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

Nobin Sarwar, Shubhashis Roy Dipta

UMBC

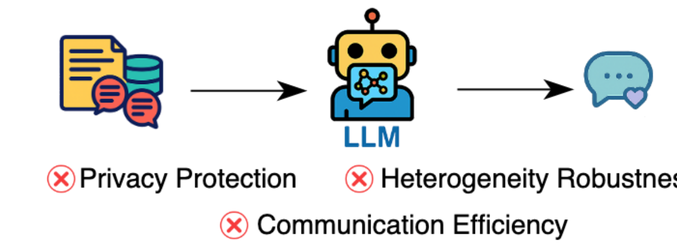
NEURAL INFORMATION PROCESSING SYSTEMS

## TL;DR

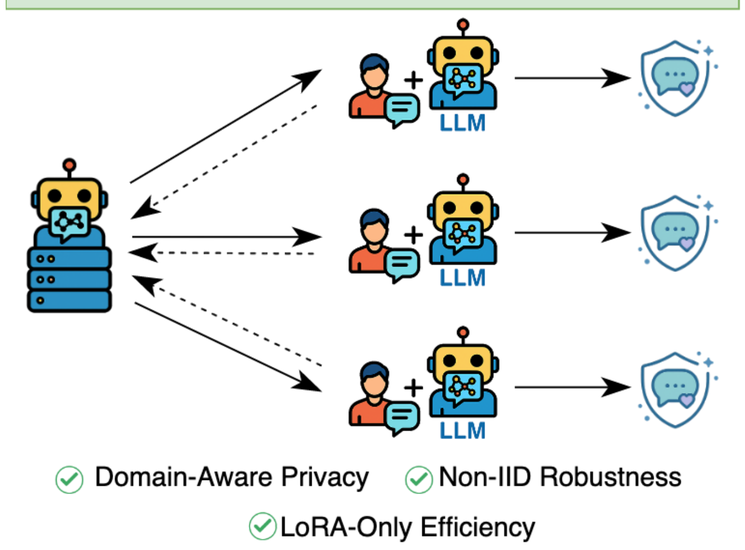
**FedMentor** combines **domain-aware DP** with efficient **LoRA** training, matching **FL utility** and improving **safety** across **heterogeneous** mental health datasets.

## Motivation

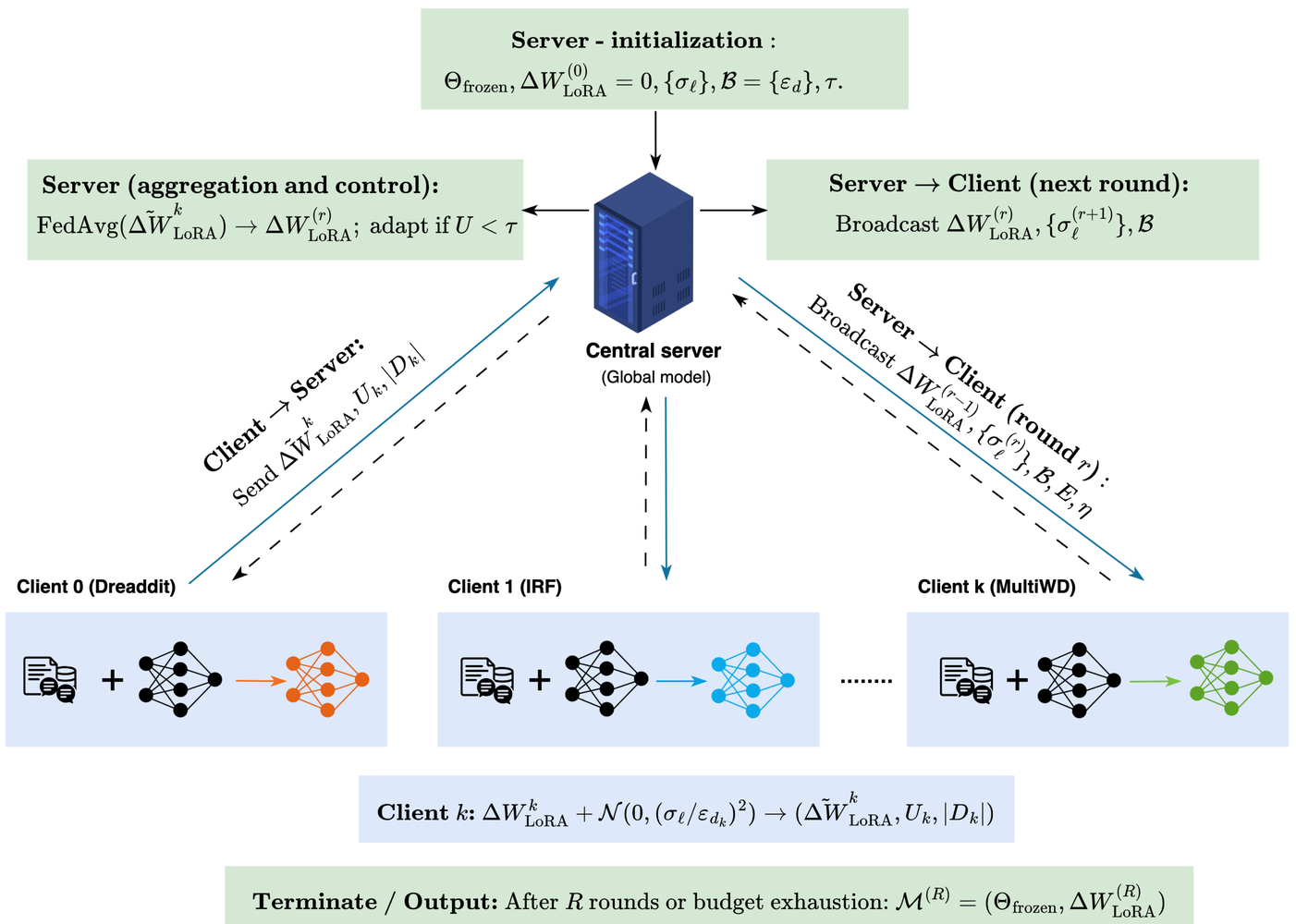
**Traditional Approach**



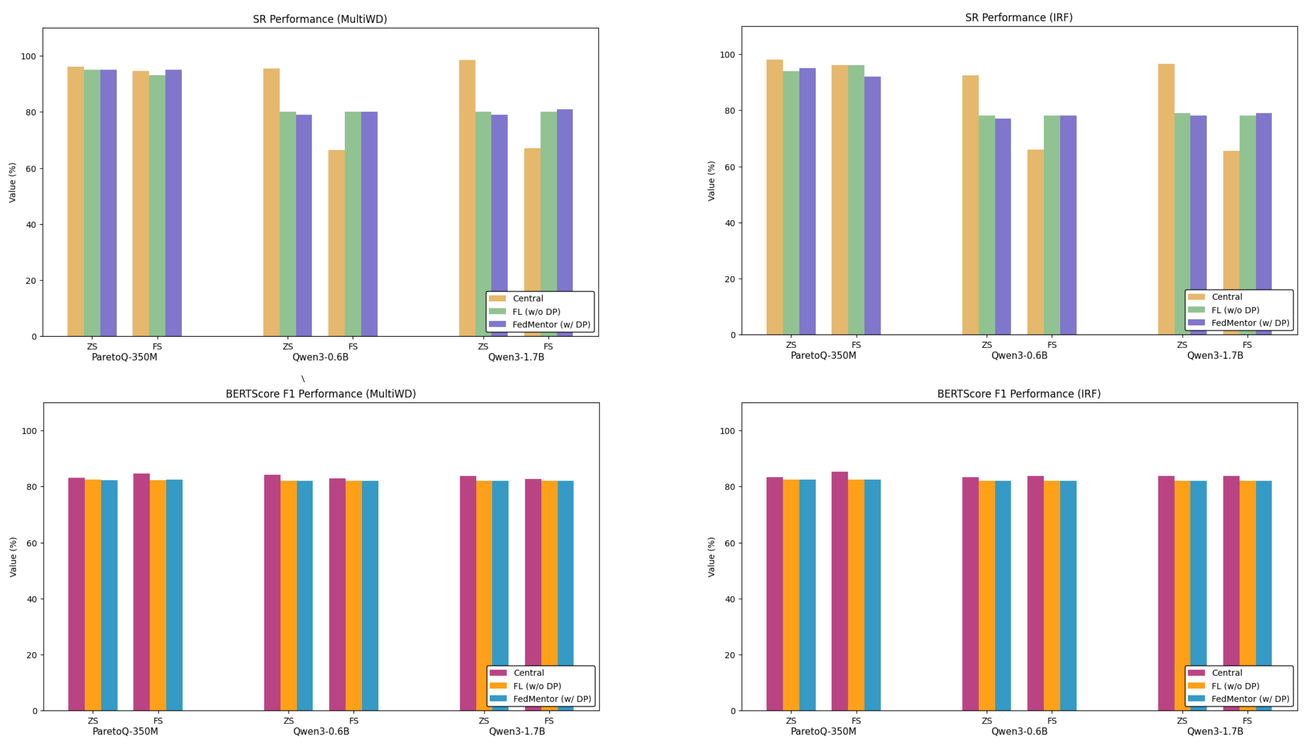

**FedMentor: Proposed Approach**




## FedMentor: Privacy-Preserving Federated LLMs




## Results: Privacy & Safety




- **FedMentor matches** FL utility; **privacy cost** is **minimal**, with safety and quality aligned

- **Safe rates** stay near FL across Dreaddit, IRF, MultiWD and reach **95%** on MultiWD

- Across Dreaddit, IRF, MultiWD, B-F1 stays within **0.2 percentage** points of FL

## Results: Practical Efficiency

- **Adapter updates** stay **small** at **17–67** MB; **communication 50–173** MB per round

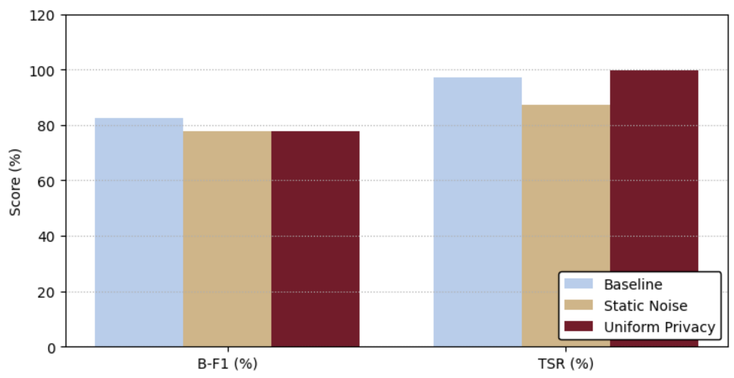
- **Peak memory** fits **80 GB**, with ~**34** GB (ParetoQ-350M) and ~**78** GB (Qwen3-0.6B/ 1.7B) footprints

| 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)** | | | | |
| **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 |

## FedMentor Stays Robust

- **Uniform privacy** maintains TSR ~ **99–100%**

- Across **ε 0.1–1.0**, B-F1 holds **78–82%** range for all backbones




## Conclusions

- We introduced **FedMentor** with **domain-aware DP**, **safe rate 95%**, **B-F1** within **0.2%** difference under heterogeneous clients.
- **LoRA-only** updates send **50–173 MB** per round and fit **80 GB** GPUs, while **uniform privacy** stabilizes safety across models and datasets.

Paper

Code