

Peer Review

Review of: "One Communication Round is All It Needs for Federated Fine-Tuning Foundation Models"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors report that federated fine-tuning of foundation models does not need many communication rounds. If each client trains for the same total number of local epochs, a single aggregation (one-shot FedAvg) gives a global model close to the one produced by multi-round FedAvg, and for the larger models, it is sometimes better. Figure 1 frames this as a size effect: ResNet-18 and an LSTM lose accuracy under one-shot training, while larger models such as GPT-2 and Llama do not. Section 3 then offers an explanation. Theorem 1 (Eq. 9) bounds the gap between one-shot and multi-round updates by $\Gamma \|w^{(0,0)}\|$ with $\Gamma = L\tau Tk$, and Figures 2 and 3, together with Table 4, argue that each factor (smoothness L , relative update size τ , total steps Tk) is smaller for pre-trained models.

The evidence is Table 1 (MMLU and ARC Challenge accuracy for TinyLlama, Gemma-2b, Llama-7b, and Llama-13b, LoRA and full fine-tuning, three training mixtures), Table 2 (MT-bench scores judged by GPT-4), a Stable Diffusion v1.5 DreamBooth run with CLIP scores of 0.3343 (one-shot) and 0.3341 (multi-round), and Section 5, which adds communication savings (3000GB to 1000GB for Llama-13b), asynchronous aggregation (Figure 5), and a privacy argument. I think the core observation is useful and cheap to test. It is also stated in a way practitioners can act on tomorrow.

2. Whether the conclusion holds on the data

Much of the empirical work is genuinely good. The authors hold the total number of local epochs fixed between the two arms, which is the right control. They run four language models across about an order of magnitude of size (1.1b to 13b), two fine-tuning methods, and a deliberately non-IID mixture (M-W, 20

clients). And they report the zero-shot baselines (Table 5) and local-only models (Table 6) that let a reader check the gains. I re-derived every AVG. value in Table 2 from its three columns; all twelve match to two decimals (for example, Llama-13b, LoRA*: $(3.02 + 4.27 + 3.26)/3 = 3.517$, reported 3.52). Table 4's Tk row also checks out (ResNet-18: $50 \times 7812 = 390600$; GPT-2: $5 \times 5625 = 28125$).

Where I think the claim runs ahead of the evidence is the size of the multi-round baseline. Table 3 sets $T = 3$ for every language-model experiment, with $k = 1$ or 2 local epochs. So "multi-round" here means three rounds. The small models, by contrast, are run with many more rounds: Table 4 lists $T = 50$ for ResNet-18 and BERT (the text gives no setting for the LSTM in Figure 1). Theorem 1 itself says the gap grows with T . The "emergent capability of FMs" in the Introduction is therefore partly confounded with the number of rounds: if Figure 1 uses the Table 4 settings, small models were compared against 50 rounds, large ones against 3. I would add one run that breaks the tie, either ResNet-18 at $T = 3$ with matched total epochs or a language model at $T = 10$ or more. Without it, the title's "One Communication Round is All It Needs" is really "one round is about as good as three".

The per-cell results are more mixed than the text says. I counted wins in Table 1. With LoRA, one-shot is higher in 5 of 12 MMLU cells and 8 of 12 ARC cells (13 of 24). With full fine-tuning, one-shot is higher in only 4 of 12 MMLU cells. Gemma-2b is the clearest counterexample under LoRA: it loses in 5 of its 6 cells, and the gaps are the largest among the LoRA rows (MMLU, M-W column, 37.69 vs 35.14; ARC, MMLU column, 43.09 vs 40.61). That model is larger than TinyLlama, whose gaps are all under 1.4 points, so the monotone size story from Figure 3 does not show up cleanly here. I would say so in Section 4 rather than leave the reader to find it.

No cell carries a standard deviation or a second seed. The standard ARC Challenge test split has about 1,170 questions, so one point is about a dozen answers, and several of the claimed wins sit well inside that (Table 1, Gemma-2b, LoRA vs LoRA*, ARC, M-W column: 42.06 vs 42.15). MT-bench scores come from GPT-4 judging, with no repeated judging reported. The CLIP difference of 0.0002 is best read as "no difference," which is fine, but it cannot support much more.

Two checks inside the paper deserve more attention. First, several full fine-tuning cells fall far below the untuned model. Llama-13b zero-shot MMLU is 46.23 (Table 5), while full fine-tuning on M-W gives 32.91 (multi-round) and 29.76 (one-shot) in Table 1. Comparing two damaged models tells us little about aggregation. The same pattern (a model below its own base) explains why the contribution "LoRA outperforms full fine-tuning in the context of one-shot federated fine-tuning" may say more about the learning rate and forgetting than about one-shot training. Second, Table 6 shows that local-only models

already reach 50.05 to 52.33 on ARC; their mean is $(50.79 + 51.02 + 50.05 + 50.43 + 52.33 + 51.22 + 51.28 + 52.30 + 51.21 + 51.11)/10 = 51.17$, against 51.88 for the one-shot global model, and two local models beat it. On an IID split of Wizard, then, a single client does almost as well as the federation. When the task barely needs collaboration, one-shot and multi-round will naturally agree. The non-IID M-W column is the fairer test, and there one-shot LoRA loses on MMLU for Gemma-2b (37.69 vs 35.14) and Llama-7b (35.37 vs 34.84).

Table 2 has a similar issue. The Base column is higher than multi-round LoRA for Gemma-2b (3.60 vs 3.43) and practically equal for Llama-13b (2.69 vs 2.71), so the multi-round baseline barely moved these models. The text also names "Gemma-7b" as one of the larger models where one-shot wins; Table 2 has no Gemma-7b row. For full fine-tuning, the AVG. column favors multi-round for Llama-13b (3.37 vs 3.06) and Gemma-2b (3.09 vs 2.90); only Llama-7b goes the other way (2.85 vs 2.78).

Now the theory. I followed the proof in Appendix D and found three problems. (a) Going from Eq. 11 to Eq. 12, the difference $\|w_i^{(0,j)} - w_i^{(t,j-kt)}\|$ is bounded by $\tau\|w^{(0,0)}\|$, but Assumption 2 only places each point within $\tau\|w^{(0,0)}\|$ of the start, so by the triangle inequality the correct bound is $2\tau\|w^{(0,0)}\|$. (b) Eq. 13 contains the factor m , while Theorem 1 sets $m = 1$; with a single client, one-shot and multi-round FedAvg (global rate 1) produce the same iterates, so ε is identically zero there, and the simplification removes the effect being studied. Keeping m , as the proof already does, would be cleaner. (c) Dropping the learning rate β is not harmless for the paper's own comparison, because LoRA uses $3e-4$ and full fine-tuning $3e-5$ (Appendix B), a tenfold factor in the bound.

The empirical estimates are also weaker than the plots suggest. The L in Figure 2(a) is a single secant ratio between $w^{(0,0)}$ and $w^{(T,k)}$ on one mini-batch, which is a lower bound on the Lipschitz constant, not an estimate of it. Figure 3 then plots $\Gamma\|w^{(0,0)}\|$ and calls it $\|\varepsilon\|$. That is the bound, and since Tk is chosen by the authors (Table 4), the ranking is partly set by design. As rendered on Qeios, the numerical values in Figures 1, 2, 3, 5, and 6 are not legible to me, so I could not check the size of the separations.

Section 5's side claims are fair in spirit but stated too strongly. The communication arithmetic is right: $2mTS$ with $m = 10$, $T = 3$, and $S = 50\text{GB}$ gives 3000GB , and $2mS$ gives 1000GB . Yet the configuration the paper itself favours is LoRA (its third contribution), where only adapters travel, so the 50GB full-model example overstates the saving in the recommended setting. The privacy paragraph says one-shot training can "eliminate the possibility of client-side privacy leakage"; server-side and API-query attacks

remain, and I would narrow the wording to client-side inversion attacks that require repeated global models.

3. Reproducibility

The good part first. Appendix B gives hardware (RTX A6000 and A100), LoRA rank 16, learning rates, batch size 64, and Tables 3 and 4 give rounds and epochs per setting. Datasets and evaluation suites are public. The following I could not find in the Qeios text:

- No code or repository link.
- No seeds and no number of repetitions per cell.
- Model versions are ambiguous. "Llama-7b" and "Llama-13b" are cited to the first LLaMA paper [21] in Section 4, while the Introduction cites Llama 2 [3]. No Hugging Face identifiers or revisions are given for any model.
- The "Wizard dataset" is cited to WizardMath [23]; which Wizard release and how many examples were used is not stated.
- The GPT-4 snapshot used as the MT-bench judge, its temperature, and the judging prompt version are not reported. MT-bench scores shift between judge versions, so this matters.
- How LoRA updates are aggregated (averaging A and B separately, or averaging the product BA) is not said. After several local epochs, these two give different models, and the choice could change the one-shot result.
- The learning-rate schedule is mentioned in Section 4 ("slower local learning rate decay") but never specified; nor is numerical precision (fp16 or bf16).
- Appendix B says full fine-tuning "set the learning rate to 8". I assume this is the batch size, but I could not verify it.
- The version of the evaluation framework [25] and the few-shot setting for MMLU and ARC are missing.

4. What specifically to fix

1. **Figure 1 and Table 4.** Add a control with matched T: ResNet-18 (or BERT) at $T = 3$, or one language model at $T \geq 10$, keeping T_k fixed. This is the single experiment that would separate "model size" from "number of rounds".

2. **Table 1 and Table 2:** run at least three seeds and report mean $\pm$ standard deviation, or give bootstrap intervals over test items. Then mark which differences exceed the noise.
3. **Appendix D, Eq. 12.** Replace τ with 2τ (triangle inequality), keep m in the statement of Theorem 1 as Eq. 13 already does, and either keep β in the bound or explain why it can be dropped when LoRA and full fine-tuning use rates differing by a factor of ten.
4. **Figure 2(a).** Describe the secant ratio as a lower-bound estimate of L , and average it over several mini-batches and point pairs.
5. **Figure 3.** Label the plotted quantity as the bound $\Gamma\|w^{(0,0)}\|$, not $\|\epsilon\|$. Better still, measure the actual distance between the one-shot and multi-round final weights, which you already have, and plot it next to the bound.
6. **Section 4, QA paragraph.** State the win counts honestly (for example, one-shot LoRA higher in 13 of 24 cells, full fine-tuning in 4 of 12 MMLU cells) and discuss Gemma-2b, which loses in 5 of 6 LoRA cells.
7. Section 4, chat paragraph: "Gemma-7b" is not in Table 2; I think Gemma-2b or Llama-7b is meant.
8. **Tables 1 and 5 together.** Flag the full fine-tuning cells that end below zero-shot (e.g., Llama-13b, MMLU, M-W: 32.91 and 29.76 vs 46.23) and either tune the full fine-tuning runs so they improve on the base model or drop the "LoRA outperforms full fine-tuning" contribution.
9. **Table 6.** Discuss what it implies: local-only models average 51.17 against 51.88 for one-shot. A harder, more heterogeneous split (Dirichlet label skew or topic-based splits) would show whether one-shot holds when clients really need each other.
10. **Section 5, communication paragraph.** Give the LoRA-adaptor numbers, since those are what the recommended setting sends.
11. Soften the privacy sentence in Section 5 and the Introduction to what one-shot actually removes.
12. **Internal counts:** Contributions say "six FMs and three tasks"; the Conclusion says "5 different models". Figure 1 evaluates on ARC Easy while Table 1 uses ARC Challenge. The text refers to the Dreambooth images as "Figure 5", but they are Figure 4.
13. Appendix C, Figure 6: the text says the multi-round score "slightly increases towards 3.90 in the 3rd round" for Llama-7b on Wizard, but Table 2 lists 3.27 (LoRA, multi-round) and 4.18 (full, multi-round); 3.90 is the one-shot LoRA* value. Please state which run Figure 6 shows.
14. Release code, configs, and the exact model and dataset identifiers listed in Section 3 of this review.

5. Recommendation

Major revisions. The observation is real, simply tested, and practically useful, and the tables are internally consistent where I could check them. The work needs a round-matched control for the size claim, seed variance for the small differences, a corrected bound in Appendix D, and a narrower reading of the privacy and communication claims. None of this changes the direction of the finding. It changes how far the title can go.

Declarations

Potential competing interests: No potential competing interests to declare.