

Mental Health Counseling LLMs in Mixed IID/Non-IID Federations

Shun Aso
shun_aso@keio.jp
Faculty of Environment
and Information Studies
Keio University
Fujisawa, Kanagawa, Japan

Satoki Hamanaka
0216hamanaka@keio.jp
Graduate School of Media
and Governance
Keio University
Fujisawa, Kanagawa, Japan

Jin Nakazawa
jin@sfc.keio.ac.jp
Faculty of Environment
and Information Studies
Keio University
Fujisawa, Kanagawa, Japan

Tadashi Okoshi
slash@sfc.keio.ac.jp
Faculty of Environment
and Information Studies
Keio University
Fujisawa, Kanagawa, Japan

Abstract

Federated Learning (FL) only shares model parameters while keeping raw data local, enabling secure fine-tuning of Large Language Models (LLMs) on sensitive clinical counseling transcripts. This offers a promising path toward accessible, high-quality mental health support. In a realistic clinical federation, general hospitals and specialized clinics coexist. The former holds topically diverse data approximating the global distribution (Identically and Independently Distributed or IID), while the latter holds topically concentrated Non-IID data. However, most existing FL methods do not account for this co-existence of IID/Non-IID clients. To address this problem, we propose two methods. (1) **Heterogeneity-Aware DualLoRA** assigns to each client global and local adapters with adaptive epochs based on heterogeneity score. (2) **Selective Federated Aggregation** routes IID clients into federated fine-tuning of global adapter and Non-IID clients into private fine-tuning of local adapter. Experiments show that both methods consistently outperform standard FL method (FedAvg) in counseling quality and personalization. Furthermore, Selective Federated Aggregation matches the strongest baseline (Dual LoRA) at 70% lower upload cost, making it highly practical for real-world clinical deployment. Our source code is publicly available at <https://github.com/shunasokeio/flm-mentalhealth/>.

CCS Concepts

• **Human-centered computing** → Ubiquitous and mobile computing systems and tools; • **Computing methodologies** → Natural language processing.

Keywords

Federated Learning, large language models, mental health, LoRA

ACM Reference Format:

Shun Aso, Satoki Hamanaka, Jin Nakazawa, and Tadashi Okoshi. 2018. Mental Health Counseling LLMs in Mixed IID/Non-IID Federations. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

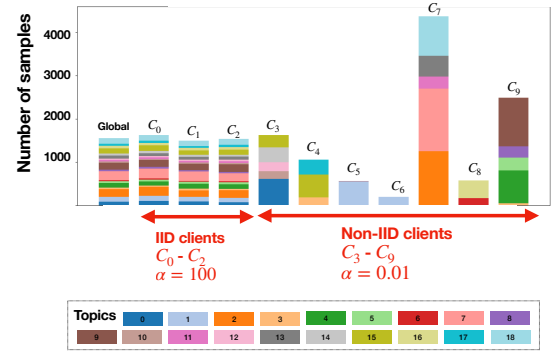


Figure 1: Client data distributions under Mixed IID/Non-IID partitioning. Topically diverse IID clients and topically concentrated Non-IID clients co-exist within the same federation

1 Introduction

Mental disorders affect approximately 1 in 7 people globally [27], yet the majority of those affected do not seek professional help. Counseling — structured dialogue-based psychotherapy — is a primary form of treatment, but remains inaccessible due to financial cost, social stigma, and geographic barriers [13]. AI-powered counseling chat bots offer a route to democratizing access, and recent work has shown that LLMs fine-tuned on real counseling transcripts can produce empathic, contextually appropriate responses [28]. However, publicly available counseling transcripts represent only a fraction of the data that exists. The richest source of training data — actual clinical records — is protected by strict privacy regulations and remains inaccessible for centralized training.

Federated Learning (FL) [14] bypasses this privacy barrier by exchanging model parameters while keeping raw data local, enabling secure training on restricted clinical datasets. Yet a realistic multi-institution federation introduces a data distribution paradigm the literature has largely overlooked: clients differ substantially in their *degree* of heterogeneity. General hospitals hold large, topically diverse cohorts approximating the global distribution (Identically and Independently Distributed or IID) [24]. In contrast, specialized clinics hold compact, topic-concentrated data (Non-IID) [31]. We term federations where both coexist **Mixed IID/Non-IID** settings (see Figure 1 for an example when random seed = 42).

The majority of existing FL methods do not explicitly account for this Mixed IID/Non-IID setting. FedAvg implicitly assumes IID for convergence, and its variants that mitigate client drift due to Non-IID clients [9, 10, 29] apply globally uniform corrections — missing the opportunity to treat IID and Non-IID clients differently. Specifically, IID clients benefit from broad global aggregation, while Non-IID clients benefit from deep local specialization. We address this gap with the following contributions:

- We conduct the first study of Mixed IID/Non-IID federated fine-tuning for mental health counseling LLMs, where IID general hospitals and Non-IID specialized clinics coexist.
- We propose **Heterogeneity-Aware (HA) DualLoRA**, a dual-adaptor architecture with a per-client entropy heterogeneity score that adapts the local and global adapter training epochs in *opposite directions*, motivated by federated convergence theory for each direction independently.
- We propose **Selective Federated Aggregation (SFA)**, which routes IID clients into global aggregation and Non-IID clients into private local training in order to achieve balanced global adapter and specialized local adapter simultaneously.
- We conduct experiments on MentalChat16K, demonstrating both methods consistently outperform FedAvg and match the strongest baseline, Dual LoRA, with reduced communication cost and justified compute overhead.

2 Related Work

2.1 Centralized Counseling LLMs

MentalChat16K [28] introduces a benchmark dataset combining synthetic counseling conversations and anonymized transcripts from interventions between behavioral health coaches and caregivers. The dataset consists of (problem, response) pairs covering a broad range of mental health concerns including depression, anxiety, and relationship issues. Additionally, it proposes a human validated seven-metric evaluation framework with GPT-4 and Gemini Pro as automated judges. The metrics include Active Listening, Empathy & Validation, Safety & Trustworthiness, Open-mindedness & Non-judgment, Clarity & Encouragement, Boundaries & Ethical, and Holistic Approach. This evaluation framework is the foundation we build on and extend. Psych8K [11] (ChatCounselor) constructs a dataset from 260 real counseling recordings provided by licensed psychologists. LLMs fine-tuned on these datasets demonstrate strong counseling ability. However, they adopt centralized training, which is impractical in real clinical settings due to privacy regulations.

2.2 FL under Non-IID data

Zhao et al. [29] demonstrated that FedAvg performance degrades under Non-IID data due to weight divergence across clients. FedNova [25] formalized a related failure mode as *objective inconsistency*, where heterogeneous local steps cause the global model to converge to an unintended objective. SCAFFOLD [9] and FedProx [10] mitigate client drift via control variates and proximal regularization, respectively. However, all these methods rely on network-wide, uniform correction frameworks. Consequently, they lack the mechanism to differentiate between coexisting IID and

Non-IID client profiles within the same federation. HiCS-FL [3] considers varying degree of heterogeneity across clients and proposes a method that preferentially samples clients with more balanced distributions. This approach misses the opportunity to exploit Non-IID clients' specialized knowledge while still benefiting from IID clients' broad coverage.

2.3 Personalized federated LoRA

FL has recently been extended to LLMs, where LoRA [7] enables efficient federated fine-tuning. Building on this, a line of work applies *personalized* FL specifically to LoRA-based fine-tuning. FD-LoRA [16] assigns each client a private personalized LoRA and a global LoRA, fusing them post-training via gradient-free scalar weight optimization. FedALT [1] keeps each client's individual LoRA un-aggregated and supplements it with a frozen Rest-of-World aggregate LoRA, mixed via a learned MoE (Mixture of Experts) gate at inference. FFA-LoRA [22] freezes A matrix and trains only B matrix, eliminating the aggregation instability caused by separately averaging jointly optimized matrices. However, all target NLP or instruction-tuning benchmarks under uniform task-level heterogeneity. None addresses varying *degrees of statistical heterogeneity* within a single federation, a realistic condition critical for deployment.

2.4 Federated LLMs for mental health

FedTherapist [21], FedMentalCare [18], and FedMentor [19] apply FL to mental health AI and target status detection, classification, and privacy compliance. However, none generates counseling-style responses to patient context — a form of direct intervention that supports users more immediately than detection or classification alone. They also do not address the Mixed IID/Non-IID heterogeneity inherent to multi-institution mental health federations. Both are critical gaps for real-world deployment.

3 Problem Formulation

We consider a federation of N clients, each holding a private local dataset $\mathcal{D}_i$ of counseling transcripts. Each entry consists of a (problem, response) pair, where the problem is a patient's message describing their mental health concern and the response is a counselor's reply. Clients differ in their degree of data heterogeneity, modeled via the Dirichlet distribution's concentration parameter α . Of the N clients, C_{IID} are IID clients holding topically diverse data approximating the global mixture ($\alpha = 100$), and $C_{\text{Non-IID}}$ are Non-IID clients holding data concentrated on a narrow topic subset ($\alpha = 0.01$), where $C_{\text{IID}} + C_{\text{Non-IID}} = N$.

The goal is to find a federated fine-tuning strategy that simultaneously (1) maximizes general Counseling Quality across all clients, (2) maximizes Personalization metric by adapting each client's model to its local topic distribution and user population, and (3) minimizes per-round communication upload and per-client compute overhead. Evaluation metrics are defined in Section 5.2 and limitations are discussed in Section 8.

4 Methodology

Figure 2 provides an overview of both proposed methods.

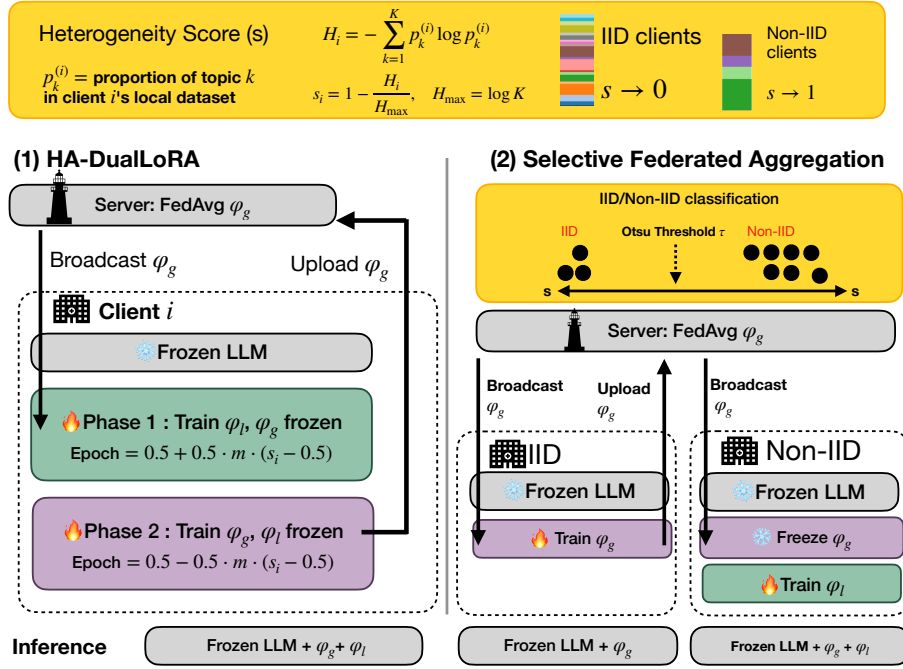


Figure 2: Overview of proposed methods. Left: HA-DualLoRA assigns each client a global adapter ϕ_g (aggregated via FedAvg) and a private local adapter ϕ_l (never uploaded), with epoch budgets adapted in opposite directions based on heterogeneity score s_i . Right: Selective Federated Aggregation classifies clients via Otsu thresholding; IID clients ($s_i < \tau$) upload ϕ_g for aggregation, while Non-IID clients ($s_i \geq \tau$) receive the broadcast ϕ_g and train only a private ϕ_l locally.

4.1 Heterogeneity Score Estimation

To differentiate IID and Non-IID clients, we estimate a per-client heterogeneity score s_i from topic proportions alone, without any raw data exchange. Clinical institutions often categorize records by condition or topic. Each client counts the number of data samples in each topic they own and shares the normalized proportions $p_k^{(i)}$ with the server (the proportion of topic k in client i 's local dataset).

Given these proportions, we compute the Shannon entropy [20] of each client's topic distribution:

$$H_i = - \sum_{k=1}^K p_k^{(i)} \log p_k^{(i)} \quad (1)$$

and normalize it to a heterogeneity score:

$$s_i = 1 - \frac{H_i}{H_{\max}}, \quad H_{\max} = \log K \quad (2)$$

where K = total number of topic clusters. $s_i \rightarrow 1$ and $s_i \rightarrow 0$ respectively indicate that client _{i} is highly concentrated (Non-IID) and uniform (IID). Entropy is preferred over gradient-based heterogeneity estimators (e.g., LOO cosine distance [9]), since it requires no warm-up training rounds, introduces zero additional communication overhead, and can be computed by each institution before the federation begins. This heterogeneity score is used for both proposed methods.

To classify clients into IID and Non-IID groups for Selective Federated Aggregation, we apply Otsu's method [15] to the set $\{s_i\}_{i=1}^N$.

Otsu's method finds a threshold τ that maximizes between-class variance, achieving parameter-free binary classification for any federation composition without manual tuning. In our experiments it achieves perfect IID/Non-IID separation (accuracy 1.0) against ground-truth Dirichlet labels.

4.2 Proposed Method 1: Heterogeneity-Aware (HA) DualLoRA

Each client holds two LoRA adapters: a *global adapter* ϕ_g aggregated via FedAvg each round, and a *local adapter* ϕ_l that accumulates private knowledge on-device and is never uploaded. Per round, training alternates between two phases: Phase 1 trains ϕ_l with ϕ_g frozen; Phase 2 trains ϕ_g with ϕ_l frozen. We train ϕ_l first (Phase 1) so that the local adapter adapts to the client's data before the global adapter is updated, consistent with the ordering used in FDLORA [16]. At inference, both adapters are combined by simple addition $\Delta W = B_g A_g + B_l A_l$, requiring no additional parameters or learned gating [8].

The key novelty is that each phase receives a *client-specific* epoch budget determined by s_i , with the two phases adapted in *opposite directions*. Let $m \in (0, 2]$ be a sensitivity multiplier controlling the degree of adaptation.

The per-client epoch allocation is:

$$\phi_l\text{-epochs}_i = 0.5 + 0.5 \cdot m \cdot (s_i - 0.5) \quad (3)$$

$$\phi_g\text{-epochs}_i = 0.5 - 0.5 \cdot m \cdot (s_i - 0.5) \quad (4)$$

so that ϕ_l -epochs_{*i*} + ϕ_g -epochs_{*i*} = 1.0 for all clients regardless of s_i and m . Both phases start from a baseline of 0.5 epochs, with heterogeneity score shifting budget between them. When $m = 0$, both phases receive equal budget (equivalent to Dual LoRA with fixed hyperparameters). When $m = 1$ (canonical), a client with $s_i = 1$ devotes 0.75 epochs to ϕ_l and 0.25 to ϕ_g . Fractional epochs are implemented as gradient steps $\lfloor \text{epoch_fraction} \times |D_i| / b \rfloor$, where $|D_i|$ is client i 's local dataset size and $b = 4$ the effective batch size. We set $m = 1$ as the canonical configuration, validated by the sensitivity ablation in Section 7.

The directions are theoretically motivated. For Phase 1 (ϕ_l): high-heterogeneity (Non-IID) clients possess concentrated, specialized local knowledge that ϕ_l should deeply internalize, thus more local steps are given. This is grounded in APFL [4], which shows that clients with greater local-global distribution divergence benefit from more local adaptation. For Phase 2 (ϕ_g): high-heterogeneity clients produce skewed gradients that cause client drift during global aggregation, thus fewer steps are given to limit this drift. This direction is supported by FedNova [25] and SCAFFOLD [9], which show that excess local steps under heterogeneity amplify objective inconsistency in the global model.

4.3 Proposed Method 2: Selective Federated Aggregation (SFA)

For deployment scenarios where communication cost is paramount, we propose Selective Federated Aggregation. Clients are classified via Otsu thresholding of s_i into IID ($s_i < \tau$) and Non-IID ($s_i \geq \tau$) groups.

IID clients train and upload ϕ_g each round for FedAvg aggregation among IID clients only. At inference, they use the frozen LLM + ϕ_g . Non-IID clients receive the broadcast ϕ_g after IID clients federation, freeze it, and train a private ϕ_l locally – never uploading. At inference, they use the frozen LLM + ϕ_g + ϕ_l .

The design follows a core insight of the Mixed IID/Non-IID setting: IID clients produce balanced gradients well-suited for global model building, while Non-IID clients' skewed gradients introduce drift without proportionate benefit. Rather than correcting for this drift (as FedProx [10] and SCAFFOLD [9] do), we exclude Non-IID clients from aggregation entirely, assigning them a distinct role: local specialization on top of a clean global foundation broadcast each round. This differs from HiCS-FL [3], which selectively samples within a unified pool rather than assigning structurally distinct roles.

5 Evaluation

5.1 Experimental Setup

Model and data. We fine-tune Qwen2.5-0.5B-Instruct [17] (494M parameters, 4-bit quantized) using LoRA on all linear layers ($r = 16$, $\alpha = 8$, $\text{dropout} = 0.075$). A lightweight, quantized base model was chosen to simulate resource-constrained deployment environments. Training uses paged AdamW [12] (optimizer states quantized to 8-bit), $\text{lr} = 2 \times 10^{-5}$, effective batch size 4. All experiments run for 10 rounds with seeds 42 and 99; results are averaged over both seeds. All experiments were conducted on NVIDIA A100 GPUs via the mdx cloud platform [23].

FL Setup. MentalChat16K dataset [28] (14,452 training examples (problem, response); 90/10 train/test split) does not provide topic-level labels, which are necessary to simulate topic heterogeneity. We therefore embed input texts by meta-llama/Llama-3.1-8B-Instruct [6] and cluster them into topics via HDBSCAN [2]. This yielded $K = 19$ topic clusters. Data is then partitioned across 10 clients (3 IID, $\alpha = 100$; 7 Non-IID, $\alpha = 0.01$) via Dirichlet sampling over these 19 clusters (per-client range 184–4,062). See Figure 1 for example with random seed = 42. Use of Dirichlet distribution follows standard FL heterogeneity benchmarks [9, 10] that use a small α for extreme Non-IID clients and a large α for IID clients. While these configurations do not precisely reflect the real clinical distributions, they provide a controlled simulation of the IID/Non-IID coexistence that characterizes multi-institutional mental healthcare federations.

5.2 Evaluation Criteria

Counseling Quality: We adopt the human validated 7-metric LLM-judge framework from MentalChat16K, scoring each response on Active Listening, Empathy & Validation, Safety & Trustworthiness, Open-mindedness & Non-judgment, Clarity & Encouragement, Boundaries & Ethical, and Holistic Approach (each 1–10, judge: GPT-4o-mini). We report the mean of these seven metrics as the *Counseling Quality* score.

Personalization: We additionally introduce an 8th standalone metric, *Personalization*. This metric assesses whether a response is tailored to the user's situation, adapts its tone and therapeutic framing, and avoids generic boilerplate advice. Personalization metric was not considered by the original seven-metric evaluation, yet is crucial for effective counseling. The metric is designed to be drop-in compatible with the MentalChat16K judge prompt, following the same 1–10 scale and GPT-4o-mini judge setup. The metric definition draws on dimensions identified across personalization evaluation frameworks including PersonaLens [30] and AuPEL [26]. The full judge prompt is available in our repository.

Communication and Compute Cost: Communication cost is a principal constraint in FL [14], and compute overhead is critical in resource-constrained clinical deployments. We therefore measure upload cost per client per round (MB) and wall-clock training time per client per round (seconds), representing communication and compute cost respectively.

5.3 Comparisons

We compare the performance of our methodology against: **Base Qwen** (no fine-tuning, zero-shot floor); **FedAvg** (single global LoRA, all clients, standard aggregation); **Local-Only** (independent per-client LoRA, no federation); **FFA-LoRA** [22] (matrix A frozen, only B aggregated, single adapter); and **Dual LoRA** (same dual-adapter architecture as HA-DualLoRA but with fixed equal epoch budgets, no heterogeneity-aware adaptation). All methods use controlled hyperparameters to ensure fair comparisons.

6 Results

Table 1 reports Counseling Quality (C), Personalization (P), and communication/compute cost for all methods.

Table 1: Main results (mean $\pm$ std over seeds 42 and 99). C = Counseling Quality (mean of 7 metrics, 1–10); P = Personalization (metric 8, standalone, 1–10); MB = upload per client per round; s/rd = compute time per client per round. Highlighted = best per column. Methods in bold = proposed methods.

Method	C-All	C-IID	C-Non-IID	P-All	P-IID	P-Non-IID	MB	s/rd
Base Qwen [17]	6.48 $\pm$ 0.047	6.45 $\pm$ 0.056	6.49 $\pm$ 0.043	5.20 $\pm$ 0.047	5.17 $\pm$ 0.104	5.21 $\pm$ 0.022	0.0	0.0
FedAvg [14]	7.07 $\pm$ 0.033	7.10 $\pm$ 0.053	7.06 $\pm$ 0.024	5.83 $\pm$ 0.040	5.86 $\pm$ 0.024	5.82 $\pm$ 0.047	35.2	308 $\pm$ 4.1
Local-Only	7.07 $\pm$ 0.073	7.09 $\pm$ 0.007	7.06 $\pm$ 0.107	5.90 $\pm$ 0.076	5.85 $\pm$ 0.009	5.91 $\pm$ 0.112	0.0	309 $\pm$ 0.0
FFA-LoRA [22]	6.98 $\pm$ 0.009	7.01 $\pm$ 0.008	6.97 $\pm$ 0.017	5.73 $\pm$ 0.040	5.81 $\pm$ 0.000	5.70 $\pm$ 0.058	19.5	290 $\pm$ 1.3
Dual LoRA	7.22 $\pm$ 0.125	7.30 $\pm$ 0.024	7.18 $\pm$ 0.169	6.08 $\pm$ 0.132	6.11 $\pm$ 0.108	6.06 $\pm$ 0.142	35.2	400 $\pm$ 4.8
HA-DualLoRA	7.21 $\pm$ 0.041	7.17 $\pm$ 0.069	7.22 $\pm$ 0.089	6.07 $\pm$ 0.095	5.97 $\pm$ 0.061	6.12 $\pm$ 0.162	35.2	394 $\pm$ 1.1
SFA	7.21 $\pm$ 0.004	7.04 $\pm$ 0.067	7.29 $\pm$ 0.034	6.09 $\pm$ 0.008	5.82 $\pm$ 0.005	6.21 $\pm$ 0.010	10.6	373 $\pm$ 5.0

Table 2: IID/Non-IID ratio ablation (seed 42). Ratios are IID/Non-IID client counts (e.g., 7/3 = 7 IID, 3 Non-IID). HA = HA-DualLoRA; Underlined = best per column.

Method	Counseling-All			Personalization-Non-IID		
	7/3	3/7	1/9	7/3	3/7	1/9
FedAvg	7.10	7.09	7.14	5.91	5.82	5.93
HA	7.12	<u>7.27</u>	7.25	5.94	6.12	6.18
SFA	<u>7.14</u>	7.22	<u>7.34</u>	<u>6.17</u>	<u>6.20</u>	<u>6.25</u>

All fine-tuned methods outperform Base Qwen (C-All: 6.98–7.22 vs. 6.48; P-All: 5.73–6.09 vs. 5.20), indicating that federated fine-tuning on counseling data provides meaningful gains regardless of method choice.

Both proposed methods outperform FedAvg on Counseling Quality and Personalization, with gains concentrated on Non-IID clients (C-Non-IID 7.22 and 7.29 vs. 7.06; P-Non-IID 6.12 and 6.21 vs. 5.82). The modest drop in SFA’s IID metrics is expected: IID clients rely on the global adapter alone, while Non-IID clients receive full dual-adapter personalization. Both methods match or reduce FedAvg’s upload cost (35.2 MB for HA-DualLoRA and 10.6 MB for SFA vs. 35.2 MB for FedAvg) but incur higher compute (394–373 seconds/round vs. 308 seconds), a tradeoff justified by the personalization gains on the clients that need it most.

SFA achieves the highest Non-IID Counseling Quality (7.29) and Personalization (6.21) of all methods, matching the strongest baseline Dual LoRA overall (C-All: 7.21 vs. 7.22; P-All: 6.09 vs. 6.08), while reducing communication cost by 70%. Excluding Non-IID clients from aggregation removes gradient drift while freeing them to specialize locally, turning a federation liability into a personalization and efficiency asset.

7 Ablations

Table 2 varies the IID/Non-IID client ratio to evaluate consistency across federation compositions. Both proposed methods outperform FedAvg on Counseling Quality and Non-IID Personalization at every ratio tested. SFA leads both metrics at most ratios, only trailing HA-DualLoRA at 3/7 Counseling Quality. This demonstrates that gains over FedAvg are not specific to the 3/7 split.

Table 3: Psych8K [11] generalization (seed 42). C = Counseling Quality; P = Personalization. Underlined = best per column. Client 7 excluded (0 training examples under Dirichlet α = 0.01 on Psych8K’s smaller dataset; production systems should enforce a minimum sample threshold per client).

Method	C-All	C-Non-IID	P-All	P-Non-IID
FedAvg	6.01	6.00	5.27	5.29
HA-DualLoRA	<u>6.02</u>	<u>6.09</u>	<u>5.39</u>	<u>5.47</u>
SFA	5.98	6.07	5.31	5.39

Table 4: HA-DualLoRA sensitivity to epoch multiplier m (seed 42; $m=1.0$ canonical). C = Counseling Quality; P = Personalization. Underlined = best per column.

Multiplier m	C-All	C-Non-IID	P-All	P-Non-IID
Low ($m = 0.5$)	7.25	7.27	6.13	6.16
Canonical ($m = 1.0$)	<u>7.27</u>	<u>7.32</u>	<u>6.18</u>	<u>6.26</u>
High ($m = 1.5$)	7.23	7.28	6.13	6.21
High ($m = 2.0$)	<u>7.27</u>	7.30	6.17	6.22

Table 3 shows that trends from MentalChat16K hold on Psych8K with narrower margins. HA-DualLoRA leads across all metrics; SFA dips slightly below FedAvg on overall Counseling Quality but gains on Non-IID metrics at lower upload cost (11.7 MB), likely due to limited per-client data under the same Dirichlet split.

Table 4 reports sensitivity to the epoch multiplier m (single-seed; deviations from Table 1 reflect seed variance). Increasing m beyond 1.0 does not consistently improve Non-IID Personalization and slightly hurts IID clients, suggesting $m = 1.0$ as a stable default.

8 Discussion

Compute and communication overhead. Proposed methods incur a 21–28% compute overhead over FedAvg, justified by consistent gains in Counseling Quality and Non-IID Personalization. Communication cost is more consequential since it scales with federation size. Thus SFA’s 70% upload reduction directly supports sustainable adoption in resource-constrained healthcare settings.

Limitations and Future Work. This study is a proof of concept under controlled conditions: 10 clients, 10 rounds, a 0.5B model, two seeds, a single LoRA rank, and Dirichlet partitioning. Larger-scale studies with more diverse heterogeneity structures are needed to validate real-world performance. We assume server-side knowledge of topic distributions, which future work should relax by exploring fully decentralized heterogeneity estimation. We rely on LLM-judge scoring, including an 8th Personalization metric not validated against clinical judgment [5]. Structured clinical evaluation protocols and human-in-the-loop safeguards are therefore essential before any deployment. This is particularly important to address hallucination and inappropriate response risks. Finally, this work is a simulation, and meaningful societal impact requires actual deployment in clinical settings. However, technical feasibility alone does not ensure adoption. It demands careful system design beyond the FL algorithm itself: how audio or text transcripts are recorded, stored, and preprocessed while respecting patient consent and regional privacy regulations. Scalable deployment therefore requires transparent governance and incentives mutually beneficial to clinicians, institutions, and users.

9 Conclusion

We presented the first systematic study of federated fine-tuning of mental health counseling LLM in Mixed IID/Non-IID federations, a setting common in healthcare but neglected in the existing FL methods. Our key finding is that treating IID and Non-IID clients differently within the same federation yields consistent gains over uniform approaches. Selective Federated Aggregation (SFA) emerges as the strongest overall method, achieving the highest Non-IID Counseling Quality and Personalization while reducing upload cost by 70%. This makes it practically compelling for resource-constrained clinical settings. HA-DualLoRA's heterogeneity-aware epoch adaptation further improves Non-IID Personalization over standard Dual LoRA, with the dual-adaptor architecture itself as the primary performance driver. This study is a step toward the deployment of federated counseling LLM systems that can extend quality mental health support to those unable to access professional care.

References

- [1] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. 2026. Fedalt: Federated fine-tuning through adaptive local training with rest-of-world lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 19728–19736.
- [2] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.
- [3] Huancheng Chen and Haris Vikalo. 2024. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Advances in Neural Information Processing Systems* 37 (2024), 65525–65561.
- [4] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. 2020. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461* (2020).
- [5] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10126–10141.
- [6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, She Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- [8] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089* (2022).
- [9] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*. PMLR, 5132–5143.
- [10] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2 (2020), 429–450.
- [11] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461* (2023).
- [12] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.
- [13] Ma Krizza V Luz. 2023. The Challenge of Accessibility and Inaccessibility of Guidance and Counseling Services in Schools based on Recent Studies: Improving Accessibility. *European Journal of Education and Pedagogy* 4, 2 (2023), 168–171.
- [14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. Pmlr, 1273–1282.
- [15] Nobuyuki Otsu et al. 1979. A threshold selection method from gray-level histograms. *Automatica* 11, 285–296 (1979).
- [16] Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. 2024. Fdlora: Personalized federated learning of large language model via dual lora tuning. *arXiv preprint arXiv:2406.07925* (2024).
- [17] A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint* (2024).
- [18] Nobin Sarwar. 2025. Fedmentalcare: towards privacy-preserving fine-tuned llms to analyze mental health status using federated learning framework. *arXiv preprint arXiv:2503.05786* (2025).
- [19] Nobin Sarwar and Shubhashis Roy Dipta. 2025. FedMentor: Domain-Aware Differential Privacy for Heterogeneous Federated LLMs in Mental Health. *arXiv preprint arXiv:2509.14275* (2025).
- [20] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.
- [21] Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho D Choi, and Sung-Ju Lee. 2023. Fedtherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 11971–11988.
- [22] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. 2024. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313* (2024).
- [23] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, et al. 2022. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, 1–7.
- [24] Maike van Niekerk, Jane Walker, Harriet Hobbs, Nicholas Magill, Mark Toynbee, Ben Steward, Eli Harriss, and Michael Sharpe. 2022. The prevalence of psychiatric disorders in general hospital inpatients: a systematic umbrella review. *Journal of the Academy of Consultation-Liaison Psychiatry* 63, 6 (2022), 567–578.
- [25] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems* 33 (2020), 7611–7623.
- [26] Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593* (2023).
- [27] World Health Organization. 2025. Mental disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>
- [28] Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5367–5378.
- [29] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandr. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582* (2018).
- [30] Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*. 18023–18055.
- [31] Hongyu Zheng, Xiaolu Jiang, Rong Yang, Shuo Wang, and Hui Zhong. 2023. Changes in major psychiatric disorders in children and adolescents from 2001 to 2020: A retrospective single-center study. *Frontiers in Psychiatry* 13 (2023), 1079456.