Title

A Federated and Parameter-Efficient Framework for Large Language Model Training in Medicine

Author list

Anran Li[1], Yuanyuan Chen[2], Wenjun Long[2], Yu Yin[3], Yan Hu[4], Hyunjae Kim[1], Weipeng Zhou[1], Yujia Zhou[1], Hongyi Peng[2], Yang Ren[1], Xuguang Ai[1], Zhenyue Qin[1], Ming Hu[5], Xiaoxiao Li[6], Han Yu[2], Yih-Chung Tham[7], Lucila Ohno-Machado[1], Hua Xu[1], Qingyu Chen[1,*]

Corresponding author:

[*]Qingyu Chen, qingyu.chen@yale.edu

Affiliations

1. Department of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, USA
2. College of Computing and Data Science, Nanyang Technological University, Singapore
3. Department of Earth Science and Engineering, Imperial College London, London, United Kingdom
4. McWilliams School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, USA
5. School of Computing and Information Systems, Singapore Management University, Singapore
6. School of Electrical and Computer Engineering, University of British Columbia, Canada
7. Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

1. Abstract

Large language models (LLMs) have demonstrated strong performance on standard medical benchmarks, including patient question answering, summarization, and diagnosis. To enable their use in clinical settings, LLMs are typically further adapted—through continued pretraining or post-training—using clinical data, thereby extending their applicability to medical tasks. However, most medical LLMs are trained on data from a single institution, as privacy and regulatory constraints limit cross-institutional data sharing. In practice, data from a single site cannot capture the substantial variability present in real-world healthcare, including differences in patient demographics, disease prevalence, and institutional documentation practices. Consequently, LLMs trained under this single-institution paradigm face critical limitations in generalizability and safety when deployed across heterogeneous healthcare systems.

Federated learning (FL) is a promising solution for enabling collaborative model development across healthcare institutions without compromising patient privacy. Yet applying FL to LLMs in medicine remains fundamentally limited. First, conventional FL requires transmitting the full model during each communication round, which becomes impractical for multi-billion–parameter LLMs given the limited computational resources available in most clinical environments. Second, many FL algorithms implicitly assume data homogeneity, whereas real-world clinical data are highly heterogeneous across patients, diseases, and institutional practices.

To address these limitations, we introduce **Fed-MedLoRA** and **Fed-MedLoRA+**, the first model-agnostic and parameter-efficient federated learning framework for adapting LLMs to medical applications. Fed-MedLoRA transmits only low-rank adapter parameters, substantially reducing communication and computation overhead, while Fed-MedLoRA+ further incorporates adaptive, data-aware aggregation to improve convergence under cross-site heterogeneity. As a case study, we apply the framework to clinical information extraction (IE)—a foundational task for transforming unstructured patient narratives into structured representations of medical entities and relations, with clinical applications in patient triage, adverse event monitoring, and decision support. Accuracy was assessed across five patient cohorts through head-to-head comparisons with domain-specific BERT models, zero-shot and fine-tuned LLaMA-3 and DeepSeek-R1, GPT-4o, and recent general-domain FL algorithms. Evaluation settings included (1) in-domain training and testing, (2) external validation on independent cohorts, and (3) a low-resource new-site adaptation scenario using real-world clinical notes from the Yale New Haven Health System. Practical feasibility was further examined under (1) incomplete or uneven task annotations across sites, (2) constrained training and inference resources, and (3) scalability across multiple institutions.

Fed-MedLoRA+ improved zero-shot LLM performance by up to 65% F1, outperformed single-site fine-tuning by approximately 25%, and exceeded domain-specific BERT models by over 40% on relation extraction. Both methods generalized robustly to external cohorts (10–70% F1 gains) and achieved strong performance in new-site adaptation (73% strict / 85% lenient F1), demonstrating their effectiveness in real-world multi-institutional deployment. Communication costs were reduced by 98.5% relative to full-model updates. Training was feasible on a single RTX 4090 (16 GB) for 8B models and on mid-range GPUs (e.g., RTX 3060 Ti) for 1B models, with

inference supported on standard laptops (e.g., Apple M3 Pro). The framework scaled to 10 participating sites with only ~2% performance degradation compared to centralized training.

Together, the results demonstrate that federated LLMs hold strong potential for medical applications while remaining feasible and resource efficient. We further provide actionable guidelines and outline future directions for advancing federated LLM development. All implementation code is publicly available at https://github.com/Yale-BIDS-Chen-Lab/FL_LLM_Med.

2. Introduction

Large language models (LLMs) are rapidly transforming the landscape of medicine, with extensive studies demonstrating their potential across a broad spectrum of applications, including outpatient management [1,2], disease diagnosis [3,4], and clinical summarization [5,6]. To support clinical adoption, growing efforts have focused on adapting general-domain LLMs through continued pretraining or post training on clinical data to improve accuracy and safety in downstream medical tasks [7,8].

However, in practice, many LLMs are trained or adapted using clinical data from a single institution, primarily due to privacy and governance constraints that make sharing patient-level data across healthcare systems challenging [9–12]. Yet clinical data vary substantially across institutions in patient demographics, disease prevalence, care pathways, documentation styles, and clinical practices, leading to domain shift and degraded model generalization [13–15]. Recent studies have highlighted that this represents a significant—and often underestimated—limitation in the development and deployment of LLMs in medicine [16,17]. For example, the accuracy of a locally fine-tuned LLM for hospital admission prediction dropped by more than 20% when evaluated on external patient populations [17].

Federated learning (FL) requires institutions to centralize and openly share raw patient data on a single server, thereby increasing regulatory and governance risks in healthcare[18]. In a standard FL workflow, models are trained locally at each site, model updates are transmitted to a central server for aggregation, and the updated global model is redistributed for further refinement [19,20]. Despite its growing adoption in medicine, however, adapting FL to LLMs remains fundamentally challenging [21–23]. First, conventional FL requires transmitting full model parameters at each communication round, which is infeasible for multi-billion–parameter LLMs given the constrained computational and networking resources available in most healthcare environments [24–26]. Second, real-world clinical data are highly heterogeneous across institutions, differing substantially in patient populations, disease prevalence, annotation availability, and documentation practices. Under such conditions, standard FL algorithms based on uniform parameter averaging often show unstable convergence and biased performance across sites [24,27].

To date, although recent perspectives have highlighted the promise of federated LLMs in medicine, concrete methodological solutions remain scarce [28,29]. Existing medical FL studies have largely focused on smaller architectures such as BERT [28] or traditional classifiers such as logistic regressions [30], which do not face the same scalability constraints as LLMs. A small

number of early works have explored FL for vision foundation models using medical imaging data [27,31]. To date, practical and model-agnostic frameworks for federated training of large language models remain largely unexplored, representing a key barrier to the development and deployment of LLMs in medicine.

To address these challenges, we introduce Fed-MedLoRA and Fed-MedLoRA+, a model-agnostic and parameter-efficient federated learning framework for training large language models in medicine. Fed-MedLoRA enables scalable federated training by transmitting only low-rank adapter parameters, substantially reducing communication and computation overhead while preserving model capacity. Building on this foundation, Fed-MedLoRA+ incorporates an adaptive, data-aware aggregation strategy that explicitly accounts for cross-site heterogeneity, thereby improving global model convergence under realistic multi-institutional conditions.

We evaluate the proposed framework using clinical information extraction (IE) as a representative downstream application. Clinical IE involves identifying and normalizing key medical entities, capturing relations among them, and mapping extracted information to standardized knowledge representations across healthcare systems [32,33]. Clinical narratives provide the most comprehensive and context-rich descriptions of patient health, including symptoms, diagnoses, procedures, medications, and social factors that are often missing from structured records [14,34]. For example, leveraging free-text clinical notes has been shown to identify over 90% more patients with adverse social determinants of health compared with structured data alone [21]. As a result, clinical IE underpins critical downstream applications such as automated cohort identification [22], adverse event monitoring [23], and clinical decision support [35]. Despite its importance, prior studies have shown that current LLMs exhibit suboptimal performance on clinical IE [10,36–38]. Early evaluations reported that even the best-performing closed-weight models (e.g., GPT-3.5 and GPT-4) achieved only 39–52% zero-shot accuracy for extracting medical problems, treatments, and tests from patient notes, with performance dropping to ~20% for more complex relational tasks such as disease–drug association extraction [37]. These limitations further motivate the need for privacy-preserving multi-institutional training strategies for LLMs.

We systematically evaluated the proposed framework for accuracy and practical feasibility. Accuracy was assessed across five independent patient cohorts totaling 42,198 entities and 41,570 relations on two core IE tasks—named entity recognition (NER) and relation extraction (RE)—under three settings: (1) in-domain training and testing to assess within-domain performance; (2) independent benchmark validation to evaluate generalization to external patient cohorts; and (3) low-resource new-site adaptation using a case study based on real-world clinical notes from the Yale New Haven Health System to simulate model deployment at new institutions with limited annotations. Both Fed-MedLoRA and Fed-MedLoRA+ were implemented using two representative open-weight LLMs—LLaMA-3 and DeepSeek-R1—as backbones and were compared head-to-head against multiple baselines: zero-shot and single-site fine-tuned LLMs (LLaMA-3, DeepSeek-R1, and zero-shot GPT-4o), fine-tuned domain-specific BERT models, and recent general-domain FL algorithms for LLMs. For practical feasibility, we further assessed (1) robustness to heterogeneous and incomplete task annotations across sites (e.g., one site with both NER and RE vs. another with only NER), (2)

computational efficiency during training and inference, and (3) scalability as the number of participating sites increased.

Across all evaluation settings, Fed-MedLoRA and Fed-MedLoRA+ consistently underperformed the baselines for both NER and RE tasks. In the in-domain benchmarks, they improved zero-shot LLM performance by up to 65% absolute F1, achieved ~25% higher F1 than single-site fine-tuned LLMs, and outperformed fine-tuned domain-specific BERT models by up to 40%. In external validation across independent cohorts, they maintained strong generalization and had 10-70% higher F1 than the baselines. In the low-resource new-site adaptation study, Fed-MedLoRA+ had the best performance, with a strict F1 score of 73% and a lenient F1 score of 85%, suggesting that a new clinical site could benefit from federated LLMs to bootstrap local model development with limited annotations. Beyond accuracy, the framework proved feasible and scalable in practical settings. Transmitting only low-rank adapters reduced communication cost by 98.5% relative to full-model fine-tuning. Both methods trained 8-billion-parameter models on a single consumer GPU (e.g., RTX 4090, 16 GB) and supported inference on widely available GPUs (e.g., RTX A6000, 3090, 4090). Using a smaller 1-billion-parameter backbone enabled training on mid-range GPUs (e.g., RTX 3060 Ti) and inference on standard laptops (e.g., Apple M3 Pro or equivalent ≥18 GB RAM) with up to 7% accuracy decrease. The framework scaled efficiently to ten participating sites, maintaining performance within 2% of centralized training, and remained robust when sites contributed uneven or incomplete task annotations, reflecting realistic multi-institutional conditions.

Overall, the results demonstrate that federated LLM training can be effective and feasible. The proposed framework highlights the strong potential of federated LLMs for medical applications. We also provide practical guidelines and identify remaining challenges and opportunities for advancing federated LLMs in medicine. To promote transparency, reproducibility, and community collaboration, all implementation code and documentation are publicly available at https://github.com/Yale-BIDS-Chen-Lab/FL_LLM_Med.

3. Results

In this section, we systematically evaluated both accuracy and practical feasibility of Fed-MedLoRA and Fed-MedLoRA+. We first present the overview detailed in Figure 1, followed by the results.

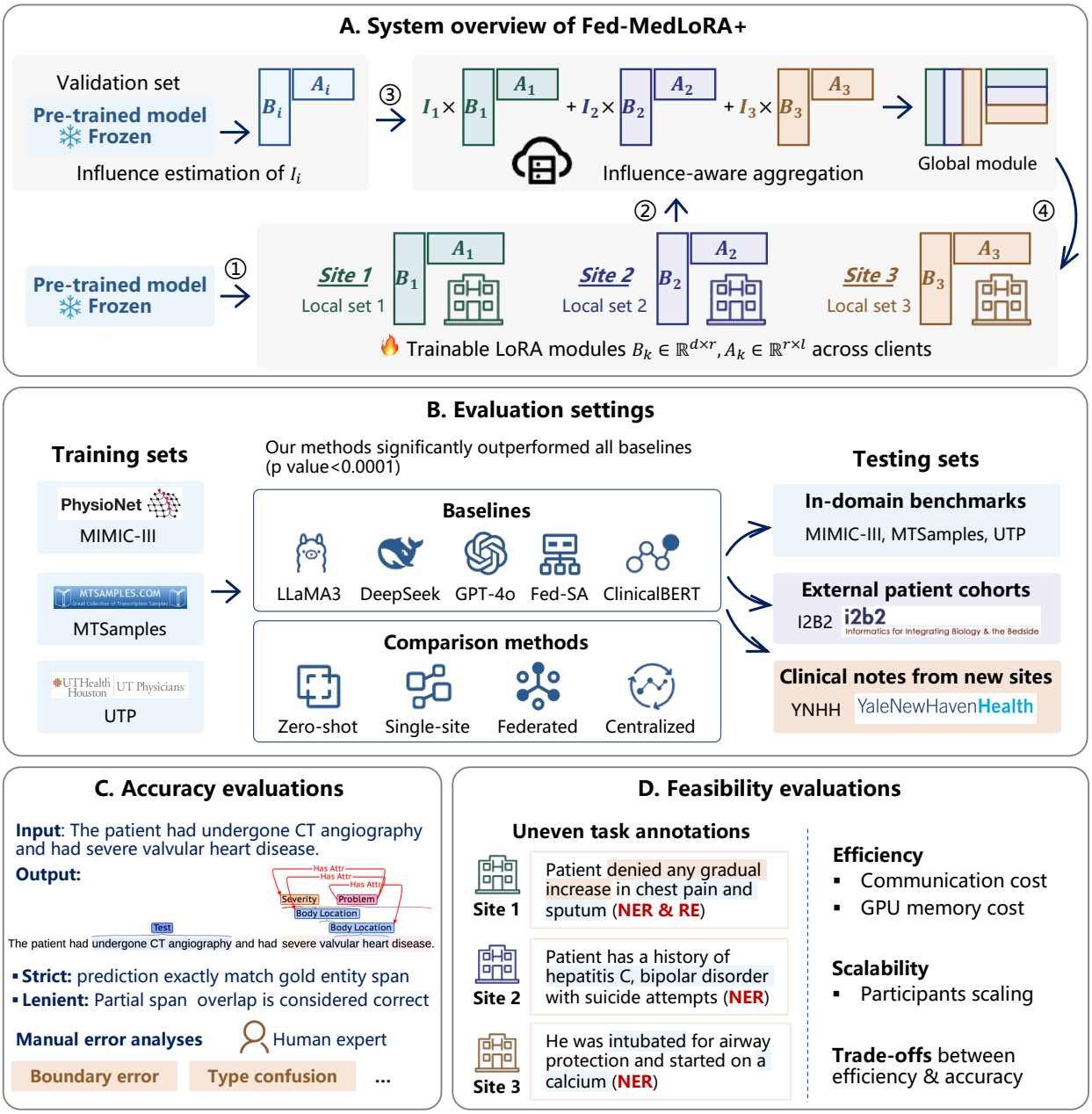

Figure 1. Overview of the study design. A. System overview of Fed-MedLoRA+: (1) each client k fine-tunes the pre-trained backbone on its local data; (2) clients upload only the fine-tuned LoRA adapters to the server; (3) the server computes an influence score for each client and performs influence-aware aggregation; (4) the server returns the updated LoRA adapters to all clients. B. Evaluation settings, including three dataset types (in-domain benchmarks, external patient cohort, and a case study on new-site adaptation using real-world clinical notes from Yale New Haven Health Systems), baselines (domain-specific BERT models, proprietary and open-weighted LLM representatives, and recent federated learning algorithms in the general domain) and comparison methods. C. Accuracy evaluations: an example input and output, evaluation metrics (strict match and lenient match), and manual error analysis. D. Feasibility evaluations: robustness to uneven task annotations (e.g., Site 1 provides both NER and RE, while Site 2, 3 provide only NER), efficiency, scalability, and efficiency-accuracy tradeoff analysis.

Table 1. Statistics of datasets, including number of documents, sentences, 4 NER entities and 16 RE modifiers across different sites of datasets.

Count	Training and testing benchmark						Independent validation set	Case study on new annotation
	Training benchmark			Testing benchmark				
	MIMIC-III	MTSamples	UTP	MIMIC-III	MTSamples	UTP	I2B2	YNHH
Documents	25	96	65	25	50	50	50	4
Sentences	4,260	5,211	4,523	4,049	2,885	5,778	4,809	291
Clinical entity categories								
Problem	2,240	4,060	4,236	2,381	2,200	3,450	1,454	242
Treatment	542	584	418	511	388	329	405	27
Test	2,302	1,529	2,609	2,542	828	2,153	1,309	175
Drug	1,047	747	710	1,021	466	568	681	44
Total	6,131	6,920	7,973	6,455	3,882	6,500	3,849	488
Relation types								
Severity	90	114	69	131	38	50	37	
Temporal	1,303	304	1,811	1,955	146	1,955	621	
Negation	467	1,519	2,339	614	800	1,876	282	
Dosage	244	60	258	228	46	227	90	
Strength	463	149	318	419	170	272	261	
Reference range	11	0	114	5	0	45	0	
Uncertain	81	62	27	58	20	15	10	
Lab value	1,579	893	1,529	1,954	397	1,327	900	
Route	346	104	351	329	89	292	245	/
Frequency	456	200	290	393	207	251	285	
Subject	19	26	128	26	12	131	0	
Form	372	54	581	336	45	526	129	
Condition	42	17	14	74	20	18	1	
Duration	67	18	20	25	10	21	27	
Body location	709	954	1,189	717	578	947	468	
Course	162	218	140	183	108	114	93	
Total	6,411	4,692	9,178	7,447	2,686	7,707	3,449	

3.1 Evaluation overview

3.1.1 Datasets

Table 1 summarizes the five cohorts used in this study. Collectively, those five cohorts consist of 42,198 entities and 41,570 relations covering four main entity categories and 16 relation types. We used four existing benchmarks manually annotated from different institutions: MIMIC-III (Medical Information Mart for Intensive Care) [39], MTSamples (Transcribed Medical Transcription Sample Reports and Examples) [40], UTP (UT Physicians notes) [15], and Partners Healthcare and Beth Israel Deaconess Medical Center notes by Informatics for Integrating Biology & the Bedside (I2B2) [41]. These datasets have been widely adopted for clinical IE method development and evaluation [15,39–41]. They reflect real-world challenges, including differences in scale (e.g., MIMIC-III notes are significantly longer than those from other datasets), entity distributions (e.g., MIMIC-III has nearly twice as many drug entities as UTP due to different patient cohorts), and missing data (e.g., MTSamples lacks reference ranges).

In addition, we manually annotated 242 problem-related entities, 27 treatment-related entities, 175 test-related entities, and 44 drug entities from 291 sentences drawn from four de-identified patient records within the Yale New Haven Health (YNHH) system for the case study.

The annotation scale was intentionally limited to reflect a realistic starting point for a new clinical site compared with established benchmarks.

3.1.2 Evaluation settings

Evaluation on accuracy. As shown in Figure 1B, we systematically evaluated the accuracy of Fed-MedLoRA and Fed-MedLoRA+ under three settings. First, standard training and testing. Models were trained on the training sets and evaluated on the corresponding test sets to assess in-domain performance. We conducted both two-site (federated learning across MIMIC-III and MTSamples) and three-site (MIMIC-III, MTSamples, and UTP) experiments under this setting. Second, independent benchmarks. To evaluate generalization to unseen distributions, models trained under the standard training and testing setting were directly tested on external datasets. In the two-site experiment (training on MIMIC-III and MTSamples), we evaluated performance on UTP and I2B2. In the three-site experiment (training on MIMIC-III, MTSamples, and UTP), we evaluated on I2B2. Third, Low-resource case study. To simulate a practical scenario where a new clinical site to do clinical IE locally, we conducted a case study using the YNHH cohort. Unlike the standard benchmarks in the previous settings, a new site typically has only minimal annotated data available for initialization and needs to adapt existing models for local development.

Evaluation metrics. As shown in Figure 1C, we reported precision, recall, and F1-score, the standard metrics for NER and RE [24,42]. Following prior work [15,24], we used two evaluation schemes. (1) Strict match: for NER, a true positive requires exact span and entity type match; for RE, this further requires the correct entity pair and relation type. (2) Lenient match: for NER, a true positive requires overlapping spans and correct entity type; for RE, relations are considered correct when constituent entities overlap with gold spans and the relation type matches.

Evaluation on feasibility. Beyond accuracy, we further evaluated the practical feasibility of Fed-MedLoRA and Fed-MedLoRA+ across three aspects. Figure 1D shows the detail. First, robustness to uneven task annotations across sites. In practice, individual sites may lack complete annotations for all tasks. We conducted both two-site and three-site experiments to examine this scenario. In the two-site experiment, Site A (MIMIC-III) included annotations for both NER and RE, whereas Site B (MTSamples) had only NER annotations. In the three-site experiment, Site C (UTP) contained only NER annotations, while Sites A and B provided both NER and RE. Second, communication and computational efficiency. We quantified memory consumption and peak GPU utilization during both training and inference to assess efficiency. Third, scalability. We conducted simulations with up to ten participating sites to evaluate the scalability under increasing federation sizes.

Baseline models. We compared Fed-MedLoRA and Fed-MedLoRA+ against three categories of baselines: (1) BERT-based models, (2) LLM-based models, and (3) recent general-domain FL algorithms for LLMs.

For BERT-based baselines, we used Bio_ClinicalBERT [43,44], initialized from BioBERT (pretrained on biomedical literature) [45] and further pretrained on clinical notes. Bio_ClinicalBERT has achieved completive performance on multiple clinical IE benchmarks and is widely adopted as a

strong baseline [44]. We reported its single-site fine-tuning performance. Single-site training reflects a common scenario in clinical practice, where models are trained on data from a single site without federated learning support. For instance, for three datasets from different sites, we trained a model on one site's training set and evaluated it on all test sets, repeating for each site. The reported results are averaged across sites.

For LLM-based baselines, we performed head-to-head comparisons using the same backbone models with and without our proposed federated learning methods. Fed-MedLoRA and Fed-MedLoRA+ are backbone-agnostic; we selected two representative open-weight LLMs LLaMA3 [46] and DeepSeek-R1-Distill [47] for evaluation. We reported both zero-shot (no fine-tuning) and single-site fine-tuned performance. Additionally, we included GPT-4o as a widely used proprietary LLM representative and reported its zero-shot performance for comparison. To complement and thoroughly assess our proposed approach, we evaluated a recently introduced FL algorithm for LLMs in the general domain named FedSA-LoRA [48] as an additional baseline. This method reduces communication by transmitting only the low rank A to the server for aggregation, under the assumption that the A matrices capture the global knowledge.

In addition, we also included centralized training for both BERT-based and LLM-based models as a reference. Centralized training pools data from all sites to train a single model, which is then evaluated on the same test sets. This configuration provides an empirical upper bound, representing an ideal but impractical setting in healthcare due to privacy and governance constraints. The implementation and hyperparameter details are provided in Section 5 Methods.

Bootstrapping and statistical analysis. For both the training/testing and independent benchmark settings, we performed bootstrapping with a sample size of 200 randomly selected instances per dataset, repeated 30 times, and reported results with 95% confidence intervals. We further conducted a two-tailed Wilcoxon rank-sum test to assess statistical significance. Both procedures followed established practices [49,50].

Table 2. Strict and lenient micro F1 scores for NER and RE test sets under *two-site experiments*. Results are reported for three backbones (LLaMA3-8B, DeepSeek-R1-Distill-8B and GPT-4o) and BERT baseline (Bio_ClinicalBERT) across five training strategies: zero-shot, single-site fine-tuning (average across sites, baseline), Fed-MedLoRA, Fed-MedLoRA+, and centralized fine-tuning (upper bound). All values are micro-averaged F1 scores (strict and lenient). Centralized (upper bound) rows are shaded in gray. For each model, the best non-centralized method is shown in bold and the second-best is underlined.

Models	Methods	NER test set				RE test set			
		Strict F1-scores		Lenient F1-scores		Strict F1-scores		Lenient F1-scores	
		MIMIC-III	MTSamples	MIMIC-III	MTSamples	MIMIC-III	MTSamples	MIMIC-III	MTSamples
Bio_ClinicalBERT	Centralized (Upper bound)	0.807	0.827	0.907	0.905	0.590	0.584	0.735	0.657
	Single site (Avg, baseline)	0.753	0.792	0.867	0.888	0.434	0.474	0.595	0.549
LLaMA3-8B	Centralized (Upper bound)	0.856	0.870	0.954	0.958	0.867	0.831	0.889	0.913
	Zero-shot (Baseline)	0.345	0.437	0.503	0.644	0.203	0.056	0.239	0.074
	Single site (Avg, baseline)	0.803	0.843	0.918	0.926	0.648	0.721	0.698	0.798
	Fed-MedLoRA (Ours)	0.847	0.858	0.942	0.929	0.850	0.817	0.860	0.849
	Fed-MedLoRA+ (Ours)	**0.850**	**0.866**	**0.949**	**0.937**	**0.860**	**0.826**	**0.868**	**0.856**
DeepSeek-R1-Distill-8B	Centralized (Upper bound)	0.852	0.862	0.952	0.963	0.863	0.829	0.842	0.911
	Zero-shot (Baseline)	0.291	0.303	0.438	0.499	0.163	0.106	0.190	0.139
	Single site (Avg, baseline)	0.797	0.836	0.902	0.921	0.589	0.719	0.644	0.793
	Fed-MedLoRA (Ours)	0.841	0.858	0.932	0.936	0.799	0.767	0.826	0.892
	Fed-MedLoRA+ (Ours)	**0.845**	**0.860**	**0.945**	**0.951**	**0.831**	**0.773**	**0.831**	**0.907**
GPT-4o	Zero-shot (Baseline)	0.556	0.602	0.815	0.834	0.124	0.096	0.289	0.178

3.2 Results on accuracy

3.2.1 Results on the training/testing setting

Table 2 summarizes model performance in the two-site experiment, where federated learning was conducted across MIMIC-III and MTSamples and evaluated on their respective test sets. Table 3 presents corresponding results for the three-site experiment involving MIMIC-III, MTSamples, and UTP under the same training/testing setting. More results including individual entity accuracy and Bootstrapping and statistical analysis are illustrated in Supplementary S2.1.

Table 3. Strict and lenient micro F1 scores for NER and RE test sets under *three-site experiments*. Results are reported for two backbones (LLaMA3-8B and DeepSeek-R1-Distill-8B) and BERT baseline (Bio_ClinicalBERT) across five training strategies: zero-shot, single-site fine-tuning (average across sites, baseline), Fed-MedLoRA, Fed-MedLoRA+, and centralized fine-tuning (upper bound). All values are micro-averaged F1 scores (strict and lenient). Centralized (upper bound) rows are shaded in gray. For each model, the best non-centralized method is shown in bold and the second-best is underlined.

Model	Method	NER test set			RE test set		
		Strict F1-score					
		MIMIC-III	MTSamples	UTP	MIMIC-III	MTSamples	UTP
Bio_ClinicalBERT	Centralized (Upper bound)	0.810	0.838	0.823	0.598	0.515	0.413
	Sigle site (Avg, baseline)	0.782	0.790	0.720	0.343	0.433	0.249
LLaMA3-8B	Centralized (Upper bound)	0.840	0.867	0.909	0.851	0.788	0.924
	Zero-shot (Baseline)	0.345	0.437	0.256	0.203	0.056	0.091
	Single site (Avg, baseline)	0.775	0.841	0.798	0.656	0.703	0.778
	Fed-MedLoRA (Ours)	**0.809**	0.861	0.897	0.744	0.759	**0.916**
	Fed-MedLoRA+ (Ours)	0.803	**0.862**	**0.898**	**0.771**	**0.762**	0.913
DeepSeek-R1-Distill-8B	Centralized (Upper bound)	0.830	0.846	0.907	0.843	0.785	0.913
	Zero-shot (Baseline)	0.291	0.303	0.251	0.163	0.106	0.110
	Single site (Avg, baseline)	0.778	**0.825**	0.806	0.605	0.725	0.756
	Fed-MedLoRA (Ours)	0.799	0.803	0.848	0.737	**0.739**	0.890
	Fed-MedLoRA+ (Ours)	**0.802**	0.818	**0.879**	**0.752**	0.736	**0.894**
GPT-4o	Zero-shot (Baseline)	0.556	0.602	0.477	0.124	0.096	0.091
		Lenient F1-score					
		MIMIC-III	MTSamples	UTP	MIMIC-III	MTSamples	UTP
Bio_ClinicalBERT	Centralized (Upper bound)	0.919	0.912	0.884	0.742	0.621	0.571
	Sigle site (Avg, baseline)	0.848	0.869	0.837	0.470	0.500	0.276
LLaMA3-8B	Centralized (Upper bound)	0.925	0.926	0.913	0.885	0.855	0.932
	Zero-shot (Baseline)	0.503	0.644	0.513	0.239	0.074	0.112
	Single site (Avg, baseline)	0.876	0.867	0.878	0.699	0.780	0.812
	Fed-MedLoRA (Ours)	0.893	0.901	0.906	0.801	0.797	0.901
	Fed-MedLoRA+ (Ours)	**0.900**	**0.914**	**0.936**	**0.814**	**0.816**	**0.910**
DeepSeek-R1-Distill-8B	Centralized (Upper bound)	0.925	0.926	0.906	0.885	0.855	0.932
	Zero-shot (Baseline)	0.438	0.499	0.433	0.190	0.139	0.128
	Single site (Avg, baseline)	0.895	0.855	0.861	0.732	0.799	0.807
	Fed-MedLoRA (Ours)	0.865	0.910	0.920	0.811	0.821	0.914
	Fed-MedLoRA+ (Ours)	**0.888**	**0.925**	**0.940**	**0.890**	**0.896**	**0.918**
GPT-4o	Zero-shot (Baseline)	0.815	0.834	0.676	0.289	0.178	0.213

Comparisons with LLM-based baselines. The proposed Fed-MedLoRA and Fed-MedLoRA+ consistently outperformed LLM-based baselines (both zero-shot and single-site fine-tuned) using the same backbone models (LLaMA3, DeepSeek-R1-Distill and GPT-4o) in head-to-head comparisons. Compared to zero-shot performance, Fed-MedLoRA and Fed-MedLoRA+ show no improvement and even decline by over 50% on both NER and RE test sets. For example, in the two-site LLaMA3 experiment on MIMIC-III, strict matching NER F1 score improves from 0.345 (zero-shot) to 0.850 (Fed-MedLoRA+, $p<0.0001$), while RE strict F1 score improves from 0.203 (zero-shot) to 0.860 (Fed-MedLoRA+, $p<0.0001$). Similar trends are observed in the three-site experiment. Although GPT-4o achieved the highest zero-shot NER performance among all evaluated LLMs on the MTSamples dataset (strict and lenient F1 scores of 0.602 and 0.834, respectively), its zero-shot performance was still significantly lower than that of federated fine-tuning models.

These results demonstrate that federated instruction tuning can significantly improve performance on NER and RE tasks.

In addition, compared to the single-site baseline—which reflects the most common clinical scenario where models are trained using data from a single institution—Fed-MedLoRA+ consistently outperform across all test sets in both two-site and three-site experiments. The largest absolute gains are observed in RE, with up to ~24.2% absolute improvement. For example, in the two-site experiment on MIMIC-III, the single-site baseline with LLaMA3 achieves 0.648 strict F1 score, compared to 0.860 for Fed-MedLoRA+ ($p<0.0001$). Similarly, with DeepSeek-R1-Distill, the single-site baseline achieves 0.589 versus 0.831 ($p<0.0001$) for Fed-MedLoRA+. This improvement may relate to the fact that RE is inherently more challenging than NER, and leveraging data from multiple sites provides greater benefit. We elaborate on this in the section Comparisons between tasks. In terms of backbone models, LLaMA3 and DeepSeek-R1-Distill demonstrate similar performance, with LLaMA3 slightly outperforming in the three-site experiments.

Comparisons with centralized learning as the estimated upper bound. Centralized learning pools data from all sites for training which is impractical for medical applications. We use it as a reference of estimated upper bound. As shown in Tables 2 and 3, Fed-MedLoRA and Fed-MedLoRA+ achieve performance competitive with centralized training. Notably, Fed-MedLoRA+ closely approach centralized learning across most datasets under both two-site and three-site experiments. For example, in the two-site experiment for NER, Fed-MedLoRA+ achieves strict F1 scores of 0.850 vs. 0.856 on MIMIC-III for centralized learning, and 0.866 vs. 0.870 ($p=0.081$) on MTSamples. In the three-site experiment, the gaps are slightly larger, e.g., 0.803 vs. 0.840 ($p=0.042$), 0.862 vs. 0.867 ($p=0.121$), and 0.898 vs. 0.909 ($p=0.076$), but Fed-MedLoRA+ remains the closest to the centralized upper bound.

Comparisons with BERT-based baselines. Fed-MedLoRA and Fed-MedLoRA+ consistently outperform BERT-based models, with gains of up to 9.7% absolute F1 score on NER (e.g., 0.850 for Fed-MedLoRA+ with LLaMA3 vs. 0.753 for single-site Bio_ClinicalBERT, $p<0.001$) and up to 42.6% absolute F1 score on RE (e.g., 0.860 for Fed-MedLoRA+ with LLaMA3 vs. 0.434 for Bio_ClinicalBERT, $p<0.0001$). Interestingly, when comparing BERT-based and LLM-based single-site baselines, LLM-based models outperform BERT across nearly all test sets, except for a marginally lower performance on MIMIC-III in the three-site experiment (0.775 vs. 0.782, $p=0.069$). We conducted a more detailed comparative analysis between BERT-based and LLM-based models, combining results from both in-domain and independent benchmark settings, as presented below.

Comparisons with general-domain FL Algorithms for LLMs. As mentioned above, we further compared the proposed methods with FedSA-LoRA [48], a recently introduced FL algorithm for LLMs in the general domain, used here as an additional baseline. Table 4 presents head-to-head comparisons from the three-site experiment using the LLaMA3-8B model. Across all evaluation datasets, Fed-MedLoRA and Fed-MedLoRA+ consistently outperformed FedSA-LoRA on both NER and RE tasks. For example, FedSA-LoRA achieved a strict F1 score of 0.298 on NER UTP, compared to 0.898 for Fed-MedLoRA+. Similarly, FedSA-LoRA obtained a strict F1 score of 0.181 for RE, compared to 0.894 for Fed-MedLoRA+. The two-site experiment results showed similar

trends (see Supplementary S2.1). We anticipate two possible reasons for these improvements. First, the proposed Fed-MedLoRA and Fed-MedLoRA+ exchange task-aware, parameter-efficient adapters that preserve task-specific information across sites, whereas FedSA-LoRA transmits only a single low-rank matrix, potentially losing important local adaptation signals. Second, FedSA-LoRA was not originally designed for clinical applications, whereas Fed-MedLoRA+ incorporates influence-aware aggregation to enhance robustness against cross-site heterogeneity.

Table 4. Strict micro F1 scores for NER and RE (LLaMA3-8B) under the three-site experiment. Methods compared: FedSA-LoRA (baseline), Fed-MedLoRA and Fed-MedLoRA+.

Method	NER test set				RE test set			
	MIMIC-III	MTSamples	UTP	I2B2	MIMIC-III	MTSamples	UTP	I2B2
FedSA-LoRA (Baseline)	0.469	0.476	0.298	0.365	0.205	0.108	0.181	0.184
Fed-MedLoRA (Ours)	**0.809**	0.861	0.897	0.833	0.737	**0.739**	0.890	0.711
Fed-MedLoRA+ (Ours)	0.803	**0.862**	**0.898**	**0.844**	**0.752**	0.736	**0.894**	**0.713**

Comparisons between Fed-MedLoRA and Fed-MedLoRA+. Across both two-site and three-site experiments, Fed-MedLoRA+ generally outperforms Fed-MedLoRA. In the two-site experiment, Fed-MedLoRA+ achieves higher scores on all four test sets (NER and RE on both datasets). In the three-site experiment, Fed-MedLoRA+ outperforms Fed-MedLoRA on four out of six test sets, with only marginal differences in the remaining two sets (e.g., 0.803 vs. 0.809 on MIMIC-III, p=0.892, and 0.913 vs. 0.916 on UTP, p=0.104). Collectively, across the two- and three-site experiments, both Fed-MedLoRA and Fed-MedLoRA+ consistently outperforms the baselines. An important observation is that adding more sites does not always guarantee better performance on every site. For instance, both Fed-MedLoRA and Fed-MedLoRA+ achieve higher test scores on MIMIC-III under the two-site experiment than under the three-site experiment. A similar pattern is observed for centralized learning, which serves as the estimated upper bound: the exact F1 score on MIMIC-III decreases from 0.856 in the two-site experiment to 0.840 in the three-site experiment. This suggests that while incorporating multiple sites generally increases robustness relative to single-site learning, it does not guarantee uniform gains on every test set. This aligns with prior findings, which highlight challenges such as data heterogeneity, domain shifts, and label distribution differences. One plausible explanation is that the third site introduces increased data heterogeneity: entity distribution differences and annotation noises (see manual error analysis results in Table 7) may bias aggregate updates and create trade-offs between sites.

Comparisons between tasks. Another observation is that results on RE test sets are generally lower than those on NER. Indeed, NER focuses on single-entity recognition, while RE requires identifying entity pairs and the semantic relation between them, which is inherently more challenging [51]. Fed-MedLoRA and Fed-MedLoRA+ demonstrate greater robustness on RE compared to baselines. For example, in the three-site experiment, the single-site LLaMA3 baseline achieves 0.775 on NER but drops to 0.656 on RE. In contrast, Fed-MedLoRA+ achieves an NER score of 0.803 and an RE score of 0.771. A similar pattern is observed with DeepSeek-R1-Distill: the single-site baseline performance drops from 0.778 on NER to 0.605 on RE, whereas Fed-MedLoRA+ achieves 0.802 on NER and 0.762 on RE. Thus, aggregating distributed

and diverse training samples via federated LoRA substantially reduces the gap of RE relative to single-site training.

Table 5. Cross-institution generalization: exact micro F1-scores for NER and RE across two-site and three-site experiments. Results compare zero-shot, single-site (average over sites), federated (Fed-MedLoRA, Fed-MedLoRA+) and centralized fine-tuning for Bio_ClinicalBERT, LLaMA3-8B and DeepSeek-R1-Distill-8B. Two-site experiment: Site A and B own MIMIC-III and MTSamples, respectively, and external test sets are UTP and I2B2. Three-site experiment: Site A, B and C own MIMIC-III, MTSamples and UTP, respectively and test set is I2B2. Centralized (upper bound) rows are shaded in gray. For each model, the best non-centralized method is shown in bold and the second-best is underlined.

Model	Method	Two-site experiment								Three-site experiment			
		NER test set				RE test set				NER test set		RE test set	
		Strict		Lenient		Strict		Lenient		Strict	Lenient	Strict	Lenient
		UTP	I2B2	UTP	I2B2	UTP	I2B2	UTP	I2B2	I2B2	I2B2	I2B2	I2B2
Bio_ClinicalBERT	Centralized (Upper bound)	0.692	0.768	0.851	0.838	0.288	0.354	0.529	0.536	0.787	0.863	0.352	0.579
	Single site (Avg, baseline)	0.678	0.747	0.846	0.808	0.270	0.232	0.441	0.419	0.752	0.816	0.219	0.495
LLaMA3-8B	Centralized (Upper bound)	0.852	0.863	0.932	0.945	0.803	0.746	0.887	0.755	0.852	0.916	0.745	0.798
	Zero-shot (Baseline)	0.256	0.330	0.513	0.496	0.091	0.096	0.116	0.114	0.330	0.496	0.096	0.114
	Single site (Avg, baseline)	0.806	0.801	0.884	0.924	0.730	0.667	0.822	0.715	0.756	0.873	0.659	0.708
	Fed-MedLoRA (Ours)	0.838	0.850	0.921	0.930	0.778	0.693	0.815	0.726	0.833	**0.896**	0.711	0.748
	Fed-MedLoRA+ (Ours)	**0.840**	**0.860**	**0.927**	**0.939**	**0.793**	**0.740**	**0.854**	**0.736**	**0.844**	0.889	**0.713**	**0.761**
DeepSeek-R1-Distill-8B	Centralized (Upper bound)	0.844	0.864	0.913	0.926	0.797	0.720	0.930	0.816	0.822	0.907	0.730	0.877
	Zero-shot (Baseline)	0.251	0.321	0.433	0.552	0.110	0.155	0.128	0.189	0.321	0.552	0.155	0.189
	Single site (Avg, baseline)	0.788	0.769	0.892	0.906	0.687	0.665	0.719	0.688	0.789	0.862	0.649	0.729
	Fed-MedLoRA (Ours)	0.819	0.845	0.901	0.915	**0.789**	0.711	0.901	0.782	0.791	0.868	**0.692**	**0.764**
	Fed-MedLoRA+ (Ours)	**0.825**	**0.855**	**0.912**	**0.917**	0.751	**0.724**	**0.928**	**0.801**	**0.799**	**0.896**	0.685	0.761
GPT-4o	Zero-shot	0.477	0.602	0.676	0.766	0.091	0.106	0.213	0.192	0.602	0.766	0.106	0.192

3.2.2 Results on the independent benchmark setting

Table 5 summarizes the results of the two-site experiment (federated learning across MIMIC-III and MTSamples, evaluated on UTP and I2B2) and the three-site experiment (federated learning across MIMIC-III, MTSamples, and UTP, evaluated on I2B2) under the independent benchmark setting. Supplementary S2.1 provides additional results on individual entity accuracy.

Comparisons with LLM-based baselines. Similar to the standard training/testing setting, both Fed-MedLoRA and Fed-MedLoRA+ consistently outperform both zero-shot and single-site fine-tuning baselines on independent benchmarks using the same LLM backbones. Note that zero-shot baselines remain unchanged across settings since no fine-tuning is applied. For NER, Fed-MedLoRA+ achieves ~58.4% absolute gain (e.g., exact F1 score of 0.256 for zero-shot vs. 0.840 for Fed-MedLoRA+ on UTP using LLaMA3) and ~70.2% absolute gain on RE (e.g., 0.091 vs. 0.793 in the same setting). Similar gains are observed with DeepSeek-R1-Distill and GPT-4o as backbones. When compared to single-site fine-tuning, the federated approach also has consistent improvements. For instance, in the three-site experiment on I2B2 with LLaMA3, Fed-MedLoRA+ achieves up to a 9.6% gain on NER (strict F1 score: 0.852 for Fed-MedLoRA+ vs. 0.756 for single-site fine-tuning) and a 5.4% gain on RE (strict F1 score: 0.713 for Fed-MedLoRA+ vs. 0.659 for single-site fine-tuning). These results suggest that federated adaption captures complementary cross-site information, leading to improved out-of-domain generalization.

Comparisons with BERT-based baselines. The performance gap between federated approaches and BERT-based baselines was more significant under independent benchmarks compared to the train/test testing. For NER, Fed-MedLoRA+ achieved ~15% higher exact F1 than single-site fine-tuned Bio_ClinicalBERT (UTP: 0.840 vs. 0.678; I2B2: 0.860 vs. 0.747) in the two-site experiment, and over 10% higher in the three-site experiment (I2B2: 0.852 vs. 0.752). For RE, Fed-MedLoRA+ achieved improvements of ~70% on UTP (0.793 vs. 0.270) and ~50% on I2B2 (0.740 vs. 0.302) in the two-site experiment, and a similar ~50% gain on I2B2 in the three-site experiment (0.713 vs. 0.219). Centralized learning results also revealed consistent patterns: centralized LLMs substantially outperformed centralized BERT models on independent benchmarks. Collectively, when comparing LLMs vs. BERT under both training/testing and independent benchmarks, the performance gap was much larger on external datasets. For instance, under the training/testing setting, centralized LLMs achieved only ~4% higher F1 than BERT on NER (e.g., 0.865 vs. 0.807 on MIMIC-III; 0.870 vs. 0.827 in the two-site experiment). In contrast, under external benchmarks, centralized LLMs had ~16% higher F1 on UTP (0.852 vs. 0.692) and ~9.5% higher on I2B2 (0.863 vs. 0.768). These results suggest that while BERT models remain competitive when training and testing distributions are similar, they generalize poorly to external datasets. This observation is consistent with prior reports in the literature[15,52,53]. Possible explanations are twofold. First, LLMs are pretrained on much larger and more diverse corpora, whereas domain-specific BERT models are pretrained on specific medical corpora (e.g., Bio_ClinicalBERT was pretrained on MIMIC-III), which might constrain their generalization ability. Second, BERT-based models may rely more heavily on lexical and annotation patterns specific to their training datasets (e.g., abbreviations, formatting, or tagging conventions), leading to reduced robustness when these patterns differ in external datasets.

Other observations. Additional observations are consistent with the training/testing results. First, the proposed federated approaches consistently achieved performance closest to centralized learning as the estimated upper bound (e.g., ~2% difference between Fed-MedLoRA+ and centralized learning on both UTP and I2B2 for NER and RE in the two-site experiment). Second, Fed-MedLoRA+ generally outperformed Fed-MedLoRA (e.g., higher scores in all 8 metrics—strict and lenient F1, NER and RE, on UTP and I2B2 in the two-site experiment). Third, LLaMA3 and DeepSeek-R1-Distill showed similar performance, with LLaMA3 slightly better for both NER and RE. In contrast, we also observed mixed effects of increasing the number of sites. Under the training/testing setting, performance in the two-site experiments was generally higher than in the three-site experiments. When applied to external benchmarks, some performance drop persisted but was less consistent. For example, with LLaMA3 on I2B2 NER, the single-site fine-tuned baseline dropped from 0.801 (two-site) to 0.756 (three-site), while Fed-MedLoRA remained robust (0.860 vs. 0.852).

3.2.3 Results on the low-resource case study

As described above, we conducted a case study using the newly annotated YNHH cohort to simulate a practical scenario in which a new site performs local clinical IE. In contrast to established benchmarks, such a site typically has only minimal annotated data available for initialization and must rely on adapting existing models rather than fine-tuning locally. Realistic options therefore include using publicly available LLMs in a zero-shot manner or leveraging

federated models trained on data from other institutions without sharing sensitive information. This case study is approved by the Yale Institutional Review Board (IRB) under protocol number 2000036010.

Table 6. Evaluation on Yale New Haven Health (YNHH) clinical notes for NER. The table reports strict and lenient micro F1 scores, precision and recall across two-site and three-site experiments. Results compare a BERT baseline (Bio_ClinicalBERT) and the LLaMA3-8B backbone under four strategies: zero-shot, Fed-MedLoRA, Fed-MedLoRA+ and centralized fine-tuning (upper bound). Centralized (upper bound) rows are shaded in gray. For each model, the best non-centralized method is shown in bold and the second-best is underlined.

Model	**Method**	**Metrics**					
		Strict			Lenient		
		F1 score	Precision	Recall	F1 score	Precision	Recall
Two-site experiment							
Bio_ClinicalBERT	Centralized (Upper bound)	0.596	0.624	0.571	0.645	0.676	0.617
LLaMA3-8B	Centralized (Upper bound)	0.732	0.778	0.690	0.823	0.881	0.772
	Zero-shot (Baseline)	0.397	0.348	0.462	0.524	0.459	0.609
	Fed-MedLoRA (Ours)	<u>0.708</u>	<u>0.759</u>	<u>0.664</u>	<u>0.794</u>	<u>0.850</u>	<u>0.745</u>
	Fed-MedLoRA+ (Ours)	**0.720**	**0.764**	**0.681**	**0.809**	**0.867**	**0.758**
Three-site experiment							
Bio_ClinicalBERT	Centralized (Upper bound)	0.606	0.631	0.583	0.646	0.669	0.624
LLaMA3-8B	Centralized (Upper bound)	0.733	0.755	0.711	0.858	0.858	0.859
	Zero-shot (Baseline)	0.397	0.348	0.462	0.524	0.459	0.609
	Fed-MedLoRA (Ours)	<u>0.722</u>	<u>0.718</u>	<u>0.726</u>	<u>0.848</u>	<u>0.843</u>	<u>0.853</u>
	Fed-MedLoRA+ (Ours)	**0.730**	**0.732**	**0.729**	**0.854**	**0.852**	**0.856**

We directly applied Fed-MedLoRA and Fed-MedLoRA+ models trained in the two-site and three-site experiments, using LLaMA3-8B as the backbone. For comparison, we also evaluated the zero-shot LLaMA3-8B baseline, along with centralized LLaMA3-8B and Bio_ClinicalBERT, which serve as empirical upper-bound references. Evaluation metrics were consistent with those described above. Table 6 summarizes the results. Overall, both Fed-MedLoRA and Fed-MedLoRA+ consistently outperformed all baselines on the new site annotations. For instance, directly applying Fed-MedLoRA+ from the three-site experiment achieved the highest performance, with a strict F1-score of 0.730 compared to 0.397 for the zero-shot baseline ($p<0.0001$), and a lenient F1-score of 0.854 compared to 0.524 for the zero-shot baseline ($p<0.0001$). While Fed-MedLoRA also demonstrated robust performance, Fed-MedLoRA+ consistently had higher scores under both strict and lenient evaluation criteria. Moreover, Fed-MedLoRA+ achieved performance comparable to the upper-bound centralized LLaMA3-8B model (e.g., 0.720 vs. 0.732 and 0.730 vs. 0.733 in the two-site and three-site experiments, $p=0.0691$, $p=0.0874$). Notably, both federated models also outperformed the centralized Bio_ClinicalBERT baseline, consistent with earlier observations on the independent benchmark setting. Results of individual entity performance are shown in Supplementary S2.2.

Table 7. NER error categories observed in 200 manually curated error instances. Count indicates the number of times each error category was observed across the sampled set (instances may be counted under multiple categories). Examples show gold annotation (Gold) and model prediction (Predict).

Categories	Count	Descriptions	Examples
Boundary/ span error (partial match)	90	Predicted span overlaps the gold but is shorter or longer	*Gold:* His urine and serum tox screens ... *Predict:* His urine and serum tox screens ...
False negative	72	Model fails to predict an entity that is present in the gold	*Gold:* He was admitted to the CMED , intubated for airway protection … *Predict:* He was admitted to the CMED , intubated for airway protection …
Type confusion	34	Correct span but wrong entity label	*Gold:* 2012-06-07 09 : 55 PM BLOOD ASA - NEG Ethanol … *Predict:* 2012-06-07 09 : 55 PM BLOOD ASA - NEG Ethanol …
False positive	18	Model predicts an entity where the gold has none	*Gold:* … above the thoracic inlet approximately 6 . 5 cm above the carina . *Predict:* … above the thoracic inlet approximately 6 . 5 cm above the carina .
Boundary & type error	16	Wrong span and wrong entity label	*Gold:* … her ST segments with evolution of Q waves . *Predict:* … her ST segments with evolution of Q waves .
Annotation error	9	The annotator annotates the wrong entity label	*Gold:* Patient had No Known Allergies to Drugs … *Predict:* Patient had No Known Allergies to Drugs …
Merged/ split entities	5	Model merges two adjacent entities into one, or splits one gold entity into multiple ones	Gold: The nasogastric tube terminates within the stomach . Predict: The nasogastric tube terminates within the stomach .

3.2.4 Manual error analysis

We further conducted a manual error analysis to characterize the common errors made by the models. From the four test corpora (MIMIC-III, MTSamples, UTP, and I2B2), we randomly sampled 200 instances in total containing NER prediction errors produced by Fed-MedLoRA+ (LLaMA3-8B). Each instance was manually reviewed and categorized consistent with prior studies [15]. Table 7 summarizes the identified error categories, their frequencies, and representative examples. Note that a single instance may contain multiple error types.

Overall, boundary errors and false negatives were the most frequent failure modes, followed by type confusion and false positives. Boundary errors typically occurred when the model predicted spans that overlapped with, but did not exactly match, the gold-standard spans—resulting in penalties under the strict F1 metric. False negatives frequently reflected omissions of shorter or nested entities.

Table 8. Model robustness results on uneven task annotations from different sites: strict micro F1 scores of Fed-MedLoRA, Fed-MedLoRA+, zero-shot, single-site and centralized learning on NER and RE tasks by fine-tuning on LLaMA3-8B model under two-site and three-site experiments. Centralized (upper bound) rows are shaded in gray. For each model, the best non-centralized method is shown in bold and the second-best is underlined.

Training set	Methods	NER test set				RE test set			
		MIMIC-III	MTSamples	UTP	I2B2	MIMIC-III	MTSamples	UTP	I2B2
Two-site experiment									
A: MIMIC-III (NER+RE) B: MTSamples (NER+RE)	Centralized (Upper bound)	0.856	0.870	0.852	0.863	0.867	0.831	0.803	0.746
	Zero-shot (Baseline)	0.345	0.437	0.256	0.330	0.203	0.056	0.091	0.096
	Single site (Avg, baseline)	0.803	0.843	0.806	0.801	0.648	0.721	0.730	0.667
	Fed-MedLoRA (Ours)	0.847	0.858	0.838	0.850	0.850	0.817	0.778	0.693
	Fed-MedLoRA+ (Ours)	**0.850**	**0.866**	**0.840**	**0.860**	**0.860**	**0.826**	**0.793**	**0.740**
A: MIMIC-III (NER+RE) B: MTSamples (NER)	Fed-MedLoRA (Ours)	0.827	0.842	0.813	0.839	0.815	0.744	0.701	0.675
	Fed-MedLoRA+ (Ours)	0.838	0.851	0.823	0.846	0.832	0.768	0.741	0.709
Three-site experiment									
A: MIMIC-III (NER+RE) B: MTSamples (NER+RE) C: UTP (NER+RE)	Centralized (Upper bound)	0.840	0.867	0.909	0.852	0.851	0.788	0.924	0.745
	Zero-shot (Baseline)	0.345	0.437	0.256	0.330	0.203	0.056	0.091	0.096
	Single site (Avg, baseline)	0.775	0.841	0.798	0.756	0.656	0.703	0.778	0.659
	Fed-MedLoRA (Ours)	**0.809**	0.861	0.897	0.833	0.744	0.759	**0.916**	0.711
	Fed-MedLoRA+ (Ours)	0.803	**0.862**	**0.898**	**0.844**	**0.771**	**0.762**	0.913	**0.713**
A: MIMIC-III (NER+RE) B: MTSamples (NER+RE) C: UTP (NER)	Fed-MedLoRA (Ours)	0.775	0.837	0.828	0.784	0.681	0.714	0.765	0.683
	Fed-MedLoRA+ (Ours)	0.781	0.843	0.816	0.797	0.742	0.729	0.772	0.694

3.3 Results on practical feasibility

3.3.1 Assessment of model robustness on uneven task annotations cross sites

Beyond accuracy, we further evaluated model robustness under a practical scenario in which different sites provided annotations for different tasks. Table 8 summarizes results from both two-site and three-site experiments. (In the two-site setting, Site A (MIMIC-III) included annotations for both Named Entity Recognition (NER) and Relation Extraction (RE), whereas Site B (MTSamples) included only NER annotations. In the three-site setting, Site C (UTP) contained only NER annotations, while Sites A and B provided both NER and RE.) Importantly, both single-site fine-tuned BERT and LLM baselines are not applicable in scenarios with uneven task annotations. We compared Fed-MedLoRA and Fed-MedLoRA+ trained under these uneven annotation conditions directly with the models trained using complete task annotations.

As shown in Table 8, both Fed-MedLoRA and Fed-MedLoRA+ maintained strong performance despite missing task annotations across sites, consistently outperforming single-site fine-tuning baselines that used complete annotations. Compared with their fully annotated counterparts, Fed-MedLoRA+ had only a slight decrease in NER performance (1.2%–1.7% reduction in F1-score across test sets). For RE, the performance drops were somewhat larger but still moderate, with absolute F1-score decreases of 2.8%–5.8%.

3.3.2 Assessment of communication costs and computational resources

In addition, we evaluated the communication overhead and computational resource requirements of Fed-MedLoRA and Fed-MedLoRA+. Specifically, we report per-round communication volumes (number of parameters transmitted and their storage size), peak GPU memory consumptions for training and inference, and total training time (in GPU-hours). To

better quantify the trade-off between model accuracy and computational efficiency, we additionally trained Fed-MedLoRA and Fed-MedLoRA+ using LLaMA3-1B as the backbone, in contrast to the 8B model used in the primary evaluations.

Communication cost. Figure 2A compares the communication overhead of Fed-MedLoRA+ against full-parameter fine-tuning. Recall that both Fed-MedLoRA and Fed-MedLoRA+ update only the low-rank adapters, rather than the full model parameters. For LLaMA3-8B model (8,030,261,248 parameters), full-parameter fine-tuning requires about 29.92 GB per site per round in float 32. This translates to a total transmission of 239 GB for the two-site experiment and 359 GB in the three-site experiment. In contrast, Fed-MedLoRA and Fed-MedLoRA+ update only 41,943,040 parameters, corresponding to a 99.48% reduction in transmission volume. The resulting transmission requirements are 1.25 GB for the two-site experiment and 1.88 GB for the three-site experiment. Similar reductions are observed for other LLM backbones.

GPU memory cost for training and inference. Figure 2B reports the peak GPU memory usage of Fed-MedLoRA+ during training. Peak GPU memory consumption exceeds the raw parameter storage because additional memory is required for gradients, activation checkpoints, optimizer states (e.g., momentum and variance estimates). For Fed-MedLoRA+ with LLaMA3-8B, peak training memory was ~14.04 GB, making it feasible to train on a single consumer-grade GPU card such as the NVIDIA RTX 4090 (16 GB)—without requiring high-end datacenter GPUs like the H100 or H200, which may not be readily available in clinical environments. For inference, since gradients and optimizer states are not present, the memory footprint was substantially lower—~6.79 GB, depending on batch size and sequence length. In practical terms, this lower inference footprint means that models trained using Fed-MedLoRA+ can be deployed on widely available GPUs such as the RTX A6000 (12–16 GB) and RTX 3090/4090, which is crucial for clinical sites that lack training infrastructure but can perform local inference [54].

Training time. Figure 2C presents the wall-clock training time in GPU-hours. For Fed-MedLoRA+ with LLaMA3-8B, the total training time was 6.89 GPU-hours for the two-site experiment and 6.98 GPU-hours for the three-site experiment. Comparable runtimes were observed for DeepSeek-R1-Distill-8B.

Additional analysis on smaller LLM backbones. Table 9 compares the performance of Fed-MedLoRA and Fed-MedLoRA+ when using LLaMA3-1B versus the larger 8B backbone models. Overall, the smaller 1B backbone had a ~3% decrease in NER accuracy and up to a 7% decrease in RE accuracy. As illustrated in Figure 2, the reduction in accuracy comes with substantial gains in efficiency. The 1B model reduced per-round communication cost (e.g., 344 MB vs. 1.25 GB in the two-site experiment), GPU memory cost during training (2.15 GB vs. 14.04 GB) and inference (1.24 GB vs. 6.83 GB), and total GPU hours (1.82 h vs. 6.89 h). In practical terms, training Fed-MedLoRA and Fed-MedLoRA+ with the 1B model can be performed on widely available GPUs such as the NVIDIA RTX 3060 Ti or RTX 4060 Ti (8 GB), while inference can be executed on standard or institution-managed laptops (e.g., devices equipped with an Apple M3 Pro–class chip or an equivalent x86 GPU with over 18 GB RAM). These results suggest the potential feasibility of FL with smaller LLM backbones in resource-constrained environments, where a modest decrease in accuracy is acceptable.

Table 9. Strict micro F1 scores of Fed-MedLoRA and Fed-MedLoRA+ on NER and RE tasks by fine-tuning on LLaMA3-1B and LLaMA3-8B models under two-site and three-site experiments.

Training set	Methods	NER test set				RE test set			
		MIMIC-III	MTSamples	UTP	I2B2	MIMIC-III	MTSamples	UTP	I2B2
Two-site experiment									
LLaMA3-1B	Fed-MedLoRA (Ours)	0.814	0.821	0.802	0.834	0.815	0.798	0.757	0.677
	Fed-MedLoRA+ (Ours)	0.821	0.829	0.817	0.831	0.831	0.802	0.774	0.728
LLaMA3-8B	Fed-MedLoRA (Ours)	0.847	0.858	0.838	0.850	0.850	0.817	0.778	0.693
	Fed-MedLoRA+ (Ours)	**0.850**	**0.866**	**0.840**	**0.860**	**0.860**	**0.826**	**0.793**	**0.740**
Three-site experiment									
LLaMA3-1B	Fed-MedLoRA (Ours)	0.779	0.819	0.854	0.819	0.734	0.735	0.886	0.704
	Fed-MedLoRA+ (Ours)	0.785	0.814	0.866	0.823	0.742	0.741	0.895	0.701
LLaMA3-8B	Fed-MedLoRA (Ours)	**0.809**	0.861	0.897	0.833	0.744	0.759	**0.916**	0.711
	Fed-MedLoRA+ (Ours)	0.803	**0.862**	**0.898**	**0.844**	**0.771**	**0.762**	0.913	**0.713**

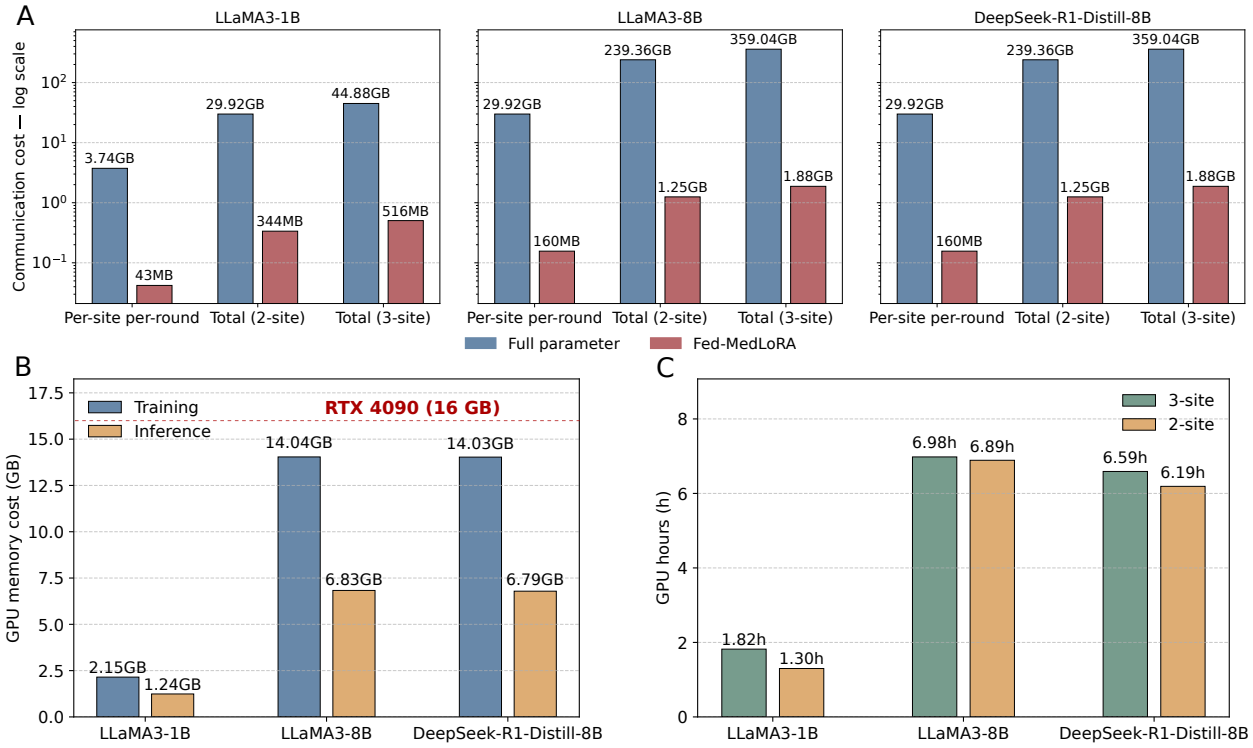

Figure 2. Resource and communication comparison across models and federated settings. (A) Communication cost (storage size) of full-parameter fine-tuning versus Fed-MedLoRA+. For each backbone, we report per-site transmission volume (bytes sent by a single client in a communication round per round, as well as the total transmission volume for two-site and three-site experiments. (B) Peak GPU memory consumption (GB) observed when training LLaMA3-1B, LLaMA3-8B, and DeepSeek-R1-Distill-8B using Fed-MedLoRA+. The horizontal dashed line marks the 16 GB of VRAM on an NVIDIA RTX 4090 for reference, demonstrating that training an 8B backbone model is feasible on high-memory consumer GPUs. (C) Training time in GPU hours for the two-site and three-site experiments.

3.6 Assessment of model scalability

Motivated by the above results, we further conducted simulations with up to k = 10 participating sites to assess scalability. For each k, the training data from MIMIC-III, MTSamples, and UTP were partitioned into k random shards, with each shard treated as an independent

site. Consistent with prior experiments, we compared the strict F1-scores of Fed-MedLoRA+ and single-site fine-tuned LLM baselines using LLaMA3-1B as the backbone for both NER and RE, with centralized learning serving as the estimated upper-bound reference under both training/testing and independent benchmark settings.

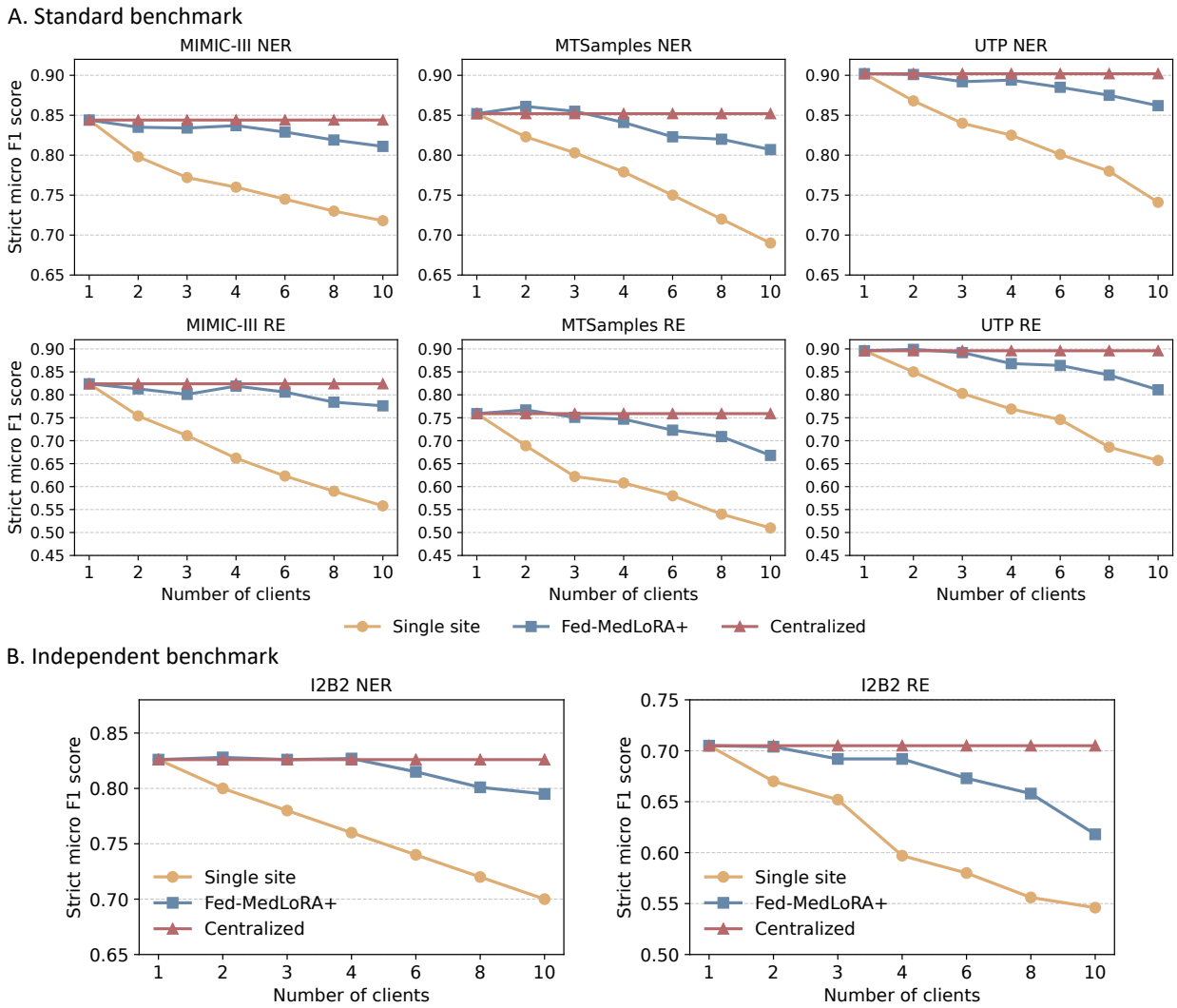

Figure 3. Scalability evaluation: exact micro F1-scores of federated fine-tuning (Fed-MedLoRA and Fed-MedLoRA+) for LLaMA3-1B on NER and RE under different numbers of participating sites (k=1, 2, 3, 4, 6, 8, 10).

Figure 3 summarizes the results. Overall, Fed-MedLoRA+ maintained performance nearly identical to centralized learning when $k \leq 5$ and had only moderate degradation as k increased to 10. The absolute performance drop at k = 10 is 3.9% for NER and 7.5% for RE compared to centralized learning. In contrast, the single-site fine-tuned LLM baseline showed sharp degradation as k increased. For instance, on the MIMIC-III dataset, Fed-MedLoRA+ achieved a 21.8% higher F1-score than the single-site baseline for RE at k =10.

We also observed that RE is inherently more challenging and more sensitive to the number of participating sites than NER, which is consistent with our findings across both training/testing and independent benchmark settings, where RE consistently achieved lower scores than NER.

In simulations with up to 10 sites, NER micro F1-scores remained above 0.79 across all test sets, whereas RE showed greater variability, with a minimum F1-score of approximately 0.618 on the I2B2 dataset.

4. Discussion

4.1 The proposed approach suggests the feasibility and potential of federated learning for large language models in medical applications.

While FL has been increasingly adopted in healthcare, most existing studies have focused on smaller models such as BERT-based or logistic regression architectures. Many current publications are survey papers that primarily discuss the potential and challenges of extending FL to LLMs in medicine [25,55]. A few pioneering works have explored adapting FL to vision foundation models, primarily for disease diagnosis based on imaging data [27,31]. This gap stems from both practical and technical barriers: (1) full-parameter federated training of LLMs is computationally expensive in terms of communication and memory requirements, often exceeding the resources available in clinical environments; (2) naïve aggregation of large parameter updates is unstable under severe client heterogeneity, which is particularly common in medical data; and (3) legal, infrastructural, and operational constraints continue to hinder multi-institutional data collaboration in healthcare. Indeed, bridging FL and LLMs remains a fundamental challenge even in the general domain, as extensively discussed in recent reviews [56,57].

In addition, we also included centralized training for both BERT-based and LLM-based models as a reference. Centralized training pools data from all sites to train a single model, which is then evaluated on the same test sets. This configuration provides an empirical upper bound, representing an ideal but impractical setting in healthcare due to privacy and governance constraints. The implementation and hyperparameter details are provided in Section 5 Methods.

Our work presents a federated, model-agnostic framework for training LLMs in medicine. The proposed Fed-MedLoRA and Fed-MedLoRA+ transmit only low-rank adapter updates, dramatically reducing communication costs, while Fed-MedLoRA+ further introduces adaptive, data-aware aggregation to mitigate clinical data heterogeneity. As a pilot study, this work demonstrates several key findings regarding the feasibility of federated LLMs in medicine. **First, federated fine-tuning of LLMs is effective.** Systematic evaluations across three settings consistently showed that both Fed-MedLoRA and Fed-MedLoRA+ outperformed existing BERT-based and LLM-based baselines. Notably, Fed-MedLoRA+ achieved performance comparable to centralized training. In the new-site case study, Fed-MedLoRA+ had the best performance with a strict F1-score of 73% and a lenient F1-score of 85%, demonstrating that a new clinical site could directly benefit from federated LLMs to bootstrap local model development even with limited annotations. The results also suggest that federated LLMs can support multi-task training, where each site contributes different available task annotations. **Second, the proposed approach improves both computational and communication efficiency.** By transmitting only low-rank adapter parameters, communication costs were reduced by 98.5% compared with full-model fine-tuning. In practice, training federated LLMs at the 8B parameter

scale using Fed-MedLoRA or Fed-MedLoRA+ requires only a single consumer-grade GPU (e.g., NVIDIA RTX 4090, 16 GB). Inference can be performed on commonly available GPUs such as the RTX A6000 (12–16 GB), RTX 3090/4090, or other high-memory workstation GPUs (≥8 GB VRAM). This efficiency improves accessibility and broadens the potential for federated LLM development in typical clinical research centers. Additional analysis on federated LLMs with the 1B backbone further showed that they can be trained with widely accessible GPUs such as the NVIDIA RTX 3060 Ti or RTX 4060 Ti (8 GB), and inference can be performed on a standard laptop (e.g., devices equipped with an Apple M3 Pro–class chip or an equivalent x86 GPU with over 18 GB RAM), with a trade-off in accuracy (for clinical IE, as the results suggested, the accuracy decrease is moderate: 3% on NER and up to 7% on RE for clinical IE)). **Third, the proposed framework demonstrates the potential of scalability and robustness.** Fed-MedLoRA+ maintained stable performance as the number of simulated sites increased to 10, with F1-score differences within ~2% of centralized training. Also, as previously discussed, the performance remained robust when participating sites provided different task annotations.

4.2 Federated learning substantially improves LLMs for clinical information extraction as a downstream case study

We evaluated the accuracy and practical feasibility of the proposed framework using clinical IE as a downstream case study. Many studies have shown that LLMs have suboptimal performance for clinical IE, often producing inconsistent predictions, incomplete extractions, and hallucinated entities or relations [15,58]. To date, fine-tuning domain-specific BERT variants remains the state-of-the-art approach for clinical IE. Recent efforts have explored LLM instruction-tuning for this purpose [15,38], but most have relied on single-institution datasets or combined multiple datasets directly—approaches that overlook the privacy and governance constraints central to real-world clinical data sharing.

The results demonstrate that both Fed-MedLoRA and Fed-MedLoRA+ consistently outperformed existing methods. Specifically, Fed-MedLoRA+ improved zero-shot LLM performance by up to 65% absolute F1 (e.g., from 0.345 to 0.850 for NER and from 0.203 to 0.860 for RE in two-site LLaMA3 experiments) and by ~25% over single-site fine-tuned LLMs (e.g., from 0.589 to 0.831 on NER with DeepSeek-R1-Distill). Compared with fine-tuned BERT baselines, Fed-MedLoRA+ also achieved large gains—improving RE by ~70% on UTP (0.793 vs. 0.270) and ~50% on i2b2 (0.740 vs. 0.302) in the two-site experiment, with comparable improvements (~50%) in the three-site experiment (0.713 vs. 0.219). Collectively, these results demonstrate that federated learning could overcome key limitations of LLMs in clinical IE, achieving robust accuracy and cross-site generalization while maintaining data privacy.

When to prefer LLMs over BERT-based models for clinical IE. While the proposed methods consistently outperformed BERT-based approaches across all settings in this study, it is important to recognize that BERT-based models remain more lightweight and deployment-friendly, with much fewer parameters and lower computational demands. Achieving superior accuracy with LLMs therefore involves trade-offs that downstream users should consider when determining when to LLMs in practice. Our head-to-head comparisons highlight three scenarios where LLMs offer distinct advantages to make their best use. **First, independent or heterogeneous patient cohorts.** When training and testing share the same distribution, the

performance gap between BERT and LLMs is modest (e.g., under the training/testing setting, Fed-MedLoRA+ was ~4% higher F1 than Bio_ClinicalBERT for NER: 0.850 vs. 0.807 on MIMIC-III; 0.866 vs. 0.827 for UTP). In contrast, on external populations, the advantage of LLMs becomes more distinct: Fed-MedLoRA+ achieved ~14.8% higher F1 on UTP (0.840 vs. 0.692) and ~9.2% higher on I2B2 (0.860 vs. 0.768). In the new-site case study, Fed-MedLoRA+ outperformed the centralized Bio_ClinicalBERT by ~13% (0.732 vs. 0.600). **Second, extractive tasks requiring beyond NER.** The results also demonstrate that more challenging tasks such as RE saw a consistent performance drop compared to NER. As mentioned, RE requires modeling cross-entity semantic relations over longer context. The results show that BERT-based models suffered a notable performance drop on RE, whereas LLMs had more robust performance. For instance, in the two-site experiment, Fed-MedLoRA+ achieved 0.860 F1 compared with 0.434 for Bio_ClinicalBERT. **Third, multi-task uses with partial task annotations**. Instruction-tuned LLMs naturally support a single model capable of performing multiple tasks (e.g., both NER and RE in this study), and prior research has shown that a single LLM can adapt to a wide range of medical tasks, from extractive to generative, through instruction tuning [59,60]. Importantly, beyond their multi-task flexibility, our results demonstrate that Fed-MedLoRA and Fed-MedLoRA+ maintained robust performance even when participating sites contributed different task annotations—for example, Site A provided NER and RE labels, while Site B included only NER. Compared with their fully annotated counterparts, Fed-MedLoRA+ had only a slight decrease in NER performance (1.2%–1.7% reduction in F1-score across test sets). For RE, the performance drops were somewhat larger but still moderate, with absolute F1-score decreases of 2.8%–5.8%. This configuration mirrors realistic clinical settings, where annotation availability varies across institutions, and it is often impractical for all sites to provide complete task labels. In contrast, standard BERT-based pipelines are typically task-specific; for instance, in this study, we followed established fine-tuning protocols to train separate BERT models for NER and RE. Although multi-task BERT variants have been proposed [61], they require customized architectures and are not readily extensible to generative tasks.

4.3 Limitations and future work

Despite the promising results, several important limitations of this study warrant discussion. Here, we highlight key aspects not addressed in the current work and call for future community efforts to translate FL for LLMs from experimental settings into real-world medical practice.

First, privacy protection within the federated LLM paradigm requires further investigation. Although our framework transmits only low-rank adapter parameters rather than full model weights or raw data, we did not explicitly incorporate additional privacy-preserving mechanisms. Prior studies have identified potential security risks of LLMs in medical applications, including the inadvertent memorization of sensitive data [62]. Systematic evaluation of standard privacy-preserving techniques—such as secure aggregation [63], differential privacy [64,65], and strict access control [66]—along with their efficiency trade-offs in federated LLM training, is essential.

Second, task selection and optimization for federated LLM training in medicine remain open questions. In this study, we used clinical IE as a representative downstream task to demonstrate the effectiveness of the proposed framework. However, LLMs can be adapted to a

wide range of clinical tasks [7,8]. Systematic evaluation across diverse medical applications—and identification of which tasks and data regimes benefit most from federated learning—will be critical for guiding future model development and prioritizing high-impact clinical use cases.

Third, deployment and evaluation in real clinical environments are necessary to inform practical adoption. While this study evaluated feasibility in terms of robustness, scalability, and efficiency, real-world deployment introduces additional challenges, such as network reliability, coordination and orchestration across institutions, hardware heterogeneity, and site-specific data governance and compliance requirements [67,68]. Addressing these factors through prospective deployments and longitudinal evaluations will be essential for translating federated LLM frameworks into operational healthcare systems [69].

5. Methods

In this section, we first describe how LLMs are fine-tuned at a single site, followed by a detailed explanation of the proposed Fed-MedLoRA and Fed-MedLoRA+ frameworks.

5.1 Supervised fine-tuning of LLMs in the single-site setting

The single-site setting is arguably the most common scenario in medical applications, where each site independently performs supervised fine-tuning of LLMs on its own local dataset [30,70]. In practice, data and models are not shared across sites due to the sensitivity of clinical data. Formally, holds a local dataset $D_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}$, where $x_{k,i}$ denotes a dataset instance and n_k is the dataset size. Site k will fine-tune a LLM backbone denoted as $W^{(0)} \in \mathbb{R}^{d\times l}$, where d and l denote the matrix dimensions.

Supervised fine-tuning of LLMs is typically performed through instruction tuning, where each dataset instance is reformulated as a unified instruction containing a task description (e.g., NER), inputs, and desired outputs with gold-standard labels (e.g., manually annotated entities) [71]. The loss function for fine-tuning all parameters of the LLM is:

$$L_k(W) = \frac{1}{n_k}\sum_{(x,y)\in D_k} l(x_{k,i}, y_{k,i}; W),$$

where l denotes the cross-entropy loss between the predicted token probability distribution and the gold-standard output tokens. Because fine-tuning all parameters of an LLM is computationally expensive, in practice, parameter-efficient instruction tuning methods such as Low-Rank Adaptation (LoRA) are adopted [72]. The main idea is to only fine-tune a subset of parameters while preserving the accuracy. In LoRA, the trainable weight update $\Delta W \in \mathbb{R}^{d\times l}$ is decomposed into two low-rank matrices, $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times l}, r \ll \min(d, l)$. During fine-tuning, only the low-rank matrices B and A are updated, while the original weight matrix W remains frozen. The updated loss function becomes:

$$L_k(W + \Delta W) = \frac{1}{n_k}\sum_{(x,y)\in D_k} l(x_{k,i}, y_{k,i}; +\Delta W),$$

5.2 Proposed Fed-MedLoRA and Fed-MedLoRA+

In contrast to the single-site setting, federated learning enables collaborative model training across multiple institutions while keeping raw data local. In each round, site k fine-tunes the model on its local dataset, transmits updates to a coordinating server, and receives aggregated

parameters, which are then used to continue local training. We proposed two federated learning variants, Fed-MedLoRA and Fed-MedLoRA+, detailed below.

5.2.1 Fed-MedLoRA

We first introduce Fed-MedLoRA, a simple yet effective algorithm inspired by the classical FedAvg [73,74]. FedAvg, which has been widely used in federated learning, aggregates full model weights from participating clients and averages them at the server [30,75]. However, for LLMs, transmitting the entire set of model weights at each round introduces significant communication overhead. To address this, Fed-MedLoRA requires clients to send only their LoRA updates. The procedure, summarized in **Algorithm 1**, consists of three phases.

Initialization. First, the server and clients store a frozen LLM backbone (e.g., LLaMA) with parameters $W^{(0)} \in \mathbb{R}^{d\times l}$. The server also initializes the global LoRA modules $B^{(0)} \in \mathbb{R}^{d\times r}$ and $A^{(0)} \in \mathbb{R}^{r\times l}$ with rank $r \ll \min(d, l)$, and prepares a unified instruction format for fine-tuning. In this study, we designed instruction formats for NER and RE tasks consistent with prior work [15] (see Supplementary S1.1 and S1.2) and distributed them to all clients.

Local updates. At communication round t, the server samples a subset M_t of m clients and transmits the global LoRA modules $B^{(t)}$, $A^{(t)}$ to them. Each selected client $k \in M_t$ initializes its local LoRA modules with the received values and performs E iterations of local training on D_k, producing updated modules $B_k^{(E)} \in \mathbb{R}^{d\times r}$ and $A_k^{(E)} \in \mathbb{R}^{r\times l}$. The client then returns these updated modules to the server.

Server aggregation. The server updates the received LoRA modules by $B_k^{(t+1)} = B_k^{(E)}$, $A_k^{(t+1)} = A_k^{(E)}$ and aggregates them to obtain the new global modules with:

$$B^{(t+1)} = \frac{1}{m}\sum_{k\in M_t}\frac{n_k}{N}B_k^{(t+1)}, A^{(t+1)} = \frac{1}{m}\sum_{k\in M_t}\frac{n_k}{N}A_k^{(t+1)},$$

where n_k is the local dataset size and $N = \sum_{k=1}^{K} n_k$ is the size of all local datasets. These steps are repeated for T communication rounds. The final global model is obtained by merging the frozen backbone with the aggregated LoRA modules $W^{(T)} = W^{(0)} + B^{(T)}A^{(T)}$.

Compared to transmitting and aggregating full model weights, which requires transmitting $d \times l$ parameters, Fed-MedLoRA reduces each round of communication to $dr + rl$ parameters per client. When $r \ll \min(d, l)$, this leads to substantial reductions in both communication cost and GPU memory footprint.

5.2.2 Fed-MedLoRA+

While Fed-MedLoRA substantially reduces the communication cost of fine-tuning LLMs in federated settings, it suffers a core limitation of FedAvg: all client updates are treated equally and aggregated via simple averaging. In clinical FL, this assumption is often violated—local datasets can vary significantly in population characteristics, class balance, annotation quality, and the prevalence of rare labels [15,30,70]. This heterogeneity (e.g., skewed label distributions, noisy or inconsistent annotations, and unrepresentative local samples) can cause simple averaging to degrade global model accuracy and amplify the influence of harmful updates [75,76].

To address this challenge, we propose Fed-MedLoRA+ (an overview is provided in Figure 1A), which dynamically estimates the influence of each client based on validation performance and performs adaptive, data-aware aggregation. Specifically, we first evaluate client updates on a small holdout validation set and compute influence scores that reflect how much each update improves or harms validation performance. These scores are then used to scale client updates during the server aggregation step. By down-weighting noisy or non-representative updates and up-weighting updates that significantly improve generalization, Fed-MedLoRA+ reduces the negative impact of local heterogeneity and builds a more robust global model on unseen sites. The procedure is summarized in **Algorithm 2**. As with Fed-MedLoRA, the process consists of three stages—initialization, local updates, and server aggregation—with the main distinction being the calculation of client influence and dynamic aggregation, described below.

Assumption. We assume the server retains a small validation set (e.g., five clinical notes) $D_v = \{(x_{v,i}, y_{v,i})\}_{i=1}^{n_v}$ with $n_v \ll n_k$ for all clients $k \in [K]$. This assumption is standard in many federated learning deployments [75,76] and is feasible in clinical contexts, where a small amount of de-identified or publicly available labeled data can be used for validation in practice.

Client influence score. During the local update stage, let M_t denote the subset of m clients selected at round t, each of which sends its updated LoRA modules $B_k^{(t+1)}$ and $A_k^{(t+1)}$ to the server. Instead of averaging, the server calculates an influence score for each client. Specifically, the server first forms the client-specific model $W_k^{(t+1)} = W^{(0)} + B_k^{(t+1)} A_k^{(t+1)}$, and evaluates it on the validation set D_v:

$$l_k^v = -\frac{1}{n_v} \sum_{j=1}^{n_v} y_{v,j} \log P(y_{v,j} | x_{v,j}; W_k^{(t+1)})$$

where $P(\cdot)$ denotes the predictive distribution. A softmax is then applied to normalize the validation losses:

$$I_k^{(t)} = \frac{\exp(-l_k^v)}{\sum_{i \in M^{(t)}} \exp(-l_k^v)},$$

This assigns larger $I_k^{(t)}$ to clients with smaller validation loss. For numerical stability, one may compute $\exp(-l_k^v - c)$ with $c = min_i(-l_i^v)$ in practice. Each influence score is then combined with the client's local dataset size n_k to form a data-aware aggregation weight:

$$C_k^{(t)} = \frac{n_k I_k^{(t)}}{\sum_{i \in M^{(t)}} n_i I_i^{(t)}}, \text{ so that } \sum_{k \in M^{(t)}} C_k^{(t)} = 1$$

Note that when all $I_k^{(t)}$ are equal, this reduces to standard FedAvg.

Influence-aware aggregation. Using the weights $\{C_k^{(t)}\}$, the server updates the global LoRA modules as:

$$B^{(t+1)} = \frac{1}{m} \sum_{k \in M^{(t)}} C_k^{(t)} B_k^{(t+1)}, \; A^{(t+1)} = \frac{1}{m} \sum_{k \in M^{(t)}} C_k^{(t)} A_k^{(t+1)},$$

and the updated global model is $W^{(t+1)} = W^{(0)} + B^{(t+1)} A^{(t+1)}$.

Because the weights $\{C_k^{(t)}\}$ form a convex probability distribution, under standard smoothness and bounded-variance assumptions, Fed-MedLoRA+ inherits the same convergence guarantees as FedAvg.

Algorithm 1: Fed-MedLoRA

Input : A pre-trained LLM $W_0 \in \mathbb{R}^{d\times l}$, $k \in [K]$, T, E, m, $D_k = \{x_{k,i}, y_{k,i}\}_{i=1}^{n_k}$, b, η

Output: The optimal LoRA modules B^*, A^*

// Server S:

Initialize global LoRA $B^{(0)} \in \mathbb{R}^{d\times r}$, $A^{(0)} \in \mathbb{R}^{r\times l}$;

for *each round* $t \in \{0, \cdots, T-1\}$ **do**

  $M_t \leftarrow$ randomly select m clients;

  for *each client* $k \in M_t$ *in parallel* **do**

    $B_k^{(t+1)}, A_k^{(t+1)} \leftarrow \texttt{LocalUpdate}(k, B^{(t)}, A^{(t)})$;

  // Update global LoRA modules

  $B^{(t+1)} = \frac{1}{m}\sum_{k\in M_t} \frac{n_k}{N} B_k^{(t+1)}$,

  $A^{(t+1)} = \frac{1}{m}\sum_{k\in M_t} \frac{n_k}{N} A_k^{(t+1)}$;

// Client $k \in [K]$:

Function `LocalUpdate`$(k, B^{(t)}, A^{(t)})$**:**

  // Client k loads global LoRA modules

  $B_k^{(1)} = B^{(t)}, A_k^{(1)} = A^{(t)}$;

  for *each local epoch* $e \in \{1, 2, \cdots, E\}$ **do**

    for *each mini-batch* $\mathcal{B}$ *in* D_k *of size* b **do**

      $W_k^{(t+1)} = W^{(0)} + B_k^{(e)} A_k^{(e)}$;

      $L_k^{(e)} = \frac{1}{b}\sum_{i\in\mathcal{B}} l(x_{k,i}, y_{k,i}; W_k^{(t+1)})$;

      // Update local LoRA modules

      $B_k^{(e+1)} \leftarrow B_k^{(e)} - \eta \nabla_B L_k^{(e)}$,

      $A_k^{(e+1)} \leftarrow A_k^{(e)} - \eta A_k^{(e)} \nabla_A L_k^{(e)}$;

  Return $B_k^{(E)}, A_k^{(E)}$;

Algorithm 2: Fed-MedLoRA+

Input : A pre-trained LLM $W_0 \in \mathbb{R}^{d\times l}$, $k \in [K]$, T, E, m, $D_k = \{x_{k,i}, y_{k,i}\}_{i=1}^{n_k}$, $D_v = \{x_{v,j}, y_{v,j}\}_{j=1}^{n_v}$, b, η

Output: The optimal LoRA modules B^*, A^*

// Server S:

Initialize global LoRA $B^{(0)} \in \mathbb{R}^{d\times r}$, $A^{(0)} \in \mathbb{R}^{r\times l}$;

for *each round* $t \in \{0, \cdots, T-1\}$ **do**

  $M_t \leftarrow$ randomly select m clients;

  for *each client* $k \in M_t$ *in parallel* **do**

    $B_k^{(t+1)}, A_k^{(t+1)} \leftarrow \texttt{LocalUpdate}(k, B^{(t)}, A^{(t)})$;

    $W_k^{(t+1)} = W^{(0)} + B_k^{(t+1)} A_k^{(t+1)}$;

    // Influence estimation

    $l_{k,v} = -\frac{1}{n_v}\sum_{j\in D_v} y_{v,j} \log P(y_{v,j}|x_{v,j}; W_k^{t+1})$;

    $I_k^{(t)} = \frac{\exp(-l_{k,v})}{\sum_{k\in M_t}\exp(-l_{k,v})}$;

  // Update global LoRA modules

  $C_k^{(t)} = \frac{n_k I_k^{(t)}}{\sum_{i\in S_t} n_i I_i^{(t)}}$;

  $B^{(t+1)} = \frac{1}{m}\sum_{k\in M_t} C_k^{(t)} B_k^{(t+1)}$,

  $A^{(t+1)} = \frac{1}{m}\sum_{k\in M_t} C_k^{(t)} A_k^{(t+1)}$;

// Client $k \in [K]$:

Function `LocalUpdate`$(k, B^{(t)}, A^{(t)})$**:**

  // Client k loads global LoRA modules

  $B_k^{(1)} = B^{(t)}, A_k^{(1)} = A^{(t)}$;

  for *each local epoch* $e \in \{1, 2, \cdots, E\}$ **do**

    for *each mini-batch* $\mathcal{B}$ *in* D_k *of size* b **do**

      $W_k^{(t+1)} = W^{(0)} + B_k^{(e)} A_k^{(e)}$;

      $L_k^{(e)} = \frac{1}{b}\sum_{i\in\mathcal{B}} l(\sigma(W_k^{(t+1)}, x_{k,i}); y_{k,i})$;

      // Update local LoRA modules

      $B_k^{(e+1)} \leftarrow B_k^{(e)} - \eta \nabla_B L_k^{(e)}$,

      $A_k^{(e+1)} \leftarrow A_k^{(e)} - \eta A_k^{(e)} \nabla_A L_k^{(e)}$;

  Return $B_k^{(E)}, A_k^{(E)}$;

Implementation and hyperparameters. The proposed Fed-MedLoRA and Fed-MedLoRA+ frameworks are backbone-agnostic. Two representative open-weight LLMs, LLaMA3-8B and DeepSeek-R1-Distill-8B, were selected as backbone models. As described in Section 4.1 Evaluation overview, we conducted head-to-head comparisons under both zero-shot and single-site fine-tuning settings using the same LLM backbones as baselines.

For single-site fine-tuning of LLMs, we applied QLoRA (4-bit quantization with low-rank adapters) to the decoder layers. Adapters were attached to all 32 decoder layers with a rank of r = 16, scaling factor α = 64, and dropout rate of 0.05. Training was performed with a learning rate of 2×10^{-4}, 2 epochs, batch size 4, warmup ratio 0.05, and a maximum input length of 800 tokens. During inference, we set the temperature to 0 to minimize randomness [77].

For Fed-MedLoRA and Fed-MedLoRA+, we adopted the same LLM fine-tuning hyperparameters for direct comparison, with aggregation rounds set to 2. For Fed-MedLoRA+, which requires a validation set, we used 5 records for validation.

In addition, we compared against fine-tuned domain-specific BERT models (Bio_ClinicalBERT). Fine-tuning used a learning rate of 2×10^{-4}, 10 epochs, batch size 64, and a maximum input length of 512 tokens, consistent with prior studies [15].

Code availability

The codes are publicly available via https://github.com/Yale-BIDS-Chen-Lab/FL_LLM_Med.

Acknowledgment

This study is supported by the National Institutes of Health grant 1R01LM014604.

Author Contribution

A.L. and Q.C. designed the research. A.L., Y.C., W.L., Y.Y., H.Y., H.K., W.Z., Y.Z., H.P., Y.R., X.A., H.Y., M.H., X.L., Y.T., L.M., H.X. and Q.C. wrote and edited the manuscript. All authors contributed to discussion and manuscript preparation.

Competing Interests Statement

None declared.

References

1. Liu, X. *et al.* Large Language Models for Outpatient Referral: Problem Definition, Benchmarking and Challenges. Preprint at https://doi.org/10.48550/arXiv.2503.08292 (2025).
2. Wan, P. *et al.* Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat. Med.* **30**, 2878–2885 (2024).
3. Liu, X. A generalist medical language model for disease diagnosis assistance | Nature Medicine. https://www.nature.com/articles/s41591-024-03416-6 (2025).
4. Zhou, S. *et al.* Large language models for disease diagnosis: a scoping review. *Npj Artif. Intell.* **1**, 9 (2025).
5. Small, W. R. *et al.* Evaluating Hospital Course Summarization by an Electronic Health Record–Based Large Language Model. *JAMA Netw. Open* **8**, e2526339 (2025).

6. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
7. Liu, F. Application of large language models in medicine | Nature Reviews Bioengineering. https://www.nature.com/articles/s44222-025-00279-5 (2025).
8. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
9. Artsi, Y. *et al.* Challenges of Implementing LLMs in Clinical Practice: Perspectives. *J. Clin. Med.* **14**, 6169 (2025).
10. Evaluation and mitigation of the limitations of large language models in clinical decision-making | Nature Medicine. https://www.nature.com/articles/s41591-024-03097-1.
11. Reddy, S. Beyond the Leaderboard: The Limitations of LLM Benchmarks and the Case for Real-World Clinical Evaluation. Preprint at https://doi.org/10.20944/preprints202511.1572.v1 (2025).
12. Rahman, S. *et al.* Generalization in Healthcare AI: Evaluation of a Clinical Large Language Model. Preprint at https://doi.org/10.48550/arXiv.2402.10965 (2024).
13. Garriga, R. *et al.* Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Rep. Med.* **4**, (2023).
14. Murff, H. J. *et al.* Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing. *JAMA* **306**, 848–855 (2011).
15. Hu, Y. *et al.* Information Extraction from Clinical Notes: Are We Ready to Switch to Large Language Models? Preprint at https://doi.org/10.48550/arXiv.2411.10020 (2025).

16. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).
17. Subasri, V. *et al.* Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. *JAMA Netw. Open* **8**, e2513685 (2025).
18. Wen, J. *et al.* A survey on federated learning: challenges and applications. *Int. J. Mach. Learn. Cybern.* **14**, 513–535 (2023).
19. Abdulrahman, S. *et al.* A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. *IEEE Internet Things J.* **8**, 5476–5497 (2021).
20. Sun, M. *et al.* Federated Learning for Large Models in Medical Imaging: A Comprehensive Review. Preprint at https://doi.org/10.48550/arXiv.2508.20414 (2025).
21. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *Npj Digit. Med.* **7**, 6 (2024).
22. JMIR Medical Informatics - Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study. https://medinform.jmir.org/2022/12/e42379/.
23. Dou, M., Tang, J., Tiwari, P., Ding, Y. & Guo, F. Drug–Drug Interaction Relation Extraction Based on Deep Learning: A Review. *ACM Comput Surv* **56**, 158:1-158:33 (2024).
24. Peng, L. *et al.* An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *Npj Digit. Med.* **7**, 1–9 (2024).
25. Maity, S. & Saikia, M. J. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **12**, 631 (2025).

26. Ren, C. *et al.* Advances and Open Challenges in Federated Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2404.15381 (2024).
27. Flow of Knowledge: Federated Fine-Tuning of LLMs in Healthcare under Non-IID Conditions. https://arxiv.org/html/2510.00543v1.
28. Jiang, W. *et al.* Federated Large Language Models: Feasibility, Robustness, Security and Future Directions. Preprint at https://doi.org/10.48550/arXiv.2505.08830 (2025).
29. Yao, Y. *et al.* Federated Large Language Models: Current Progress and Future Directions. Preprint at https://doi.org/10.48550/arXiv.2409.15723 (2024).
30. Kuo, T.-T., Gabriel, R. A., Koola, J., Schooley, R. T. & Ohno-Machado, L. Distributed cross-learning for equitable federated models - privacy-preserving prediction on data from five California hospitals. *Nat. Commun.* **16**, 1371 (2025).
31. Feng, L. & Chen, S. Taming Vision-Language Models for Federated Foundation Models on Heterogeneous Medical Imaging Modalities. in *Proceedings of the 2025 International Conference on Multimedia Retrieval* 303–311 (Association for Computing Machinery, New York, NY, USA, 2025). doi:10.1145/3731715.3733441.
32. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022 | npj Digital Medicine. https://www.nature.com/articles/s41746-022-00730-6.
33. Wei, C.-H. *et al.* PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res.* **52**, W540–W546 (2024).
34. Keloth, V. K. *et al.* Social determinants of health extraction from clinical notes across institutions using large language models. *NPJ Digit. Med.* **8**, 287 (2025).

35. Levra, A. G. *et al.* A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. *Eur. J. Intern. Med.* **131**, 113–120 (2025).
36. Chen, Q. Benchmarking large language models for biomedical natural language processing applications and recommendations | Nature Communications. https://www.nature.com/articles/s41467-025-56989-2 (2025).
37. Hu, Y. *et al.* Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc.* **31**, 1812–1820 (2024).
38. Liu, L. *et al.* Human-level information extraction from clinical reports with fine-tuned language models. 2024.11.18.24317466 Preprint at https://doi.org/10.1101/2024.11.18.24317466 (2024).
39. Johnson, A., Pollard, T. & Mark, R. MIMIC-III Clinical Database. PhysioNet https://doi.org/10.13026/C2XW26 (2015).
40. Transcribed Medical Transcription Sample Reports and Examples - MTSamples. https://mtsamples.com/.
41. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc. JAMIA* **18**, 552 (2011).
42. Farmaan, M. Understanding Named Entity Recognition Evaluation Metrics with Implementation in Scikit-Learn. *featurepreneur* https://medium.com/featurepreneur/understanding-named-entity-recognition-evaluation-metrics-with-implementation-in-scikit-learn-d94adbdfeb62 (2024).

43. emilyalsentzer/Bio_ClinicalBERT · Hugging Face. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT.
44. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. Preprint at https://doi.org/10.48550/arXiv.1904.03323 (2019).
45. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
46. Grattafiori, A. *et al.* The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
47. deepseek-ai/DeepSeek-R1-Distill-Llama-8B · Hugging Face. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B (2025).
48. Guo, P. *et al.* Selective Aggregation for Low-Rank Adaptation in Federated Learning. Preprint at https://doi.org/10.48550/arXiv.2410.01463 (2025).
49. Chen, Q. *et al.* Multimodal, multitask, multiattention (M3) deep learning detection of reticular pseudodrusen: Toward automated and accessible classification of age-related macular degeneration. *J. Am. Med. Inform. Assoc.* **28**, 1135–1148 (2021).
50. Lenskjold, A. Artificial intelligence tools trained on human-labeled data reflect human biases: a case study in a large clinical consecutive knee osteoarthritis cohort | Scientific Reports. https://www.nature.com/articles/s41598-024-75752-z (2024).
51. A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers | ACM Computing Surveys. https://dl.acm.org/doi/10.1145/3674501.
52. Sheikhalishahi, S. *et al.* Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med. Inform.* **7**, e12239 (2019).

53. Wang, Y. *et al.* Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
54. Teo, Z. L. *et al.* Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Rep. Med.* **5**, 101419 (2024).
55. Hu, Y., Xu, C., Lin, B., Yang, W. & Tang, Y. Y. Medical multimodal large language models: A systematic review. *Intell. Oncol.* **1**, 308–325 (2025).
56. Bytez.com *et al.* FLoRA: Federated Fine-Tuning Large Language Models with ... https://bytez.com/docs/neurips/95025/paper (2024).
57. Cho, Y. J., Liu, L., Xu, Z., Fahrezi, A. & Joshi, G. Heterogeneous LoRA for Federated Fine-tuning of On-Device Foundation Models. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 12903–12913 (Association for Computational Linguistics, Miami, Florida, USA, 2024). doi:10.18653/v1/2024.emnlp-main.717.
58. Nagar, A. *et al.* LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction. in *The Sixth Workshop on Insights from Negative Results in NLP* (eds Drozd, A., Sedoc, J., Tafreshi, S., Akula, A. & Shu, R.) 106–120 (Association for Computational Linguistics, Albuquerque, New Mexico, 2025). doi:10.18653/v1/2025.insights-1.11.
59. Han, W. *et al.* MedINST: Meta Dataset of Biomedical Instructions. in *Findings of the Association for Computational Linguistics: EMNLP 2024* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 8221–8240 (Association for Computational Linguistics, Miami, Florida, USA, 2024). doi:10.18653/v1/2024.findings-emnlp.482.

60. Tran, H., Yang, Z., Yao, Z. & Yu, H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *J. Am. Med. Inform. Assoc.* **31**, 1821–1832 (2024).
61. Tong, Y., Chen, Y. & Shi, X. A Multi-Task Approach for Improving Biomedical Named Entity Recognition by Incorporating Multi-Granularity information. in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (eds Zong, C., Xia, F., Li, W. & Navigli, R.) 4804–4813 (Association for Computational Linguistics, Online, 2021). doi:10.18653/v1/2021.findings-acl.424.
62. Li, A. *et al.* Memorization in Large Language Models in Medicine: Prevalence, Characteristics, and Implications. Preprint at https://doi.org/10.48550/arXiv.2509.08604 (2025).
63. Fereidooni, H. *et al.* SAFELearn: Secure Aggregation for private FEderated Learning. in *2021 IEEE Security and Privacy Workshops (SPW)* 56–62 (2021). doi:10.1109/SPW53761.2021.00017.
64. Li, X., Tramèr, F., Liang, P. & Hashimoto, T. Large Language Models Can Be Strong Differentially Private Learners. Preprint at https://doi.org/10.48550/arXiv.2110.05679 (2022).
65. Charles, Z. *et al.* Fine-Tuning Large Language Models with User-Level Differential Privacy. Preprint at https://doi.org/10.48550/arXiv.2407.07737 (2024).
66. Kerl, M., Bodin, U. & Schelén, O. Privacy-preserving attribute-based access control using homomorphic encryption. *Cybersecurity* **8**, 5 (2025).

67. Tham, Y. C. *et al.* Building the world's first truly global medical foundation model. *Nat. Med.* **31**, 3580–3585 (2025).
68. Matta, S. S. & Bolli, M. FEDERATED LEARNING FOR PRIVACY-PRESERVING HEALTHCARE DATA SHARING: ENABLING GLOBAL AI COLLABORATION. *Am. J. Sch. Res. Innov.* **4**, 320–351 (2025).
69. Wirth, F. N., Meurers, T., Johns, M. & Prasser, F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Med. Inform. Decis. Mak.* **21**, 242 (2021).
70. Federated target trial emulation using distributed observational data for treatment effect estimation | npj Digital Medicine. https://www.nature.com/articles/s41746-025-01803-y.
71. Zhang, S. *et al.* Instruction Tuning for Large Language Models: A Survey. Preprint at https://doi.org/10.48550/arXiv.2308.10792 (2025).
72. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).
73. McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
74. Li, A., Zhang, L., Wang, J., Han, F. & Li, X.-Y. Privacy-Preserving Efficient Federated-Learning Model Debugging. *IEEE Trans. Parallel Distrib. Syst.* **33**, 2291–2303 (2022).
75. Li, A. *et al.* FedCSS: Joint Client-and-Sample Selection for Hard Sample-Aware Noise-Robust Federated Learning. *Proc ACM Manag Data* **1**, 212:1-212:24 (2023).

76. Wang, J., Zhang, L., Li, A., You, X. & Cheng, H. Efficient Participant Contribution Evaluation for Horizontal and Vertical Federated Learning. in *2022 IEEE 38th International Conference on Data Engineering (ICDE)* 911–923 (2022). doi:10.1109/ICDE53745.2022.00073.
77. Patel, D. *et al.* Exploring Temperature Effects on Large Language Models Across Various Clinical Tasks. 2024.07.22.24310824 Preprint at https://doi.org/10.1101/2024.07.22.24310824 (2024).