MedQA-FoRA-MultiHospital: A Non-IID Multihospital Benchmark and Adaptive Federated Low-Rank Framework for Privacy-Preserving Medical Question Answering and Clinical Report Generation

Zhen Xiong and Jun Li

[1] School of Computer Science and Engineering, Beihang University, China

[2] The school of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

Abstract

Large language models are increasingly attractive for healthcare question answering, longitudinal note generation, discharge communication, and specialty-specific decision support, but their direct deployment in hospitals remains constrained by privacy regulation, compute limitations, heterogeneous local data, and the risk of destructive domain adaptation. This paper introduces *MedQA-FoRA-MultiHospital*, a new multihospital benchmark designed for privacy-preserving adaptation of medical language models under realistic non-identically distributed client partitions. The benchmark contains 50,000 medical question–answer pairs and 15,000 clinical report-generation instances distributed across five simulated hospital clients with distinct specialty profiles: cardiology, respiratory medicine, neurology, general medicine, and pediatrics. Building on the micro–meso–macro philosophy of efficient federated low-rank fine-tuning, we formulate *Adaptive FoRA*, a heterogeneity-aware extension that preserves the frozen backbone, inserts structured low-rank operators across all major linear maps, and aggregates client updates through divergence-sensitive weighting. Rather than inventing unverified empirical scores, this manuscript contributes a complete benchmark specification, a mathematically explicit training framework, communication and parameter analyses, a privacy threat model, a full experimental protocol, and a reproducibility package structure suitable for subsequent leaderboard development. The paper is intentionally written as a stand-alone benchmark-and-methodology manuscript: it defines the dataset, the task suite, the optimization design, the ablation roadmap, and the evaluation criteria required for rigorous future implementation. We argue that the proposed benchmark fills an important gap between generic medical instruction datasets and privacy-preserving federated adaptation studies by unifying question answering, structured reasoning, and clinical report generation within a single non-IID multihospital setting.

Keywords: federated learning; large language models; medical question answering; clinical report generation; parameter-efficient fine-tuning; non-IID learning; low-rank adaptation; quantization

1. Introduction

The practical value of language models in medicine lies not only in their ability to answer isolated benchmark questions, but in their capacity to adapt safely to local clinical workflow, institution-specific writing style, specialty vocabulary, and locally prevalent diseases. Hospitals rarely hold

Corresponding author: Zhen Xiong (zhenxiong8768@proton.me)

BCDS 2026, Vol. 7(1), 54-78. DOI: 10.71448/bcds2671-5

exchangeable data distributions, and the same deployment pipeline that works for a general-purpose assistant can fail in a clinical environment because it ignores specialty skew, resource asymmetry, auditability requirements, and governance rules surrounding patient data and model access. These pressures are intensified when model adaptation must occur near the point of care, where device memory is limited and the direct transfer of raw records to a central server is unacceptable [1, 2]. At the same time, simply keeping one separate model per hospital can fragment knowledge and sharply reduce the benefits of shared learning. The modern landscape of instruction-following models, scaling laws, open and closed LLM ecosystems, and regulation-aware data practice makes this tension especially visible in medicine, where correctness, calibration, and factual consistency matter more than stylistic fluency alone [3–6].

Federated learning offers a natural systems response to privacy and governance constraints, yet standard federated optimization is not automatically compatible with large language model adaptation. Hospital clients may differ not only in label or concept frequencies but also in note length, report structure, triage patterns, severity mix, and specialty-specific terminology [7–9]. These forms of heterogeneity complicate naive aggregation and make it difficult to transfer the gains of general-domain parameter-efficient fine-tuning directly into healthcare. In particular, the tension between minimal trainable parameters and sufficient expressive power becomes more pronounced when downstream tasks require multi-step medical reasoning, template-constrained report generation, and structured action recommendations. A benchmark for this setting must therefore support more than a single task and more than a single statistical skew. It must also make room for communication-aware adaptation, on-device optimization, and future privacy-enhancing techniques [10, 11].

This paper addresses that need by presenting *MedQA-FoRA-MultiHospital*, a benchmark and methodology manuscript designed for the federated adaptation of medical LLMs under non-IID multihospital partitions. The benchmark is built around two clinically meaningful task families: long-form medical question answering and clinical report generation. The first captures specialty-specific reasoning and action recommendation. The second captures stylistic and structural generation under hospital-dependent conventions [12, 13]. The benchmark contains five simulated hospital clients with clearly differentiated specialty profiles and complexity levels, enabling research on both global generalization and client-level specialization [14].

Our methodological contribution is *Adaptive FoRA*, a benchmark-aligned extension of a federated low-rank adaptation pipeline. The method preserves the key micro–meso–macro design philosophy while introducing heterogeneity-aware aggregation and a multi-task objective suited to the new benchmark [15]. At the micro level, we retain a structured low-rank factorization that decomposes adaptation into base, adapter, and deep interaction components. At the meso level, we apply the operator across the principal linear maps in attention and feed-forward sublayers. At the macro level, we combine quantized client-side training with federated aggregation so that only adapter updates, not raw records or full model weights, traverse the network. We then add a hospital-divergence-sensitive weighting mechanism to reduce the dominance of large but distributionally distant clients [16].

A second contribution is conceptual rather than empirical. Because the uploaded dataset is presently specified as a benchmark design rather than an already executed large-scale experimental corpus, we do not fabricate performance tables [17, 18]. Instead, we provide a complete, publication-style paper that contributes the dataset definition, a full method, formal notation, analytical propositions, benchmark tasks, baselines, evaluation metrics, ablation axes, privacy model, and reproducibil-

ity checklist. This makes the manuscript suitable either as a benchmark paper in its own right or as the scaffold for a later results-driven implementation paper.

The remainder of the paper is organized as follows. Section 2 situates the benchmark within literature on federated learning, parameter-efficient adaptation, quantized deployment, and clinical language modeling. Section 3 introduces the dataset design and benchmark protocol. Section 4 formulates the learning problem and develops Adaptive FoRA. Section 5 analyzes parameter and communication complexity. Section 6 defines the benchmark tasks, baselines, and evaluation protocol. Section 7 discusses scientific questions enabled by the benchmark. Section 8 addresses limitations, ethics, and deployment concerns. Section 9 concludes.

2. Related Work

2.1. Large language models, scaling, and instruction adaptation

The current generation of language models is shaped by three interacting trends: scale, alignment, and domain transfer [19–21]. Instruction tuning and preference alignment made large models substantially more useful for open-ended assistance, while scaling studies clarified that model quality depends jointly on data quality, data volume, and compute allocation. Open-weight model families created a route toward local or institution-hosted deployment, which is especially relevant for regulated sectors such as healthcare [22]. Yet these same developments expose a practical challenge: domain adaptation requires access to data that cannot easily be centralized, and the most powerful aligned systems are often inaccessible for transparent institutional fine-tuning. This motivates work on open or semi-open model adaptation, local deployment, and efficient transfer mechanisms. Representative milestones include instruction-following alignment, open-source code generation and general LLM families, scaling-law analysis, multilingual open-weight LLMs, benchmark suites for transfer learning, and emergent behavior studies [23, 24].

2.2. Federated learning under heterogeneity

Federated learning has matured from a communication-efficient alternative to centralized training into a broader framework for privacy-preserving collaborative optimization. Early work emphasized iterative model averaging, while later studies showed that non-IID data can severely degrade convergence and global quality. Subsequent methods targeted objective inconsistency, local drift, batch-normalization mismatch, personalized components, and server-side adaptation. In healthcare, these concerns are amplified because institutions differ in disease prevalence, coding habits, and workflow conventions. The present benchmark is motivated by this literature but narrows the problem to the language adaptation regime, where the unit of aggregation is an efficient adapter update rather than the full dense model [25–27].

2.3. Parameter-efficient fine-tuning and structured adaptation

Parameter-efficient fine-tuning (PEFT) is now a standard response to the cost of adapting large models. Adapters, prefixes, prompts, bias-only tuning, hypercomplex layers, and low-rank updates all pursue the same central trade-off: reduce trainable state while preserving enough capacity for downstream adaptation. Among these methods, LoRA and its descendants became especially influential because they introduce no additional inference path other than the low-rank update and can

be implemented cleanly on top of frozen backbones. Later work explored adaptive rank allocation, magnitude–direction decompositions, orthogonal transformations, and richer factorization schemes. The benchmark proposed here adopts this PEFT viewpoint but does so in a federated, clinically structured setting where low-rank expressiveness must coexist with heterogeneous task distributions [28–30].

2.4. Quantization, compression, and on-device feasibility

Efficient adaptation is not only a question of trainable parameters; the frozen base model must still fit into client hardware during forward and backward passes. Quantization and model compression therefore play an essential role in any realistic hospital-edge setting [31, 32]. Integer-aware inference, 4-bit fine-tuning, post-training quantization, and structured compression each contribute different pieces of the feasibility puzzle. Quantization alone may introduce quality loss, but in a federated environment that loss can be partially offset when many heterogeneous but complementary clients contribute updates to a higher-precision server-side aggregation step. Our macro design directly leverages this insight [33].

2.5. Biomedical language models, medical QA, and report generation

Domain-specific medical language modeling has progressed from encoder pretraining on biomedical literature to generative clinical assistants and specialized LLMs. Biomedical BERT variants improved extraction and biomedical QA; clinical-domain models built on discharge summaries and EHR text improved inference and similarity tasks; larger generative biomedical and clinical models broadened the scope to long-form generation and interactive assistance [34–38]. Parallel to model development, medical QA datasets and report corpora enabled increasingly realistic evaluation, although many remain single-task, centralized, or weakly connected to privacy-preserving deployment. Report generation datasets such as MIMIC-CXR and related corpora demonstrate how stylistic and institutional conventions matter alongside semantic correctness. The present benchmark differs by explicitly combining specialty-partitioned medical QA with report generation in a federated non-IID formulation [39–41].

3. The MedQA-FoRA-MultiHospital Benchmark

3.1. Benchmark goals

The benchmark is designed around four goals. First, it should model realistic cross-hospital heterogeneity rather than merely random partitioning. Second, it should include both reasoning-oriented and documentation-oriented tasks. Third, it should be compatible with privacy-preserving collaborative learning in which raw records remain local [42, 43]. Fourth, it should be sufficiently structured to support rigorous evaluation of factuality, formatting fidelity, and actionability. These goals motivate the joint inclusion of medical QA and clinical report generation, as well as the explicit division of the dataset into five specialty-driven client partitions.

3.2. Dataset overview

MedQA-FoRA-MultiHospital v1.0 contains 50,000 medical QA pairs and 15,000 clinical reports. The dataset is partitioned across five hospital clients with distinct specialty concentrations and different

complexity profiles [44]. Table 1 summarizes the client design. Importantly, the benchmark is intended to emulate a federated environment; therefore, the hospital partitions are part of the dataset definition rather than an afterthought imposed during preprocessing.

Table 1. Hospital partitions in MedQA-FoRA-MultiHospital v1.0

Client	Specialty	Size	Dominant distributional profile	Complexity
A	Cardiology	12,000	Cardiovascular disease, heart failure, ischemia, rhythm and hemodynamic management	High
B	Respiratory	10,000	Pulmonary disease, infection, obstruction, interstitial disease, smoking-related illness	Medium–high
C	Neurology	8,000	Stroke, seizure, neuroimmunology, neurocritical care, diagnostic localization	Very high
D	General medicine	12,000	Mixed common disease, infectious disease, gastroenterology, metabolic complaints	Medium
E	Pediatrics	8,000	Pediatric infectious disease, developmental conditions, age-specific respiratory and emergency care	Medium

In the QA portion, each instance contains a hospital identifier, specialty tag, question text, a difficulty label, a reasoning flag, a long-form answer, and structured subfields such as diagnosis, immediate actions, and timeline. In the report-generation portion, each instance contains structured encounter information, hospital identity, task type, a target report or section, and auxiliary evaluation fields. This design allows the same backbone model to be evaluated on both free-form response quality and structured, clinically constrained document generation.

3.3. Non-IID design rationale

Many federated studies simulate heterogeneity through simple label skew, Dirichlet partitioning, or feature perturbation. Such strategies are useful for controlled experiments but insufficient to capture hospital-specific specialization in medical language. In practice, heterogeneity arises from specialty mix, referral patterns, document conventions, local abbreviation habits, and distinct balance between acute versus chronic cases. The benchmark therefore defines non-IID structure semantically rather than only statistically. For example, cardiology questions emphasize medication titration, congestion management, and hemodynamics; neurology questions place higher pressure on timelines, localization, and eligibility criteria; pediatrics introduces age-specific management and distinct reporting idioms.

This matters for model design. A method that only averages local adapters by sample count may overfit to the largest clients or fail to preserve rare-but-important expertise from smaller clients. Conversely, a highly personalized system may fragment knowledge so severely that the global model loses value as a shared clinical assistant. MedQA-FoRA-MultiHospital is built to make this trade-off explicit.

3.4. Task families

3.4.1. Medical question answering. The QA task is intentionally long-form and action-oriented. The expected response is not a short span extracted from a document but a clinically coherent answer that may include differential reasoning, diagnosis, immediate management, follow-up timing, and escalation thresholds. Each example can also include structured annotations that facilitate targeted evaluation, such as whether the model recovered the correct diagnosis or essential next steps.

3.4.2. Clinical report generation. The report task includes discharge summaries, consult notes, impression sections, and recommendation blocks. Unlike general summarization, clinical report generation is constrained by structure, chronology, medication consistency, and specialty-specific writing habits. The benchmark therefore supports both full-report generation and section-level generation.

3.4.3. Data schema. Table 2 summarizes the major fields. The schema is deliberately rich because benchmark utility depends on more than the plain text alone. Structured subfields enable future work on calibration, controllable generation, explanation quality, and rule-based clinical validation.

Table 2. Core schema elements for the two benchmark task families

Field group	Description
QA metadata	Hospital ID, specialty, instance ID, source department, verification role, date
QA question	Clinical scenario text, question type, difficulty label, reasoning requirement flag
QA answer	Long-form answer text with structured diagnosis, immediate actions, and timeline
Report context	Patient context, encounter context, structured findings, medications, vital trends
Report target	Expected generation, gold-standard impression, recommendations, follow-up fields
Evaluation auxiliaries	Space for QA grading rubrics, factuality checks, report similarity, and clinician review

3.4.4. Recommended splits. Because each hospital is a client, splits must preserve within-client chronology or at least institutional integrity. We recommend a hospital-local split of 70% training, 10% validation, and 20% test for each client, with no leakage of parallel rewritten reports or near-duplicate QA templates across splits. For cross-hospital transfer evaluation, we further recommend leave-one-hospital-out experiments in which one client is completely excluded from training and used solely for transfer assessment. This creates a stringent test of specialty generalization.

3.4.5. Quality control and governance assumptions. The uploaded benchmark specification describes the corpus as a designed multihospital dataset with simulated clients. Accordingly, this manuscript treats the resource as a benchmark design and not as an already audited de-identified clinical repository. For publication-grade release, we recommend three explicit validation layers: (i) clinician review of a stratified subset of QA answers and gold reports; (ii) duplicate and contradiction screening; and (iii) privacy review for any fields that could preserve latent identifying information. This benchmark is therefore best viewed as a strong, structured foundation for a transparent and extensible release pipeline.

4. Problem Formulation and Adaptive FoRA

4.1. Learning setting

Let $\mathcal{C} = \{1, \ldots, n\}$ denote hospital clients. Client i holds local dataset $\mathcal{D}_i = \mathcal{D}_i^{QA} \cup \mathcal{D}_i^{RG}$, where the first component corresponds to medical question answering and the second to report generation. A pretrained backbone model with frozen weights W_0 is shared across clients. The objective is to learn a small set of trainable adaptation parameters ΔW_i at each client without transmitting raw records or full dense model states.

The central challenge is that each $\mathcal{D}_i$ is drawn from a different distribution $P_i(X, Y)$ due to specialty, complexity, writing style, and task mix. We therefore seek a federated training objective that preserves the efficiency benefits of low-rank adaptation while explicitly acknowledging client divergence.

4.2. Efficient fine-tuning view

Standard full fine-tuning updates all model parameters:

$$h = W_0 x. \tag{1}$$

Parameter-efficient tuning instead keeps W_0 fixed and learns a correction term:

$$h = W_0 x + \Delta W x. \tag{2}$$

For conventional LoRA, $\Delta W = BA$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$ with $r \ll d$.

4.3. Micro-level FoRA operator

To increase expressive power without abandoning low-rank efficiency, we define a structured update operator

$$\Delta W_{\text{FoRA}} = E(D + CB + I_r)A, \tag{3}$$

where $A \in \mathbb{R}^{d\times r}$, $E \in \mathbb{R}^{r\times d}$, $D \in \mathbb{R}^{r\times r}$, $B \in \mathbb{R}^{s\times r}$, and $C \in \mathbb{R}^{r\times s}$ with $s \ll d$. The term EA captures a base low-rank update, EDA captures an adapter-style interaction in the rank space, and $ECBA$ introduces a deeper low-dimensional composition. The resulting layer output is

$$h = W_0 x + E(D + CB + I_r)Ax. \tag{4}$$

Moreover, Eq. (3) admits the explicit decomposition

$$\Delta W_{\text{FoRA}} = EA + EDA + ECBA, \tag{5}$$

which makes clear that FoRA preserves the low-rank outer interface while enriching the internal rank-space geometry. The first term is the familiar base adaptation, the second term introduces a learnable interaction within the rank-r latent space, and the third term introduces a deeper bottlenecked composition through the scale rank s. Because every summand factors through A on the right and E on the left, the update remains low-rank in the ambient space even though its internal coupling is more expressive than conventional two-factor adaptation.

Proposition 1. *For any matrices with compatible dimensions, the FoRA update in Eq. (3) satisfies*

$$\operatorname{rank}(\Delta W_{FoRA}) \le r. \tag{6}$$

Consequently, FoRA increases expressivity through richer rank-space interactions rather than by abandoning the low-rank regime.

Proof. Let $M = D + CB + I_r$. Then $\Delta W_{\text{FoRA}} = EMA$ with $M \in \mathbb{R}^{r\times r}$. Since $\operatorname{rank}(EMA) \le \min\{\operatorname{rank}(E), \operatorname{rank}(M), \operatorname{rank}(A)\} \le r$, the claim follows. □

4.4. Meso-level insertion strategy

A low-rank operator is only as useful as its placement. In transformer-based language models, clinically relevant adaptation is distributed across both self-attention and feed-forward pathways, so we attach FoRA to the query, key, and value projections, to the attention output projection when memory allows, to the feed-forward input and output maps, and optionally to lightweight task heads

for QA or report generation. We refer to this complete placement strategy as *Meso-FoRA*. Relative to attention-only or sparse placement strategies, this meso view is designed for tasks where both retrieval-style reasoning and output formatting matter. If $\mathcal{M}$ denotes the set of adapted linear maps and $\ell \in \mathcal{M}$ has width d_ℓ, rank r_ℓ, and scale rank s_ℓ, then the total trainable parameter budget is

$$P_{\text{tot}} = \sum_{\ell \in \mathcal{M}} (2d_\ell r_\ell + r_\ell^2 + 2r_\ell s_\ell), \tag{7}$$

which will later connect directly to memory and communication accounting.

4.5. Macro-level federated optimization with quantized clients

At the macro level, the goal is collaborative adaptation under client-side resource limits. Let $Q_b(W_0)$ denote a b-bit quantized version of the frozen backbone deployed locally, where in our recommended protocol $b = 4$ and the quantization family is NF4 with BF16 compute for adapter operations. Client i solves

$$\min_{\Theta_i} \mathcal{L}_i(\Theta_i; Q_b(W_0), \mathcal{D}_i), \tag{8}$$

where Θ_i is the collection of all FoRA parameters attached to the selected linear layers. Only Θ_i or its delta relative to the previous round is uploaded. The server never needs local records and need not expose the full-precision base weights to clients beyond the chosen protected deployment interface.

4.6. Adaptive aggregation for non-IID hospitals

A central novelty of this paper is the adaptation of macro aggregation to client heterogeneity. Let $\Delta\Theta_i^{(t)}$ denote the client update at round t. Standard federated averaging uses sample-size weights:

$$\Delta\Theta_{\text{avg}}^{(t)} = \sum_{i=1}^{n} \frac{|\mathcal{D}_i|}{\sum_j |\mathcal{D}_j|} \Delta\Theta_i^{(t)}. \tag{9}$$

This is often suboptimal when the largest client is not the most representative or when a small specialty client carries clinically crucial expertise. We therefore define divergence-aware weights

$$\alpha_i^{(t)} = \frac{|\mathcal{D}_i|^\beta \exp(-\lambda\, \delta_i^{(t)})}{\sum_{j=1}^{n} |\mathcal{D}_j|^\beta \exp(-\lambda\, \delta_j^{(t)})}, \tag{10}$$

where $\delta_i^{(t)}$ measures client divergence and $\beta, \lambda \geq 0$ control the influence of client size and divergence penalty. A practical choice is

$$\delta_i^{(t)} = \eta_1\, \text{JS}(p_i^{task}, p^{task}) + \eta_2\, \text{JS}(p_i^{spec}, p^{spec}) + \eta_3\, \text{drift}_i^{(t)}, \tag{11}$$

where the first term compares local versus global task mixture, the second compares specialty composition, and the third captures update drift, for example via cosine dissimilarity or norm ratio to the current global direction.

The server update then becomes

$$\Theta^{(t+1)} = \Theta^{(t)} + \sum_{i=1}^{n} \alpha_i^{(t)} \Delta\Theta_i^{(t)}. \tag{12}$$

Eq. (12) preserves the communication profile of adapter-only aggregation but biases the global model toward clients that are both informative and distributionally compatible.

Proposition 2. *If $\lambda = 0$, then Eq. (10) reduces to size-aware aggregation with weights proportional to $|\mathcal{D}_i|^\beta$. If, in addition, $\beta = 1$, then Eq. (12) reduces exactly to classical FedAvg on the transmitted adapter deltas.*

Proof. Setting $\lambda = 0$ removes the divergence penalty so that the exponential factor in Eq. (10) is constant across clients. The normalization therefore yields $\alpha_i^{(t)} = |\mathcal{D}_i|^\beta / \sum_j |\mathcal{D}_j|^\beta$. The special case $\beta = 1$ is the standard sample-size weighting used by FedAvg. □

To quantify how far the adaptive rule may deviate from sample-size weighting, let $\pi_i = |\mathcal{D}_i| / \sum_j |\mathcal{D}_j|$ and assume $\|\Delta\Theta_i^{(t)}\|_2 \leq G$ for all i. Then

$$\left\| \sum_{i=1}^{n} \alpha_i^{(t)} \Delta\Theta_i^{(t)} - \sum_{i=1}^{n} \pi_i \Delta\Theta_i^{(t)} \right\|_2 \leq G \, \|\alpha^{(t)} - \pi\|_1, \tag{13}$$

which shows that the aggregation perturbation is controlled directly by the distance between the adaptive weight vector and the sample-size baseline. This bound is useful because it separates statistical design from optimization magnitude: as long as the weight shift remains moderate, adaptive aggregation cannot arbitrarily distort the global direction.

Each client optimizes a joint objective

$$\mathcal{L}_i = \lambda_{qa} \mathcal{L}_i^{QA} + \lambda_{rg} \mathcal{L}_i^{RG} + \lambda_{str} \mathcal{L}_i^{struct} + \lambda_{cal} \mathcal{L}_i^{cal}. \tag{14}$$

Here $\mathcal{L}_i^{QA}$ is the token-level generation loss for medical QA, $\mathcal{L}_i^{RG}$ is the token-level generation loss for reports, $\mathcal{L}_i^{struct}$ penalizes structural violations such as missing required sections or action fields, and $\mathcal{L}_i^{cal}$ is an optional calibration term when confidence targets are available. Writing the local QA set as $\mathcal{S}_i^{QA}$ and the report set as $\mathcal{S}_i^{RG}$, one convenient instantiation is

$$\mathcal{L}_i^{QA} = -\frac{1}{|\mathcal{S}_i^{QA}|} \sum_{(x,y) \in \mathcal{S}_i^{QA}} \sum_{t=1}^{|y|} \log p_\Theta(y_t \mid y_{<t}, x), \tag{15}$$

$$\mathcal{L}_i^{RG} = -\frac{1}{|\mathcal{S}_i^{RG}|} \sum_{(x,z) \in \mathcal{S}_i^{RG}} \sum_{t=1}^{|z|} \log p_\Theta(z_t \mid z_{<t}, x), \tag{16}$$

$$\mathcal{L}_i^{struct} = \frac{1}{|\mathcal{S}_i|} \sum_{u \in \mathcal{S}_i} \sum_{k=1}^{K} \omega_k \, \ell_k(g_k(\hat{y}_u), g_k(y_u)), \tag{17}$$

$$\mathcal{L}_i^{cal} = \frac{1}{|\mathcal{S}_i|} \sum_{u \in \mathcal{S}_i} (c_\Theta(u) - s(u))^2, \tag{18}$$

where $g_k(\cdot)$ extracts the kth structured field, such as diagnosis, immediate action, or follow-up timing, ℓ_k is the corresponding mismatch penalty, and $c_\Theta(u)$ denotes a confidence estimate aligned to a supervision target $s(u)$. The presence of structured diagnosis and action fields in the benchmark makes $\mathcal{L}_i^{struct}$ especially useful because it penalizes clinically important omissions that can remain hidden under fluent free-text generation.

To stabilize local optimization under non-IID drift, we further recommend the proximal local objective

$$\widetilde{\mathcal{L}}_i^{(t)}(\Theta) = \mathcal{L}_i(\Theta) + \frac{\mu}{2} \|\Theta - \Theta^{(t)}\|_2^2, \tag{19}$$

which penalizes overly aggressive movement away from the current global adapter state. A client using stochastic gradients $g_{i,b}^{(t,e)}$ at local step e and minibatch b may then update according to

$$\Theta_i^{(t,e+1)} = \Theta_i^{(t,e)} - \eta\Big(g_{i,b}^{(t,e)} + \mu(\Theta_i^{(t,e)} - \Theta^{(t)})\Big), \tag{20}$$

with $\Theta_i^{(t,0)} = \Theta^{(t)}$ and uploaded delta $\Delta\Theta_i^{(t)} = \Theta_i^{(t,E)} - \Theta^{(t)}$ after E local epochs.

At each communication round, the server selects a subset of hospitals and transmits the current adapter state. Every participating client loads the quantized frozen backbone together with the trainable FoRA parameters, optimizes the proximal multi-task objective in Eq. (19) for E local epochs, and uploads only the adapter delta together with optional divergence summaries. The server then computes the adaptive weights, aggregates the received deltas using Eq. (12), and redistributes the updated global adapter state for the next round. If m_t clients participate in round t, the resulting global step can be written compactly as

$$\Theta^{(t+1)} = \Theta^{(t)} + \sum_{i\in\mathcal{S}_t} \alpha_i^{(t)}(\Theta_i^{(t,E)} - \Theta^{(t)}), \tag{21}$$

where $\mathcal{S}_t$ is the active client set. The benchmark specification recommends five clients, full participation per round, batch size 8, three local epochs, 100 rounds, learning rate 2×10^{-4}, rank $r = 8$, scale rank $s = 2$, dropout 0.1, 4-bit NF4 quantization, BF16 compute, and AdamW with cosine scheduling. These settings are intended as a starting reference rather than an immutable rule.

The benchmark assumes an honest-but-curious server and hospital clients that do not share raw records. The primary threat surfaces are: (i) leakage through transmitted adapter updates, (ii) memorization of sensitive strings in local fine-tuning, and (iii) model extraction or misuse of full backbone weights. The macro design reduces risk by keeping records local and transferring only small adapter states. In practice, additional defenses such as secure aggregation, client-level differential privacy, and audit logging can be layered on top. The benchmark is therefore designed to support privacy research rather than claim privacy by architecture alone.

5. Complexity and Communication Analysis

Suppose a square linear map $W \in \mathbb{R}^{d\times d}$ is adapted by a single FoRA module with rank r and scale rank s. Then the trainable parameter count is

$$P_{\text{FoRA}} = dr + rd + r^2 + rs + sr = 2dr + r^2 + 2rs. \tag{22}$$

For LoRA, the comparable count is

$$P_{\text{LoRA}} = 2dr. \tag{23}$$

Proposition 3. *For a fixed layer width d, the relative overhead of FoRA over LoRA for one adapted matrix is*

$$\rho = \frac{P_{FoRA}}{P_{LoRA}} = 1 + \frac{r + 2s}{2d}. \tag{24}$$

Hence, when $d \gg r, s$, the additional trainable cost is modest.

Proof. Substituting P_{FoRA} and P_{LoRA} gives

$$\rho = \frac{2dr + r^2 + 2rs}{2dr} = 1 + \frac{r + 2s}{2d}.$$

The asymptotic claim follows immediately because the additive term vanishes as d dominates the low-rank dimensions. □

This proposition formalizes why richer rank-space interactions can be added without losing the overall efficiency advantage of PEFT.

Let M denote the number of linear maps adapted across the model. Then one client upload per round is

$$\mathcal{O}(M(2dr + r^2 + 2rs)), \tag{25}$$

rather than $\mathcal{O}(Md^2)$ for dense fine-tuning. If secure aggregation is used, the same asymptotic upload size remains, with additional cryptographic overhead determined by the security protocol rather than the model architecture.

Proposition 4. *When only adapter deltas are communicated, the round-wise communication reduction relative to full dense adaptation is approximately*

$$\gamma \approx \frac{2dr + r^2 + 2rs}{d^2}, \tag{26}$$

per adapted matrix, which is small for $r, s \ll d$.

Proof. The numerator is the communicated adapter state and the denominator is the dense weight size. Dividing yields the stated ratio. □

If b_Δ denotes the number of bits used to store each transmitted adapter parameter and m_t clients participate at round t, then a more explicit round-wise traffic model is

$$\mathcal{C}_{\text{up}}^{(t)} = m_t\, b_\Delta \sum_{\ell \in \mathcal{M}} (2d_\ell r_\ell + r_\ell^2 + 2r_\ell s_\ell), \qquad \mathcal{C}_{\text{down}}^{(t)} = m_t\, b_\Delta \sum_{\ell \in \mathcal{M}} (2d_\ell r_\ell + r_\ell^2 + 2r_\ell s_\ell), \tag{27}$$

so that the total bidirectional communication over T rounds is $\sum_{t=1}^{T} (\mathcal{C}_{\text{up}}^{(t)} + \mathcal{C}_{\text{down}}^{(t)})$. This expression emphasizes that communication scales with adapter design and participation rate rather than with the dense backbone size.

The macro design separates memory into the quantized frozen backbone, the optimizer and activations associated with the trainable adapters, and the temporary activations required during local sequence processing. Because the backbone remains frozen and quantized, optimizer state is needed only for the adapters. If b_W is the storage precision for frozen weights, b_Θ is the adapter precision, and A_{act} denotes activation memory, then an abstract memory budget is

$$\mathcal{M}_{\text{client}} \approx b_W |W_0| + b_\Theta P_{\text{tot}} + A_{\text{act}}, \tag{28}$$

which makes clear why reducing P_{tot} and quantizing W_0 are the dominant levers for single-device feasibility.

The benchmark is deliberately configured so that communication and memory are first-class outcomes, not incidental implementation details. Any method evaluated on MedQA-FoRA-MultiHospital should therefore report not only answer quality and generation fidelity but also adapter size, upload volume, and device memory footprint. A strong method for this benchmark is one that preserves specialty knowledge without letting the cost of collaboration erase the practical advantages of PEFT.

6. Benchmark Tasks, Baselines, and Evaluation Protocol

We define four benchmark tasks:

- *Task A: In-hospital medical QA*

  Train and evaluate within each hospital partition. This tests local adaptation quality and establishes per-client ceilings.

- *Task B: Global federated medical QA*

  Train jointly through federated adaptation and evaluate both globally and per client. This measures whether collaboration improves answer quality without erasing specialty-specific competence.

- *Task C: In-hospital report generation*

  Train report generation heads or generative decoders locally to measure documentation performance within specialty-specific conventions.

- *Task D: Cross-hospital transfer*

  Train federated models on four hospitals and evaluate on the held-out fifth hospital. This is the most stringent generalization setting and directly probes robustness to unseen specialty composition.

A benchmark is only useful when it supports meaningful comparison. The baseline suite should therefore include local full fine-tuning as a conceptual upper bound for small-scale experiments, local LoRA or QLoRA to measure hospital-specific adaptation without collaboration, federated LoRA with FedAvg as the direct adapter-only baseline, federated LoRA with heterogeneity-aware optimizers such as FedProx, SCAFFOLD, or FedNova, offsite-style or partial-layer adaptation as reduced-transmission alternatives, and finally the proposed Adaptive FoRA method. When compute permits, both Baichuan2-7B and Baichuan2-13B should be run as primary backbones so that scale sensitivity can be measured rather than assumed.

Table 3 summarizes the core metrics.

Table 3. Recommended evaluation metrics for MedQA-FoRA-MultiHospital

Setting	Metrics
Medical QA	Exact match on structured diagnosis where available; token-level F1; clinician rubric for correctness, safety, completeness, and actionability; calibration score if confidence is generated
Report generation	ROUGE-L, BERTScore, section completeness, medication consistency, timeline consistency, structured recommendation match, clinician factuality score
Federated systems	Round-wise communication volume, adapter size, wall-clock time per round, peak device memory, client drift statistics
Cross-hospital transfer	Held-out hospital macro averages, worst-client score, inter-client variance, specialty-specific breakdown

Because open-ended medical answers can be semantically correct despite surface variation, automatic metrics should be supplemented with clinician review or structured-field matching whenever possible. The benchmark already anticipates this by storing structured answer components.

A robust benchmark paper should define the axes along which future implementations must be tested. The essential ablations are the adapter rank r and scale rank s, attention-only versus full meso placement, sample-size versus divergence-aware aggregation, full precision versus 4-bit and 8-bit client backbones, single-task versus multi-task training, the inclusion or removal of structural losses, full versus partial client participation, and personalized local refinement after federated training. Together these factors map the accuracy–efficiency frontier rather than reporting a single isolated operating point.

For all future benchmark submissions, we recommend reporting at least three random seeds, hospital-level confidence intervals, both macro and micro averages, explicit worst-client performance, and paired significance testing for the key method comparisons. This is important because a method that improves average score while sharply harming one specialty client may be clinically unacceptable.

A minimal but rigorous implementation should tokenize QA and report tasks with explicit task-prefix formatting, preserve hospital-local train/validation/test splits, initialize FoRA matrices with Gaussian initialization for A, C, and D and zeros for B and E, quantize the backbone locally to NF4 while computing adapter paths in BF16, optimize with AdamW and cosine scheduling for the prescribed number of rounds, evaluate both automatic metrics and structured-field recovery, and release scripts for split construction, local training, aggregation, and scoring.

7. Benchmark Instantiation and Dataset-Grounded Evaluation Configuration

The benchmark specification already fixes the central training recipe closely enough to support a publication-grade methods section. The primary backbone is Baichuan2-13B, with Baichuan2-7B retained as a lower-resource comparison point. The federated configuration uses five fixed hospital clients, full client participation in each round, three local epochs, batch size 8, one hundred communication rounds, and client learning rate 2×10^{-4}. Optimization uses AdamW with cosine scheduling, warmup ratio 0.03, and weight decay 0.01. The adapter configuration sets rank $r = 8$, scale rank $s = 2$, LoRA-style scaling factor $\alpha = 32$, and dropout 0.1, with FoRA modules attached across the principal attention and feed-forward linear maps. Client-side training uses 4-bit NF4 quantization with double quantization and BF16 compute. Table 4 summarizes the dataset-grounded configuration used throughout this manuscript.

Table 4. Dataset-grounded experimental configuration instantiated from the uploaded benchmark specification

Item	Instantiated value
Backbone model	Baichuan2-13B (primary), with Baichuan2-7B as the lower-resource comparison model
Tokenizer and max length	Baichuan2 tokenizer family; sequence length 1024 for the baseline reproducibility profile
Hospital clients	5 fixed clients (A–E), full participation in each communication round
Rounds and local epochs	100 rounds, 3 local epochs, batch size 8
Optimization	AdamW, cosine scheduler, warmup ratio 0.03, weight decay 0.01, learning rate 2×10^{-4}
Adapter configuration	FoRA with $r = 8$, $s = 2$, $\alpha = 32$, dropout 0.1, meso-level insertion into attention and FFN linear layers
Quantization	4-bit NF4, double quantization enabled, BF16 compute dtype
Aggregation	Divergence-aware Adaptive FoRA, with FedAvg on adapter deltas as the baseline comparator
Seeds	Minimum of 3 reporting seeds recommended for final benchmark submissions
Hardware reporting rule	Report client device memory, server aggregation environment, and end-to-end wall-clock time explicitly in the final empirical study

The medical QA portion of MedQA-FoRA-MultiHospital contains 50,000 examples distributed across the five hospitals in a deliberately non-IID fashion. The benchmark specification also reports hospital-level complexity ratios, which are useful for interpreting client drift and specialization. To remove ambiguity from future experimental setup, Table 5 converts the recommended 70%/10%/20% training, validation, and test protocol into exact per-hospital counts.

Table 5. Hospital-wise QA composition and exact split counts under the recommended 70%/10%/20% protocol

Client	Specialty	QA pairs	Complexity profile	Train	Val.	Test	Avg. report words
A	Cardiology	12,000	72% high / 28% medium	8,400	1,200	2,400	342
B	Respiratory	10,000	58% high / 42% medium	7,000	1,000	2,000	298
C	Neurology	8,000	85% high / 15% medium	5,600	800	1,600	389
D	General medicine	12,000	40% high / 60% medium	8,400	1,200	2,400	256
E	Pediatrics	8,000	50% high / 50% medium	5,600	800	1,600	267
Total	—	50,000	Non-IID across specialty and complexity	35,000	5,000	10,000	—

These counts make the benchmark immediately executable. They also show that the most difficult client is C, where the neurology partition carries the highest proportion of high-complexity cases, while D provides the broadest mixed-distribution anchor for the global model. The report-length column, taken directly from the uploaded dataset statistics, anticipates that report-generation difficulty will not align perfectly with QA difficulty: C has the longest average reports even though it is not the largest client, whereas D contains the shortest reports despite being one of the largest QA partitions.

The benchmark is designed so that each client differs not only by label frequency but by semantic workload. Cardiology emphasizes hemodynamics, heart-failure optimization, and cardiovascular co-morbidity. Respiratory medicine mixes chronic obstruction, infectious disease, and interstitial patterns. Neurology concentrates more heavily on timeline-sensitive and localization-sensitive cases. General medicine has the widest spread across infectious, gastrointestinal, metabolic, and residual categories, while pediatrics brings age-specific management and different urgency profiles. Table 6 records the hospital-level condition distributions used by the benchmark.

Table 6. Hospital-level condition distributions supplied by the benchmark specification

Client	Condition distribution
A	65% cardiac, 15% endocrine, 10% renal, 10% other
B	55% respiratory, 20% infectious, 15% allergy, 10% other
C	70% neurological, 15% psychiatric, 10% pain, 5% other
D	30% infectious, 25% gastrointestinal, 20% metabolic, 25% other
E	60% pediatric infectious, 20% developmental, 10% genetic, 10% other

This condition mix is precisely what motivates the adaptive weighting mechanism in Eq. (10). A purely sample-size-driven server could easily overweight A and D because they are the largest clients, while under-protecting the highly specialized but smaller neurology and pediatrics partitions. In this benchmark, therefore, the global model should be judged not only by its pooled average but by its ability to preserve minority specialty competence under aggregation.

The uploaded dataset specification provides two fully worked report-generation samples, one from cardiology and one from respiratory medicine, together with gold automatic evaluation fields. These

values are not experimental system results; they are part of the benchmark definition and therefore function as reference targets for future scoring pipelines. Table 7 records the available sample-level report metadata exactly as provided.

Table 7. Report-generation exemplars included in the uploaded benchmark specification. The scores are sample-level gold evaluation fields embedded in the dataset design, not benchmark-wide model results

Client	Report type	Primary gold label	ROUGE-L	BERTScore	Clinical accuracy
A	Discharge summary generation	Acute decompensated HFrEF improved with diuresis and GDMT optimization	0.89	0.93	0.95
B	Pulmonary consult note generation	COPD GOLD Stage 2, Group E	0.86	0.90	0.92

Although only two report exemplars are fully expanded in the current benchmark text, they are useful for defining the expected evaluation style. The first emphasizes chronology, medication reconciliation, and discharge instructions, while the second emphasizes consult structure, pulmonary function interpretation, and plan formatting. This supports the claim that report generation in the benchmark is not generic summarization but clinically structured document synthesis.

The evaluation protocol is fully specified in the uploaded dataset. Rather than leaving the metrics as a future choice, the benchmark aligns each task family with an explicit metric set. Medical QA uses token-level F1, exact match where structured answers permit it, and GPT-4-assisted or clinician-assisted grading through a CA-style rubric. Report generation uses ROUGE-L, BERTScore, and clinical accuracy, supplemented by section completeness and consistency checks. Systems evaluation uses peak GPU memory, training time, throughput, and communication volume. Table 8 consolidates these benchmark-defined metrics.

Table 8. Task-specific and systems-specific metrics instantiated from the benchmark specification

Evaluation block	Instantiated metric set
Medical QA	Token-level F1, exact match where structured labels allow, CA-style assisted grading, diagnosis match, immediate-action recovery, timeline recovery
Clinical reports	ROUGE-L, BERTScore, clinical accuracy, section completeness, recommendation match, medication and chronology consistency
Held-out hospital transfer	Macro average on unseen client, worst-client score, inter-client variance, specialty-level breakdown
Systems and deployment	Peak memory usage (GB), training time, throughput (tokens/s), communication per round, adapter-size footprint
Manual review	Safety-oriented error review, omission analysis, specialty confusion audit, escalation-trigger failures

Since the benchmark is explicitly designed for non-IID federated evaluation, leave-one-hospital-out transfer is not optional but central. Table 9 gives the exact five-fold held-out schedule. These folds can be executed without additional dataset design work.

This schedule is particularly informative because each held-out client removes a different kind of medical competence from the training federation. Holding out C tests the model's ability to transfer timeline-critical neurological reasoning from non-neurology clients, while holding out E tests whether age-specific pediatric management can be reconstructed from adult-dominated training data.

Table 9. Exact leave-one-hospital-out transfer schedule defined for MedQA-FoRA-MultiHospital

Fold	Held-out client	Training clients
F1	A (Cardiology)	B, C, D, E
F2	B (Respiratory)	A, C, D, E
F3	C (Neurology)	A, B, D, E
F4	D (General medicine)	A, B, C, E
F5	E (Pediatrics)	A, B, C, D

The benchmark is also specific enough to define the ablation grid concretely. Instead of leaving ablations as empty result shells, Table 10 fixes the planned rank-scale configurations, and Table 11 fixes the component-removal study corresponding directly to the mathematical design in Section 4. These tables specify what must be run; they do not fabricate what the outcomes will be.

Table 10. Rank and scale-rank ablation grid recommended by the benchmark design

Setting	r	s	Adapter scale	Purpose
A1	4	1	Small	Tests the lowest-capacity FoRA regime under strong efficiency pressure
A2	8	2	Baseline	Matches the default benchmark configuration
A3	12	3	Medium-large	Tests whether richer rank-space structure improves minority-specialty retention
A4	16	4	Large	Tests diminishing returns and communication overhead under higher adapter capacity

Table 11. System and design ablation plan instantiated from the benchmark method

Configuration	Controlled change relative to the full method
Full Adaptive FoRA	Meso-level FoRA, divergence-aware aggregation, multi-task objective, NF4 client quantization
Attention-only placement	Removes FFN adaptation while preserving FoRA in attention projections
FedAvg instead of adaptive aggregation	Replaces Eq. (12) with sample-size averaging on adapter deltas
No quantization	Keeps the frozen backbone in higher precision to isolate the effect of NF4 client compression
Single-task training only	Trains QA or report generation separately to test the value of the joint objective in Eq. (14)

The source methodology and the uploaded dataset jointly emphasize that memory and communication are core scientific outputs. For that reason, the final empirical paper built on this benchmark should always report the systems quantities in Table 12. The benchmark already provides reference resource targets from the parent FoRA methodology for Macro-FoRA operation: approximately 10 GB client memory for a 7B model and 14 GB for a 13B model under the quantized setting, together with explicit reporting of training time and throughput.

Finally, the benchmark supports a clinically meaningful error analysis without any need for placeholder counts. A proper qualitative section should review at least one success case and one failure case from each task family, and it should categorize errors into omission of key actions, unsafe escalation advice, chronology inconsistency, medication mismatch, shallow template compliance, and specialty confusion. These categories are already recoverable from the structured answer fields and the report templates supplied by the benchmark. In this sense, the dataset is not merely large enough for training; it is also structured enough for interpretable post hoc evaluation.

Table 12. Systems quantities that must be reported for a complete empirical benchmark submission

Systems quantity	Reporting rule in the final empirical paper
Peak memory (GB)	Report per client and per backbone size; include 7B and 13B when both are used
Communication per round (MB)	Report upload and download traffic for adapter deltas separately
Round time (s)	Report end-to-end wall-clock time including local training and server aggregation
Throughput (tokens/s)	Report decoder-side or end-to-end throughput under the stated sequence length
Reference Macro-FoRA targets	Use 10 GB (7B) and 14 GB (13B) as methodology-consistent reference resource targets for quantized clients

8. Discussion

Most existing medical instruction datasets are centralized and do not confront privacy-preserving adaptation directly [5, 45–47]. Most federated learning benchmarks, by contrast, are not built around long-form language tasks that mix reasoning, action recommendation, and style-constrained generation [48]. MedQA-FoRA-MultiHospital occupies the intersection of these two needs. Its design makes it possible to study not only whether a model can answer a question, but whether collaborative training can preserve specialty knowledge, documentation structure, and deployability under hospital-like constraints [49].

A second reason the benchmark matters is methodological clarity. The field increasingly reports adapter-based or quantization-based improvements without adequately specifying how non-IID partitions were created, how model quality was judged beyond lexical overlap, or whether resource savings were achieved on the client or only in aggregate. By encoding clients, tasks, schema, and evaluation at the dataset-definition level, the benchmark forces these choices into the open.

Single-source medical QA datasets are useful for measuring medical knowledge, but they do not reflect the institutional specialization that drives real federated deployment [50–55]. Similarly, report-generation datasets often focus on one document type or one modality, such as radiology, without connecting generation to question answering or collaborative adaptation. The proposed benchmark does not replace these resources; rather, it provides a systems-aware layer above them by organizing task families into client-specific partitions and aligning them with a realistic efficient fine-tuning workflow [56, 57].

Several failure modes are foreseeable and should be part of benchmark reporting, including specialty collapse in which the global model overfits to the largest clients and loses niche expertise, template overfitting in which report generation becomes stylistically rigid while factuality deteriorates, quantization fragility that harms clinically critical tokens, answer inflation in which long responses appear confident but omit key immediate actions, and aggregation instability under highly divergent client updates [58]. By explicitly defining structured fields and worst-client reporting, the benchmark makes these problems easier to detect [59].

A recurring weakness in rapidly assembled LLM papers is the presentation of unverified or poorly documented experimental numbers. This manuscript deliberately avoids that failure mode [60]. Its scientific contribution lies in defining the benchmark and a complete adaptive federated method with explicit formulas, complexity analysis, recommended baselines, and evaluation design. Once the benchmark is instantiated with audited splits and trained models, a follow-up paper can attach empirical leaderboards without changing the conceptual core [61–65].

9. Limitations, Ethics, and Reproducibility

The most important limitation is that the current benchmark originates from a structured specification rather than an already released audited clinical corpus. The five hospitals are simulated clients, which is an appropriate design choice for benchmark construction but not a substitute for institution-approved de-identified real-world release. A second limitation is that question and report distributions are manually specified at the client level; although this creates meaningful non-IID structure, it still simplifies the full complexity of healthcare systems, where clients may differ simultaneously by geography, patient socioeconomic profile, coding system, and documentation platform. A third limitation is that no empirical leaderboard is claimed here.

Medical language modeling raises risks that extend beyond ordinary benchmark design. A model can be articulate and still clinically unsafe. It can memorize sensitive spans, hallucinate contraindicated advice, or hide its uncertainty behind confident prose. For these reasons, benchmark use should never treat automatic metrics as sufficient evidence of medical usefulness, should include clinician review for any high-stakes claims, should report worst-client and worst-specialty outcomes rather than only averages, and should treat the benchmark as research infrastructure rather than as proof of clinical readiness.

To support transparent future implementation, benchmark releases should include split-construction scripts and hospital partition manifests, tokenization and prompt-format definitions, local and server training code, aggregation logs and divergence estimates, a hardware profile for each reported experiment, automatic and human evaluation rubrics, and documentation of excluded or corrected examples.

10. Conclusion

This paper introduced MedQA-FoRA-MultiHospital, a benchmark for privacy-preserving medical LLM adaptation under non-IID multihospital partitions, together with Adaptive FoRA, a heterogeneity-aware extension of structured low-rank federated adaptation. The benchmark unifies two clinically meaningful task families—medical question answering and clinical report generation—across five specialty-driven hospital clients [66–68]. The methodology preserves the efficiency advantages of PEFT, expands adaptation across transformer linear maps, and incorporates divergence-aware aggregation for collaborative training under statistical heterogeneity [69–72]. Rather than presenting invented performance numbers, the manuscript contributes a complete benchmark-and-method framework: dataset design, formal problem statement, analytical propositions, evaluation protocol, privacy model, and reproducibility guidance. We hope this paper serves as a rigorous foundation for future empirical implementations and leaderboard-driven studies in federated medical language modeling.

References

[1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, 35*, 27730-27744.

[2] Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., ... & Tang, J. (2023, August). Codegeex: A

pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 5673-5684).

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

[4] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556, 10.*

[5] Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325, 1*(1).

[6] Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing, 10*(3152676), 10-5555.

[7] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* (pp. 1273-1282). Pmlr.

[8] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning, 14*(1-2), 1-210.

[9] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429-450.

[10] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020, November). Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning* (pp. 5132-5143). PMLR.

[11] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems, 33*, 7611-7623.

[12] Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623.*

[13] T Dinh, C., Tran, N., & Nguyen, J. (2020). Personalized federated learning with moreau envelopes. *Advances in neural information processing systems,* 33, 21394-21405.

[14] Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., ... & McMahan, H. B. (2020). Adaptive federated optimization. *arXiv preprint arXiv:2003.00295.*

[15] Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., & Saligrama, V. (2021). Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263.*

[16] Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818.*

[17] Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L., & Yang, Q. (2023). Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049.*

[18] Xiao, G., Lin, J., & Han, S. (2023). Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870.*

[19] Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., ... & Zhou, J. (2024, August). Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5260-5271).

[20] Wang, J., & Li, X. (2024). Position paper: Assessing robustness, privacy, and fairness in federated learning integrated with foundation models. arXiv preprint arXiv:2402.01857.

[21] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *Iclr, 1*(2), 3.

[22] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems, 36*, 10088-10115.

[23] Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., ... & Zhao, T. (2023). Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512.*

[24] Liu, S. Y., Wang, C. Y., Yin, H., Molchanov, P., Wang, Y. C. F., Cheng, K. T., & Chen, M. H. (2024, July). Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning.*

[25] Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., ... & Schölkopf, B. (2023). Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243.*

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

[27] Li, X. L., & Liang, P. (2021, August). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582-4597).

[28] Lester, B., Al-Rfou, R., & Constant, N. (2021, November). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3045-3059).

[29] Zaken, E. B., Goldberg, Y., & Ravfogel, S. (2022, May). Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1-9).

[30] Karimi Mahabadi, R., Henderson, J., & Ruder, S. (2021). Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems, 34*, 1022-1035.

[31] Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems, 35*, 30318-30332.

[32] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323.*

[33] Lin, J., Tang, J., Tang, H., Yang, S., Chen, W. M., Wang, W. C., ... & Han, S. (2024). Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems, 6*, 87-100.

[34] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023, July). Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning* (pp. 38087-38099). PMLR.

[35] Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., & He, Y. (2022). Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems, 35*, 27168-27183.

[36] Xia, M., Gao, T., Zeng, Z., & Chen, D. (2023). Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694.*

[37] Gu, Y., Dong, L., Wei, F., & Huang, M. (2023). Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543.*

[38] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.*

[39] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[40] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67.

[41] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., ... & Tang, J. (2022). Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414.*

[42] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682.*

[43] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., ... & Wu, Z. (2023). Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305.*

[44] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353-355).

[45] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383-2392).

[46] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240.

[47] Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 72-78).

[48] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH), 3*(1), 1-23.

[49] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6), bbac409.

[50] Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., ... & Wu, Y. (2022). Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540.*

[51] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172-180.

[52] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus, 15*(6).

[53] Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences, 11*(14), 6421.

[54] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022, April). Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning* (pp. 248-260). PMLR.

[55] Kim, Y., Wu, J., Abdulle, Y., & Wu, H. (2024, August). MedExQA: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing* (pp. 167-181).

[56] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*(1), 1-9.

[57] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data, 6*(1), 317.

[58] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 590-597).

[59] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74-81).

[60] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675.*

[61] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318).

[62] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics, 9*, 391-409.

[63] Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., ... & Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097.*

[64] Hsu, T. M. H., Qi, H., & Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335.*

[65] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582.*

[66] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557.*

[67] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175-1191).

[68] Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts et al. "Extracting training data from large language models." In *30th USENIX Security Symposium (USENIX Security 21),* pp. 2633-2650. 2021.

[69] Nasr, M., Shokri, R., & Houmansadr, A. (2019, May). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 739-753). IEEE.

[70] Nguyen, A. T., Torr, P., & Lim, S. N. (2022). Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems, 35,* 38831-38843.

[71] Kevin, I., Wang, K., Zhou, X., Liang, W., Yan, Z., & She, J. (2021). Federated transfer learning based cross-domain prediction for smart manufacturing. *IEEE Transactions on Industrial Informatics, 18*(6), 4088-4096.

[72] Wang, X., Hu, J., Lin, H., Liu, W., Moon, H., & Piran, M. J. (2022). Federated learning-empowered disease diagnosis mechanism in the internet of medical things: From the privacy-preservation perspective. *IEEE Transactions on Industrial Informatics, 19*(7), 7905-7913.

[73] Sultana, K., Ahmed, K., Gu, B., & Wang, H. (2023). Elastic optimization for stragglers in edge federated learning. *Big Data Mining and Analytics, 6*(4), 404-420.

[74] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2021, April). Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 487-503).

[75] Sun, X., Ji, Y., Ma, B., & Li, X. (2023). A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*

[76] Bao, Z., Chen, W., Xiao, S., Ren, K., Wu, J., Zhong, C., ... & Wei, Z. (2023). Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346.*

A. Appendix

A.1. Medical QA fields

Each medical QA record should include a unique `id`; a `hospital_id` from A–E; a hospital-aligned `specialty` tag; the free-form clinical prompt in `question.text`; a task descriptor in `question.type`; a difficulty label in `question.difficulty`; the Boolean reasoning flag `question.requires_reasoning`; the long-form gold answer in `answer.text`; structured supervision fields such as `answer.structured.diagnosis`, `answer.structured.immediate_actions`, and `answer.structured.timeline`; and provenance-style metadata such as `metadata.source`, `metadata.verified_by`, and a de-identified `metadata.date`.

A.2. Clinical report fields

Each report-generation record should include `id` and `hospital_id`, a `task_type` such as discharge summary or consult note generation, the structured inputs grouped under `input_encounter_data`, the report target in `expected_generation`, gold reference fields such as `ground_truth_report.impression` and `ground_truth_report.recommendations`, and optional automatic score placeholders reserved for development workflows rather than for final evaluation.

Adaptive FoRA Training Procedure

The training procedure is concise when written in state-space form. First initialize the frozen backbone W_0 and the shared FoRA state $\Theta^{(0)}$. At round t, select participating clients $\mathcal{S}_t$, distribute $\Theta^{(t)}$, and let each client load the quantized backbone $Q_b(W_0)$ with FoRA modules inserted into the chosen attention and FFN layers. Local optimization proceeds by repeated application of Eq. (20) for E epochs, after which each client computes its delta $\Delta\Theta_i^{(t)}$ together with divergence summary $\delta_i^{(t)}$. The server converts these summaries into adaptive weights through Eq. (10), aggregates the received updates through Eq. (12), broadcasts the new global state, and finally evaluates the resulting model on federated validation sets and any held-out hospital used for transfer testing.

A future empirical paper built on this manuscript should report, at minimum, the dataset version and split checksum, backbone model size and tokenizer, the number of adapted layers, rank r, scale rank s, and dropout, the quantization type and compute dtype, the aggregation rule and divergence metric, the number of local epochs and communication rounds together with the participation rate, all QA and report metrics, the size of the clinician-reviewed subset, the total communication volume and device-memory footprint, and an error analysis containing at least ten representative failure cases.

How to cite this article: Zhen Xiong and Jun Li (2026). MedQA-FoRA-MultiHospital: A Non-IID Multihospital Benchmark and Adaptive Federated Low-Rank Framework for Privacy-Preserving Medical Question Answering and Clinical Report Generation. *Bulletin of Computer and Data Sciences*, 7(1), 54-78. DOI: 10.71448/bcds2671-5

Received: 22/1/2026 **Revised:** 25/2/2026 **Accepted:** 18/3/2026 **Published:** 31/03/2026

Bulletin of Computer and Data Sciences is a peer-reviewed open access journal.