Junkyard Pre-Training: Federated LLM Pre-Training on Decentralized Edges

DONGQI CAI[1,2], SHANGGUANG WANG[1], NICHOLAS D. LANE[2], MENGWEI XU[1],

[1]Beijing University of Posts and Telecommunications, [2]University of Cambridge

Large Language Models (LLMs) have significantly advanced natural language understanding and enabled multimodal applications on edge devices [5]. However, training LLMs demands extensive computational resources and large-scale data. During the past decades, nearly 50% of global internet data flows through edge servers [1], posing a huge burden to communication bandwidth, yet their computational capabilities remain underutilized—typical CPU utilization is often below 10%. The deprecation of edge computing is a big waste, because over 60% of the carbon emissions from the computing comes from the production of devices. We propose the first federated LLM pre-training framework targeting decentralized edge devices to leverage those 'junkyard' computing resources. Our system is implemented on Photon [3], a flexible FL pre-training architecture. The edge training recipe is written in raw C for efficiency and broad hardware compatibility.

We evaluate our system on commercial edge chipsets. To handle memory constraints, momentum computation for gradient updating is offloaded to the server-side aggregator, with a local batch size of 1 and a maximum sequence length of 128, reducing peak memory usage to 1.3 GB.

Experimental results show that: (1) Optimal local step selection is critical. communication dominates training time, contributing over 93%. Increasing the number of local steps from 1 to 8 reduces total training time by up to 3×. However, while more local steps reduce communication cost, they may cause the per round training to diverge, requiring more steps to reach the same loss and thereby offsetting the communication savings. (2) Computing cost can be amortized losslessly. Although occasional spikes appear—indicating efforts to synchronize client updates—the overall training performance using 50 clients per round is nearly equivalent to sequential training with a single client. This demonstrates that a full batch can be processed in parallel by distributing it across many clients without loss in training performance. (3) Mobile pre-training is feasible through continual parallelism. We estimate the performance of mobile pre-training at potentially larger client scales. Unlike centralized training, where parallelism gains saturate beyond a critical batch size, mobile pre-training can continue to benefit from parallelism due to the large number of available clients. We anticipate that with thousands of mobile devices, it is possible to achieve performance comparable to centralized pre-training.

This design enables lightweight deployment and offers the potential for sustainable model pre-training. For example, increasing CPU usage from 0% to 80% on an NVIDIA Jetson board raises power consumption by only 18% [2], enabling pre-training with minimal additional energy on edge servers that are already powered on for communication tasks.

References

[1] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. 2020. An overview on edge computing research. *IEEE access* 8 (2020), 85714–85728.

[2] Erol Gelenbe. 2025. Minimizing delay and power consumption at the edge. *Sensors* 25, 2 (2025), 502.

[3] Lorenzo Sani, Alex Iacob, Zeyu Cao, Royson Lee, Bill Marino, Yan Gao, Dongqi Cai, Zexi Li, Wanru Zhao, Xinchi Qiu, et al. 2025. Photon: Federated LLM Pre-Training. *to appear at Proceedings of Machine Learning and Systems (MLSys 25)*.

[4] Jennifer Switzer, Gabriel Marcano, Ryan Kastner, and Pat Pannuto. 2023. Junkyard computing: Repurposing discarded smartphones to minimize carbon. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 400–412.

[5] Mengwei Xu*, Dongqi Cai*, Wangsong Yin*, Shangguang Wang, Xin Jin, and Xuanzhe Liu. 2025. Resource-efficient algorithms and systems of foundation models: A survey. *Comput. Surveys* 57, 5 (2025), 1–39.

Author's Contact Information: dc912@cam.ac.uk.