requires iterative exchanges: an agent's initial impression may be refined when confronted with a contradictory differential diagnosis from a specialist agent, and this refinement might in turn trigger a re-evaluation by the evidence agent. If every such interaction is mediated through a human clinician for review, the system forfeits much of its efficiency gain. Conversely, if agents autonomously update their internal states and converge on a collective recommendation, the absence of human oversight raises questions of accountability. Architectural decisions about the degrees of autonomy, the quorum mechanisms for consensus, and the extent of human-in-the-loop intervention become central governance parameters. The deployment of a multi-agent system thus requires careful modeling of information flow topologies—whether hierarchical, peer-to-peer, or hybrid—and each topology carries distinct implications for security, fairness, and the propagation of errors. For instance, a strictly hierarchical topology might accelerate decision-making but create a single point of adversarial influence, whereas a fully peer-to-peer topology distributes risk but complicates convergence guarantees.

Furthermore, multi-agent clinical systems introduce a socio-technical dimension in which the division of cognitive labor must reflect the existing jurisdictional boundaries of medical practice. An agent representing radiology should not autonomously override the assessment of an agent representing pathology without explicit conflict resolution rules, and these rules must be encoded in both the software middleware and the institutional policies that govern the system's operation. The architecture must therefore expose policy enforcement points that are inspectable by clinical governance committees and regulators. The use of blockchain as a policy enforcement and event-sourcing layer provides a mechanism to transform these architectural requirements into cryptographically verifiable state transitions. The multi-agent architecture itself remains agnostic to the underlying consensus mechanism, but the choice of coordination substrate profoundly affects the system's ability to enforce role-based constraints and to resist Byzantine behaviors among agents.

3. Blockchain-Enabled Coordination and Trust

3.1 The Blockchain as a Coordination Substrate

Blockchain technology offers a decentralized append-only ledger whose most critical property for multi-agent clinical collaboration is not monetary tokenization but tamper-evident state replication among mutually distrustful stakeholders. In a clinical multi-agent deployment, each participating institution—be it a hospital, a reference laboratory, or a public health authority—operates validating nodes that collectively maintain a ledger recording agent-generated clinical assertions, provenance metadata, and access authorization changes. This design transforms the coordination problem from one of pairwise secure channels and centralized orchestration to one of distributed consensus on a shared history of clinical reasoning. Approaches such as proof-of-authority or practical Byzantine fault tolerance, already explored in healthcare blockchain literature [7, 10], offer throughput and permissioning models that align with the closed-participant nature of clinical consortia, circumventing the energy and latency profiles of public proof-of-work networks.

The blockchain's role as a coordination substrate extends beyond mere logging. Smart contracts, executed deterministically on all validating nodes, can encode the conditions under which one agent's output may be integrated into the collective clinical assessment. For example, a smart contract might specify that the medication recommendation agent's output must be cross-validated by at least two specialist agents before it is appended to the permanent clinical record and forwarded to the electronic health record system. Because smart contracts run on every node, their execution provides a transparent and auditable means of automating governance rules that might otherwise be opaque or inconsistently applied. This introduces, however, a notable trade-off: the granularity of on-chain logic directly affects the performance and complexity of the system. Fine-grained contractual logic capturing detailed clinical pathways can create smart contracts that are expensive to validate and hard to update, while coarse-grained logic risks leaving critical coordination gaps.

3.2 Audit Trails, Provenance, and Confidentiality

Clinical decisions must be accompanied by rich provenance records that enable retrospective review in the context of medical liability, quality improvement, and regulatory compliance. A blockchain ledger provides a natural mechanism for constructing immutable audit trails, whereby each contribution from an LLM agent is timestamped and linked to the cryptographic identifiers of the agent version, the fine-tuning dataset, and the institutional operator. Such provenance records are essential for attributing responsibility in the event of an adverse outcome and for satisfying regulatory frameworks such as the FDA's proposed approaches for AI-based software as a medical device. However, the permanence of on-chain data directly conflicts with established privacy principles, including the right to erasure under data protection regulations. This tension has motivated hybrid architectures where only hashes, zero-knowledge proofs, or anonymized references to off-chain data stores are placed on the ledger [25, 21]. In the multi-agent LLM context, this approach can be extended by having agents exchange encrypted reasoning traces through decentralized storage networks while the blockchain records attestations of their exchange and the outcomes of consensus validation steps. The system thus gains the integrity and non-repudiation benefits of the ledger without exposing clinical content to permanent replication across all nodes.

The design of the confidential audit trail necessitates a clear separation between on-chain metadata and off-chain content, along with carefully managed encryption key hierarchies. Each clinical encounter could be associated with a temporary session key that is made available to authorized agents via distributed key generation schemes, ensuring that decryption rights are time-limited and tied to the specific care context. After the clinical episode concludes, the off-chain content can be moved to long-term archival storage under the control of the patient's designated data steward, leaving only a compact verifiable credential on the blockchain that attests to the existence and integrity of the complete record without exposing its substance. Such hybrid architectures illustrate the nuanced infrastructural decisions that distinguish a practical blockchain-enabled multi-agent system from a purely theoretical one.

4. Security and Adversarial Robustness in Clinical Multi-Agent Systems

Clinical decision support systems, by virtue of their influence on patient care, present an attractive target for adversaries ranging from financially motivated fraudsters to actors seeking to undermine public trust in health institutions. Adversarial attacks against LLMs have been extensively studied in the general domain, covering input perturbations, data poisoning, and model extraction [12, 13]. In a multi-agent clinical setting, the attack surface

expands considerably because an adversary may compromise a single agent—either by subverting its fine-tuning pipeline or by crafting adversarially perturbed prompts—and use it as a vector to poison the collective reasoning process. The interdependent nature of multi-agent deliberation means that a corrupted recommendation from a specialist agent can cascade through the iterative consensus protocol, biasing the outputs of other agents even if those agents remain individually uncompromised. This cascade effect is particularly dangerous when the adversarial modifications are subtle and exploit the well-known overconfidence and hallucination tendencies of LLMs.

Recent research aimed at strengthening the security posture of LLM agents in medical contexts has explored adversarial training, defensive distillation, and input sanitization techniques specifically tailored to clinical language [11, 14, 15]. The integration of these defensive measures into a multi-agent system raises a governance question: which entities are responsible for certifying the robustness of each agent, and how is that certification continuously validated in a decentralized consortium? One viable model involves periodic adversarial robustness audits whose results are recorded on the blockchain as part of the agent's identity credential. Before another agent incorporates a peer's reasoning output, a smart contract can verify that the peer's credential has not expired and that its recent audit score exceeds a consortium-defined threshold. This machinery transforms adversarial robustness from a static property of a model artifact into a dynamic, collectively enforced system property. Yet the scheme introduces its own risks: if the audit process itself relies on a specific adversarial testing toolkit, any vulnerability in that toolkit could be exploited to grant fraudulent credentials to a compromised agent.

Differential privacy and federated learning techniques further shape the security and privacy posture of the collaboration. Federated learning permits multiple institutions to collaboratively fine-tune their LLM agents without sharing raw clinical data, a paradigm that has gained traction in digital health [20, 24, 19]. However, gradient leakage and membership inference attacks remain practical concerns [23, 18]. A blockchain can organize the rounds of federated learning, log contribution weights, and enforce secure aggregation protocols through smart contracts that require multiple validators to attest to the integrity of each update. In such a setup, the ledger records not the gradients themselves but commitments and zero-knowledge proofs of correct aggregation, allowing the consortium to detect and attribute poisoning attempts without exposing intermediate model information. This approach aligns with broader visions of confidential computing in healthcare consortia and exemplifies how blockchain infrastructure can serve a security coordination function rather than a data storage function.

5. Governance, Fairness, and Ethical Dimensions

The deployment of a blockchain-enabled multi-agent LLM system for clinical decisions necessarily crosses multiple jurisdictional and institutional boundaries, raising profound governance challenges. Governance in this context encompasses the rules for admitting new agents, the allocation of voting power in consensus protocols, the processes for updating smart contract logic, and the mechanisms for resolving disputes when an agent's recommendation is later found to have contributed to patient harm. Because the blockchain enforces rules algorithmically, the initial design of governance parameters carries outsized long-term consequences. A consortium that grants disproportionate voting weight to large academic medical centers, for example, risks entrenching systemic biases and marginalizing

community hospitals that serve underrepresented populations, thereby compromising the fairness of the collective intelligence.

Fairness analysis in multi-agent clinical systems must examine both procedural fairness in agent interactions and outcome fairness across patient demographics. Procedural fairness can be partially addressed through transparent allocation of block validation rights and through diversity requirements encoded in the smart contracts that select which agents participate in a given clinical deliberation. Outcome fairness is more elusive, as it depends on the composition of the training data for each agent, the demographic representativeness of the cases discussed within the consortium, and the feedback loops created when clinical recommendations influence subsequent training data collection. The immutability of the blockchain ledger provides a permanent record of which agents participated in which decisions and what data they referenced, enabling retrospective fairness audits that would be difficult to conduct in a purely ephemeral coordination system. However, the same immutability raises concerns about the perpetuation of biased historical patterns if on-chain governance mechanisms are not explicitly designed to facilitate iterative policy evolution. Mechanisms such as versioned smart contracts and time-bound governance charters, modifiable through a predetermined supermajority vote, offer a pathway to balance stability with adaptability.

Ethical dimensions extend beyond fairness to encompass informed consent, the right to opt out of AI-mediated decision pathways, and the delineation of responsibility between human clinicians and machine agents. When a multi-agent deliberation produces a treatment recommendation that a physician overrides, the blockchain records both the agent-generated recommendation and the physician's final decision, together with the rationale entered by the clinician. This dual record creates a robust evidentiary trail that can support both quality assurance and medicolegal analysis. Designing consent interfaces that adequately explain to patients the nature of multi-agent AI deliberation constitutes a further socio-technical challenge that intersects with user experience design, health literacy, and regulatory requirements for explainability. The system architecture does not itself solve these ethical tensions, but it must be built to accommodate evolving ethical standards without requiring complete infrastructural overhauls, a design property that argues for modular and parameterized governance layers.

6. Sustainability, Infrastructure, and Deployment Realities

Long-term sustainability of a blockchain-enabled clinical collaboration platform depends on economic, technical, and organizational factors that extend well beyond initial prototype deployment. The computational overhead of maintaining a distributed ledger across multiple clinical sites, combined with the inference cost of large language models and the communication overhead of multi-agent deliberation, creates a resource profile that must be carefully managed to avoid disproportionate energy consumption and operating expense. While permissioned blockchain architectures avoid the extreme energy demands of proof-of-work networks, the aggregate cost of running secure enclaves, performing zero-knowledge proof generation, and storing on-chain attestations remains significant. Economic sustainability models may involve membership fees, transaction-based micro-accounting internal to the consortium, or cost-sharing calibrated to the volume of patient encounters processed. These models must be transparent to regulatory auditors and equitable across participants, avoiding a situation where smaller providers are priced out of collaboration.

From an infrastructural standpoint, the platform must interface with existing hospital information systems, laboratory information management systems, and electronic health record platforms. Such integration demands standardized APIs, robust identity and access management, and failover mechanisms that can sustain clinical operations during network partitions or blockchain consensus delays. The technical community has developed middleware solutions that bridge HL7 FHIR-based health data systems with blockchain networks, yet the incorporation of LLM agents introduces additional latency and reliability requirements. During a partial network outage, a multi-agent system must either degrade gracefully, temporarily transferring authority to local fallback reasoning pipelines, or halt decision support until quorum can be reestablished. Neither option is ideal, and the design choice must be documented in the system's risk management file and approved by the relevant regulatory bodies.

Deployment realities also include the heterogeneous legal and regulatory landscapes across jurisdictions. A multi-state or multi-national consortium must reconcile differing data localization laws, liability standards for AI-based medical devices, and rules for cross-border health data flows. The blockchain's role as a global state machine can conflict with requirements that certain data never leave a national territory. Consequently, practical deployments frequently adopt a network-of-networks model where multiple regional blockchain instances interoperate through notary schemes or inter-ledger protocols, with the LLM agents themselves orchestrated at a higher abstraction layer. Achieving interoperability without centralization of trust remains a formidable research and engineering challenge, but one that must be confronted if blockchain-enabled clinical collaboration is to mature beyond isolated pilot projects. The governance structures established in the early phases of such consortia will heavily influence the ability to navigate these cross-jurisdictional complexities over the system's lifecycle.

7. Conclusion

This paper has presented an integrative analysis of blockchain-enabled secure collaboration for multi-agent large language model systems in clinical decision support, examining the architecture, security, governance, and sustainability dimensions through a system-level lens. The proposed framework locates the blockchain not as a panacea but as a specialized coordination and trust layer that addresses the fragmented ownership, adversarial risk, and auditability requirements peculiar to multi-stakeholder clinical intelligence. By structuring the collaboration around a consortium ledger, the system gains tamper-evident provenance, automated policy enforcement through smart contracts, and decentralized resilience against single points of compromise. The inclusion of adversarial robustness mechanisms as a continuously verified credential, recorded and enforced on-chain, represents a promising direction for defending the emergent decision surface of interacting language agents. Nevertheless, the analysis underscores persistent tensions: the trade-off between ledger transparency and patient confidentiality, the governance complexity of updating distributed clinical logic, the fairness implications of institutional power imbalances encoded in consensus protocols, and the long-term economic viability of computationally intensive hybrid architectures. Addressing these tensions requires not only engineering advances but also sustained interdisciplinary collaboration among computer scientists, clinicians, ethicists, regulators, and health system administrators. Future work should empirically evaluate latency, throughput, and adversarial resilience metrics in realistic multi-agent clinical simulations, while also engaging with the policy community to develop adaptive regulatory frameworks

that can keep pace with the rapid co-evolution of large language models and decentralized infrastructure.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

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

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

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

5. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

6. Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. In 2016 2nd International Conference on Open and Big Data (OBD) (pp. 25–30). IEEE.

7. Kuo, T. T., Kim, H. E., & Ohno-Machado, L. (2017). Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association, 24(6), 1211–1220.

8. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (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).

9. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. In Proceedings of Machine Learning and Systems, 1, 374–388.

10. Li, X., Jiang, P., Chen, T., Luo, X., & Wen, Q. (2020). A survey on the security of blockchain systems. Future Generation Computer Systems, 107, 841–853.

11. 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.

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

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

14. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE Symposium on Security and Privacy (pp. 582–597). IEEE.

15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.

16. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2022). Adversarial training for language models. arXiv preprint arXiv:2207.14287.

17. Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D., & Goldstein, T. (2019). Adversarially robust transfer learning. In International Conference on Learning Representations.

18. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (pp. 3–18). IEEE.

19. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322–1333).

20. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

21. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318).

22. Zhang, K., Tian, Z., & Li, T. (2018). Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. IEEE Transactions on Industrial Informatics, 16(6), 4177–4186.

23. Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., & Wang, F. (2021). Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5(1), 1–19.

24. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

25. Chen, Y., Ding, S., Xu, Z., Zheng, H., & Yang, S. (2020). Blockchain-based medical records secure storage and medical service framework. Journal of Medical Systems, 44(1), 1–9.