FedMentor: Domain-Aware Differential Privacy for Heterogeneous Federated LLMs in Mental Health

Anonymous ACL submission

Abstract

Privacy-preserving adaptation of Large Language Models (LLMs) in sensitive domains (e.g., mental health) requires balancing strict confidentiality with model utility and safety. We propose **FedMentor**, a federated fine-tuning framework that integrates Low-Rank Adaptation (LoRA) and domain-aware Differential Privacy (DP) to meet per-domain privacy budgets while maintaining performance. Each client (domain) applies a custom DP noise scale proportional to its data sensitivity, and the server adaptively reduces noise when utility falls below a threshold. In experiments on three mental health datasets, we show that FedMentor improves safety over standard Federated Learning (FL) without privacy, raising safe output rates by up to three points and lowering toxicity, while maintaining utility (BERTScore F1 and ROUGE-L) within 0.5% of the non-private baseline and close to the centralized upper bound. The framework scales to backbones with up to 1.7B parameters on single-GPU clients, requiring < 173 MB of communication per-round. FedMentor demonstrates a practical approach to privately fine-tune LLMs for safer deployments in healthcare and other sensitive fields.

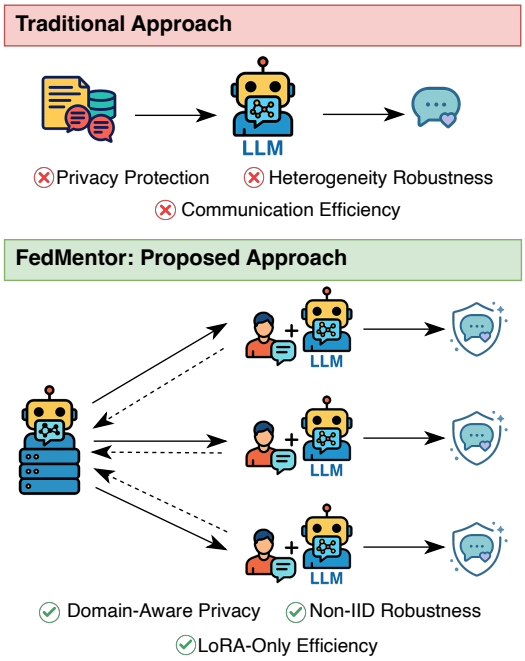


Figure 1: Comparison between traditional LLM-based methods for mental health and **FedMentor**. The baseline lacks privacy protection, robustness to heterogeneity, and communication efficiency, whereas FedMentor introduces domain-aware privacy, achieves robustness under non-IID data, and improves efficiency through LoRA-only updates.

1 Introduction

Mental health arises from interacting cognitive, affective, and behavioral processes that shape individual functioning and societal stability. Demand for scalable support has accelerated interest in LLMs for conversational assistance (Yang et al., 2023a; Xu et al., 2024; Yang et al., 2024). Deployment remains challenging due to strict privacy requirements, limited interpretability, and legal constraints under HIPAA and GDPR (May and Denecke, 2022; Nicholas et al., 2020; United States Congress, 1996; European Parliament and Council of the European Union, 2016). User inputs may include explicit self harm ideation or other clinical signals that require strong confidentiality (e.g., *[self-harm ideation example]*). These constraints make traditional centralized training challenging for sensitive user data, highlighting the need for techniques that adapt models in a privacy-preserving and communication-efficient manner for healthcare settings. As shown in Figure 1, our approach addresses these needs by combining domain-aware privacy, robustness under non-IID client distributions, and parameter-efficient updates.

There is growing interest across the AI and healthcare communities in developing trustworthy and scalable mental health chatbots that can

provide rapid, accessible, and confidential psychological support. In 2019, mental disorders affected about 970 million people worldwide, or one in eight individuals (Institute for Health Metrics and Evaluation (IHME), 2023). Conversational agents powered by large language models have since emerged as promising tools for supporting mental well-being at scale (Towards Healthcare, 2023). The global market for mental health chatbots, valued at $0.99 billion in 2022, is projected to grow to $6.51 billion by 2032, underscoring strong clinical and commercial demand (Towards Healthcare, 2023). At the same time, deploying LLMs in this setting raises critical challenges in privacy protection, legal compliance, and efficient communication and computation for personalized care (Lai et al., 2023; United States Congress, 1996; European Parliament and Council of the European Union, 2016). These challenges have motivated research into privacy-preserving learning, parameter-efficient adaptation, and federated methods tailored for high-risk domains such as mental health, where user trust and confidentiality are essential.

To investigate privacy and communication challenges in adapting LLMs to sensitive mental health domains, we present **FedMentor**, a federated framework that assigns domain-aware Differential Privacy budgets and fine-tunes only LoRA adapters, aggregated with FedAvg (McMahan et al., 2017), to achieve private adaptation while preserving utility and fairness under heterogeneity. The overall architecture is shown in Figure 2. FedMentor is designed for scenarios that demand distinct privacy guarantees across domains and where non-IID data can introduce drift and unequal outcomes. We evaluate the framework on three datasets (Dreaddit (Turcan and McKeown, 2019), IRF (Garg et al., 2023), MultiWD (Garg et al., 2024)) and five backbones (MobileLLM-ParetoQ-350M (Liu et al., 2024b), SmolLM2 (360M and 1.7B) (Allal et al., 2025), and Qwen3 (0.6B and 1.7B) (Yang et al., 2025)). FedMentor improves safety with minimal loss in utility: for instance, on MultiWD the Toxicity Safe Rate (TSR) rises by two points while toxicity decreases from 2.92 to 1.98 (Table 1), and BERTScore F1 (Zhang et al., 2020) and ROUGE-L (Lin, 2004) remain close to standard FL and near the centralized upper bound. Client-level relevance remains consistent across domains, and efficiency is achieved by communicating only adapter weights through LoRA, keeping computation and bandwidth feasible for single-GPU clients. To the best of our knowledge, this is the first study that combines domain-aware DP (Dwork et al., 2014; Abadi et al., 2016) with federated LoRA fine-tuning for large language models in mental health.

We summarize the main contributions of this work as follows:

- **Domain-aware private adaptation for mental health.** We formulate a federated LoRA approach with per-domain (ϵ, δ) that protects sensitive mental health text while preserving task utility. FedMentor raises safety (e.g., +2 percentage points TSR with lower toxicity) and keeps BERTScore F1 and ROUGE-L close to FL and near centralized training.
- **Robustness to non-IID heterogeneity.** We demonstrate stable utility and fairness across Dreaddit, IRF, and MultiWD, with per-client relevance closely aligned and no domain collapse. An adaptive noise mechanism maintains performance under strict budgets for sensitive domains.
- **Practical efficiency via adapters.** By exchanging only LoRA adapters, FedMentor reduces communication and memory enough to train backbones up to 1.7B parameters on single-GPU clients, enabling deployment in resource-constrained healthcare settings.

2 FedMentor Foundations: Problem and Notation

We consider a Federated Learning setting with K distributed clients, where each client $k \in \{1, \ldots, K\}$ holds a private mental health dataset $\mathcal{D}_k$ from domain $d_k \in \{\text{IRF}, \text{Dreaddit}, \text{MultiWD}\}$. The objective is to collaboratively fine-tune a global language model $\mathcal{M}_{\text{global}}$ while ensuring privacy, leveraging domain heterogeneity, and maintaining computational efficiency.

The optimization problem seeks optimal LoRA adapters that minimize the weighted empirical risk:

$$\Delta W^*_{\text{LoRA}} = \arg\min_{\Delta W_{\text{LoRA}}} \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \mathcal{L}_k(\Theta_{\text{frozen}} + \Delta W_{\text{LoRA}}; \mathcal{D}_k)\,, \tag{1}$$

where $|\mathcal{D}| = \sum_{j=1}^{K} |\mathcal{D}_j|$ and Θ_{frozen} denotes the frozen backbone parameters.

This optimization operates under three critical constraints:

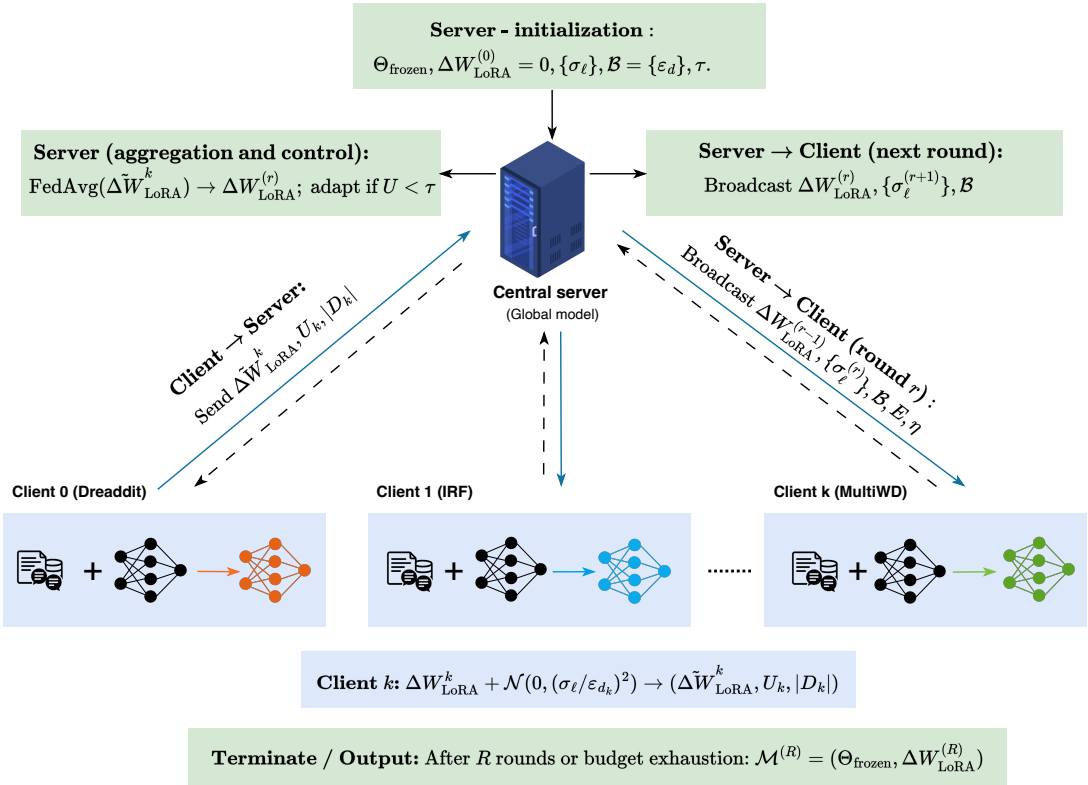


Figure 2: **FedMentor** pipeline. The server freezes the backbone and initializes LoRA adapters, layer scales, domain privacy budgets, and a utility threshold. Each round it broadcasts the current adapters; clients train LoRA on local data, add Gaussian noise per budget ε_d, and return noised adapters with a utility signal. The server aggregates with FedAvg and reduces noise when utility $< \tau$. After R rounds the model is the frozen backbone plus learned LoRA adapters.

Privacy Constraint. Each client update must satisfy Differential Privacy (DP) (Dwork et al., 2014):

$$\Delta\tilde{W}^k_{\text{LoRA}} = \Delta W^k_{\text{LoRA}} + \mathcal{N}\left(0, \left(\frac{\sigma_l}{\varepsilon_{d_k}}\right)^2 I\right), \quad \text{satisfying } (\varepsilon_{d_k}, \delta)\text{-DP}, \tag{2}$$

where $\varepsilon_{d_k} \in \{0.5, 1.5, 2.0\}$ with IRF $= 0.5$ (high sensitivity), Dreaddit $= 2.0$ (medium), and MultiWD $= 1.5$ (low to medium); smaller ε indicates stronger privacy for more sensitive domains (e.g., interpersonal risk factors).

Utility and Heterogeneity Constraints. The global model must maintain clinical viability:

$$\mathcal{U}_m(\mathcal{M}_{\text{global}}) \geq \tau_m, \quad \forall m \in \{\text{BERTScore}, \text{SafeRate}, \text{Relevance}, \text{Perplexity}\}, \tag{3}$$

where τ_m represents minimum acceptable thresholds for safe deployment. At the same time, heterogeneity arises from domain shifts:

$$P_{d_i}(x, y) \neq P_{d_j}(x, y), \quad \forall i \neq j, \quad d_i, d_j \in \{\text{IRF}, \text{Dreaddit}, \text{MultiWD}\}. \tag{4}$$

Instead of treating heterogeneity as a limitation, our formulation leverages complementary distributions to improve robustness and cross-domain generalization.

Efficiency Constraint. Low-Rank Adaptation enables scalable deployment by reducing computation and communication costs:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d\times r}, \quad A \in \mathbb{R}^{r\times k}, \quad r \ll \min\{d, k\}. \tag{5}$$

This decomposition reduces trainable parameters from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d + k))$. For communication, if N denotes the total backbone parameters, the transmission cost is reduced by a factor of $\frac{r(d_{\text{in}}+d_{\text{out}})}{N} \approx 10^{-3}$, which enables practical deployment on resource-constrained clinical devices while maintaining model expressiveness.

3 FedMentor Framework

FedMentor combines Federated Learning (FL) (McMahan et al., 2017), Low-Rank Adaptation (LoRA) (Hu et al., 2021), and domain-aware

Differential Privacy (DP) (Dwork et al., 2014; Acharya et al., 2020; Noble et al., 2022; Charles et al., 2024) to enable scalable, privacy-preserving fine-tuning of LLMs for mental health applications. Unlike traditional FL approaches that transmit full model weights, FedMentor leverages lightweight LoRA updates with client-side noise injection. Algorithm 1 outlines the training process.

3.1 Client Training

Each client k maintains dataset $\mathcal{D}_k = \mathcal{D}_k^{\text{train}} \cup \mathcal{D}_k^{\text{val}}$ where $\mathcal{D}_k^{\text{train}} \cap \mathcal{D}_k^{\text{val}} = \emptyset$. During local training, client k receives global LoRA weights $\Delta W_{\text{LoRA}}^{(r-1)}$ and attaches them to the frozen backbone. For each epoch $e \in \{1, \ldots, E\}$ and minibatch $b \subset \mathcal{D}_k^{\text{train}}$, parameters update via:

$$\Delta W_{\text{LoRA}}^k \leftarrow \Delta W_{\text{LoRA}}^k - \eta \cdot \nabla_{\Delta W} \mathcal{L}\Big(\Theta_{\text{frozen}} + \Delta W_{\text{LoRA}}^k;\, b\Big) \tag{6}$$

where η is the learning rate and $\mathcal{L}$ is the cross-entropy loss. Notably, only the LoRA weights (typically $< 1\%$ of model parameters) are trained, while the backbone remains frozen.

Client-Side Privacy. To guarantee user-level privacy, clients apply noise before transmission. After local training, for every parameter $p \in \Delta W_k^{(r)}$:

$$\tilde{W}_{k,p}^{(r)} = \Delta W_{k,p}^{(r)} + \mathcal{N}\left(0, \left(\frac{\sigma_l(p)}{\varepsilon_{d_k}}\right)^2 I\right), \tag{7}$$

where $\sigma_l(p)$ is parameter-specific noise scale and ε_{d_k} follows domain sensitivity from Equation 2. This implements a Gaussian mechanism with $(\varepsilon_{d(k)}, \delta)$-Differential Privacy, ensuring all raw weights and local data remain private.

Layer- and Adapter-Aware Noise Calibration. Noise scales are adapted to reflect the relative sensitivity of parameters to perturbation. We classify LoRA weights by their network position, setting $\sigma_l(w) = 0.01 \cdot \alpha(w)$ for early layers, $\sigma_l(w) = 0.008 \cdot \alpha(w)$ for middle layers, and $\sigma_l(w) = 0.005 \cdot \alpha(w)$ for late layers, where $\alpha(w) = 1.2$ for LoRA-A matrices and $\alpha(w) = 0.8$ for LoRA-B matrices. This scaling reflects empirical observations that early layers and LoRA-A adapters are more sensitive to perturbations, while late layers and LoRA-B adapters are more robust.

Domain-Specific Privacy Implementation. Combining layer-aware noise calibration with domain sensitivity (from Equation 2), the complete privatized update becomes:

$$\Delta \tilde{W}_{\text{LoRA}}^k = \Delta W_{\text{LoRA}}^k + \mathcal{N}\Big(0, \big(\tfrac{\sigma_l(w)}{\varepsilon_{d_k}}\big)^2 I\Big) \tag{8}$$

where ε_{d_k} follows the domain-specific budgets defined in the privacy constraint.

Utility-Aware Adjustment. To maintain task quality, FedMentor tracks proxy metrics (e.g., BERTScore-F1, safety, relevance). If aggregate utility at round r falls below threshold τ, noise scales are reduced:

$$\sigma_l(w) \;\leftarrow\; \alpha \cdot \sigma_l(w), \quad \alpha \in (0, 1). \tag{9}$$

This privacy–utility feedback loop allows FedMentor to dynamically trade off noise magnitude against task-specific clinical utility.

3.2 Server Aggregation

The server aggregates noised client updates using dataset-weighted FedAvg (McMahan et al., 2017):

$$\{\Delta \tilde{W}_{\text{LoRA}}^k\}_{k=1}^K. \tag{10}$$

$$\Delta W_{\text{LoRA}}^{(r)} = \sum_{k=1}^{K} \alpha_k \, \Delta \tilde{W}_{\text{LoRA}}^k, \qquad \alpha_k = \frac{|\mathcal{D}_k^{\text{train}}|}{\sum_{j=1}^{K} |\mathcal{D}_j^{\text{train}}|} \tag{11}$$

where ΔW_{LoRA} represents the collection of all LoRA matrices $\{W_{\text{lora_A}}^{(l)}, W_{\text{lora_B}}^{(l)}\}_{l=1}^L$ across L adapted layers, with each matrix aggregated independently.

The global model update combines the frozen quantized backbone with aggregated LoRA weights:

$$\mathcal{M}_{\text{global}}^{(r)} = \Theta_{\text{backbone}} + \Delta W_{\text{LoRA}}^{(r)}. \tag{12}$$

3.3 Communication Efficiency

FedMentor transmits only LoRA adapter weights rather than full model parameters. The communication cost per client is:

$$\text{CommCost} = \mathcal{O}(r \cdot \sum_l (d_{\text{in}}^{(l)} + d_{\text{out}}^{(l)})) \tag{13}$$

where r is the LoRA rank and the sum is over all adapted layers. With typical LoRA configurations ($r \in \{8, 16\}$) applied to models with billions of parameters, this achieves over 99% compression. For instance, adapting a 1.7B parameter model with rank-16 LoRA requires transmitting only ∼2.3 MB versus ∼6.8 GB for the full model (Hu et al., 2021).

Algorithm 1 FEDMENTOR: Domain-aware DP LoRA for heterogeneous Federated LLMs

Input: Datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, domains $\{d_k\}_{k=1}^{K}$, rounds R, epochs E, learning rate η, privacy budgets $\{\varepsilon_d\}$, thresholds $\{\tau_m\}$

Output: Global model $\mathcal{M}^{(R)} = (\Theta_{\text{frozen}}, \Delta W_{\text{LoRA}}^{(R)})$

Server init
$\mathcal{M}^{(0)} := (\Theta_{\text{frozen}}, \Delta W_{\text{LoRA}}^{(0)})$
$\sigma_l :=$ ClassifyLayers$(\Delta W_{\text{LoRA}}^{(0)})$
$\mathcal{B} := \{\text{IRF} : 0.5,\ \text{Dreaddit} : 2.0,\ \text{MultiWD} : 1.5\}$
for round $r = 1$ to R **do**
  Broadcast LoRA weights $\Delta W_{\text{LoRA}}^{(r-1)}$ to all clients ▷ **Server update**
  for client $k \in \{1, \ldots, K\}$ **in parallel do** ▷ **Client update**
    Attach $\Delta W_{\text{LoRA}}^{(r-1)}$ to frozen Θ_{frozen}
    for epoch $e = 1$ to E **do**
      for batch $b \subset \mathcal{D}_k^{\text{train}}$ **do**
        Update:
$$\Delta W_{\text{LoRA}}^{k} \leftarrow \Delta W_{\text{LoRA}}^{k} - \eta \nabla_{\Delta W} \mathcal{L}\big(\Theta_{\text{frozen}} + \Delta W_{\text{LoRA}}^{k};\, b\big)$$
      end for
    end for
    $\varepsilon_k := \mathcal{B}_{d_k}$
    Add noise:
$$\Delta \tilde{W}_{\text{LoRA}}^{k} \leftarrow \Delta W_{\text{LoRA}}^{k} + \mathcal{N}\big(0, (\sigma_l/\varepsilon_k)^2 I\big)$$
    Send noised LoRA weights $\Delta \tilde{W}_{\text{LoRA}}^{k}$ to server
  end for
  Collect $\{\Delta \tilde{W}_{\text{LoRA}}^{k}\}_{k=1}^{K}$ from clients ▷ **Client aggregation**
  $\Delta W_{\text{LoRA}}^{(r)} \leftarrow \sum_{k=1}^{K} \frac{|\mathcal{D}_k^{\text{train}}|}{\sum_j |\mathcal{D}_j^{\text{train}}|} \Delta \tilde{W}_{\text{LoRA}}^{k}$
  If $\mathcal{U}_m < \tau_m$ for any m: $\sigma_l \leftarrow 0.8\,\sigma_l$
  Update budget:
$$\mathcal{B}_d \leftarrow \mathcal{B}_d - 0.1\,\varepsilon_d \text{ for each domain } d$$
end for
return $\mathcal{M}^{(R)}$ with final LoRA weights $\Delta W_{\text{LoRA}}^{(R)}$

4 Experiment Settings

Datasets. We evaluate FedMentor on three mental health datasets: **Dreaddit** (Turcan and McKeown, 2019) for stress detection, **Interpersonal Risk Factors (IRF)** (Garg et al., 2023) for Thwarted Belongingness (TBe) and Perceived Burdensomeness (PBu), and **MultiWD** (Garg et al., 2024) for multi-label wellness prediction. Together, these corpora cover complementary tasks and form a natural non-IID benchmark across domains. Regulatory requirements under HIPAA and GDPR restrict centralized collection, and lengthy institutional approvals further limit access, producing small, fragmented datasets (Act et al., 1996; Regulation, 2018; Nicholas et al., 2020). Prior FL work mitigates performance degradation from heterogeneity with methods such as SCAFFOLD (Karimireddy et al., 2020), FedNova (Wang et al., 2020), and Adaptive FedOpt (Reddi et al., 2021). We adopt a domain-aware FL setting where each dataset defines a client domain and use this setting to evaluate FedMentor under heterogeneous conditions. App. B.1 provides dataset statistics and preprocessing details.

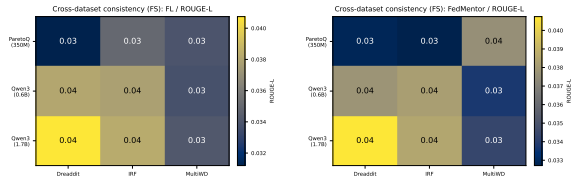
Models. FL places tight limits on client compute, memory, and upload bandwidth; therefore, we adopt compact LLM backbones. We fine-tune five lightweight models from 350M to 1.7B parameters: **MobileLLM-ParetoQ-350M** (Liu et al., 2024b), **SmolLM2** (360M and 1.7B) (Allal et al., 2025), and **Qwen3** (0.6B and 1.7B) (Yang et al., 2025). These backbones integrate quantization and distillation to enhance efficiency, enabling deployment on edge devices and GPUs with limited memory. Their compact size further supports faster client-level fine-tuning and inference, which is critical under the resource and communication constraints of FL (Ro et al., 2022; Lin et al., 2022; Wang et al., 2024b). To ensure fairness, FedMentor and all baselines are evaluated on identical model architectures so that observed differences arise solely from training strategies. Additional architectural details are provided in App. B.2.

Baselines. We compare three setups under identical backbones and LoRA ranks (here, *w/o* denotes *without*). **(i) Centralized (w/o FL, w/o DP):** pool all datasets and fine-tune a single LoRA adapter on the combined corpus, providing an optimistic upper bound. **(ii) Federated with LoRA (w/o DP):** run FedAvg with frozen backbones and client-specific LoRA adapters; the server aggregates adapter parameters using data-size weighting to isolate decentralization effects. **(iii) Federated LoRA under domain-aware DP (FedMentor):** clients allocate domain-specific privacy budgets with layer- and adapter-specific noise scaling; the server performs FedAvg on noised adapters and reduces noise when utility proxies fall below thresholds. Full baseline configurations appear in App. B.3.

Evaluation Metrics. We evaluate centralized models with toxicity mean and max, safe rate (Gehman et al., 2020), BERTScore F1 (Zhang et al., 2020), relevance (Reimers and Gurevych, 2019), ROUGE-L (Lin, 2004), and perplexity (Jelinek et al., 1977). In federated settings we report toxicity mean, safe rate, BERTScore F1, relevance, and ROUGE-L. With DP we additionally track domain sensitivity, training time, final losses, memory, adapter size, and round-level averages including communication overhead measured by LoRA updates. App. B.4 provides details on the evaluation metrics.

Setting	Method	SR (%) ↑		Tmn (%) ↓		B-F1 (%) ↑		R-L (%) ↑		REL (%) ↑	
		ZS	FS	ZS	FS	ZS	FS	ZS	FS	ZS	FS
Dreaddit											
Central	ParetoQ-350M	98.0	98.5	1.16	0.70	83.2	86.1	6.25	13.3	23.5	18.2
	Qwen3-0.6B	99.5	91.0	0.67	5.04	83.0	84.3	5.83	9.30	38.3	91.6
	Qwen3-1.7B	99.5	90.0	0.49	5.07	83.4	84.3	6.38	8.83	36.0	91.3
FL (w/o DP)	ParetoQ-350M	99.0	99.0	0.76	0.90	82.4	82.4	3.30	3.12	22.5	22.5
	Qwen3-0.6B	92.0	93.0	2.43	2.20	82.5	82.5	3.78	3.81	65.8	65.5
	Qwen3-1.7B	93.0	93.0	2.35	2.52	82.4	82.4	4.03	4.07	62.9	62.4
FedMentor (FL w/ DP)	ParetoQ-350M	94.0	98.0	2.32	1.28	82.5	82.4	2.91	3.29	22.6	23.7
	Qwen3-0.6B	92.0	93.0	2.23	2.19	82.5	82.5	3.82	3.78	66.3	66.1
	Qwen3-1.7B	93.0	92.0	2.29	2.66	82.4	82.4	4.05	4.07	62.9	62.7
IRF											
Central	ParetoQ-350M	98.0	96.0	1.04	2.21	83.2	85.3	4.89	11.9	26.7	31.0
	Qwen3-0.6B	92.5	66.0	4.25	17.0	83.2	83.7	4.29	6.72	43.5	92.6
	Qwen3-1.7B	96.5	65.5	1.00	16.8	83.8	83.8	4.90	7.75	28.4	92.2
FL (w/o DP)	ParetoQ-350M	94.0	96.0	1.27	2.52	82.5	82.5	3.36	3.50	23.9	25.4
	Qwen3-0.6B	78.0	78.0	8.36	8.44	82.1	82.1	3.80	3.78	70.6	71.3
	Qwen3-1.7B	79.0	78.0	7.97	8.33	82.0	82.0	3.81	3.83	68.5	68.9
FedMentor (FL w/ DP)	ParetoQ-350M	95.0	92.0	1.71	3.47	82.4	82.4	3.03	3.27	25.3	24.0
	Qwen3-0.6B	77.0	78.0	8.53	8.44	82.1	82.1	3.79	3.82	69.8	71.4
	Qwen3-1.7B	78.0	79.0	8.37	8.18	82.0	82.0	3.82	3.85	68.0	68.2
MultiWD											
Central	ParetoQ-350M	96.0	94.5	2.25	2.78	83.0	84.6	5.00	7.06	25.5	27.0
	Qwen3-0.6B	95.5	66.5	2.78	17.9	84.1	82.8	5.77	5.21	27.0	88.5
	Qwen3-1.7B	98.5	67.0	1.31	17.2	83.7	82.7	6.50	4.61	23.4	88.1
FL (w/o DP)	ParetoQ-350M	95.0	93.0	2.26	2.92	82.4	82.3	3.57	3.39	24.3	24.2
	Qwen3-0.6B	80.0	80.0	5.91	5.86	82.0	82.0	3.34	3.38	70.9	70.2
	Qwen3-1.7B	80.0	80.0	5.57	5.66	81.9	81.9	3.43	3.42	68.3	68.3
FedMentor (FL w/ DP)	ParetoQ-350M	95.0	95.0	1.68	1.98	82.3	82.4	3.31	3.76	23.5	22.9
	Qwen3-0.6B	79.0	80.0	6.03	6.25	82.0	82.0	3.39	3.37	69.8	70.7
	Qwen3-1.7B	79.0	81.0	6.30	6.03	81.9	81.9	3.39	3.44	68.7	68.1

Table 1: Performance of LLM backbones under different training setups on Dreaddit, IRF, and MultiWD. The table reports Zero-Shot (ZS) and Few-Shot (FS) results for Safe Rate (SR ≡ TSR), Toxicity mean (Tmn), BERTScore F1 (B-F1), ROUGE-L (R-L), and Relevance (REL). Centralized denotes fine-tuning on combined data (upper bound), FL (w/o DP) refers to FL without DP, and FedMentor applies domain-aware DP. In the FL setting, each dataset is treated as a single client to reflect domain heterogeneity. All metrics are in % (↑ higher is better; ↓ lower is better).



(a) FL (b) FedMentor

Figure 3: Cross-dataset consistency (FS, ROUGE-L) under FL (left) and FedMentor (right).

Prompt Construction. We use two prompting protocols to ensure consistent evaluation across datasets and models. Zero-shot evaluation (Kojima et al., 2022) tests instruction following without the use of in-context examples, relying only on task instructions. Building on this, few-shot evaluation (Brown et al., 2020; Sanh et al., 2022) performs supervised LoRA adaptation on each client under domain-aware DP with FedAvg aggregation, and inference reuses the same wrapper employed in the zero-shot setting. Full templates, wrappers, and examples appear in App. B.5.

5 Main results

Privacy and Safety. FedMentor enforces strict per-domain DP while achieving safer outputs and nearly the same utility as non-private FL (Table 1). Across Dreaddit, IRF, and MultiWD, TSR increases and toxicity decreases, while BERTScore F1 and ROUGE-L change slightly. On MultiWD (few-shot, ParetoQ-350M), TSR improves from 93.0% to 95.0% and mean toxicity drops from 2.92 to 1.98. On IRF, the strictest privacy domain, FedMentor achieves 92% safe outputs versus 96% for FL and keeps BERTScore F1 and ROUGE-L within 0.1 to 0.2 of the no-DP baseline. Scaling up, FedMentor with Qwen3-0.6B matches the FL safe rate (≈93%) with identical BERTScore F1, and

Model	Adpt (MB)	Comm (MB)	Mem (GB)	Time (min)
Global summary (aggregated)				
ParetoQ-350M	16.56	49.69	33.74	7.64
Qwen3-0.6B	38.50	110.00	77.86	10.78
Qwen3-1.7B	66.50	172.90	77.86	11.37
Per-dataset breakdown (Client centric)				
Dreaddit				
ParetoQ-350M	16.56	49.69	33.51	4.86
Qwen3-0.6B	38.50	110.00	77.28	7.22
Qwen3-1.7B	66.50	172.90	77.86	7.80
IRF				
ParetoQ-350M	16.56	49.69	33.62	3.96
Qwen3-0.6B	38.50	110.00	77.86	5.85
Qwen3-1.7B	66.50	172.90	77.86	6.31
MultiWD				
ParetoQ-350M	16.56	49.69	33.74	7.64
Qwen3-0.6B	38.50	110.00	76.70	11.37
Qwen3-1.7B	66.50	172.90	77.86	12.25

Table 2: Efficiency comparison of LLM backbones under FedMentor. Columns report adapter size (Adpt), communication per-round (Comm), peak GPU memory (Mem), and training time per-round (Time).

with Qwen3-1.7B the utility gap narrows further. Overall, FedMentor provides per-domain private training that improves safety over vanilla FL and approaches the centralized upper bound; App. C reports consistent trends for SmolLM2-360M/1.7B.

Utility under Heterogeneity. FedMentor sustains competitive utility on non-IID data, and few-shot fine-tuning consistently outperforms zero-shot prompting. In Table 1, under few-shot training FedMentor's performance closely matches FL without DP: for example, on Dreaddit the FedMentor BERTScore F1 and ROUGE-L are within 0.1 and 0.2 of the no-DP FL values, and safe rate remains high (98% vs 99%). Few-shot adaptation yields clear gains over zero-shot across all models. In particular, relevance and BERTScore improve by roughly 1–3 points with few-shot training compared to zero-shot (e.g., IRF relevance rises from 69.8% to 71.4% under FedMentor with Qwen3-0.6B). We observe no domain collapse: Figure 3 shows only modest variation across datasets. Furthermore, at the client level FedMentor exhibits consistent behavior with FL: Figure 5 (App. C) shows per-client relevance scores overlapping (within $\pm 1\%$), indicating uniform utility across clients. Although absolute metrics are lower without fine-tuning, the relative gap between FedMentor and FL remains stable, reinforcing that strong privacy comes at minimal utility cost under heterogeneous data.

Practical Efficiency. FedMentor satisfies practical efficiency requirements in communication, memory, and speed on single-GPU clients. Table 2 reports compact LoRA updates: 16.56 MB for ParetoQ-350M, 38.50 MB for Qwen3-0.6B, and 66.50 MB for Qwen3-1.7B. With three clients, this translates to about 49.7 MB, 110.0 MB, and 172.9 MB communicated per-round, since only adapters, not full weights, are exchanged. Peak memory per client remains about 33.7 GB for the 350M model and 77.9 GB for the Qwen3 models, within a single 80 GB GPU; per-round Qwen3-1.7B memory for the IRF ε sweep in Table 5 (App. C) shows all clients below the device limit. Average round time (two local epochs, three clients) is 7.6, 10.8, and 11.4 minutes for the 350M, 0.6B, and 1.7B backbones. Dataset trends follow size: IRF is fastest (3.96 min at 350M), Dreaddit is intermediate, and MultiWD is slowest (up to 12.3 min at 1.7B). Communication cost is constant across datasets (50–173 MB) as it depends on model size only. Together with Table 1, these results indicate low overhead and practical training times.

6 Ablation Studies

Uniform Privacy. Uniform Privacy applies a single global privacy budget to all domains. In Figure 4a–b and Table 3, this choice markedly improves fairness in safety: with a uniform ϵ=1.0, the TSR concentrates at ∼99–100% and the client spread contracts to 1% (vs. 6% under the domain-specific baseline). In contrast, the static noise variant shows lower average safety (87.33% TSR) and a larger spread (18%). The utility-gated variant with τ=0 yields the lowest safety (74.67% TSR). BERTScore F1 and REL remain close to the baseline across settings, indicating that the fairness gains from a uniform budget come with only modest utility change. All panels in Figure 4 report the final global model after 8 rounds.

Static Noise. We ablate adaptive noise by fixing the noise scale across rounds. Table 3 shows that Static Noise reduces safety and increases disparity: TSR drops to 87.33% and the client spread widens to 18% (baseline 97.33%, spread 6%). The rightward bars in Figure 4a–b align with this trend, reflecting weaker safety and utility than the domain-aware and uniform settings. The result indicates that dynamic privacy control is beneficial for cross-domain balance, whereas a fixed noise schedule

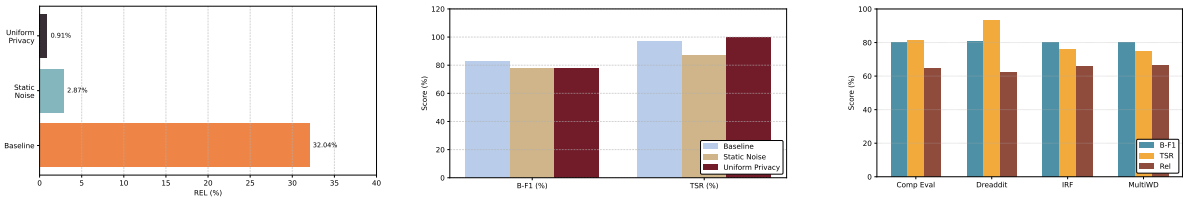


Figure 4: ParetoQ-350M: (a) REL and (b) B-F1 with TSR under Baseline, Static, and Uniform. Qwen3-1.7B: (c) ε ablation of B-F1, TSR, and REL on global evaluation and on Dreaddit, IRF, and MultiWD. All panels report the final global model after 8 rounds (2 local epochs per-round).

Privacy Strategy	Target Scope	$\varepsilon_{\text{glob}}$	Budgets (D/I/M)	σ	τ	TSR mean	TSR min	TSR max	TSR std	Spread
Baseline	Domain Spec.	–	**2.0 / 0.5 / 1.5**	derived	–	97.33	94	100	2.49	6
IRF Sensitivity	Domain Spec.	fixed	**2.0 / 0.1 / 1.5**	derived	–	78.33	66	88	9.18	22
		fixed	**2.0 / 0.5 / 1.5**	derived	–	75.33	57	90	13.72	33
		fixed	**2.0 / 1.0 / 1.5**	derived	–	77.33	70	86	6.60	16
Static Noise	Global	–	–	σ=0.008	–	87.33	80	98	7.72	18
Uniform Privacy	Global	1.0	–	derived	–	99.67	99	100	0.47	1
Utility Threshold	Global	–	–	–	B-F1 τ=0	74.67	65	85	8.18	20

Table 3: Client fairness after 8 federated rounds (ParetoQ-350M). Budgets are reported as D/I/M where D = Dreaddit, I = IRF, and M = MultiWD. IRF Sensitivity varies only the IRF budget; Dreaddit and MultiWD are fixed at 2.0 and 1.5. Metrics in % (↑ higher is better; ↓ lower is better).

amplifies inter-client gaps without yielding compensatory gains.

App. D.1 covers the IRF ϵ sweep; App. D.2 the utility-threshold ablation; App. D.3 per-client memory scaling across rounds.

7 Related Works

FL for Mental Health. Federated Learning (FL) has demonstrated performance comparable to centralized training while satisfying strict privacy regulations such as HIPAA and GDPR (Lieftink et al., 2024; Peng et al., 2024; Woisetschläger et al., 2024). By training directly on decentralized data, FL is well-suited for mobile health and multi-clinic settings. In mental health, FedTherapist (Shin et al., 2023) leverages on-device FL with speech and keyboard signals to track depression, stress, and mood, and FedMood (Xu et al., 2021) introduces a multi-view framework for depression diagnosis using heterogeneous mobile health data. These studies, along with work on mobile sensing, electronic health records, and multi-institutional collaboration, demonstrate the potential of FL for sensitive mental health applications (Rauniyar et al., 2023).

Building on these foundations, recent work has extended FL to LLMs for conversational and diagnostic support. FedMentalCare (Sarwar, 2025) combines FL with LoRA to fine tune lightweight LLMs while reducing communication costs. However, many methods assume homogeneous data and only partly address scalability and efficiency. We introduce FedMentor, which enables heterogeneous FL through explicit domain modeling, LoRA-based communication reduction, and domain-aware DP for practical and privacy-preserving LLM fine tuning in mental health.

App. A discusses related work on LLMs in Mental Health and Differential Privacy in FL.

8 Conclusion

We introduced FedMentor, a federated fine-tuning framework that pairs domain-aware Differential Privacy with LoRA adapters for sensitive mental health dialogue. Across heterogeneous domains, FedMentor raises safety while keeping utility near non-private Federated Learning and near centralized training. Practicality follows from communicating adapter weights only, which keeps memory and bandwidth within single GPU budgets. Ablations confirm that domain-specific budgets and adaptive noise are essential, since removing either reduces accuracy and amplifies client disparities. The study utilizes a limited set of datasets and automated metrics without clinician review. Future work will scale federations, incorporate clinician-in-the-loop approaches, and conduct multilingual evaluation, as well as extend fairness analyses.

Limitations

Several factors limit the present study. First, the evaluation utilizes three English social-media datasets, treating each dataset as a single client, which limits external validity for larger cross-silo federations and clinical conversations. Second, assessment relies on automatic quality and safety proxies, including detector-based toxicity and embedding similarity, without clinician review or user studies. Third, the work reports no audit of demographic or linguistic subgroups and does not provide round-level composition accounting or attack-based testing for privacy loss. Fourth, system results reflect an A100 80 GB environment and a small number of clients, so hardware and scale shifts may alter efficiency. Finally, comparisons do not include alternative differentially private fine-tuning strategies beyond LoRA-based adapters.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Jayadev Acharya, Kallista Bonawitz, Peter Kairouz, Daniel Ramage, and Ziteng Sun. 2020. Context aware local differential privacy. In *International Conference on Machine Learning*, pages 52–62. PMLR.

Accountability Act and 1 others. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Miguel Baidal, Erik Derner, and Nuria Oliver. 2025. Guardians of trust: Risks and opportunities for llms in mental health. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 11–22.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self-states from social media posts. In *The 10th Workshop on Computational Linguistics and Clinical Psychology*, page 256.

Zachary Charles, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Nicole Mitchell, Krishna Pillutla, and Keith Rush. 2024. Fine-tuning large language models with user-level differential privacy. *arXiv preprint arXiv:2407.07737*.

Olivia Choudhury, Aris Gkoulalas-Divanis, Theodoros Salonidis, Issa Sylla, Yoonyoung Park, Grace Hsu, and Amar Das. 2019. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910.02578*.

Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.

Cynthia Dwork, Aaron Roth, and 1 others. 2014. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407.

Hamid Ebadi, David Sands, and Gerardo Schneider. 2015. Differential privacy: Now it's getting personal. *Acm Sigplan Notices*, 50(1):69–81.

European Parliament and Council of the European Union. 2016. General Data Protection Regulation (GDPR). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. Accessed August 2025.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can ai relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221.

Muskan Garg, Xingyi Liu, MSVPJ Sathvik, Shaina Raza, and Sunghwan Sohn. 2024. Multiwd: Multi-label wellness dimensions in social media posts. *Journal of biomedical informatics*, 150:104586.

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. 2023. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of ACL 2023*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. 2021. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR.

Batool Haider, Atmika Gorti, Aman Chadha, and Manas Gaur. 2025. Mental health equity in llms: Leveraging multi-hop question answering to detect amplified and silenced perspectives. *arXiv preprint arXiv:2506.18116*.

Amey Hengle, Atharva Kulkarni, Shantanu Deepak Patankar, Madhumitha Chandrasekaran, Sneha D'silva, Jemima S. Jacob, and Rashmi Gupta. 2024. Still not quite there: Evaluating large language models for comorbid mental health diagnosis. In *Proceedings of EMNLP 2024*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Institute for Health Metrics and Evaluation (IHME). 2023. Global Health Data Exchange (GHDx). Accessed August 2025.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Zach Jorgensen, Ting Yu, and Graham Cormode. 2015. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*, pages 1023–1034. IEEE.

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*, pages 5132–5143. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Vivek Kumar, Pushpraj Singh Rajwat, Giacomo Medda, Eirini Ntoutsi, and Diego Reforgiato Recupero. 2024. Unlocking llms: Addressing scarce data and bias challenges in mental health and therapeutic counselling. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 238–251.

Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. 2025. Exploring user security and privacy attitudes and concerns toward the use of {General-Purpose}{LLM} chatbots for mental health. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6007–6024.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Natalie Lieftink, Carolina dos S Ribeiro, Mark Kroon, George B Haringhuizen, Albert Wong, and Linda HM van de Burgwal. 2024. The potential of federated learning for public health purposes: a qualitative analysis of gdpr compliance, europe, 2021. *Eurosurveillance*, 29(38):2300695.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Kush R Varshney. 2023. Towards healthy ai: Large language models need therapists too. *arXiv preprint arXiv:2304.00416*.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2022. Fednlp: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. 2021. Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment*, 15(4):828–840.

Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. 2024a. Cross-silo federated learning with record-level personalized differential privacy. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 303–317.

Ken Liu, Shengyuan Hu, Steven Z Wu, and Virginia Smith. 2022. On privacy and personalization in cross-silo federated learning. *Advances in neural information processing systems*, 35:5925–5940.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2025. Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*, 16(2):1–24.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, and 1 others. 2024b. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.

Richard May and Kerstin Denecke. 2022. Security, privacy, and healthcare-related conversational agents: a scoping review. *Informatics for Health and Social Care*, 47(2):194–210.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *International Conference on Learning Representations*.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. Welldunn: On the robustness and explainability of language models and large language models in identifying wellness dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J Soled, Michael L Birnbaum, Srijan Kumar, and 1 others. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7488–7511.

Jennifer Nicholas, Sandersan Onie, and Mark E Larsen. 2020. Ethics and Privacy in Social Media Research for Mental Health. *Current psychiatry reports*, 22:1–7.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. 2022. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 10110–10145. PMLR.

Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Ju Sun, and Rui Zhang. 2024. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *NPJ Digital Medicine*, 7(1):127.

Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. 2023. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 11(5):7374–7398.

Sashank Reddi, Zachary Burr Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Brendan McMahan. 2021. Adaptive federated optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Protection Regulation. 2018. General data protection regulation. *Intouch*, 25:1–5.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Jae Hun Ro, Theresa Breiner, Lara McConnaughey, Mingqing Chen, Ananda Theertha Suresh, Shankar Kumar, and Rajiv Mathews. 2022. Scaling language model size in cross-device federated learning. *arXiv preprint arXiv:2204.09715*.

Rob Romijnders, Christos Louizos, Yuki M. Asano, and Max Welling. 2024. Protect your score: Contact-tracing with differential privacy guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14829–14837.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, and 1 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

SM Sarwar. 2025. Fedmentalcare: towards privacy-preserving fine-tuned llms to analyze mental health status using federated learning framework. *arXiv preprint arXiv:2503.05786*.

Ramit Sawhney, Atula Neerkaje, Ivan Habernal, and Lucie Flek. 2023. How much user context do we need? privacy by design in mental health nlp applications. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 766–776.

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*, pages 11971–11988.

Towards Healthcare. 2023. Chatbots for mental health and therapy market size envisioned at USD 6.51 billion by 2032. Accessed August 2025.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of LOUHI 2019*.

United States Congress. 1996. Health Insurance Portability and Accountability Act of 1996. Public Law No. 104–191, 110 Stat. 1936. Accessed August 2025.

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*, 33:7611–7623.

Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An evaluation of generative pre-training model-based therapy chatbot for caregivers. *arXiv preprint arXiv:2107.13115*.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate V. Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024a. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of EMNLP 2024*.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. 2024b. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533.

Herbert Woisetschläger, Alexander Erben, Bill Marino, Shiqiang Wang, Nicholas D Lane, Ruben Mayer, and Hans-Arno Jacobsen. 2024. Federated learning priorities under the european union artificial intelligence act. *CoRR*.

Xiaohang Xu, Hao Peng, Lichao Sun, Md Zakirul Alam Bhuiyan, Lianzhong Liu, and Lifang He. 2021. Fedmood: Federated learning on mobile health data for mood detection. *arXiv preprint arXiv:2102.09342*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023a. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023b. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xinzhe Zheng, Sijie Ji, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. 2025. ProMind-LLM: Proactive mental health care via causal reasoning with sensor data. In *Findings of ACL 2025*.

A Additional Related Works

LLMs in Mental Health. LLMs are rapidly adopted for mental health applications, supporting tasks such as condition detection, diagnosis, and therapeutic dialogue (Wang et al., 2021; Yang et al., 2023b,a; Xu et al., 2024; Wang et al., 2024a; Mohammadi et al., 2024; Zheng et al., 2025; Chan et al., 2025; Hengle et al., 2024; Haider et al., 2025). While these systems demonstrate strong potential for supportive interaction, most current approaches are centralized and depend on aggregating sensitive dialogue data. Such data are scarce due to confidentiality concerns, and strict regulations, including HIPAA and GDPR, further constrain sharing (Kwesi et al., 2025; Baidal et al., 2025; Nguyen et al., 2025). As a result, datasets are often small and biased, which limits the robustness and generalization of models.

These limitations reveal a critical privacy-utility trade-off that existing LLM methods rarely address, reducing their feasibility in real-world clinical contexts (Lin et al., 2023; Chung et al., 2023; Gabriel et al., 2024; Kumar et al., 2024). FedMentor addresses this challenge by integrating Federated Learning with domain-aware Differential Privacy, enabling collaborative training without centralizing data. This framework allows models to learn from sensitive mental health texts while preserving confidentiality, thereby enhancing both trust and clinical applicability.

Differential Privacy in FL. Differential privacy (DP) has become a standard approach for protecting sensitive data in ML, including FL (Dwork et al., 2014; Abadi et al., 2016; Girgis et al., 2021; Noble et al., 2022). Applying DP in practice, however, remains difficult, particularly in mental health settings. Traditional methods that uniformly clip gradients and add noise to updates often degrade performance in healthcare, where datasets are small and signals are fine-grained (McMahan et al., 2018). Although DP has been studied in areas such as electronic health records and medical imaging, its application in mental health remains limited, with only a few recent studies exploring privacy-by-design strategies (Choudhury et al., 2019; Sawhney et al., 2023; Romijnders et al., 2024). Recent research further highlights that not all model components or data are equally sensitive, motivating personalized and domain-aware DP strategies that adapt privacy to data characteristics (Ebadi et al., 2015; Jorgensen et al., 2015; Liu et al., 2021, 2022, 2024a). In LLMs, parameter-efficient methods such as DP-LoRA (Liu et al., 2025) reduce risk and communication costs by perturbing only low-rank adapter updates. However, most approaches still adopt a one-size-fits-all noise and overlook the unique sensitivities of mental health text. To overcome this limitation, we introduce FedMentor, a framework that integrates domain-aware DP with heterogeneous federated learning to provide strong privacy guarantees while preserving clinical utility.

B Additional Experiment Settings

B.1 Dataset Details

Dreaddit (Turcan and McKeown, 2019) is a Reddit corpus for stress detection with 3,553 text segments labeled as stressed or not stressed. We use the official splits, add a 10% stratified validation set, and frame the task as supportive-response generation with stress-sensitive instructions. Labels are retained for monitoring and prompt-target selection.

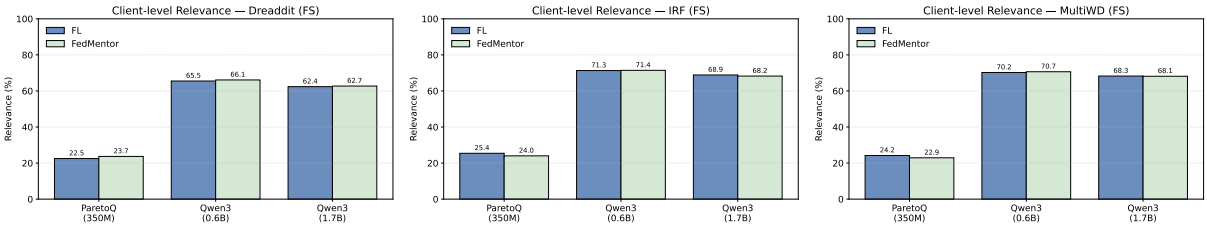
IRF (Garg et al., 2023) contains 3,522 posts annotated for Thwarted Belongingness and Perceived Burdensomeness. We follow the original splits and map the multi-label annotations to a binary indicator of interpersonal risk, aligning with our generation objective. Labels are used for utility tracking and evaluation.

MultiWD (Garg et al., 2024) provides 3,281 Reddit posts annotated with six wellness dimensions. We form a 10% validation set, apply minority upsampling to balance classes, and reduce the labels to a binary indicator for supportive-response monitoring while retaining per-dimension information for conditioning and evaluation.

Each dataset is assigned to a distinct client, forming a natural non-IID configuration with domain-specific shifts. Preprocessing removes empty rows, standardizes text fields, and coerces labels to $0, 1$, supporting a uniform generation-centric framing across stress, interpersonal risk, and wellness.

B.2 Model Specification

MobileLLM-ParetoQ (350M) (Liu et al., 2024b). MobileLLM targets on device efficiency with a deep thin Transformer, embedding tying, grouped query attention, and blockwise weight sharing. ParetoQ adds sub 4 bit quantization with quantization aware training and learnable scaling. In *FedMentor*, we load 4 bit NF4 with bf16 compute



(a) Dreaddit (b) IRF (c) MultiWD

Figure 5: Client level relevance on three datasets. Bars compare FL and FedMentor for ParetoQ 350M, Qwen3 0.6B, and Qwen3 1.7B.

and fine-tune only LoRA adapters (rank 8), producing very small adapter states. Domain aware DP is applied directly to adapters, which keeps privacy cost and per round communication low.

SmolLM2 (360M and 1.7B) (Allal et al., 2025). SmolLM2 uses a data centric multi stage pipeline and efficient components such as RMSNorm, rotary embeddings, and gated feed forward layers. In *FedMentor*, we freeze the backbone, attach LoRA adapters (rank 16), quantize to 4 bit NF4 with bf16 compute, and aggregate noised adapters with FedAvg. Utility-guided noise reduction maintains clinical metrics while preserving low-bandwidth through compact adapter updates.

Qwen3 (0.6B and 1.7B) (Yang et al., 2025). Qwen3 provides dense variants trained with instruction tuning and reinforcement learning for reasoning and remains memory efficient. In *FedMentor*, Qwen3 integrates with our prompt manager for zero shot and few shot generation. We quantize the frozen backbone to 4 bit NF4 with bf16 compute, fine-tune LoRA adapters (rank 16), and apply domain specific DP budgets with layer and adapter aware noise scaling before FedAvg, achieving privacy preserving updates without logit sharing or knowledge distillation.

B.3 Baseline Specification

Centralized (w/o FL, w/o DP). Pool all datasets and fine-tune a single LoRA adapter on the combined corpus without privacy constraints. Serves as an optimistic upper bound when central aggregation is hypothetically allowed.

Federated with LoRA (w/o DP). Standard FedAvg with LoRA adapters. Each client fine-tunes its local adapter while the backbone remains frozen. The server aggregates adapter parameters via data-weighted averaging, isolating the effect of decentralization.

Federated LoRA under Domain-Aware DP (FedMentor, Ours). Each client applies domain specific privacy budgets to its LoRA parameters with layer and adapter aware noise scaling, assigning larger noise to more sensitive components and tighter budgets for higher risk domains. The server performs FedAvg over noised adapters, monitors clinical utility proxies during rounds, and applies utility guided noise reduction when thresholds are violated. This design preserves privacy, reduces communication by sharing only adapters, and stabilizes training under non IID client distributions.

Dataset-Specific Instruction Prompts

Dreaddit. Analyze this Reddit post for signs of psychological stress. Provide a brief, supportive response in 1–2 sentences. If stress is present, acknowledge feelings and suggest one helpful action. If no stress, provide encouragement. Be empathetic but avoid medical advice.
Post: {post}

IRF. Analyze this post for signs of interpersonal risk (thwarted belongingness or perceived burdensomeness). Provide a brief, supportive response in 1–2 sentences. Be empathetic but avoid medical advice.
Post: {post}

MultiWD. Analyze this post for social and emotional wellness concerns. Social wellness involves relationships, connection, and community. Emotional wellness involves managing feelings, stress, and mental health. Provide a brief, supportive response in 1–2 sentences. Be empathetic and suggest practical coping strategies.
Post: {post}

Figure 6: Dataset-specific instructions

B.4 Detailed Evaluation Metrics

We report metrics for quality, safety, and systems efficiency that match the implementation.

Centralized (w/o FL, w/o DP). We evaluate toxicity mean and toxicity max of generated responses using Detoxify, safe rate (fraction of responses

Zero-Shot Model Wrappers

Qwen

```
<|im_start|>user
{INSTRUCTION}
Post: {post}
<|im_end|>
<|im_start|>assistant
```

SmolLM2

```
### Input:
{INSTRUCTION}
Post: {post}

### Response:
```

MobileLLM

```
{INSTRUCTION}
Post: {post}

Response:
```

Figure 7: Prompt templates: Zero-shot wrappers for **Qwen**, **SmolLM2**, and **MobileLLM**.

below a fixed toxicity threshold), BERTScore F1 for semantic similarity, relevance mean via embedding based cosine similarity, ROUGE-L for lexical overlap, and perplexity mean as a fluency proxy. These quantify response quality and safety for single model generation.

Federated with LoRA (w/o DP). For each client and in aggregate we report toxicity mean, safe rate, BERTScore F1, relevance mean, and ROUGE-L. In accordance with the code, toxicity max and perplexity mean are not computed in the federated evaluator.

Federated with LoRA (w/ DP) (System Metrics). To characterize privacy–utility tradeoffs and overhead, we log domain sensitivity, training time, final train loss, final eval loss, memory used (MB), and LoRA weights size (MB) per client, along with per round aggregates: average train loss, average eval loss, total training time, average memory usage, and communication overhead (MB). Communication overhead reflects the size of adapter payloads, since only LoRA adapters are shared rather than full model weights.

B.5 Prompt Construction for Zero-Shot and Few-Shot

Zero-Shot Prompting. We follow standard zero-shot evaluation, supplying a single instruction plus the post, wrapped in the backbone native template: Qwen uses chat role tags, SmolLM2 uses *Input*/*Response* headers, and MobileLLM uses plain text with a trailing *Response:*. Instructions are domain aligned: stress identification and supportive coping for **Dreaddit**; empathetic interpersonal risk screening for **IRF** (thwarted belongingness or perceived burdensomeness); and brief practical guidance for social and emotional wellness cues in **MultiWD**. This uniform yet domain aware setup isolates instruction following while keeping comparisons consistent across datasets and models. Exact wrappers and examples are shown in Figure 7, with dataset specific instructions in Figure 6.

Few-Shot and Train-Time Shape

Few-shot wrapper (generic k).

```
[template prefix]
{INSTRUCTION}

Example 1
Post: {post_1}
Response: {response_1}
...
Example k
Post: {post_k}
Response: {response_k}

Post: {post}
Response:
[template suffix]
```

Exact train-time prompt shape.

```
[template prefix] +
{INSTRUCTION} + "\n\nPost: {post}" +
[template suffix] +
{supervised target response} +
[template end]
```

Figure 8: Prompt templates: Few-shot wrapper and exact train-time prompt shape.

Few-Shot Prompting. We implement few-shot as supervised LoRA adaptation on each client rather than in context exemplars. Only adapters are trained on a quantized backbone; updates are privatized with domain aware Gaussian noise and aggregated with FedAvg. Short supportive targets align tone and structure to each domain, which increases safe rate, reduces toxicity mean, and improves BERTScore F1. At inference, we reuse the zero-shot wrapper for a clean comparison. The generic train time wrapper and target concatenation appear in Figure 8.

Setting	Method	SR (%) ↑		Tmx (%) ↓		B-F1 (%) ↑		R-L (%) ↑		Rel (%) ↑	
		ZS	FS	ZS	FS	ZS	FS	ZS	FS	ZS	FS
Dreaddit											
Central	SmolLM2-360M	98.0	98.0	76.44	79.45	82.6	83.3	6.35	7.79	23.5	28.1
	SmolLM2-1.7B	98.0	98.0	74.79	75.52	82.7	83.6	5.75	6.60	23.2	27.2
FedMentor (FL w/ DP)	SmolLM2-360M	98.0	98.0	0.59	0.65	82.7	82.7	2.12	2.17	26.9	44.3
	SmolLM2-1.7B	97.0	98.0	0.54	0.68	82.2	82.0	2.02	2.49	31.6	0.5
IRF											
Central	SmolLM2-360M	98.0	98.0	36.00	62.42	84.0	86.3	2.52	11.85	27.1	27.1
	SmolLM2-1.7B	98.0	97.5	50.28	62.07	82.9	85.1	2.92	7.39	28.7	28.1
FedMentor (FL w/ DP)	SmolLM2-360M	78.0	79.0	7.96	7.80	82.9	81.6	3.78	3.67	70.3	71.6
	SmolLM2-1.7B	77.0	79.0	9.09	8.23	83.0	81.3	3.38	3.55	67.1	0.0
MultiWD											
Central	SmolLM2-360M	97.5	98.5	47.74	61.66	84.4	86.7	6.23	11.57	39.8	26.8
	SmolLM2-1.7B	97.5	98.0	70.80	74.85	83.4	85.2	6.27	7.86	42.9	27.9
FedMentor (FL w/ DP)	SmolLM2-360M	97.0	96.0	1.12	1.91	83.0	82.8	3.29	3.17	38.3	43.5
	SmolLM2-1.7B	97.0	96.0	1.59	2.43	81.7	78.9	2.99	3.64	41.8	0.1

Table 4: Performance of SmolLM2 backbones under centralized and FedMentor (FL w/ DP) training on Dreaddit, IRF, and MultiWD. Columns report Zero-shot (ZS) and Few-shot (FS) for Safe Rate (SR), Toxicity max (Tmx), BERTScore F1 (B-F1), ROUGE L (R-L), and Relevance (Rel). All metrics are in %.

Round	Client 0 (GB)	Client 1 (GB)	Client 2 (GB)
0	18.85	20.04	21.23
1	22.42	23.60	24.79
2	25.98	27.16	28.35
3	29.54	30.72	31.91
4	33.10	34.29	35.47
5	36.66	37.85	39.03
6	40.22	41.41	42.59
7	43.78	44.97	46.16

Table 5: Per-round client memory usage for Qwen3-1.7B under the IRF ϵ-sensitivity test with $\epsilon = 1.0$. The sweep varies only the IRF domain privacy budget, while Dreaddit and MultiWD remain fixed at 2.0 and 1.5, respectively.

B.6 Hardware Details

All experiments were conducted on a server with NVIDIA A100 GPUs, each with 80 GB of memory. Mixed-precision computation with bfloat16 (bf16) was used for both training and inference. Models were loaded in bf16 and executed with 4-bit NF4 quantization and bf16 compute. Up to two A100 GPUs were available, and each federated round allocated client training to a single GPU, enabling parallel execution across clients. This setup allowed efficient LoRA fine-tuning of models with up to 1.7B parameters and concurrent simulation of multiple clients. All training times, memory usage, and communication costs reported in this work were measured in this environment.

C Supplementary Results

Table 4 reports centralized and FedMentor performance for SmolLM2 on Dreaddit, IRF, and MultiWD. The pattern mirrors the main models: FedMentor keeps B-F1 and ROUGE close to baselines while improving safety.

Table 5 presents per-round client memory usage for Qwen3-1.7B under the IRF ϵ-sensitivity test with $\epsilon = 1.0$. The sweep varies only the IRF domain privacy budget, while Dreaddit and MultiWD are held fixed at 2.0 and 1.5, respectively. These results complement the main experiments by showing how memory demands evolve across training rounds when privacy constraints differ across domains.

Figure 5 compares FL and FedMentor for ParetoQ 350M, Qwen3 0.6B, and Qwen3 1.7B; per-client relevance remains aligned, indicating stable utility under non-IID data and no domain collapse.

D Additional Ablations and Efficiency Analyses

D.1 IRF ϵ Sweep

We vary only the IRF domain budget while holding Dreaddit and MultiWD fixed (Figure 4c; Table 6). For Qwen3-1.7B, tightening privacy from ϵ=1.0 to ϵ=0.1 increases safety (TSR $81.33\% \rightarrow 92.00\%$)

Method	TSR	REL	B-F1
$\varepsilon = 0.1$			
ParetoQ-350M	78.3	5.2	78.0
Qwen3-0.6B	83.7	69.4	82.2
Qwen3-1.7B	92.0	62.7	82.4
$\varepsilon = 0.5$			
ParetoQ-350M	75.3	4.9	78.1
Qwen3-0.6B	83.7	69.4	82.2
Qwen3-1.7B	92.0	62.7	82.4
$\varepsilon = 1.0$			
ParetoQ-350M	77.3	5.1	78.0
Qwen3-0.6B	83.7	69.4	82.2
Qwen3-1.7B	81.3	64.8	80.3

Table 6: IRF ε sweep with other DP hyperparameters fixed. Global summary of Toxicity Safe Rate (TSR), Relevance (REL), and BERTScore F1 (B-F1) after 8 federated rounds across 3 datasets. Metrics in % (↑ higher is better; ↓ lower is better).

with a small B-F1 change ($80.33\% \rightarrow 82.40\%$) and a modest REL shift ($64.78\% \rightarrow 62.70\%$). For ParetoQ-350M, REL remains low across the sweep (about 5%), and TSR varies between 75.33% and 78.33%. These trends indicate that stronger IRF privacy can raise safety for larger models with limited utility loss, whereas smaller backbones display nearly constant utility under the same adjustments.

D.2 Utility Threshold Calibration (τ)

We evaluate a utility-driven noise adjustment where the server reduces noise only if B-F1 falls below a threshold τ. Table 3 shows that the τ=0 setting yields the lowest safety, with TSR at 74.67% and a spread of 20%, compared to the baseline TSR of 97.33% and spread of 6%. The observation is consistent with the ParetoQ-350M trends in Figure 4a–b: aggressive utility gating undermines safety, indicating the need for a calibrated threshold.

D.3 Per-Client Memory Scaling across Rounds

Table 5 reports memory usage per client for Qwen3-1.7B under the IRF ϵ=1.0 setting. Memory increases smoothly over rounds: at round 0, clients use about 18.85, 20.04, and 21.23 GB; by round 7, these reach 43.78, 44.97, and 46.16 GB. The monotonic rise indicates predictable scaling during training while keeping per-client usage within a narrow band across clients.

E Use of AI-Assisted Tools

During development, we used GitHub Copilot to suggest code completions and refactorings. For writing, we utilized ChatGPT to proofread and refine phrasing. All inputs to these tools originated with the authors, and every suggestion was reviewed and edited by the authors, who remain responsible for the final content.