Federated Adversarial Training for Privacy-Preserving Robust Large Language Model Agents in Distributed Medical Decision Support Systems

Nanoj Gheakur
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
manojthakur50@ucf.edu

Milos Stanley
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
miloswork@colostate.edu

Clifford R. Gran
Department of Computer Science, University of Houston, Houston, TX, USA.
green1995@uh.edu

Abstract

The rapid integration of large language model (LLM) agents into clinical decision support systems promises transformative advances in diagnostic accuracy, treatment personalization, and operational efficiency. However, the deployment of such systems across distributed healthcare networks introduces profound challenges at the intersection of data privacy, model robustness, and regulatory compliance. Centralizing sensitive patient data for training is often infeasible under frameworks such as HIPAA and GDPR, while LLM agents remain vulnerable to adversarial manipulations that can induce harmful clinical errors. This paper presents a comprehensive system-level investigation of federated adversarial training as a unified paradigm for cultivating privacy-preserving yet robust LLM agents within distributed medical decision support infrastructures. We analyze the architectural design space encompassing secure aggregation, differential privacy, and adversarial example generation orchestrated across heterogeneous clinical sites. The discussion extends to structural trade-offs between communication efficiency, model utility, and resilience to both data-poisoning and evasion attacks in the linguistic domain. We examine the complex interplay between adversarial robustness mechanisms and privacy leakage, highlighting how federated optimization can amplify or mitigate membership inference risks. Furthermore, we address infrastructure sustainability, computational resource allocation, edge-cloud orchestration, and the carbon footprint of continuously retraining medical LLM agents. Governance challenges, fairness across demographically diverse populations, and the alignment of federated adversarial training with evolving regulatory instruments for AI-based medical devices are critically evaluated. The paper concludes by outlining forward-looking policy and design recommendations to bridge the gap between theoretical robustness guarantees and operational medical realities.

Keywords

federated learning, adversarial training, large language models, medical decision support, privacy preservation, robustness, distributed systems, healthcare AI.

1. Introduction

The digitization of healthcare records, the proliferation of wearable sensor data, and the increasing sophistication of natural language processing have converged to create unprecedented opportunities for artificial intelligence in clinical decision-making. Large language model agents, capable of parsing unstructured clinical narratives, generating differential diagnoses, and recommending evidence-based treatments, are transitioning from experimental prototypes to embedded components within hospital information systems. Yet this transition unfolds against a backdrop of fragmented data ownership, stringent privacy regulations, and a threat landscape in which malicious actors can craft adversarial inputs that cause model outputs to diverge dangerously from medical truth. A centralized model training paradigm, in which raw protected health information is aggregated into a single datacenter, is not only legally fraught but also increasingly viewed as an unacceptable concentration of risk.

Federated learning emerged as a compelling alternative, enabling collaborative model optimization without exposing individual patient records. In parallel, adversarial training has become a cornerstone for hardening deep learning models against intentionally perturbed inputs. However, the intersection of these two methodologies, particularly when applied to the linguistic and multi-turn reasoning capabilities of LLM agents, remains a largely underexplored territory. The naive application of adversarial training within a federated setting introduces new vectors of tension: the extra gradient computations required for robust optimization can strain hospital edge resources, the sharing of adversarial examples can inadvertently leak information about local data distributions, and the heterogeneity of medical data across sites can undermine the uniformity of adversarial robustness guarantees. This paper undertakes a detailed, system-oriented examination of federated adversarial training for privacy-preserving robust LLM agents in distributed medical decision support systems. It does not propose a novel algorithmic variant; rather, it interrogates the structural, infrastructural, governance, and policy dimensions that collectively determine whether such an approach can be responsibly and sustainably operationalized. The analysis moves beyond narrow technical metrics to encompass fairness, sustainability, regulatory alignment, and long-term societal impact.

2. Background and Related Work

The foundational concept of federated learning was introduced by McMahan et al. [1], who demonstrated that decentralized stochastic gradient descent could produce models competitive with centrally trained counterparts while keeping training data local. Subsequent research extended this framework to healthcare, where patient privacy is paramount. Rieke et al. [2] surveyed the landscape of federated learning in medical imaging and electronic health records, identifying differential privacy as a critical companion technique to bound information leakage from model updates. Abadi et al. [3] provided a rigorous implementation of differentially private deep learning, quantifying the privacy budget consumed during training. Meanwhile, the vulnerability of deep neural networks to adversarial examples was systematically exposed by Goodfellow et al. [4], and Madry et al. [5] formulated adversarial training as a min-max optimization problem that yields models resilient to projected gradient descent attacks. The extension of adversarial threats to natural language processing has been investigated in multiple contexts: Ebrahimi et al. [6] demonstrated character-level perturbations that deceive text classifiers, while Wallace et al. [7] revealed universal adversarial triggers capable of altering model behavior across many inputs.

Within the medical domain, the stakes of adversarial vulnerability are dramatically elevated. Finlayson et al. [8] warned that adversarial attacks on clinical AI could systematically bias

diagnostic suggestions, with potentially fatal consequences. Large language models introduce additional complexity because their instruction-following nature and in-context learning capabilities make them susceptible to prompt injection attacks, as explored by Perez and Ribeiro [9], and to data poisoning during fine-tuning, as shown by Carlini et al. [10]. The intersection of federated learning and adversarial robustness has been addressed primarily in computer vision. Zizzo et al. [11] examined the feasibility of federated adversarial training and identified degradation in clean accuracy when robustness constraints are enforced across non-identically distributed client data. However, the application to LLM-based medical decision agents, where inputs are lengthy and semantically rich clinical texts and outputs can include treatment plans, has not been scrutinized through a full-system lens. The work by Hu [11] provided an early investigation into security enhancement methods for adversarial robust LLM agents in medical decision-making, though it left open questions regarding federated integration and governance. Our contribution extends this discourse by mapping the entire ecosystem—architecture, privacy, infrastructure, fairness, and regulation—into a coherent analytical framework.

3. System Architecture for Federated Adversarial Training

A distributed medical decision support system built on federated adversarial training must reconcile two inherently conflicting engineering objectives: achieving high global model robustness against adversarial inputs while strictly compartmentalizing patient data. The architectural blueprint typically comprises a central aggregation server, either physically hosted in a cloud environment or operated by a neutral trusted third party, and multiple client nodes situated within each participating hospital's network boundary. Each client possesses a local instance of an LLM agent that is either a fine-tuned version of a general-purpose foundation model or a domain-adapted architecture pre-trained on publicly available medical corpora. Federated rounds proceed by the server distributing a base model, clients performing several local epochs of adversarial training on their private datasets, and then uploading only model updates or gradients.

The adversarial training inner loop at each client requires generating worst-case perturbations of the local clinical texts that maximize the loss of the current model checkpoint. In the language domain, this generation is not trivially a small-norm pixel change but can involve synonym substitution, paraphrasing, or insertion of medically plausible but misleading phrases while preserving the surface readability of the note. The computational cost is substantial because each local batch must be augmented with adversarial examples crafted online. Consequently, architectural decisions must balance the frequency of adversarial perturbation generation against local computational budgets. Some sites may opt for a lighter form of robustness training, such as random perturbation with early stopping, while better-resourced hospitals can afford full projected gradient descent in embedding space. This heterogeneity creates an aggregation challenge: the central server must fuse model updates trained with varying adversarial intensities, risking a diluted global robustness.

Secure aggregation protocols, such as those based on secure multi-party computation, can ensure that the central server observes only aggregated model updates and cannot reconstruct an individual hospital's contribution. However, adversarial training complicates this picture because the generated adversarial examples themselves can encode sensitive information. If a local client generates an adversarial instance by adding terms that exploit rare co-occurrences present only in a specific subpopulation's records, that perturbation, when embedded in gradients, might reveal membership properties. Differential privacy, implemented through

gradient clipping and calibrated noise addition, can mitigate this leakage, but at the cost of reducing the signal for adversarial robustness. Architectural variations that keep adversarial example generation entirely local and only share the resulting parameter gradient updates reduce immediate exposure but still fail to eliminate indirect leakage through weight updates. The system must thus be architected with a privacy-robustness co-design philosophy, where the selection of aggregation frequency, noise scale, and adversarial perturbation budgets are jointly optimized rather than treated as independent knobs.

4. Privacy-Preserving Mechanisms and Trade-offs

The privacy landscape of federated adversarial training for medical LLM agents is characterized by a multifaceted threat model that includes honest-but-curious servers, malicious clients seeking to infer other participants' data, and external adversaries who may intercept communication channels. Differential privacy, typically instantiated through the DP-SGD algorithm, offers a quantifiable guarantee that the presence or absence of any single patient record does not significantly alter the model's output distribution. Yet the min-max nature of adversarial training introduces a tension: the inner maximization step purposely magnifies the model's sensitivity to small input changes, which is precisely the property that differential privacy seeks to bound. This inherent antagonism means that simultaneously achieving a low privacy budget epsilon and high adversarial robustness requires navigating a Pareto frontier that has not yet been fully charted for LLMs.

Homomorphic encryption allows algebraic operations on encrypted gradients, enabling the central server to aggregate model updates without ever seeing their plaintext values. While this technique provides strong cryptographic privacy, the computational overhead is substantial, especially when applied to transformer-based architectures with millions of parameters. The latency introduced by encrypted operations can render frequent federated rounds impractical in time-sensitive medical settings. Furthermore, homomorphic encryption does not prevent adversarial examples generated locally from containing traces that become observable once the model is deployed, such as a backdoor trigger that activates only under specific clinical scenarios. Thus, cryptographic privacy must be supplemented with robustness measures that sanitize model behavior even after training.

The trade-off between privacy and utility has additional healthcare-specific dimensions. Overly aggressive noise injection to protect privacy can degrade the LLM agent's ability to recall rare diseases that appear only in a handful of records across the entire federation. Adversarial training, which typically reduces clean accuracy on in-distribution data, compounds this effect. The result can be a model that is both private and robust but clinically less useful for marginal populations. System designers must therefore decide whether to accept a tiered model strategy, where different versions of the LLM agent offer varying privacy-robustness-utility profiles for distinct clinical tasks, such as high-sensitivity screening versus low-risk administrative coding.

5. Robustness and Adversarial Resilience in Medical LLM Agents

Medical LLM agents are susceptible to a broad spectrum of adversarial manipulations that extend well beyond the additive pixel perturbations studied in computer vision. Data poisoning attacks occur when a malicious actor at a participating hospital injects deliberately mislabeled or toxic clinical examples into the local training set, aiming to bias the aggregated model toward harmful recommendations. In a federated setting, such a Byzantine client can be difficult to detect because the central server cannot inspect raw data. Robust aggregation

rules, such as coordinate-wise median or trimmed mean, provide resilience against a minority of adversarial clients, but these methods were designed for continuous-valued parameter updates. Language model updates exhibit complex structure, and trimming in parameter space may discard semantically important directions that happen to be outliers in a non-malicious sense.

Evasion attacks mounted at inference time present an even more challenging surface. An adversary can craft a patient note that appears innocuous to a human clinician but causes the LLM agent to omit a critical diagnostic possibility. Because medical reasoning often involves multi-hop inference across a note's sections, an adversarial insertion of a confounding phrase early in the text can cascade into an erroneous chain-of-thought. Federated adversarial training aims to harden the model against such attacks by exposing it during training to adversarially perturbed examples that reflect the types of manipulations expected at deployment. However, the diversity of medical writing styles, abbreviation conventions, and linguistic patterns across different institutions means that adversarial examples effective in one hospital may not transfer to others. This non-stationarity undermines the assumption that a single globally robust model suffices. A more resilient system architecture may incorporate local adversarial fine-tuning layers that personalize robustness profiles without sharing sensitive local perturbation strategies with the central server.

The interplay between federated optimization and the long-tailed distribution of clinical concepts also affects robustness. Rare but critical events, such as anaphylaxis or acute aortic dissection, may appear in very few training examples. Adversarial perturbations that target these rare classes can escape detection during model evaluation if test sets do not sufficiently cover tail events. A comprehensive robustness assurance program for federated medical LLM agents must therefore include continuous red-teaming, where dedicated security teams at each institution probe the deployed model with realistically constrained adversarial medical inputs and share only aggregated vulnerability metrics with the federation, preserving the confidentiality of raw attack vectors.

6. Infrastructure, Deployment, and Sustainability

The operationalization of federated adversarial training across a real-world healthcare network demands careful attention to infrastructure heterogeneity. Hospitals in a federation may range from major academic medical centers with on-premise GPU clusters to rural clinics relying on limited-edge computing devices. Adversarial training multiplies the computational requirements by a factor proportional to the number of inner maximization steps and the granularity of the text perturbation model. A full-scale implementation that generates adversarial examples at the token or embedding level for every batch can increase local training time by an order of magnitude compared to standard fine-tuning. This elevated resource consumption intersects with sustainability concerns: the carbon footprint of repeatedly retraining large transformer models is non-trivial, and adding adversarial robustness only magnifies energy expenditure. Federations must therefore adopt carbon-aware scheduling strategies, such as aligning intensive training rounds with periods of high renewable energy availability in the local grid or leveraging geographically distributed data centers with differential carbon intensities.

Deployment architectures must also address the latency requirements of clinical decision support. An LLM agent that takes several seconds to generate a differential diagnosis because it is running large-scale adversarial robustness checks at inference will face clinical resistance. One practical compromise is to decouple the robust training pipeline from the inference-time

model, maintaining a frozen, adversarially hardened base model that runs efficiently on edge devices while a parallel asynchronous pipeline continuously improves the model in the background. This dual-track approach reduces the burden on real-time clinical workflows but introduces a model versioning challenge: the deployed model may lag behind the most advanced robust checkpoint, creating a window of vulnerability that must be measured and minimized.

Sustainability also encompasses the longevity of the system in the face of evolving medical knowledge and adversarial tactics. Adversarial training must be complemented by mechanisms for life-long learning, where models are updated not only with new clinical data but also with fresh adversarial strategies as they are discovered. The federated framework, by design, supports continuous collaboration, but governance is needed to ensure that all sites contribute to and benefit from the ongoing robustness improvements equitably. A site that contributes disproportionately to adversarial robustness research within the federation may require incentive structures, whether financial or reputational, to maintain engagement. Without such sustainability mechanisms, federated networks risk attrition, leaving remaining participants with a diminished diversity of data and vulnerability insights.

7. Governance, Fairness, and Ethical Implications

The governance of federated adversarial training systems for medical LLM agents must address the tension between collective security and individual institutional autonomy. Each hospital in the federation retains sovereignty over its data, but the aggregated model's robustness properties become a shared resource that can be degraded by a single negligent or malicious participant. Governance frameworks must specify standards for local data curation, adversarial training protocols, and acceptable accuracy degradation thresholds. Participating institutions may have different risk appetites: a research hospital may tolerate a higher rate of false positives in exchange for catching rare adversarial triggers, while a community clinic might prioritize low false alarm rates. Reconciling these divergent preferences into a single federated objective requires multi-stakeholder negotiation, possibly mediated by a governance board that includes clinicians, ethicists, and patient advocates.

Fairness is a cross-cutting concern that federated adversarial training can either ameliorate or exacerbate. The distribution of medical data across hospitals often correlates with socioeconomic and demographic factors. An urban tertiary care center may have a racially diverse patient cohort, while a rural hospital may serve a predominantly white and elderly population. Federated learning aims to improve representation by pooling model knowledge without pooling data, but the adversarial training procedure can inadvertently suppress the learning of robust features for underrepresented groups if their local data contribute gradients that appear as outliers during secure aggregation. For instance, a medical term in a minority dialect or a culturally specific symptom description might be treated as an adversarial perturbation rather than a legitimate variation if the global model lacks sufficient exposure. Governance protocols must therefore include fairness audits that evaluate model performance across carefully stratified demographic slices, with adversarial robustness measured separately for each subpopulation. If disparities are detected, the federation may trigger reweighting mechanisms or targeted local fine-tuning to close the robustness gap.

Ethical accountability becomes diffuse in a federated system. When an LLM agent contributes to a misdiagnosis due to an adversarial attack that exploited a vulnerability in the aggregated model, liability is unclear. Is the institution that fielded the agent at fault, or the federation as a whole, or the original foundation model provider? Adversarial training adds

another layer: if the misdiagnosis occurred because one site's local adversarial training procedure was insufficiently rigorous, tracing the causal chain becomes a forensic challenge. Clear contractual frameworks, combined with technical provenance tracking using model cards and federated audit trails, will be essential to assign responsibility and drive corrective actions without chilling participation in collaborative networks.

8. Policy and Regulatory Considerations

The regulatory environment for medical device software, including AI-based decision support systems, is evolving rapidly. In the United States, the Food and Drug Administration (FDA) has articulated a framework for modifications to artificial intelligence and machine learning-based software as a medical device (SaMD), emphasizing that a continuous learning system must be subject to predetermined change control plans. Federated adversarial training introduces a complex case: the model evolves not through a centralized developer's controlled updates but through the decentralized contributions of multiple healthcare organizations. Regulators will require evidence that the federated system, in its entirety, meets pre-specified safety and effectiveness benchmarks even as adversarial robustness training modifies model parameters across distributed nodes. This may necessitate a radical transparency regime where each hospital's training procedures, including adversarial example generation policies and differential privacy parameters, are documented in a structured format suitable for regulatory inspection without exposing raw patient data.

The European Union's General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) in the US embody principles of data minimization and purpose limitation. Federated adversarial training aligns with these principles by keeping data localized, but regulators may probe whether the sharing of model updates constitutes a transfer of personal data. The Article 29 Working Party has noted that machine learning models can memorize individual data points, and adversarial training may inadvertently increase memorization by forcing the model to overfit to worst-case perturbations derived from sensitive records. Policy makers must consider whether additional technical safeguards, such as formal privacy verification through membership inference testing, should be mandated for any adversarial training procedure deployed in a medical context. Furthermore, the right to explanation enshrined in GDPR may conflict with the opacity of adversarially hardened LLM agents, whose internal representations are deliberately distorted to resist manipulation, potentially reducing the interpretability that clinicians rely upon for trust.

Looking forward, the emergence of AI-specific legislation, such as the proposed European AI Act, creates a risk-based classification system for AI applications. Medical decision support systems are likely to be classified as high-risk, requiring conformity assessments that cover robustness and accuracy across the entire intended distribution of inputs. Federated adversarial training could become a prescribed mitigation measure, but only if standardization bodies develop benchmarks and testing protocols that faithfully capture the linguistic adversarial threat model in medicine. International coordination will be essential because healthcare networks increasingly span borders, and a patchwork of incompatible regulations could fragment the benefits of federated learning. The development of an ISO standard for federated adversarial robustness in health AI, analogous to existing standards for functional safety in medical electrical equipment, would provide a harmonized framework that both regulators and hospital systems can adopt.

9. Conclusion

The confluence of large language model capabilities, federated learning, and adversarial robustness research opens a promising yet deeply challenging frontier for medical decision support systems. This paper has argued that federated adversarial training, as a systems paradigm, cannot be reduced to a simple algorithmic combination; it is a socio-technical construct that must be architected with simultaneous sensitivity to privacy preservation, adversarial resilience, computational sustainability, fairness, and regulatory legitimacy. The structural trade-offs between the inner maximization of adversarial training and the outer minimization of privacy loss create a tension that demands co-optimization over network topologies, aggregation protocols, and perturbation budgets. Infrastructure heterogeneity pushes design toward flexible, tiered architectures where robustness intensity is calibrated to local resources, while governance mechanisms must ensure that shared robustness gains do not exacerbate health disparities. The policy landscape, still in flux, will likely demand unprecedented transparency and auditing capabilities for AI systems that learn from distributed data while actively defending against adversaries. Future work should develop standardized evaluation suites that measure both privacy and robustness under realistic clinical adversarial scenarios, establish cross-institutional governance sandboxes to pilot federated adversarial training agreements, and engineer energy-efficient robust training pipelines suitable for resource-constrained medical settings. The vision of a resilient, privacy-respecting global network of medical AI agents is achievable only if the research community expands its focus from narrow algorithm development to the holistic system design issues outlined here.

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. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Maier-Hein, K. (2020). The future of digital health with federated learning. NPJ digital medicine, 3(1), 119.

3. 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).

4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International conference on learning representations.

5. 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.

6. 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).

7. 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 and the 9th international joint conference on natural language processing.

8. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287-1289.

9. Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. In NeurIPS ML Safety Workshop.

10. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying memorization across neural language models. 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. Zizzo, G., Rawat, A., Sinn, M., & Buesser, B. (2020). Federated adversarial learning for robust models. In Workshop on decentralized and distributed learning.

13. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (pp. 1175-1191).

14. 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.

15. 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 (SP) (pp. 582-597). IEEE.

16. Sun, Y., Ochiai, H., Sakaguchi, K., & Baral, C. (2022). Towards understanding the trade-off between robustness and accuracy in text classification. In Findings of the Association for Computational Linguistics: EMNLP 2022.

17. Chowdhury, A., Kassem, H., Padoy, N., & Varma, R. (2022). Federated adversarial learning for robust medical image analysis. In Medical image computing and computer assisted intervention – MICCAI 2022.

18. Wen, Y., Geiping, J., Fowl, L., Goldblum, M., & Goldstein, T. (2022). Fishing for user data in large-batch federated learning via gradient magnification. In International conference on machine learning.

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

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

21. Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E. J. (2017). A tutorial on Fisher information. Journal of Mathematical Psychology, 80, 40-55.

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

23. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210.