Federated Adversarial Training of Large Language Model Agents for Distributed Healthcare Systems

Trevor D. Douglas
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
trevor1981@uab.edu

Ole Lopez
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
ole98@buffalo.edu

Abstract

The integration of large language models into distributed healthcare infrastructures introduces unprecedented capabilities for clinical decision support, patient communication, and administrative automation, yet simultaneously exposes these systems to adversarial vulnerabilities that can undermine patient safety and data integrity. This paper proposes a federated adversarial training framework specifically designed for large language model agents operating across decentralized healthcare networks. The framework orchestrates collaborative adversarial training among multiple institutional nodes while preserving strict data locality constraints imposed by healthcare privacy regulations. We examine the architectural trade-offs between model robustness, communication efficiency, and governance compliance within a federated paradigm, emphasizing the role of adversarial example generation and distribution as a shared public good. The paper further analyzes the structural implications of deploying such agents in real-world clinical environments, including the need for continuous monitoring, fairness across heterogeneous patient populations, and the mitigation of distribution shift from participating institutions. Policy-oriented considerations regarding certification of robustness, liability for adversarial failures, and the socioeconomic barriers to participation are discussed. Through a cross-domain comparison with federated learning in other sensitive domains, we identify the unique challenges posed by the generative and interactive nature of language models. The paper concludes with a forward-looking perspective on sustainable adversarial defense mechanisms that balance utility, privacy, and equity in federated healthcare systems. This work contributes a system-level blueprint for the robust and responsible deployment of large language model agents in distributed critical-care environments.

Keywords

federated learning, adversarial training, large language models, healthcare systems, distributed infrastructure, robustness, governance, privacy, clinical decision support, socio-technical systems.

1. Introduction

The adoption of large language models (LLMs) in healthcare promises to transform clinical workflows by enabling natural language interfaces for electronic health record summarization, diagnostic reasoning support, and patient-facing chatbots [1], [2]. However, the very capabilities that make LLMs powerful also render them susceptible to adversarial

manipulation, where carefully crafted perturbations to input prompts or training data can induce harmful outputs [3], [4]. In a distributed healthcare system comprising multiple hospitals, clinics, and research institutions, the attack surface expands significantly: each node may host a locally deployed LLM agent that communicates with a central orchestrator or with peer agents, creating opportunities for cross-node poisoning, evasion, and model extraction [5]. Centralized adversarial training, while effective in controlled settings, becomes infeasible when data cannot be aggregated due to privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe [6].

Federated learning offers a natural solution by enabling collaborative model improvement without raw data exchange [7]. Extending this paradigm to adversarial training, however, introduces novel challenges: adversarial examples must be generated locally, their distributions may differ across institutions due to population heterogeneity, and the communication overhead of sharing gradient information for adversarial robustness can be prohibitive. Moreover, the generative nature of LLMs—where outputs are sequences of tokens rather than scalar predictions—complicates the definition of a universal adversarial objective. This paper develops a federated adversarial training framework tailored to LLM agents in healthcare, addressing the interplay between robustness, privacy, and system governance. We analyze architectural choices such as synchronous versus asynchronous aggregation, differential privacy budgets for adversarial updates, and the role of a central coordinator in certifying robustness at the system level. By situating the technical discussion within the socio-technical reality of healthcare delivery, we highlight the governance mechanisms necessary to translate algorithmic robustness into clinical trustworthiness.

2. Background and Related Work

Federated learning has been extensively studied in medical imaging and genomics, where data cannot be centralized due to privacy and regulatory constraints [8], [9]. Standard approaches like FedAvg aggregate model weights from participating clients after local training, but they assume i.i.d. data distributions and benign clients—both assumptions rarely hold in practice. Adversarial training, originally developed for image classifiers, augments the training set with adversarial examples to improve robustness against small perturbations [10]. Extending adversarial training to LLMs has proven challenging because of the discrete nature of text tokens and the high cost of generating semantically meaningful perturbations [11], [12]. In the healthcare domain, recent work has demonstrated that LLMs used for clinical decision-making can be fooled by adversarial prompts that alter a few words in a patient history, leading to incorrect diagnoses or dangerous treatment recommendations [11]. This vulnerability is especially concerning in a distributed setting where multiple agents might be compromised through a single poisoned model update.

Several prior studies have attempted to combine federated learning with adversarial training. These efforts have largely focused on classification tasks and have not addressed the unique properties of generative language models [13], [14]. Furthermore, existing frameworks often assume that all clients share a common adversarial threat model, which is unrealistic in healthcare where different institutions serve populations with distinct demographic profiles and disease prevalence. The work by Hu [11] is particularly relevant as it investigates security enhancements for adversarial robust LLM agents in medical decision-making tasks, highlighting the need for tailored defenses that account for domain-specific adversarial patterns. Our framework builds on these insights by embedding adversarial training within a

federated architecture that respects data locality and institutional autonomy. We also draw inspiration from the broader literature on trustworthy AI, which emphasizes the importance of transparency, auditability, and fairness alongside robustness [15], [16].

3. System Architecture for Federated Adversarial Training

The proposed architecture consists of a central coordination server and multiple client nodes, each representing a healthcare institution that hosts a local LLM agent. The server maintains a global model that is periodically distributed to clients for local fine-tuning. In each communication round, every client generates a set of adversarial examples using its local patient data, applies adversarial training to update its local model, and then sends the updated model parameters (or gradient updates) back to the server. The server aggregates these updates using a robust aggregation rule designed to tolerate Byzantine behavior and data heterogeneity [17]. A critical design choice is the method for generating adversarial examples locally. Because healthcare data are high-dimensional and contain sensitive structured and unstructured elements, we advocate for using a combination of gradient-based attacks (e.g., fast gradient sign method adapted for token embeddings) and task-specific heuristics that inject plausible clinical misspellings or paraphrases [11], [12]. The adversarial examples must remain clinically valid—they should not alter the medical meaning beyond the perturbation necessary to test robustness.

The server also plays a role in certifying the overall robustness of the aggregated model. This certification can be achieved through formal verification techniques that bound the impact of input perturbations on model outputs, or through empirical evaluation using a held-out adversarial test set that is centrally generated from a small public corpus [18]. The frequency of communication and the amount of data shared represent a fundamental trade-off. Frequent updates improve convergence and robustness but increase communication costs and the risk of gradient inversion attacks, where a malicious server could reconstruct patient-level information from the parameter updates. To mitigate this, we incorporate differential privacy by clipping and adding noise to the local updates before transmission [19]. The noise level is calibrated to the desired privacy budget, which must be acceptable to clinical regulatory bodies. Additionally, institutions may have varying computational capabilities and data volumes; asynchronous aggregation protocols allow each node to participate at its own pace, reducing the risk of excluding resource-constrained hospitals.

4. Training Dynamics and Robustness Guarantees

The federated adversarial training process is inherently non-stationary: each client's local training landscape shifts as the global model evolves and as new adversarial examples are generated. Convergence analysis under such conditions is complex, but we can draw on results from non-convex optimization and robust statistics. The key insight is that the adversarial objective—a minimax optimization between the model weights and the worst-case perturbation—can be decomposed into a sum of local objectives, each representing a different data distribution [14]. The global robustness of the aggregated model depends on the diversity and coverage of the adversarial examples contributed by all clients. Institutions serving rare diseases or minority populations may generate adversarial examples that are crucial for achieving equitable robustness across the entire federated system. Therefore, the aggregation rule must be weighted not only by data size but also by the diversity of adversarial patterns [20].

A major challenge is that adversarial training can degrade performance on benign inputs, a phenomenon known as the robustness-accuracy trade-off [10]. In healthcare, this trade-off is particularly acute because even a small drop in diagnostic accuracy on clean data can have severe clinical consequences. The framework addresses this by using a mixed training objective that combines the standard cross-entropy loss on benign examples with an adversarial loss on perturbed examples, with a hyperparameter that controls the relative importance. This hyperparameter can be tuned per client based on the local prevalence of adversarial threats; for example, a node that has historically experienced targeted attacks might assign higher weight to adversarial loss. Furthermore, the server can periodically evaluate the global model on a held-out validation set that includes both benign and adversarial examples to monitor the balance and adjust the training schedule accordingly.

5. Governance and Policy Implications

Deploying federated adversarial training in healthcare raises governance questions that extend beyond technical optimization. The first concerns the certification of robustness. Regulatory bodies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) require that medical software demonstrate safety and effectiveness before approval. Robustness against adversarial inputs must be a component of that certification, but there is currently no standardized testing protocol for LLM-based clinical systems [15]. The federated nature of our framework complicates certification because the deployed model at a given institution may differ from the globally aggregated model due to local fine-tuning or asynchronous updates. One approach is to require each institution to publish a robustness certificate that quantifies the worst-case behavior of its locally deployed agent under a set of predefined adversarial constraints. The central coordinator can then aggregate these certificates to provide a system-wide assurance.

The second governance dimension is accountability. If an adversarial attack on a federated LLM agent causes patient harm, who is liable? The institution that deployed the agent? The developer of the global model? The clients that contributed adversarial examples? The lack of clear legal frameworks creates uncertainty that may hinder adoption. We argue for a shared-responsibility model where each participant is responsible for the security of its local data and the integrity of its updates, while the central coordinator assumes liability for systemic failures resulting from the aggregation mechanism [16]. This model aligns with existing practices in medical device networks where manufacturers and hospitals share risk.

Third, fairness across institutions with unequal resources must be addressed. Wealthier hospitals with larger datasets and more powerful computing infrastructure can generate more effective adversarial examples and benefit more from federated training, potentially leaving under-resourced clinics with less robust models. A federated system can mitigate this by implementing a fairness-aware aggregation scheme that gives higher weight to updates from underrepresented nodes, or by providing compute credits to smaller institutions through a public funding mechanism. Additionally, the adversarial examples themselves can be viewed as a public good; creating a shared repository of clinically realistic adversarial patterns, anonymized to protect privacy, could help all participants improve robustness [11].

6. Case Studies and Deployment Scenarios

We consider two deployment scenarios to illustrate the practical implications of the framework. The first is a consortium of regional hospitals collaborating to train a clinical LLM for discharge summarization. Each hospital has its own electronic health record system

and patient demographics. Under the federated adversarial training framework, each hospital generates adversarial examples by perturbing key clinical entities such as medication names, dosages, and diagnosis codes. The central server aggregates updates using a policy that requires at least three hospitals to agree on the direction of weight updates before incorporating them, protecting against a compromised node. Over several rounds, the global model becomes robust to perturbations that cause it to omit critical discharge instructions or hallucinate contraindicated medications. The consortium also establishes an adversarial oversight board that reviews the generated adversarial examples for clinical plausibility and ensures that no patient identifiable information is leaked through the gradient updates.

The second scenario involves a national telehealth network that uses LLM agents for triage and patient education. Here, the threat model includes adversarial prompts from end users attempting to bypass triage rules or extract confidential advice. Because telehealth patients interact directly with the LLM, the system must be robust to both intentional attacks and unintentional misinputs. Federated adversarial training is uniquely suited because different telehealth providers serve distinct populations with varying language patterns and cultural contexts. A provider serving a large non-native English speaking population, for instance, might generate adversarial examples involving common grammatical errors that could trigger misinterpretations. The central coordinator can use a meta-learning approach to adapt the global robustness strategy to the linguistic diversity observed across providers, ensuring that the model does not become overfitted to a dominant dialect [21].

7. Discussion and Future Directions

The framework presented here is not without limitations. The communication cost of transmitting full model parameters for LLMs with billions of parameters is currently prohibitive for many healthcare institutions with limited bandwidth. Parameter-efficient fine-tuning methods, such as adapters or low-rank updates, can reduce this overhead but may also limit the capacity for adversarial robustness because they modify only a small portion of the model [22]. Another open challenge is the dynamic nature of adversarial threats: new attack strategies emerge as the model is deployed, requiring continuous adversarial training that must be orchestrated across the federated network without disrupting clinical operations. Online adversarial learning, where the model updates in response to observed attacks in real time, could be integrated but introduces additional privacy risks because the attack patterns may reveal sensitive patient information.

Future research should explore the use of zero-shot and few-shot robustness transfer across domains. For instance, if a hospital has trained a robust LLM for radiology report generation, can that robustness be transferred to a different hospital without sharing data? This would require disentangling robustness features from domain-specific patterns, a challenging but potentially valuable direction. Additionally, the intersection of federated adversarial training with explainability is promising: clinicians need to understand why a model is robust to certain perturbations but not others, and explanations could inform trust calibration [23]. Finally, the ethical dimension of adversarial training itself warrants scrutiny. Generating adversarial examples that deliberately cause a model to err can be considered a form of automated red-teaming; however, if these examples are not properly anonymized, they could inadvertently expose vulnerabilities that might be exploited by malicious actors. A governance protocol for the ethical generation, storage, and deletion of adversarial examples must be part of the framework.

8. Conclusion

Federated adversarial training of large language model agents offers a pathway to deploy robust, privacy-preserving AI assistants across distributed healthcare systems. By combining the strengths of federated learning with adversarial training techniques adapted for generative models, the framework addresses the critical need for resilience against malicious inputs while respecting the regulatory and ethical constraints of clinical environments. The structural trade-offs between robustness, communication efficiency, and governance complexity require careful orchestration, but the potential benefits—safer clinical decisions, equitable robustness across institutions, and sustained trust in AI-enabled healthcare—justify the investment. As LLMs become more deeply integrated into the fabric of healthcare delivery, the principles articulated here can serve as a foundation for building systems that are not only intelligent but also secure and just.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

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

3. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 2153–2162).

4. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8018–8025.

5. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics (pp. 2938–2948). PMLR.

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

7. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

8. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Bakas, S. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1–7.

9. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 1–12.

10. 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.

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. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 31–36).

13. Shah, A., Bhagoji, A. N., Chaterji, S., & Huang, F. (2022). Federated adversarial training with multi-objective optimization. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (pp. 2721–2734).

14. Zizzo, G., Rawat, A., & Sinn, M. (2020). Federated adversarial training for robust machine learning. In International Workshop on Federated Learning for User Privacy and Data Confidentiality at ICML.

15. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

16. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

17. Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 30.

18. Wong, E., & Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In International Conference on Machine Learning (pp. 5283–5292). PMLR.

19. 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).

20. Mohri, M., Sivek, G., & Suresh, A. T. (2019). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615–4625). PMLR.

21. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning (pp. 1126–1135). PMLR.

22. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790–2799). PMLR.

23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

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

25. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. In International Conference on Learning Representations.