Duet: Federated SLM–LLM Co-Inference for Mobile AI

Chunlin Tian[1,2], Filippo Serafini[2], Li Li[*1], and Nicholas D. Lane[*2]
[1] *University of Macau* [2] *University of Cambridge*

Abstract. Deploying large language models (LLMs) on mobile devices is challenging: they are too large to run locally, and sending personal context to the cloud threatens privacy. We present Duet, a hybrid federated framework that pairs a cloud-hosted black-box LLM (e.g., ChatGPT) with on-device small language models (SLMs, e.g., Gemma-2B). Personal data and the fine-tuning process stay on the device; the cloud LLM contributes general knowledge without exposing its weights. Duet introduces (i) a heterogeneity-aware adaptive-rank LoRA strategy that fits each device's memory and runtime budget, (ii) a parameter-free Mixture-of-Experts router that absorbs cross-user data heterogeneity, and (iii) a logit-level fusion step that lets the edge SLM and cloud LLM jointly produce each token in real time. On fifteen NVIDIA Jetson devices, Duet consistently outperforms both SLM-only and cloud LLM baselines on the BBH reasoning benchmark, while substantially reducing communication overhead.

Motivation

Mobile users increasingly rely on LLMs, but the dominant deployment paths are unsatisfying. Sending every query to a cloud LLM leaks personal context and is constrained by network latency; running an open-source LLM locally is infeasible (LLaMA-65B alone needs 130 GB in FP16 [2]); and shrinking to an on-device SLM sacrifices reasoning capabilities that make LLMs useful [1]. Closed proprietary LLMs (ChatGPT, Claude) further rule out methods that need access to model internals. The mobile setting therefore needs a design where *data stay on the device*, *LLM weights stay in the cloud*, and the two still cooperate within a real-time budget.

Approach

Duet co-trains and co-serves an SLM/LLM pair across the edge–cloud boundary (Fig. 1). *Federated specialization.* Each device fine-tunes only LoRA [3] adapters on a shared SLM base; a server aggregates adapters across devices, enlarging the effective training distribution without centralizing data [4]. *Heterogeneity-aware adaptation.* Duet estimates available memory and per-token latency on each device and selects a per-layer LoRA rank under those constraints, so a flagship phone and a low-end Jetson Nano can participate in the same round. *Edge–cloud logit fusion.* At inference, both the on-device SLM (personalized to the user) and the cloud LLM (general knowledge) each produce next-token logits; Duet fuses them on the device, so the cloud sees only abstracted queries and never the user's local context.

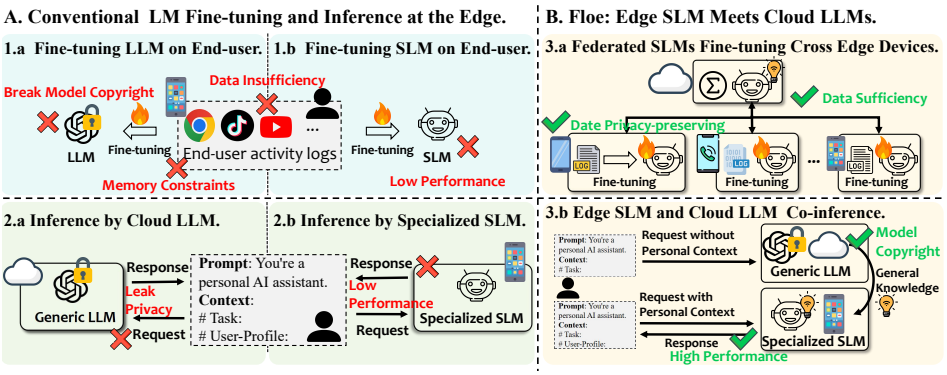

Figure 1: Duet co-locates a personalized edge SLM with a black-box cloud LLM.

Results

We evaluate Duet on 15 NVIDIA Jetson devices with a Gemma-2B edge SLM and a Gemma-7B cloud LLM. On BBH [5], Duet lifts average accuracy to 46.4%, beating SLM-only (32.2%), LLM-only (42.8%), and the strongest federated baseline LLM-FedMoE (45.1%), without exposing local data or LLM weights. Adaptive-rank LoRA admits low-memory Nanos that static-rank baselines drop, while the MoE router reaches 87.6% top-1 expert alignment across task families. Transmitting only LoRA adapters cuts per-round traffic by 99% versus full-model federation, and a two-stage privacy detector (F1 = 94.3%) routes 93% of sensitive prompts entirely on-device. The recipe also generalizes: Llama-3-70B/3B gives +5.6% over the SLM, and the closed Gemini-1.5-Flash API gives +2.6%.

References

[1] Google. Gemma: Open Models Based on Gemini Research and Technology. 2024.
[2] Touvron et al. LLaMA: Open and Efficient Foundation Language Models. 2023.
[3] Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.
[4] Kuang et al. FederatedScope-LLM: A Comprehensive Package for Fine-tuning LLMs in Federated Learning. *KDD*, 2024.
[5] Suzgun et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *ACL*, 2023.

*Corresponding author