Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sequential Federated Fine-Tuning of Large Language Models for Clinical Outcome Prediction in Intensive Care Units

Yifan Gao
The University of Texas at San Antonio
Tongnian Wang
The University of Tennessee at Chattanooga
Yuanxiong Guo
yuanxiong.guo@utsa.edu

The University of Texas at San Antonio
Yanmin Gong
Texas A&M University

Research Article

Keywords: Large Language Model (LLM), Intensive Care Units, Sequential Federated Fine-Tuning, Clinical Outcome Prediction

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11602066/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sequential Federated Fine-Tuning of Large Language Models for Clinical Outcome Prediction in Intensive Care Units

Yifan Gao, BS[1], Tongnian Wang, PhD[2], Yuanxiong Guo, PhD[1], Yanmin Gong, PhD[3]

[1]Department of Information Systems and Cybersecurity, The University of Texas at San Antonio, San Antonio, TX 78249, United States

[2]Gary W. Rollins College of Business, The University of Tennessee at Chattanooga, Chattanooga, TN 37403, United States

[3]School of Engineering Medicine, Texas A&M University, Houston, TX 77030, United States

[*]Corresponding author: Yuanxiong Guo, PhD, Department of Information Systems and Cybersecurity, The University of Texas at San Antonio, 1 UTSA Circle, TX 78249, United States (yuanxiong.guo@utsa.edu)

Abstract

Large language models (LLMs) show promise for clinical outcome prediction in the intensive care unit (ICU), but collaborative fine-tuning across hospitals is constrained by privacy and data-governance concerns. Federated learning (FL) enables collaborative training without exchanging raw data, but conventional parallel FL requires coordinated participation and centralized aggregation, which can be challenging when institutions differ in training availability. We develop and evaluate a novel framework called sequential federated fine-tuning (SFF) framework that enables efficient and privacy-aware collaborative clinical outcome prediction in ICUs across multiple hospitals. By transferring low-rank adapters sequentially from one hospital to the next, SFF supports asynchronous learning, allowing institutions to remain offline until their turn while another institution performs local training. Using 45,149 ICU stays from 27 hospitals in the eICU Collaborative Research Database (eICU-CRD), we applied low-rank adaptation (LoRA) to Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct and compared SFF with centralized fine-tuning, local fine-tuning, zero-shot prompting, and federated averaging (FedAvg) for in-hospital mortality and three-class length-of-stay (LoS) prediction. SFF achieved performance close to centralized fine-tuning while outperforming local training and FedAvg. For mortality prediction, SFF reached AUROCs of 86.08% and 89.24% for Qwen-2.5-7B and LLaMA-3.1-8B, respectively; for LoS classification, it reached 75.32% and 76.57%. The advantage over FedAvg was observed across most hospital cohorts and remained evident when local training data were limited. Our work advances the practical, privacy-aware deployment of LLMs in multi-institutional clinical settings for collaborative clinical outcome prediction in ICUs.

Keywords: Large Language Model (LLM), Intensive Care Units, Sequential Federated Fine-Tuning, Clinical Outcome Prediction

1 Introduction

The intensive care unit (ICU) is a data-intensive clinical environment in which patients generate frequent measurements from vital signs, laboratory tests, and other bedside assessments[1]. Integrating these data is essential for clinical prediction, but conventional machine-learning (ML) models are typically developed for individual tasks[2]. In contrast, large language models (LLMs) offer a reusable pretrained backbone and a natural-language interface that can be adapted to different clinical prediction tasks through fine-tuning[3]. Smaller open-source models (e.g., Qwen-2.5-7B and LLaMA-3.1-8B) are attractive due to computational, privacy, and data-governance considerations. Moreover, these smaller models can benefit from fine-tuning on local clinical data for downstream prediction tasks[4–6].

However, fine-tuning these models requires access to clinical data, which may contain sensitive personal health information (PHI) distributed across different institutions[7–9]. Existing research on privacy in healthcare AI has emphasized that the sensitivity of PHI, together with the need to coordinate data access across healthcare institutions, makes data protection a central concern [10]. Conventional centralized training, which pools data across institutions, raises privacy risks and introduces additional regulatory and data-governance requirements under regulations such as the General Data Protection Regulation (GDPR)[11] and the Health Insurance Portability and Accountability Act (HIPAA)[12]. One way to avoid cross-institutional data sharing is the local approach, in which each institution fine-tunes an LLM solely on its own clinical data[13]. A key limitation of this approach is that institutions cannot leverage knowledge from other institutions[14]. Differences in patient populations and clinical practice across hospitals can lead

to distinct local data distributions, limiting the generalizability of locally trained models to other institutions[15]. Models trained on limited local data may reflect existing biases or demographic imbalances, potentially affecting predictive performance across patient populations[16, 17]. Together, these challenges reinforce the need for collaborative fine-tuning that enables knowledge sharing across institutions to improve robustness while keeping patient data local.

Federated learning (FL) enables collaborative model training without exchanging raw data[13]. However, conventional parallel FL requires coordinated participation across institutions, which can be challenging when institutions differ in network connectivity and training availability [18]. Sequential federated learning (SFL) provides an alternative that reduces the need for simultaneous institutional participation[19]. However, the effectiveness of sequential federated fine-tuning for clinical outcome prediction in ICUs using structured EHR data comprising static patient characteristics and multivariate physiological trajectories remains unclear.

Building on this emerging line of work, we investigate sequential federated fine-tuning (SFF) of LLMs for ICU outcome prediction across healthcare institutions. In-hospital mortality and LoS are widely regarded as key indicators of patient outcomes and resource utilization in ICUs[20–22]. Both are influenced by many interacting clinical factors and remain difficult to predict from EHR data, which comprise time-series measurements (e.g., vital signs and laboratory results) and tabular variables (e.g., demographics and admission details)[4]. In this study, we focus on two ICU prediction tasks using clinical data from the first 24 hours after ICU admission: in-hospital mortality before hospital discharge and LoS classification, which support timely intervention for high-risk patients and ICU bed management, respectively. We explore the following research questions (RQs):

RQ1: How effectively does SFF enable collaborative fine-tuning of LLMs for ICU outcome prediction across institutions without centralizing patient-level data?

RQ2: How does SFF perform across individual hospital cohorts and under varying levels of local training-data availability?

RQ3: What is the effect of fixed cyclic versus stochastic hospital ordering on SFF predictive performance?

To address these questions, this study makes three main contributions. First, we develop and evaluate an SFF framework for collaborative adaptation of open-source LLMs for in-hospital mortality and LoS prediction using structured EHR data from ICU stays.-Second, we benchmark SFF on the multicenter eICU database against centralized fine-tuning, local fine-tuning, zero-shot prompting, and federated averaging (FedAvg) baselines, and evaluate its performance across individual hospital cohorts and varying levels of local training data. Third, we compare fixed cyclic and stochastic hospital-ordering strategies to examine their effects on SFF predictive performance. Across both prediction tasks, SFF achieved performance close to centralized fine-tuning while outperforming local fine-tuning and FedAvg. Our findings offer practical insights for privacy-aware LLM fine-tuning in multi-hospital clinical outcome prediction in ICUs.

2 Related Work

2.1 LLMs for Clinical Outcome Prediction

Large language models (LLMs) have shown increasing promise across healthcare applications[4, 5, 23–25]. Their reusable pretrained backbones and natural-language interfaces enable adaptation to a range of downstream clinical tasks through prompting or fine-tuning. Although frontier proprietary models (e.g., GPT-4o) achieve strong zero-shot performance[4], their reliance on

external APIs can limit real-world clinical deployment[26]. Smaller open-source models, such as Qwen-2.5-7B and LLaMA-3.1-8B, provide an alternative that can be locally deployed and fine-tuned on clinical data for downstream prediction tasks[5, 6]. In this context, recent studies have explored the use of open-source LLMs for clinical prediction tasks using both unstructured clinical text and structured EHR data[4, 27]. For example, Zhou et al.[27] fine-tuned LLMs on clinical notes for early epilepsy detection, and Zhu et al.[4] demonstrated the feasibility of using prompted and fine-tuned LLMs for structured EHR prediction tasks, including in-hospital mortality and length-of-stay (LoS) prediction. However, existing studies have primarily focused on centralized or institution-specific settings, and collaborative LLM fine-tuning for clinical outcome prediction across multiple hospitals remains underexplored.

2.2 Federated Learning for LLM Finetuning

Federated learning (FL) is a distributed learning paradigm in which institutions collaboratively train models without exchanging raw data, and has emerged as a promising approach for reducing direct data sharing in healthcare[28–30]. For example, Dayan et al.[28] applied FL to predict clinical outcomes for COVID-19 patients across 20 institutions worldwide without centralizing patient-level data. Other studies have leveraged FL for medical imaging tasks such as brain tumor segmentation[29] and drug discovery in pharmaceutical collaborations[30].

Common FL architectures include center-to-peer, parallel FL, where a central server aggregates updates from all clients, and peer-to-peer, serial FL, where clients exchange updates directly[31, 32]. Center-to-peer FL enables parallel local training with periodic centralized aggregation, providing an efficient approach to collaborative model optimization[33]. However, parallel FL requires training to take place simultaneously across all participating institutions, which can be challenging when institutions differ in network connectivity and training availability[18].

Additionally, in center-to-peer FL, the central aggregation server can represent a single point of failure or control[34]. In contrast, sequential federated learning (SFL), a peer-to-peer FL method, transfers model updates sequentially across institutions, allowing one institution at a time to be online and update the model while the other institutions remain offline[19].

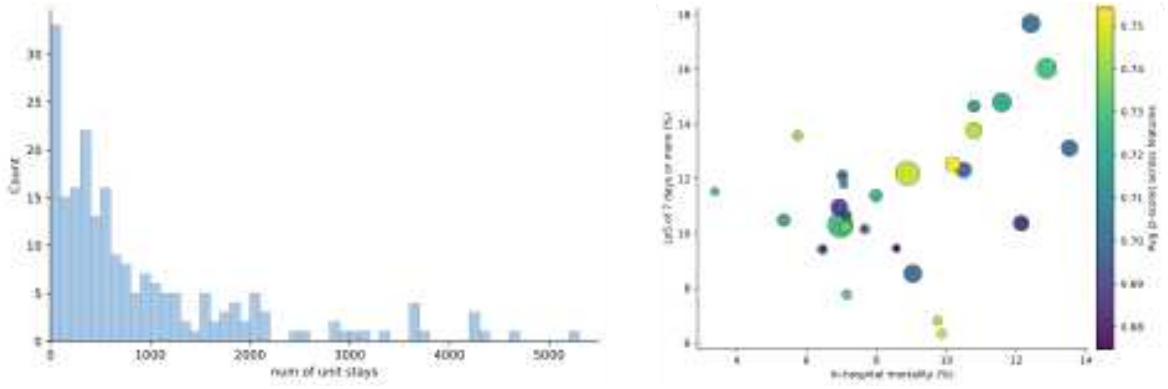
Extending FL to LLM fine-tuning also faces additional scalability constraints, as the size of modern LLMs substantially increases both communication and computation costs[35, 36]. To address this, recent work has combined FL with parameter-efficient fine-tuning such as low-rank adaptation (LoRA), transferring only low-rank adapter parameters to reduce communication and training overhead. For example, Li et al.[37] proposed Fed-MedLoRA for clinical information extraction, and Chen et al.[38] introduced a federated fine-tuning approach for medical question answering following the federated averaging (FedAvg) algorithm. However, these approaches require participating institutions to remain online during parallel training, which can delay the LLM fine-tuning process when an institution is temporarily unavailable. More recently, Liu et al.[39] evaluated sequential multi-site fine-tuning of LLMs for mobility functional status extraction from unstructured clinical notes. However, the study was limited to three institutions and a single fixed site order, without comparison with parallel FL methods.

3 Methods

3.1 Data Sources and Study Population

We use the eICU Collaborative Research Database (eICU-CRD)[40]. Spanning 335 ICU units in 208 hospitals across the United States, it records over 200,000 ICU admissions between 2014 and 2015. These 208 hospitals differ in patient populations, clinical workflows, and documentation practices, contributing to substantial variation across institutions[41]. As shown

in Fig. **1(a)**, the number of ICU stays varies widely across the 208 hospitals. For each hospital admission, we retained only the first ICU stay. We excluded stays with a discharge time at or before ICU admission and those lasting less than 24 hours to ensure a complete 24-hour observation window. For patients with multiple hospital admissions, we retained only the earliest eligible ICU stay to prevent patient overlap across data splits. Finally, we include only hospitals with at least 1,000 qualifying stays to ensure sufficient local data for fine-tuning. This yields the final cohort of 45,149 stays from 27 hospitals. As shown in Fig. **1(b)**, the retained hospitals remain heterogeneous in cohort size, clinical outcomes, and feature distributions.



(a) ICU Stays Distributions (b) Distributions of Hospitals

Fig. 1 Hospital distributions in the eICU-CRD. (a) ICU stay counts across the 208 hospitals. (b) Clinical heterogeneity across the 27 hospitals retained for federated fine-tuning.

For each stay in this cohort, we consider basic demographics together with a set of 17 clinical variables as the input to our model [4]. These demographic and physiologic variables differ between the outcome groups, as summarized in Table **1**. In-hospital mortality, defined as death before hospital discharge[20, 21], occurred in 4,132 of 45,149 stays (9.2%). For LoS, 39,806 (88.2%), 3,750 (8.3%), and 1,593 (3.5%) stays were in the less-than-one-week, one-to-two-week, and more-than-two-week LoS groups[42], respectively. Compared with survivors, non-survivors

were older (median 69 vs 64 years), had higher heart rates (90.3 vs 82.4 bpm), and lower GCS scores (10.0 vs 14.8). A similar gradient appeared across the LoS groups, where longer stays showed lower GCS and higher inspired-oxygen fraction. All reported differences were statistically significant (*p < 0.05).

Table 1. Characteristics of the cohort (N = 45,149 ICU stays across 27 hospitals) presented as medians for continuous variables and n (%) for categorical variables, stratified by in-hospital mortality and LoS. *p value < 0.05. P values for continuous variables are based on pairwise Wilcoxon rank sum test. P values for categorical variables are based on pairwise Pearson's chi-squared tests for proportions. GCS: Glasgow Coma Scale. LoS: Length-of-Stay. BP: Blood Pressure

Variable	In-hospital Mortality		LoS		
	Non-mortality	Mortality	< 1 week	1-2 weeks	> 2 weeks
Patients, n	41,017	4,132	39,806	3,750	1,593
Demographics					
Age (years)	64.0 (52.0–75.0)	69.0 (59.0–79.0)*	65.0 (53.0–76.0)	64.0 (52.0–74.0)*	61.0 (50.0–71.0)*
Sex (Female), n (%)	18,389 (44.8%)	1,843 (44.6%)	18,000 (45.2%)	1,581 (42.2%)*	651 (40.9%)*
Height (cm)	170 (163–178)	170 (163–178)	170 (163–178)	170 (163–178)*	170 (163–180)*
Weight (kg)	81.6 (68.0–98.4)	79.0 (65.0–95.7)*	81.0 (67.6–97.7)	82.7 (68.0–100.0)*	84.5 (70.3–103.5)*
Vital Signs					
Heart rate (bpm)	82.4 (72.1–94.0)	90.3 (77.2–103.3)*	82.4 (72.0–94.0)	88.0 (75.6–100.9)*	88.6 (76.9–102.0)*
Systolic BP (mmHg)	121 (109–134)	114 (104–128)*	120 (109–134)	116 (106–130)*	117 (106–130)*
Diastolic BP (mmHg)	63.8 (56.8–72.0)	60.3 (53.8–68.2)*	63.8 (56.6–72.0)	61.9 (55.5–69.5)*	61.6 (55.5–68.9)*
Mean BP (mmHg)	80.1 (72.6–89.2)	75.8 (68.9–84.4)*	80.0 (72.4–89.2)	77.6 (70.7–86.0)*	77.6 (71.4–85.8)*
Respiratory rate (bpm)	18.4 (16.1–21.3)	20.5 (17.0–24.3)*	18.4 (16.1–21.5)	19.2 (16.3–22.7)*	19.0 (16.1–23.1)*
Temperature (°C)	36.8 (36.6–37.1)	36.8 (36.4–37.2)*	36.8 (36.6–37.1)	36.9 (36.6–37.3)*	37.0 (36.6–37.4)*
Oxygen saturation (%)	97.0 (95.6–98.4)	97.1 (95.2–98.8)	97.0 (95.6–98.4)	97.2 (95.6–98.8)*	97.4 (95.6–98.9)*
Fraction inspired oxygen (%)	40.0 (31.0–50.0)	50.9 (40.0–67.7)*	40.0 (30.5–50.0)	49.3 (40.0–62.0)*	50.0 (40.1–65.7)*
Neurological (GCS)					
GCS total score	14.8 (12.7–15.0)	10.0 (6.0–14.1)*	14.8 (12.9–15.0)	11.5 (8.5–14.7)*	10.0 (7.5–14.0)*

eye opening	4.00 (3.50–4.00)	3.00 (1.47–4.00)*	4.00 (3.50–4.00)	4.00 (3.50–4.00)	3.00 (2.00–3.85)*
motor response	6.00 (6.00–6.00)	5.25 (3.50–6.00)*	6.00 (6.00–6.00)	5.88 (4.90–6.00)*	5.40 (4.33–6.00)*
verbal response	5.00 (3.57–5.00)	2.00 (1.00–4.50)*	5.00 (3.67–5.00)	3.00 (1.00–5.00)*	1.65 (1.00–4.33)*
Lab tests					
Glucose (mg/dL)	128 (107–154)	139 (113–173)*	128 (107–155)	134 (112–163)*	135 (115–161)*
pH	7.37 (7.32–7.41)	7.35 (7.28–7.42)*	7.37 (7.32–7.41)	7.36 (7.31–7.42)*	7.36 (7.30–7.42)*
Other					
Capillary refill rate	0.00 (0.00–0.00)	0.00 (0.00–0.60)*	0.00 (0.00–0.00)	0.00 (0.00–0.17)*	0.00 (0.00–0.30)*

3.2 Supervised Fine-tuning with LoRA

To support multi-institutional collaboration without centralizing raw patient data, we design an SFF framework for LLMs. The framework applies LoRA in the serial federated setting, enabling parameter-efficient fine-tuning across multi-hospital ICU cohorts while keeping raw patient data local. We adopt open-source, instruction-tuned LLMs as the backbones, which can be deployed locally at each hospital without external API requests. To adapt LLMs to clinical prediction, we serialize each ICU stay into a structured text prompt X_p, as illustrated in Fig. **2**. Each prompt includes patient demographics, the 24-hour trajectories of the 17 clinical variables with units and reference ranges, and a task-specific question with predefined response options for in-hospital mortality or LoS prediction. Missing values are represented as "NaN" for numerical variables and "unknown" for categorical variables. We apply supervised fine-tuning with quantized low-rank adaptation (QLoRA)[43], a parameter-efficient fine-tuning method widely used for adapting LLMs to downstream tasks. Specifically, during training, the pretrained weight matrix $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$ is frozen, and only two low-rank trainable matrices $A \in \mathbb{R}^{r \times d_{in}}$ and $B \in \mathbb{R}^{d_{out} \times r}$ are updated, where $r \ll \min(d_{in}, d_{out})$. For a linear projection $h = W_0 x$, the modified forward pass yields:

$$h = W_0 x + \Delta W x = W_0 x + \alpha \cdot BAx,$$

where α is a scaling factor. In the federated setting, only low-rank adapter parameters A, B are trained and transferred, reducing training and communication overhead. At each hospital i, the adapter A, B is updated by minimizing the standard token-level cross-entropy loss on the label tokens over the local dataset $\mathcal{D}_i$ of hospital H_i, while the backbone W_0 remains frozen. We consider N hospitals, where each hospital i holds $\mathcal{D}_i$ with a local loss $\mathcal{L}_i(\theta)$, where θ denotes the collection of trainable LoRA parameters A and B, with A initialized using Kaiming uniform and B initialized to zeros.

[System message]
You are an experienced critical care physician working in an Intensive Care Unit (ICU), skilled in interpreting longitudinal patient data and predicting clinical outcomes.

[User message]
I will provide you with the patient's clinical data from the first 24 hours of their ICU admission. Each feature is presented as a list of 24 hourly values (Hour 1 = first hour after ICU admission, Hour 24 = end of the first day). Missing values are `NaN` for numerical and "unknown" for categorical. Units and reference ranges are listed alongside each feature.

Patient Background:

- Sex: `{{SEX}}`
- Age: `{{AGE}}`

Clinical Features Over First 24 Hours:

- `{{Feature name}}` *(Unit: {{unit}}. Reference: {{range}})*: [`{{h1}}`, `{{h2}}`, ..., `{{h24}}`]
- *...(17 features in total)*

`{{QUESTION}}`

Task 1 – In-hospital mortality
Question: Based on the data above, will this patient die during the current hospital stay?
Answer with only one word: "Yes" or "No".

Task 2 – ICU length-of-stay (3 classes)
Question: Based on the data above, how long will this ICU stay last in total?
Respond with one of the three exact phrases: "Short stay (less than one week)", "Medium stay (one to two weeks)", or "Long stay (more than two weeks)".

Fig. 2 Prompt templates for ICU prediction tasks.

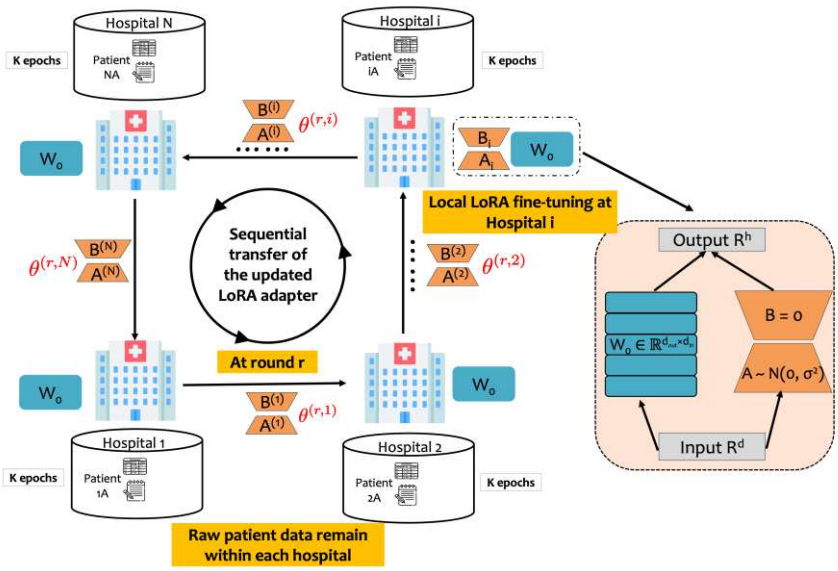


Fig. 3 Workflow of the Sequential Federated LLM Fine-tuning Framework.

3.3 SFF Workflow

The Fig. **3** illustrates the overall workflow of the SFF framework. In SFF, each round consists of sequential updates from N hospitals, where $\theta^{(r,i)}$ denotes the state of the LoRA adapter θ after the hospital H_i completes its local update in round r. During each round, the hospital H_i fine-tunes the LoRA adapter on its local data $\mathcal{D}_i$ for K epochs using the AdamW optimizer and forwards $\theta^{(r,i)}$ directly to the next hospital in a cyclic order (cyclic weight transfer or CWT), while the LLM backbone W_0 remains frozen. We denote this local fine-tuning step by $LocalUpdate(\theta, D_i; K)$ in Algorithms 1 and 2. We consider two hospital-ordering variants: CWT-fixed, which follows a fixed cyclic order across hospitals, and CWT-stochastic (also referred to as stochastic weight transfer, STWT) [44, 45], which randomizes the hospital order at the start of each round. The pseudocodes for both methods are given in Algorithms **1** and **2**.

Algorithm 1: Cyclic weight transfer

Data: Hospitals $\{H_1, \ldots, H_N\}$ with local datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$
Result: Cyclically updated adapter $\theta^{(R,N)}$ after R rounds
Input: Rounds R, local epochs K

for $r = 1$ **to** R **do**
  for $i = 1$ **to** N **do**
    $\theta^{(r,i)} \leftarrow \text{LocalUpdate}\left(\theta^{(r,i-1)}, \mathcal{D}_i; K\right)$
  $\theta^{(r+1,0)} \leftarrow \theta^{(r,N)}$

Algorithm 2: Stochastic weight transfer

Data: Hospitals $\{H_1, \ldots, H_N\}$ with local datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$
Result: Stochastically updated adapter $\theta^{(R,N)}$ after R rounds
Input: Rounds R, local epochs K

for $r = 1$ **to** R **do**
  Sample a random permutation π_r of $\{1, \ldots, N\}$
  for $j = 1$ **to** N **do**
    $i \leftarrow \pi_r(j)$
    $\theta^{(r,j)} \leftarrow \text{LocalUpdate}\left(\theta^{(r,j-1)}, \mathcal{D}_i; K\right)$
  $\theta^{(r+1,0)} \leftarrow \theta^{(r,N)}$

4 Evaluation

We evaluate our method on the real-world eICU dataset and compare our SFF method with the baselines briefly described below:

- **Centralized** fine-tuning, where data from all hospitals are pooled to fine-tune LLMs.
- **Local** fine-tuning, where each hospital fine-tunes a base LLM on its own local data independently.
- **Zero-shot** prompting, where each base LLM is evaluated directly on the test set using structured prompts, without fine-tuning.
- **FedAvg**, a center-to-peer federated method where each hospital fine-tunes the current LoRA adapters locally and a central server averages the updated adapter matrices A and B each round, weighted by each hospital's training dataset size, before sending the averaged adapters back to the hospitals for the next round.

4.1 Evaluation metrics

We evaluate predictive performance using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC), where higher values

indicate better discriminative performance. AUROC summarizes a model's ability to distinguish between outcome classes across decision thresholds, while AUPRC reflects the trade-off between precision and recall and is particularly informative when outcome classes are imbalanced. Because LoS classification is a three-class task, AUROC and AUPRC are macro-averaged across classes using a one-vs-rest scheme, whereas for binary mortality they are computed directly. For each prediction, class probabilities were obtained by applying softmax to the model logits corresponding to the candidate response labels. These probabilities were used to compute AUROC and AUPRC.

4.2 Implementation Details and Statistical Analysis

We fine-tune two base LLMs, Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct, with QLoRA, freezing the 4-bit-quantized backbones and updating only LoRA adapters (r = 8, = 32, dropout 0.05) applied to the query, key, and value projection matrices in multi-head attention (MHA) and linear transformation matrices in feed-forward network (FFN) modules. All models are trained with AdamW at a learning rate of 1×10^{-4}, an effective batch size of 8, and a maximum sequence length of 2048, and the best checkpoint is chosen on the validation set. Both centralized and local fine-tuning are performed for 3 epochs, with the local baseline training a separate adapter per hospital and reporting the mean across the 27 hospitals. For the federated methods (FedAvg, CWT, and STWT), we use the same federated setup. Given the limited number of hospital cohorts, all of them participate in every communication round without subsampling. Because a large number of local epochs can hinder the convergence of FedAvg [33], we set the local epochs per round to = 1 and the number of communication rounds to = 3. Finally, all experiments were performed on 4 NVIDIA A6000 GPUs. Patient data from each hospital are randomly split into 80% for training, 10% for validation, and 10% for testing.

For statistical evaluation, we report bootstrap 95% confidence intervals [46] for the performance. We assess significance against the centralized baseline using two-sided paired bootstrap tests. The local baseline consists of 27 per-hospital models, for which we use a one-sample *t-test*. For the fine-tuning data efficiency analysis, significance between SFF and FedAvg is assessed using the Wilcoxon signed-rank test. Significance levels are denoted *p < 0.05, **p < 0.01, ***p < 0.001.

5 Results

5.1 Overall Model Performance

Table **2** reports in-hospital mortality and LoS prediction performance for two LLMs on the pooled test set from all 27 hospitals. For local training, the results are averaged across the 27 local models, each evaluated on its corresponding hospital's test set. For in-hospital mortality prediction, the SFF methods achieved performance close to centralized fine-tuning and outperformed local training, zero-shot prompting, and FedAvg across both base LLMs. For Qwen-2.5-7B-Instruct, CWT-stochastic reached the highest AUROC (86.08, 95% CI: 81.99-89.93), close to centralized fine-tuning (85.04, 95% CI: 80.72-89.13) and outperforming local training (73.58, 95% CI: 71.51-79.65) and FedAvg (80.59, 95% CI: 69.27-87.75). For LLaMA-3.1-8B-Instruct, CWT-fixed achieved an AUROC of 89.24 (95% CI: 85.30-93.25), close to centralized fine-tuning (88.93, 95% CI: 84.81-94.02) and outperforming local training (78.69, 95% CI: 70.40-83.23) and FedAvg (83.43, 95% CI: 78.09-91.28), while CWT-stochastic obtained the highest AUPRC (51.79, 95% CI: 39.55-63.67). Compared with FedAvg, SFF improved AUROC by 6.8% (86.08 vs 80.59) on Qwen-2.5-7B-Instruct and 7.0% (89.24 vs 83.43) on LLaMA-3.1-8B-Instruct.

Table 2. Performance comparison on in-hospital mortality and LoS prediction. The best result within each model is shown in **bold.** Asterisks indicate statistically significant differences from centralized fine-tuning: *p < 0.05, **p < 0.01, and ***p < 0.001.

Model	Method	In-hospital Mortality		LoS	
		AUROC(↑)	AUPRC(↑)	AUROC(↑)	AUPRC(↑)
Qwen-2.5-7B-Instruct	Centralized (Upper bound)	85.04[80.72,89.13]	45.36[34.92,56.16]	75.56[72.96,78.14]	42.07[40.66,44.32]
	Local (Avg, baseline)	73.58[71.51,79.65]	24.44[22.18,26.70]	60.93[56.97,62.91]***	36.23[35.28,37.88]***
	Zero-shot (baseline)	71.50[66.26,76.64]	21.01[15.96,27.64]	52.33[49.31,55.31]	34.11[33.50,35.33]
	FedAvg	80.59[69.27,87.75]*	28.51[17.76,31.89]	64.94[60.98,66.83]**	37.60[36.50,39.38]**
	CWT-fixed	85.37[81.22,89.36]	44.82[35.05,55.94]	75.22[72.56,77.92]	42.08[40.74,44.30]
	CWT-stochastic (STWT)	**86.08[81.99,89.93]**	**46.03[36.10,56.69]**	**75.32[72.65,78.01]**	**42.17[40.81,44.35]**
LLaMA-3.1-8B-Instruct	Centralized (Upper bound)	88.93[84.81,94.02]	48.26[36.79,61.21]	76.99[75.04,80.31]	43.53[41.96,45.86]
	Local (Avg, baseline)	78.69[70.40,83.23]***	30.27[19.98,37.65]***	63.76[58.85,64.67]***	36.60[35.69,38.21]***
	Zero-shot (baseline)	65.52[59.94,71.28]	18.48[13.82,25.35]	55.92[52.87,58.99]	36.94[35.20,39.50]
	FedAvg	83.43[78.09,91.28]***	36.85[25.76,49.91]***	69.70[63.88,71.47]***	38.31[37.22,40.40]***
	CWT-fixed	**89.24[85.30,93.25]**	49.10[37.82,62.14]	**76.57[73.85,80.08]**	43.45[41.82,46.09]
	CWT-stochastic (STWT)	88.99[85.03,93.30]	**51.79[39.55,63.67]**	76.42[74.44,79.64]	**43.75[42.47,46.24]**

For LoS prediction, CWT-stochastic achieved an AUROC of 75.32 (95% CI: 72.65-78.01) for Qwen-2.5-7B-Instruct, close to centralized fine-tuning (75.56, 95% CI: 72.96-78.14) and outperforming local training (60.93, 95% CI: 56.97-62.91) and FedAvg (64.94, 95% CI: 60.98-66.83). For LLaMA-3.1-8B-Instruct, CWT-fixed achieved an AUROC of 76.57 (95% CI: 73.85-80.08), close to centralized fine-tuning (76.99, 95% CI: 75.04-80.31) and outperforming local training (63.76, 95% CI: 58.85-64.67) and FedAvg (69.70, 95% CI: 63.88-71.47), while CWT-stochastic obtained the highest AUPRC (43.75, 95% CI: 42.47-46.24). Compared with FedAvg, SFF improved AUROC by 16.0% (75.32 vs 64.94) on Qwen-2.5-7B and 9.9% (76.57 vs 69.70) on LLaMA-3.1-8B.

5.2 Robustness of Model Performance Across Hospitals

To examine whether the overall performance gains were consistent across hospitals, we compared FedAvg with the two SFF variants separately at each of the 27 hospitals. As shown in Fig. **4**, SFF showed consistent performance gains over FedAvg at most hospitals for both mortality and LoS prediction. For in-hospital mortality prediction, the improvement was most consistent with Qwen-2.5-7B-Instruct, where the win rates of CWT-fixed and CWT-stochastic reached 92.6% (25 of 27 hospitals) and 96.3% (26 of 27 hospitals) for AUROC, and 96.3% and 100% (27 of 27 hospitals) for AUPRC, respectively.

With LLaMA-3.1-8B-Instruct, the corresponding win rates were lower, particularly for AUROC, indicating that the advantage over FedAvg was less consistent across hospitals. For LoS prediction, SFF showed more consistent gains across hospitals for both base models. Both variants outperformed FedAvg in AUROC at 100% of hospitals with Qwen-2.5-7B and at 96.3% of hospitals with LLaMA-3.1-8B-Instruct. For AUPRC, the corresponding proportions were 96.3% and 100% for Qwen-2.5-7B-Instruct, and 92.6% and 96.3% for LLaMA-3.1-8B-Instruct. These results show that SFF improved performance across a broad range of hospitals, with the most consistent gains observed for LoS prediction.

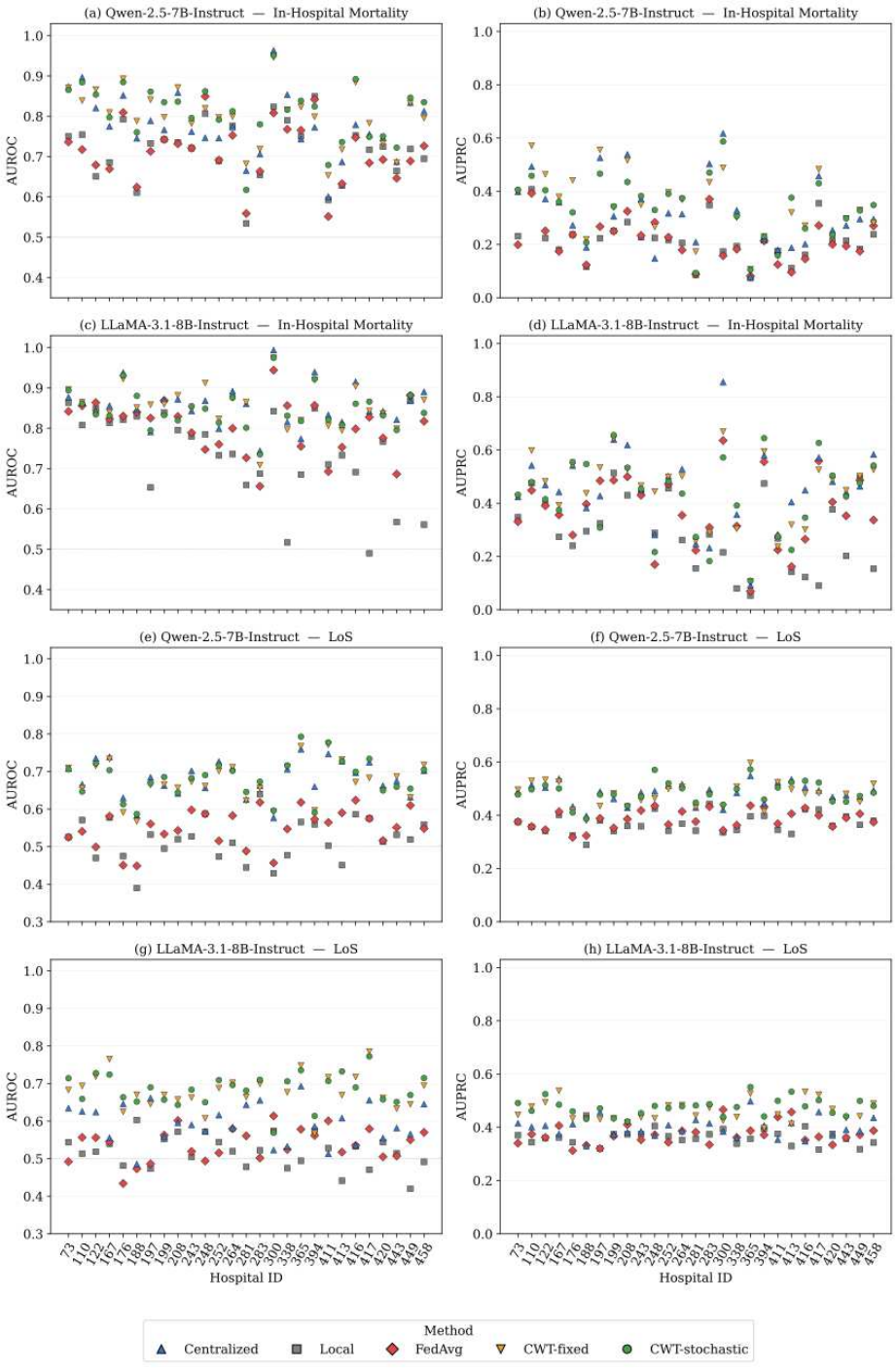


Fig. 4 Panels (a)–(h) show per-hospital AUROC and AUPRC for in-hospital mortality and LoS prediction using Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct. Results are shown for centralized, local, FedAvg, CWT-fixed, and CWT-stochastic across the 27 hospitals. Each point shows performance within one hospital; Table 2 reports performance on the pooled test set.

5.3 Fine-Tuning Data Efficiency Analysis

We further examined the influence of local training data size on model performance by fine-tuning the LLMs on randomly sampled subsets of each hospital's training data, ranging from 10% to 100%. As shown in Fig. **5**, predictive performance generally improved as more local training data became available. Across the evaluated data ratios, SFF remained close to centralized fine-tuning while generally outperforming FedAvg and local training in both AUROC and AUPRC, indicating that the performance advantage of SFF was maintained across varying levels of local data availability.

For LLaMA-3.1-8B-Instruct, with only 10% of the local training data, CWT-fixed and CWT-stochastic achieved relative AUROC improvements over FedAvg of 12.5% (87.28 vs. 77.57) and 10.1% (85.39 vs. 77.57) for in-hospital mortality prediction, and 11.4% (71.51 vs. 64.18) and 13.6% (72.88 vs. 64.18) for LoS classification, respectively. At 100% of the local training data, the corresponding relative improvements were 7.0% (89.24 vs. 83.43) and 6.7% (88.99 vs. 83.43) for in-hospital mortality prediction and 9.9% (76.57 vs. 69.70) and 9.6% (76.42 vs. 69.70) for LoS prediction. Similarly, for Qwen-2.5-7B-Instruct, both SFF variants remained close to centralized fine-tuning across the evaluated data ratios.

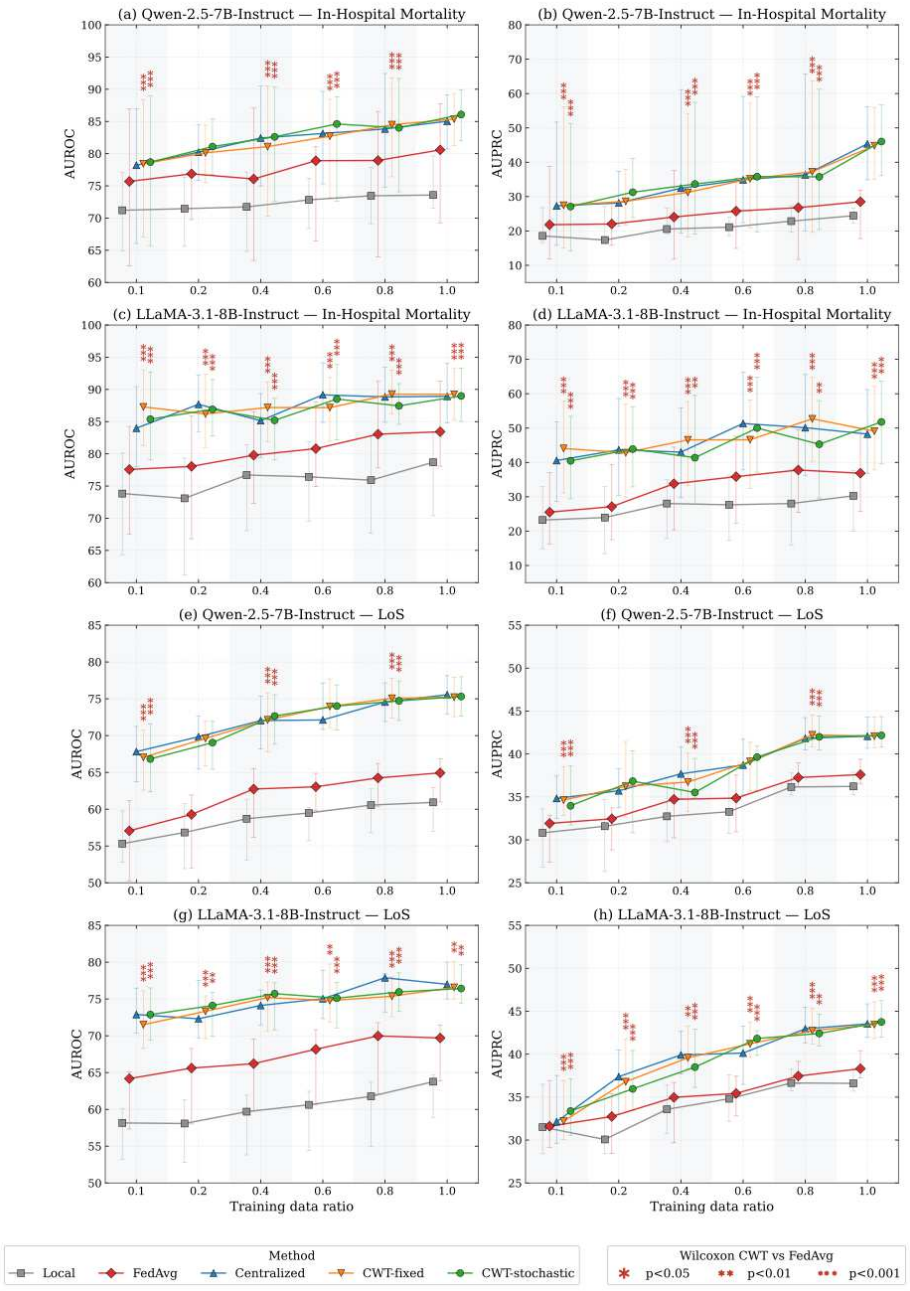
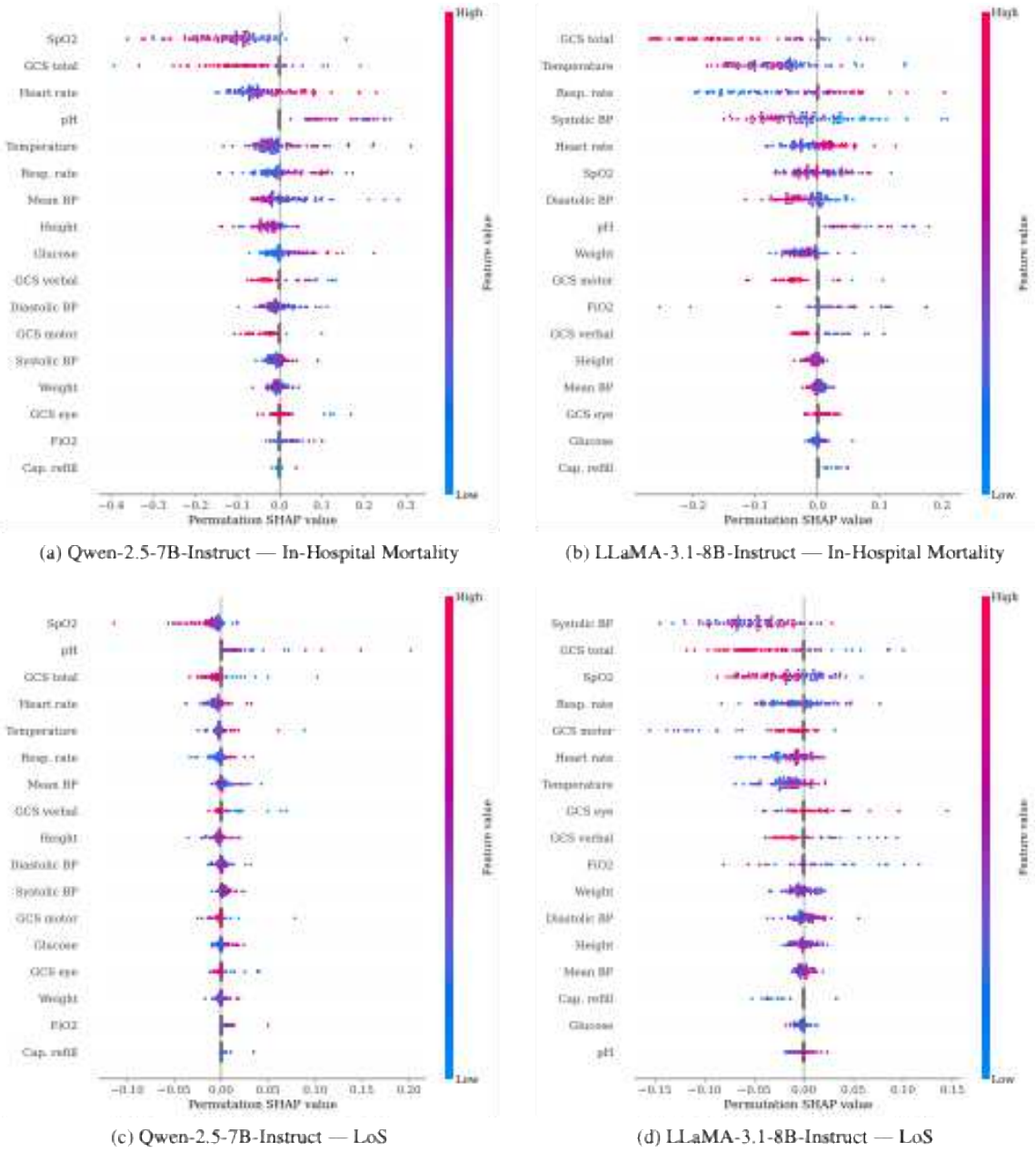


Fig 5. Performance across fine-tuning data ratios. Panels (a)–(h) present AUROC and AUPRC across different proportions of local training data for in-hospital mortality and LoS prediction. Asterisks indicate statistically significant differences between each SFF variant and FedAvg based on Wilcoxon signed-rank tests: *p < 0.05, **p < 0.01, and ***p < 0.001.



(a) Qwen-2.5-7B-Instruct — In-Hospital Mortality

(b) LLaMA-3.1-8B-Instruct — In-Hospital Mortality

(c) Qwen-2.5-7B-Instruct — LoS

(d) LLaMA-3.1-8B-Instruct — LoS

Fig. 6. Feature attribution results for in-hospital mortality and LoS prediction using Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct. Features are ranked by mean absolute attribution value, with color indicating the corresponding feature value.

5.4 Interpretability Analysis with Permutation SHAP

To interpret how clinical variables influenced the predictions of the SFF-based models, we applied Permutation SHapley Additive exPlanations (SHAP) [47], a model-agnostic Shapley-value attribution method that has also been used in LLM explainability studies [48, 49]. We focused the analysis on the CWT-stochastic variant. Feature attributions were computed for the 17 clinical variables included in the input prompt. For each ICU stay, Permutation SHAP started from a baseline prompt in which all values from the first 24 hours were masked as missing and sequentially revealed each variable according to a random ordering. Attributions were estimated using 10 permutations per stay for a stratified subsample of 100 ICU stays for each combination of model and prediction task, drawn from the shared held-out test set. Fig. **6** presents the global feature attribution results for in-hospital mortality and ICU LoS prediction. Positive attribution values indicate that revealing the feature from the all-masked baseline increased the predicted probability of the target class, whereas negative values indicate that revealing the feature decreased the predicted probability.

For in-hospital mortality prediction, GCS total score, respiratory rate, heart rate, and oxygen saturation consistently showed strong influence across both models, although their relative importance varied. For LoS prediction, GCS total score and oxygen saturation were consistently among the most influential variables, while Qwen-2.5-7B-Instruct placed greater importance on pH and heart rate and LLaMA-3.1-8B-Instruct emphasized systolic blood pressure and respiratory rate. Across the two prediction tasks, GCS total score and oxygen saturation emerged as common influential features. Overall, the feature attribution patterns highlighted clinically relevant variables related to neurological status and oxygenation across both ICU prediction tasks. Identifying the features that contribute most strongly to mortality and LoS predictions can provide clinicians with additional context for interpreting model predictions. This information

may help clinicians understand why a patient is assigned a higher mortality risk or a particular ICU LoS category and compare the model's prediction with bedside assessments of neurological status, respiratory function, and overall clinical severity.

6 Discussion

This study evaluated whether sequential federated fine-tuning can support collaborative adaptation of LLMs for ICU outcome prediction when patient-level data cannot be centrally pooled across hospitals. Across the 27 hospitals represented in the eICU cohort, SFF achieved performance close to centralized fine-tuning and outperformed local training and FedAvg for both ICU prediction tasks. These gains were observed across most hospitals and were maintained even when local training data were limited. Given the substantial heterogeneity in patient populations, clinical practices, and EHR systems across institutions [15, 37], these findings suggest that sequential collaboration can support clinical prediction across diverse hospital cohorts.

The difference between SFF and FedAvg may be particularly relevant in multi-hospital settings, where patient populations and clinical data distributions can vary substantially across institutions [31, 37, 38]. FedAvg relies on independently optimized local updates that are aggregated at a central server, and this aggregation can be sensitive to cross-site heterogeneity when local training objectives differ [31, 33]. In contrast, SFF transfers a shared adapter sequentially from one hospital to the next, allowing each institution to continue fine-tuning from knowledge accumulated at previously visited hospitals. One possible explanation for the robustness of SFF across hospitals, particularly when local training data were limited, is that sequential transfer may better preserve knowledge accumulated across institutions. Furthermore, SFF reduces the need for synchronized institutional participation because training at the active hospital does not

require the remaining hospitals to be online. Compared with conventional parallel FL, this design requires less cross-site coordination and may make collaborative fine-tuning better suited to settings with intermittent connectivity and uneven access to local computing resources.

These findings also have important implications for clinical prediction in multi-institutional ICU settings. Mortality prediction can support early identification of high-risk patients, whereas LoS prediction can inform ICU bed management and resource planning [20, 21]. The observed performance gains when local training data were limited may be particularly relevant for institutions with smaller local datasets, which could benefit from knowledge accumulated through multi-hospital collaboration. Beyond predictive performance, the interpretability analysis further identified clinically relevant variables, including neurological status and oxygenation-related measures, among the most influential variables across the two ICU tasks. This alignment with established clinical indicators provides additional clinical context for the model predictions and supports their interpretability.

This study has several important limitations. First, we used clinical information from only the first 24 hours after ICU admission and focused on two outcomes, which may not capture later changes in patient condition or the full variation in ICU length of stay. Second, all hospital cohorts were derived from the eICU-CRD, which may limit generalizability to other health systems. Third, although SFF keeps raw patient data local, transferred adapter parameters may still be vulnerable to information leakage, and formal privacy mechanisms were not evaluated in this study. Finally, we did not systematically assess sensitivity to different hospital sequences or potential forgetting during sequential fine-tuning.

7 Conclusion

In this retrospective study, we examined whether SFF can support collaborative adaptation of LLMs for ICU outcome prediction without centralizing patient-level data. Using the multicenter eICU-CRD, we evaluated SFF for in-hospital mortality and ICU LoS prediction across 27 hospital cohorts. SFF achieved performance close to centralized fine-tuning and consistently outperformed FedAvg and local training. These findings suggest that sequential federated fine-tuning can support cross-institutional LLM adaptation without centralized model aggregation. Future work should evaluate privacy-preserving mechanisms, systematically assess sensitivity to hospital ordering, and prospectively validate the framework across independent health systems.

Funding The work of Y. Guo and Y. Gong was partially supported by the U.S. National Science Foundation.

Conflicts of interest The authors declare no competing interests.

Data availability The data underlying this article are available at https://physionet.org/content/eicu-crd/2.0/.

References

1. Sauer CM, Dam TA, Celi LA, et al (2022) Systematic review and comparison of publicly available ICU data sets—A decision guide for clinicians and data scientists. Crit Care Med 50:e581–e588. https://doi.org/10.1097/CCM.0000000000005517
2. Wornow M, Xu Y, Thapa R, et al (2023) The shaky foundations of large language models and foundation models for electronic health records. npj Digit Med 6:135. https://doi.org/10.1038/s41746-023-00879-8
3. Jiang LY, Liu XC, Nejatian NP, et al (2023) Health system-scale language models are all-purpose prediction engines. Nature 619:357–362. https://doi.org/10.1038/s41586-023-06160-y
4. Zhu Y, Gao J, Wang Z, et al (2026) ClinicRealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. npj Digit Med 9:319. https://doi.org/10.1038/s41746-026-02539-z

5. Wang Y, Dai Y, Wang R, et al (2026) Integrating large language models for enhanced predictive analytics in healthcare. npj Digit Med 9:416. https://doi.org/10.1038/s41746-026-02572-y

6. K. Rostam ZR, Kertész G (2025) Advances in pre-trained Language models for domain-specific text classification: A systematic review. ACM Trans Intell Syst Technol 16:124:1-124:41. https://doi.org/10.1145/3763002

7. Hager P, Jungmann F, Holland R, et al (2024) Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 30:2613–2622. https://doi.org/10.1038/s41591-024-03097-1

8. Artsi Y, Sorin V, Glicksberg BS, et al (2025) Challenges of implementing LLMs in clinical practice: perspectives. J Clin Med 14:6169. https://doi.org/10.3390/jcm14176169

9. Schaye V, DiTullio D, Guzman BV, et al (2025) Large language model–based assessment of clinical reasoning documentation in the electronic health record across two institutions: Development and validation study. J Med Internet Res 27:e67967. https://doi.org/10.2196/67967

10. Murdoch B (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 22:122. https://doi.org/10.1186/s12910-021-00687-3

11. Voigt P, Von Dem Bussche A (2017) The EU general data protection regulation (GDPR). Springer International Publishing

12. Department of Health and Human Services Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the health information technology for economic and clinical health act and the genetic information nondiscrimination act; other modifications to the HIPAA rules.

13. Rieke N, Hancox J, Li W, et al (2020) The future of digital health with federated learning. npj Digit Med 3:119. https://doi.org/10.1038/s41746-020-00323-1

14. Yang J, Soltan AAS, Clifton DA (2022) Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. npj Digit Med 5:69. https://doi.org/10.1038/s41746-022-00614-9

15. Kiser AC, Eilbeck K, Ferraro JP, et al (2022) Standard vocabularies to improve machine learning model transferability with electronic health record data: Retrospective cohort study using health care–associated infection. JMIR Med Inform 10:e39057. https://doi.org/10.2196/39057

16. Mahajan A, Obermeyer Z, Daneshjou R, et al (2025) Cognitive bias in clinical large language models. npj Digit Med 8:428. https://doi.org/10.1038/s41746-025-01790-0

17. Ji Y, Ma W, Sivarajkumar S, et al (2025) Mitigating the risk of health inequity exacerbated by large language models. npj Digit Med 8:246. https://doi.org/10.1038/s41746-025-01576-4

18. Chang K, Balachandar N, Lam C, et al (2018) Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association 25:945–954. https://doi.org/10.1093/jamia/ocy017

19. Bo Z-H, Guo Y, Lyu J, et al (2023) Relay learning: a physically secure framework for clinical multi-site deep learning. npj Digit Med 6:204. https://doi.org/10.1038/s41746-023-00934-4

20. Garber S, Okhrin Y (2026) Machine learning for intensive care unit length-of-stay prediction: A simulation-based approach to bed capacity management. Med Decis Making 46:355–370. https://doi.org/10.1177/0272989X251406639

21. Rocheteau E, Liò P, Hyland S (2021) Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: Proceedings of the Conference on Health, Inference, and Learning. Association for Computing Machinery, pp 58–68

22. Hang Y, Liu L, Chen R, et al (2025) Predicting mortality risk in the intensive care unit using a hierarchical inception network for heterogeneous time series. Biomedical Signal Processing and Control 107:107759. https://doi.org/10.1016/j.bspc.2025.107759

23. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al (2023) Large language models in medicine. Nat Med 29:1930–1940. https://doi.org/10.1038/s41591-023-02448-8

24. Bedi S, Cui H, Fuentes M, et al (2026) Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med 32:943–951. https://doi.org/10.1038/s41591-025-04151-2

25. Singhal K, Tu T, Gottweis J, et al (2025) Toward expert-level medical question answering with large language models. Nat Med 31:943–950. https://doi.org/10.1038/s41591-024-03423-7

26. Chen L-C, Zack T, Mandair D, et al (2025) From zero-shot to bedside: A practical playbook for adapting open-source large language models to clinical symptom extraction. Proceedings of Machine Learning Research 297:1023

27. Zhou S, Yu K, Zhan Z, et al (2026) Early epilepsy detection from electronic health records with large language models. npj Digit Med. https://doi.org/10.1038/s41746-026-03346-2

28. Dayan I, Roth HR, Zhong A, et al (2021) Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 27:1735–1743. https://doi.org/10.1038/s41591-021-01506-3

29. Pati S, Baid U, Edwards B, et al (2022) Federated learning enables big data for rare cancer boundary detection. Nat Commun 13:7346. https://doi.org/10.1038/s41467-022-33407-5

30. Heyndrickx W, Mervin L, Morawietz T, et al (2024) MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model 64:2331–2344. https://doi.org/10.1021/acs.jcim.3c00799

31. Kairouz P, McMahan HB, Avent B, et al (2021) Advances and open problems in federated learning. arXiv 10.48550/arXiv.1912.04977

32. Nguyen QN, Wu H, Pontikos N, Wang SY (2026) Addressing generalizability in clinical named entity recognition: Federated learning or large language models? AMIA Annual Symposium Proceedings

33. McMahan B, Moore E, Ramage D, et al (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, pp 1273–1282

34. Kuo T-T, Gabriel RA, Koola J, et al (2025) Distributed cross-learning for equitable federated models privacy-preserving prediction on data from five california hospitals. Nat Commun 16:1371. https://doi.org/10.1038/s41467-025-56510-9

35. Sun Y, Li Z, Li Y, Ding B (2024) Improving LoRA in privacy-preserving federated learning. In: In International conference on learning representations. pp 17978–17994

36. Wu Y, Tian C, Li J, et al (2026) A survey on federated fine-tuning of large language models. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2503.12016

37. Li A, Chen Y, Long W, et al (2026) Toward federated large language models in medicine: a parameter-efficient framework for privacy-preserving, multi-institutional adaptation. npj Digit Med. https://doi.org/10.1038/s41746-026-03064-9

38. Chen Z, Ji Y, Wang B, et al (2025) Flow of knowledge: Federated fine-tuning of LLMs in healthcare under Non-IID conditions. arXiv 10.48550/arXiv.2510.00543

39. Liu X, Garg M, Jeon E, et al (2026) Sequential multi-site fine-tuning for incremental deployment of large language models for mobility functional status extraction. JAMIA Open 9:ooag139. https://doi.org/10.1093/jamiaopen/ooag139

40. Pollard TJ, Johnson AEW, Raffa JD, et al (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. Scientific Data 5:180178

41. Zhu H, Bai J, Li N, et al (2025) FedWeight: Mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. npj Digit Med 8:286. https://doi.org/10.1038/s41746-025-01661-8

42. Chen C, Yu J, Chen S, et al (2026) ClinicalBench: Can LLMs beat traditional ML models in clinical prediction? In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp 10567–10577

43. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023) QLORA: Efficient finetuning of quantized LLMs. In: 37th Conference on Neural Information Processing Systems. pp 10088–10115

44. Nguyen QN, Wu H, Pontikos N, Wang SY (2026) Addressing generalizability in clinical named entity recognition: Federated learning or large language models? AMIA Annu Symp Proc 2025:949–958

45. Darzi E, Sijtsema NM, Van Ooijen PMA (2024) A comparative study of federated learning methods for COVID-19 detection. Sci Rep 14:3944. https://doi.org/10.1038/s41598-024-54323-2

46. Carpenter J, Bithell J (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 19:1141–1164

47. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing System

48. Contreras M, Kapoor S, Zhang J, et al (2025) A large language model for delirium prediction in the intensive care unit using structured electronic health records. Sci Rep 15:38890. https://doi.org/10.1038/s41598-025-22634-7

49. Paes LM, Wei D, Do HJ, et al (2025) Multi-level explanations for generative language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp 32291--32317