2024 USENIX Annual Technical Conference

FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences

Mengwei Xu, **Dongqi Cai***, Yaozong Wu, Xiang Li, and Shangguang Wang

Beijing University of Posts and Telecommunications (BUPT)

*Corresponding email: cdq@bupt.edu.cn

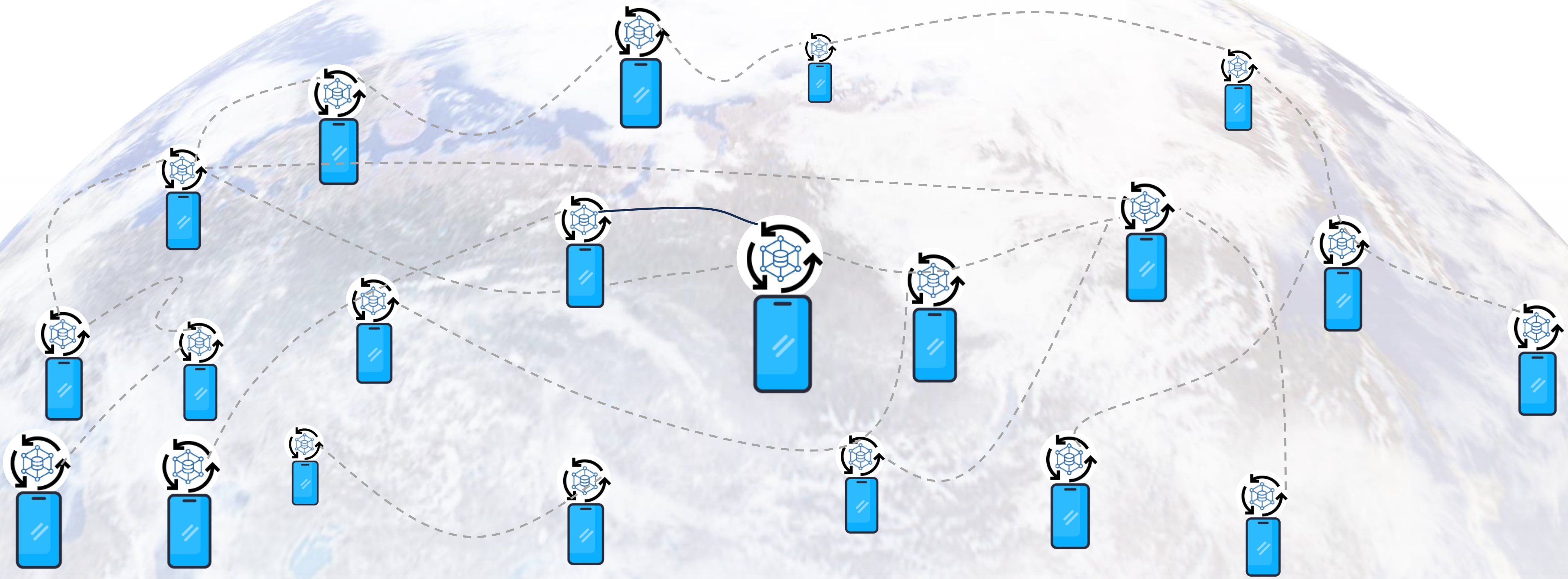
ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

Background: Federated LLM (FedLLM)

[Reference: FlowerLLM, Flower Ltd.]



(1) Democratizing LLMs (2) Stronger LLMs

Motivation: FedLLM unique challenge

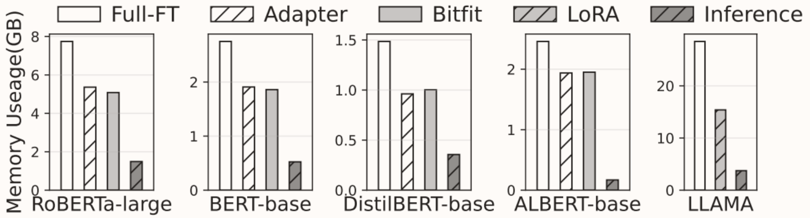


Figure 1: Peak memory footprint of different training methods and inference. Batch size: 8.

- Huge **memory** footprint

Motivation：FedLLM unique challenge

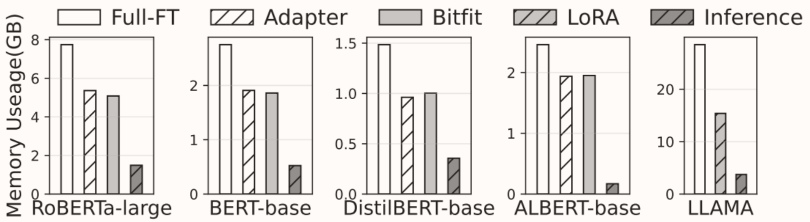
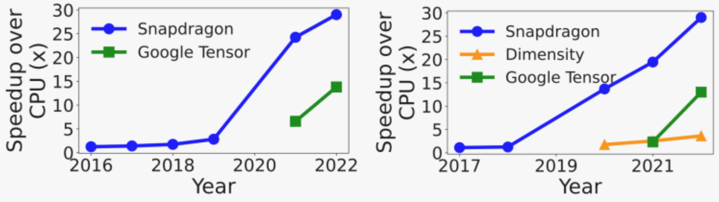


Figure 1: Peak memory footprint of different training methods and inference. Batch size: 8.



(a) ALBERT (b) MobileBERT

Figure 2: Performance evolvement of mobile NPU. Numbers are from AI Benchmark [2].

- Huge **memory** footprint
- Incompatible with mobile **accelerators**

Motivation：FedLLM unique challenge

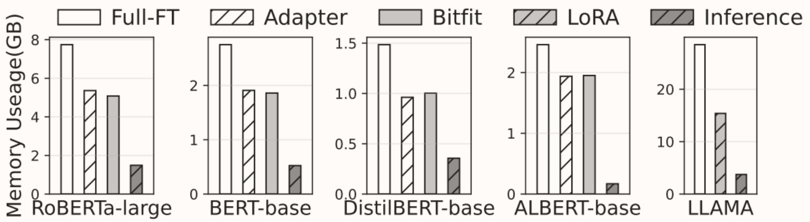


Figure 1: Peak memory footprint of different training methods and inference. Batch size: 8.

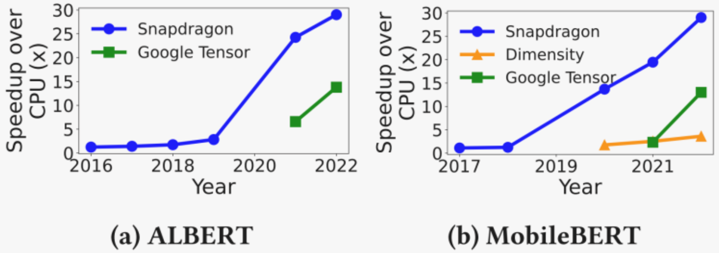


Figure 2: Performance evolvement of mobile NPU. Numbers are from AI Benchmark [2].

- Huge **memory** footprint
- Incompatible with mobile **accelerators**
- Limited device **scalability**

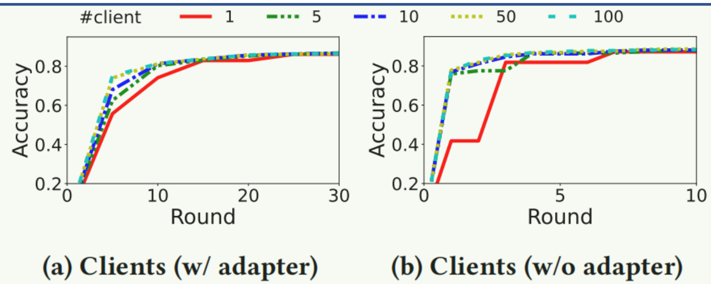


Figure 3: Backpropagation-based FL has low device scalability.

Root: Backpropagation (BP)

They can all be attributed to BP-based gradient computing.

Algorithms	Trainable Parameters	Memory Footprint (GB)			
		Weights	Activations	Gradients	Total
FT-full	354.3M (100%)	1.3	5.1	1.3	7.7
FT-adapter	3.2M (9.0%)	1.3	3.9	0.02	5.2
FT-bitfit	0.3M (0.8%)	1.3	3.8	0.009	5.1
FT-lora	0.8M (2.2%)	1.3	3.8	0.01	5.1
Inference	/	1.3	0.2	0	1.5

Alternatives: BP-free Training

Backpropagation-Free Training

Charles, 1988

Estimation of the mean of a multivariate normal distribution.

Zero-order opt.

1. HSIC
2. BP-free algo.
3. …

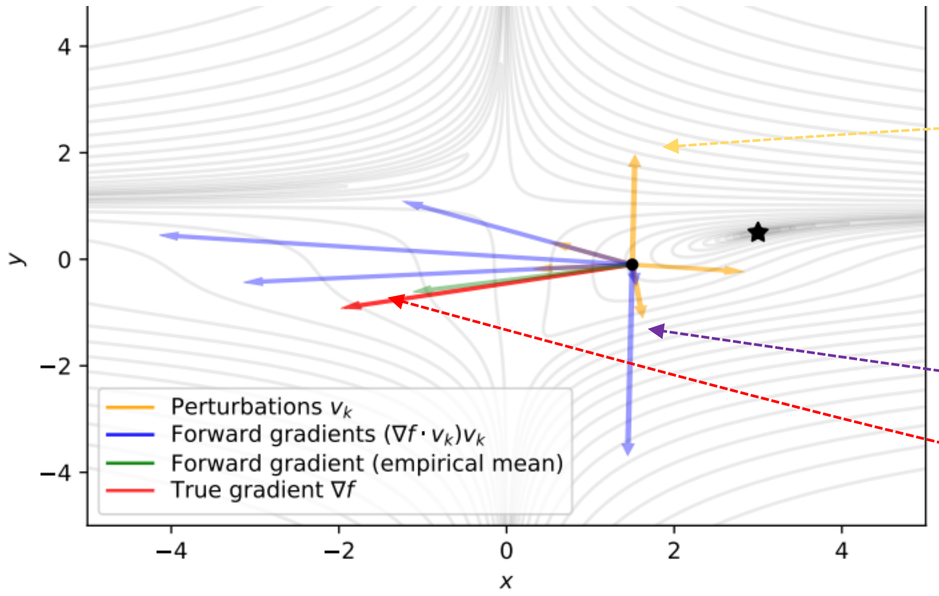
Hinton, 2022

The Forward-Forward Algorithm: Some Preliminary Investigations

Concurrent work

1. Forward gradient
2. BBT (for LLM)
3. Preprint (for FL)

Design: Forward Gradient



Baydin A G, Pearlmutter B A, Syme D, et al.
Gradients without backpropagation

Perturbations

$$\boxed{\nabla_v f(\theta) = \lim_{h\to 0} \frac{f(\theta + h\cdot v) - f(\theta)}{h},}$$

$$g_v(\theta) := \nabla_v f(\theta) v = (\nabla f(\theta) \cdot v)v,$$

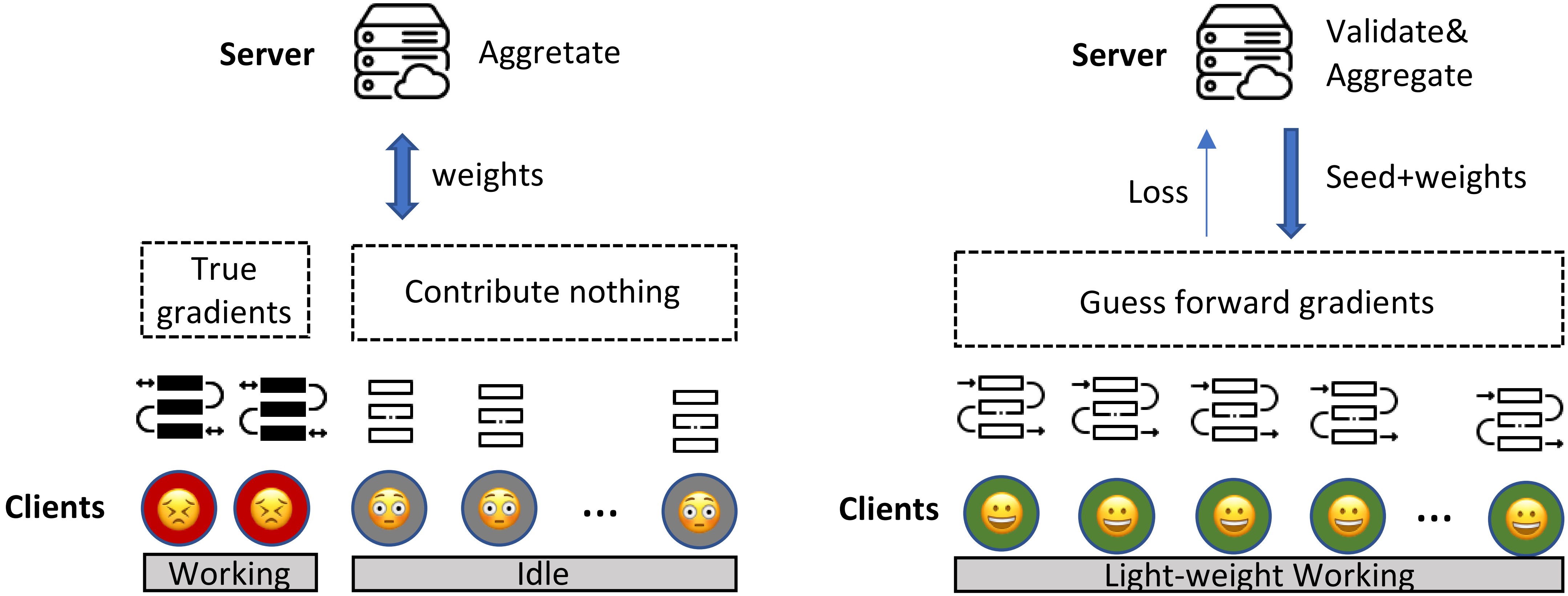
Forward gradients

True (BP-based) gradients

$$\nabla f(\boldsymbol{\theta}) = \left[\frac{\partial f}{\partial \theta_1}, \dots, \frac{\partial f}{\partial \theta_n}\right]^{\mathrm{T}}.$$

- Forward gradient: unbiased estimation of BP-based gradient

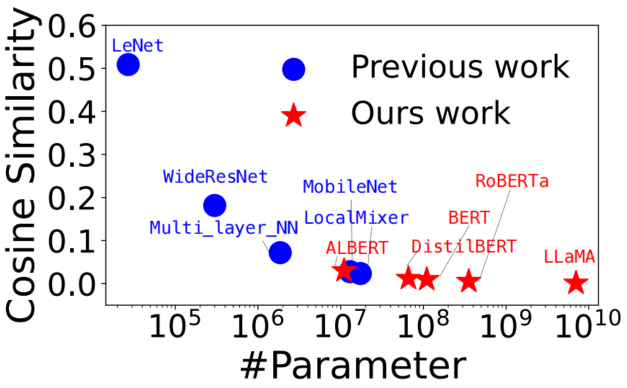
Design: System Overview



(a). Classic Backward-FL

(b). Our Forward-FL

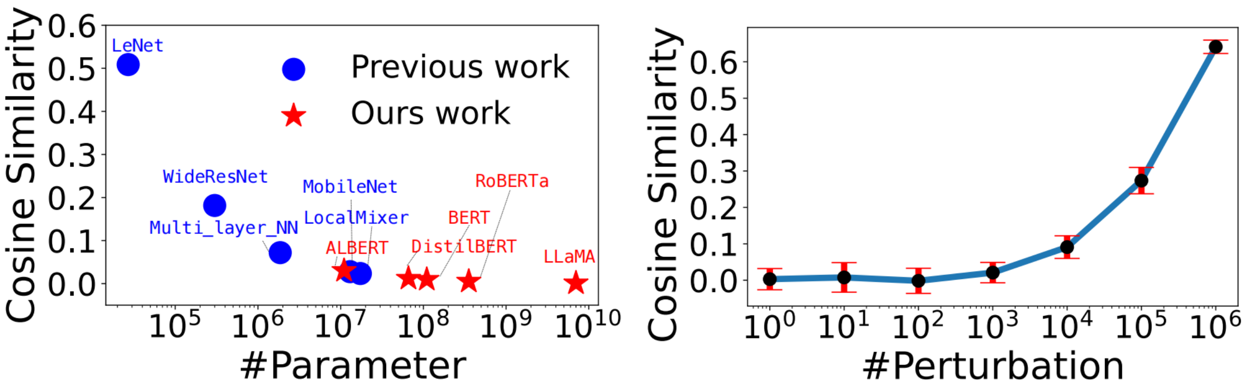
Design #1: Parameter-efficient BP-Free



(a) Effect of model size.

- Previous BP-Free Literatures only apply to tiny models.

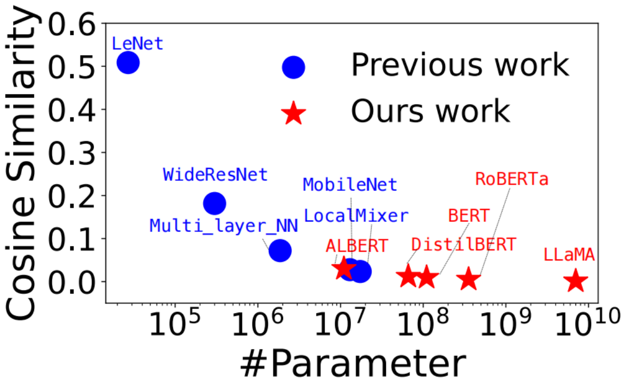
Design #1: Parameter-efficient BP-Free



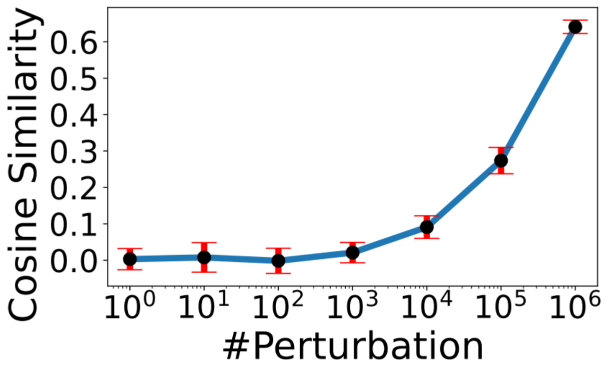
(a) Effect of model size. **(b) Effect of perturbation.**

- Previous BP-Free Literatures only apply to tiny models.
- Reason: Number of perturbations are huge.

Design #1: Parameter-efficient BP-Free



(a) Effect of model size.



(b) Effect of perturbation.

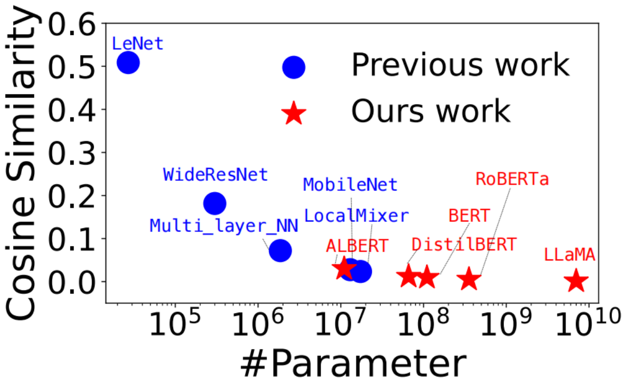
Number of perturbations

Hessian rank of Loss

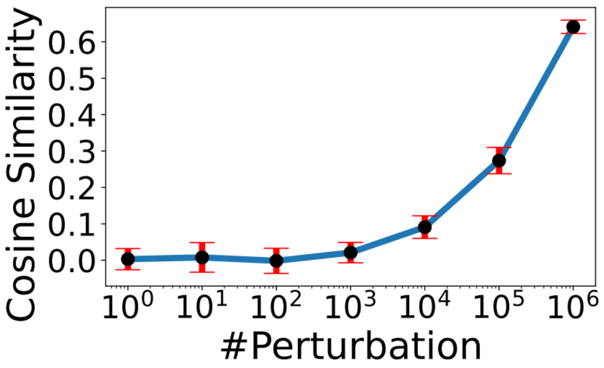
Model size

- Previous BP-Free Literatures only apply to tiny models.
- Reason: Number of perturbations are huge.

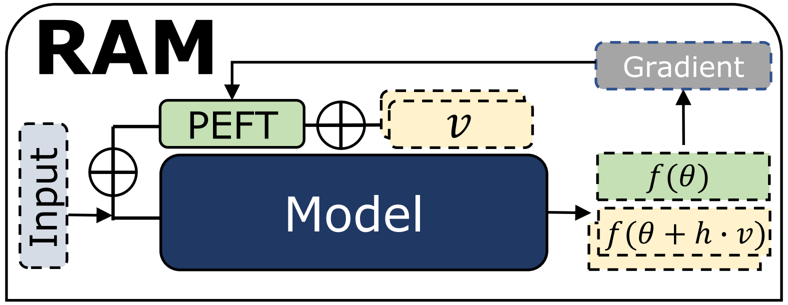
Design #1: Parameter-efficient BP-Free

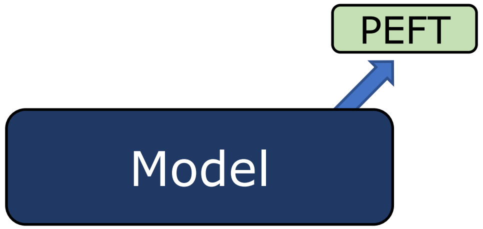


(a) Effect of model size.



(b) Effect of perturbation.





- Previous BP-Free Literatures only apply to tiny models.
- Reason: Number of perturbations are huge.

Design #2: Client Workloads Adaptation

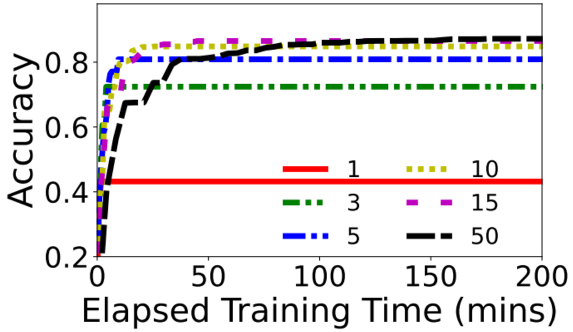


Figure 7: Optimal Global-PS varies across training.

- How many perturbations?

Design #2: Client Workloads Adaptation

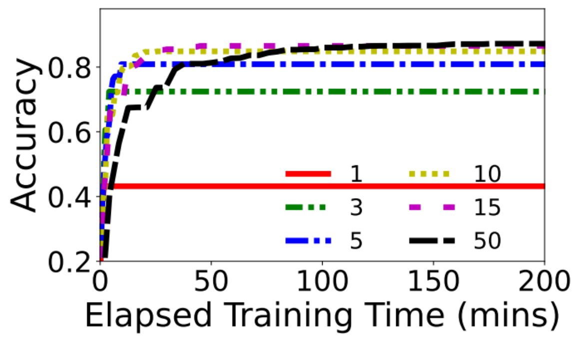


Figure 7: Optimal Global-PS varies across training.

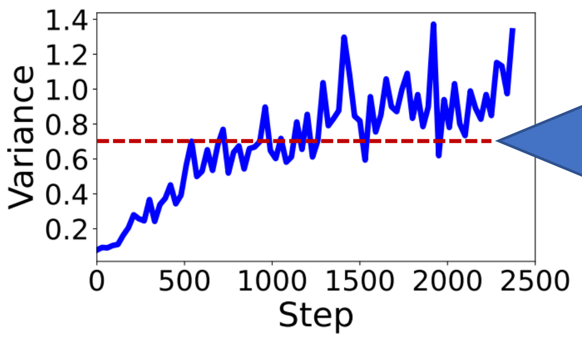
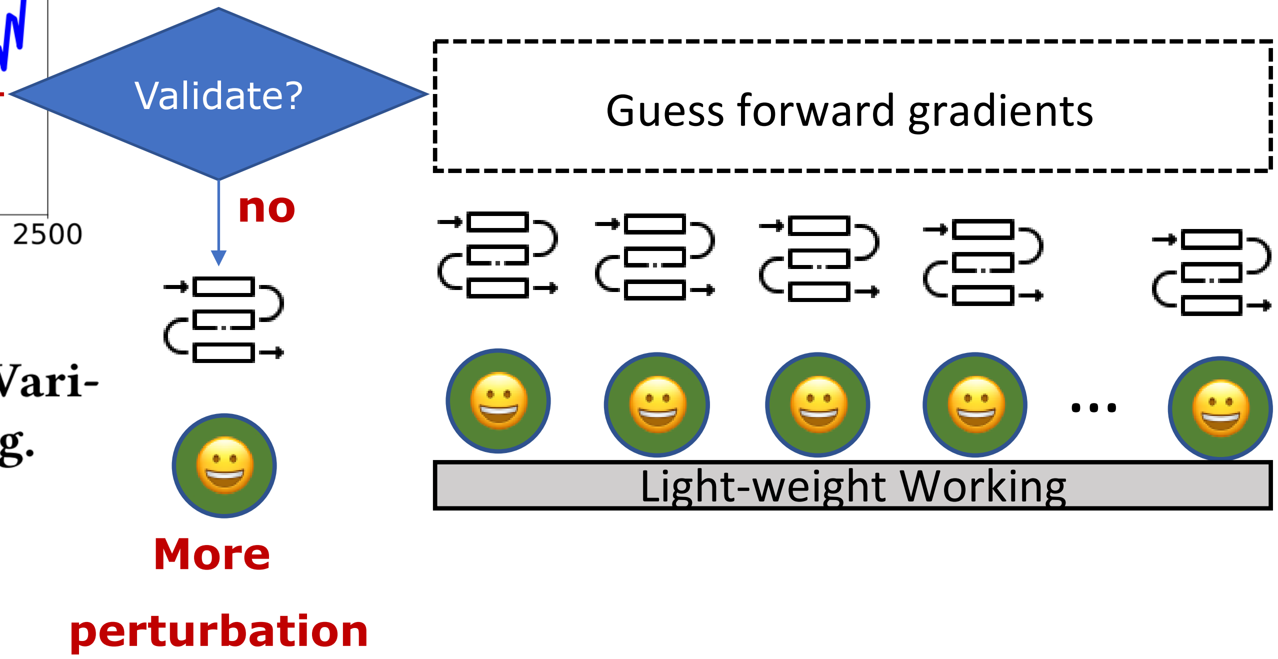
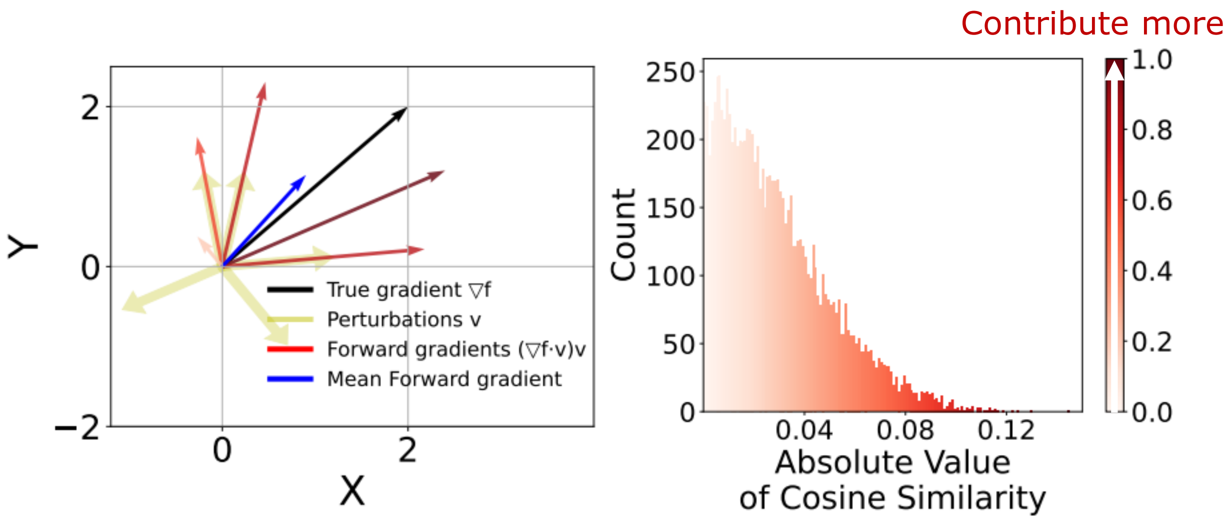


Figure 8: Gradient Variance throught training.



- How many perturbations?
- We decide on the gradient variance.

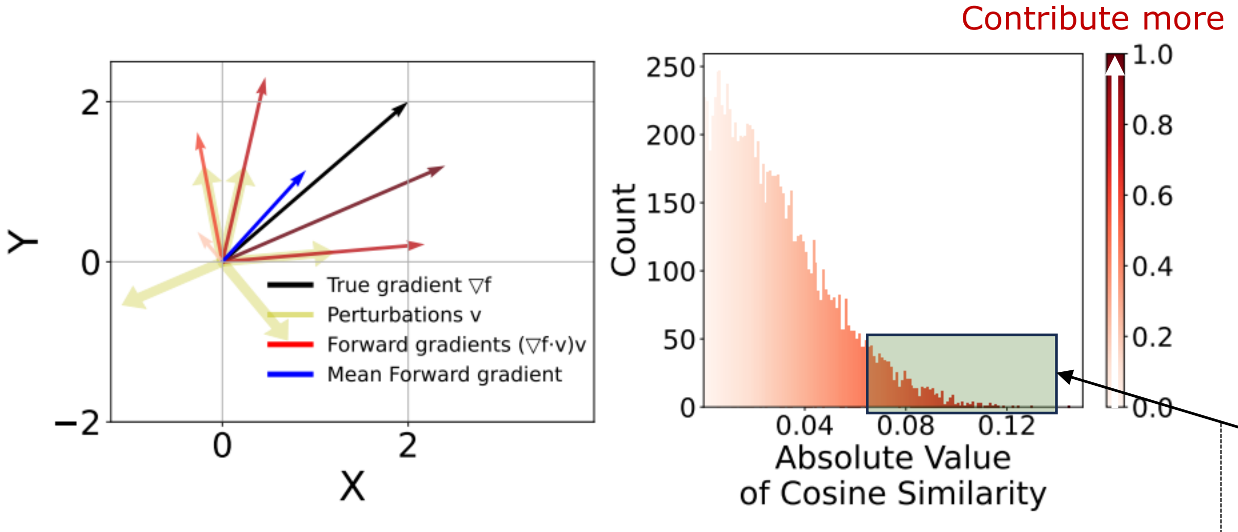
Design #3: Discriminative sampler



(a) Samples (b) Statistics

- Over 60% of computed forward gradients contribute to less than 30% final aggregated gradient.

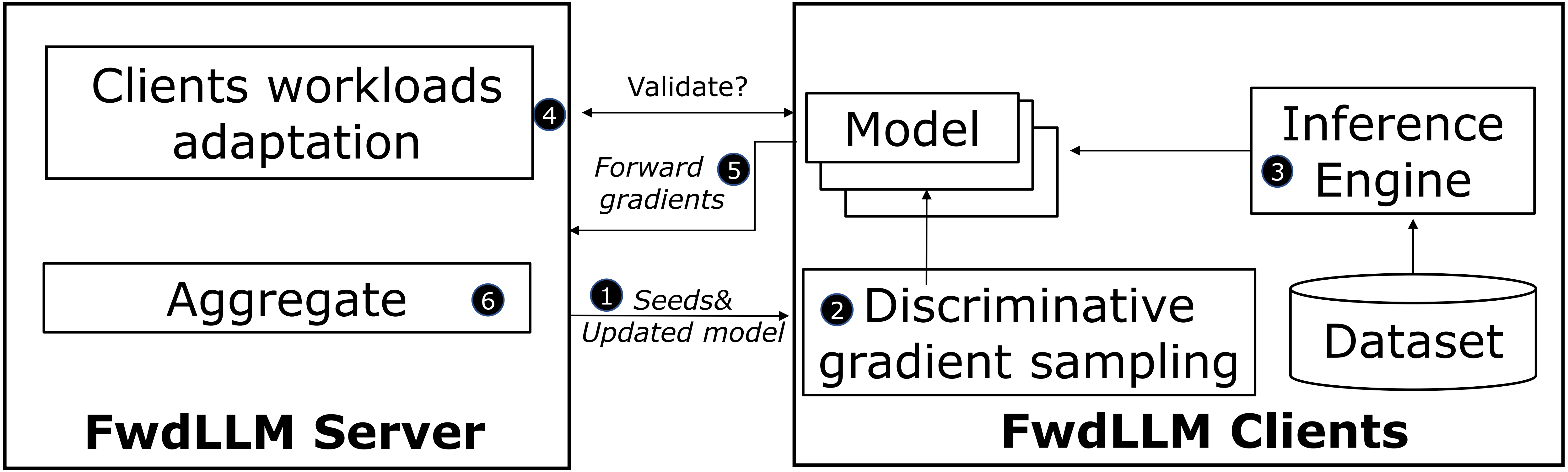
Design #3: Discriminative sampler



(a) Samples (b) Statistics

- Over 60% of computed forward gradients contribute to less than 30% final aggregated gradient.
- We propose to filter out those more valuable perturbations.

Design: Holistic Workflow



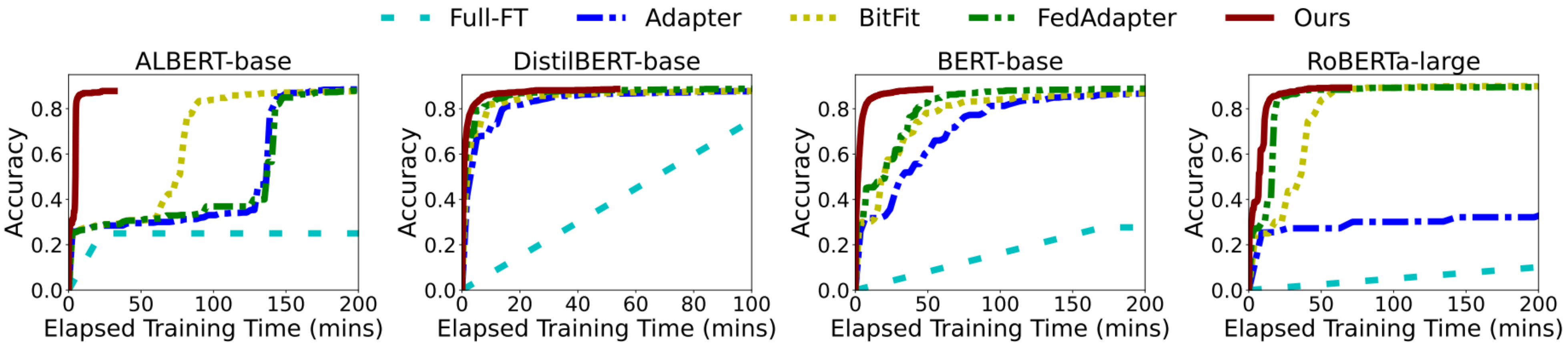
Evaluation: Setup

- Model:

Models	Arch.	Params.	PEFT	Infer. Libs
ALBERT-base [46]	Encoder-only	12M	BitFit	TFLite [5]
DistilBERT-base [77]	Encoder-only	66M	Adapter	TFLite [5]
BERT-base [27]	Encoder-only	110M	Bitfit	TFLite [5]
RoBERTa-large [63]	Encoder-only	340M	Bitfit	TFLite [5]
LLaMA [85]	Decoder-only	7B	LoRA	llama.cpp [6]

- Dataset:
  - Discriminative (YAHOO, AGNEWS, YELP-P)
  - Generative (SQUAD)
- Baselines:
  - Vanilla Backpropagation-based Federated LLM Fine-tuning (Full-FT)
  - Parameter-efficient FedLLM Fine-tuning (Adapter, BitFit, LoRA)
  - Optimized Parameter-efficient FedLLM Fine-tuning (FedAdapter)

Evaluation: End-to-end Performance

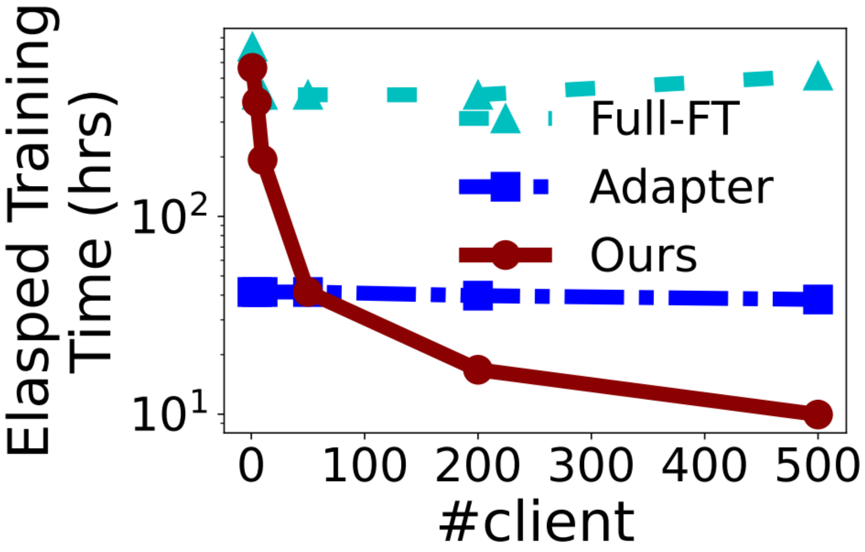


FwdLLM achieves **significant** improvements with mobile **NPU**. (**up to 132x**)

Convergence Time (mins)	ALBERT-base			DistilBERT-base			BERT-base			RoBERTa-large		
	AGNEWS	YAHOO	YELP-P	AGNEWS	YAHOO	YELP-P	AGNEWS	YAHOO	YELP-P	AGNEWS	YAHOO	YELP-P
Full-FT	4598.3	1076.0	5871.3	721.0	651.4	892.7	1535.2	1090.9	2217.4	3833.6	Err	Err
Adapter	168.3	509.9	948.3	84.7	115.3	119.6	250.1	311.8	370.8	860.0	132.7	1319.3
Adapter (FedAvg)	1325.6	2147.9	1119.6	136.9	485.7	141.2	595.2	1718.6	704.6	298.1	1067.0	410.4
Bitfit	174.8	350.5	367.0	76.4	134.8	116.7	272.8	366.3	307.2	58.9	131.4	196.3
FedAdapter	187.8	303.1	293.2	29.5	59.9	52.5	89.5	176.2	212.7	27.0	45.9	123.1
Ours (CPU)	227.1	315.9	271.6	61.5	110.5	92.2	200.7	462.7	242.8	194.3	277.3	95.3
Ours (GPU)	53.2	73.0	63.5	28.1	32.5	42.0	31.1	57.5	37.5	49.1	60.4	24.1
Ours (NPU)	22.7	30.4	27.0	21.9	18.1	32.7	27.6	49.0	33.2	28.9	30.1	14.1

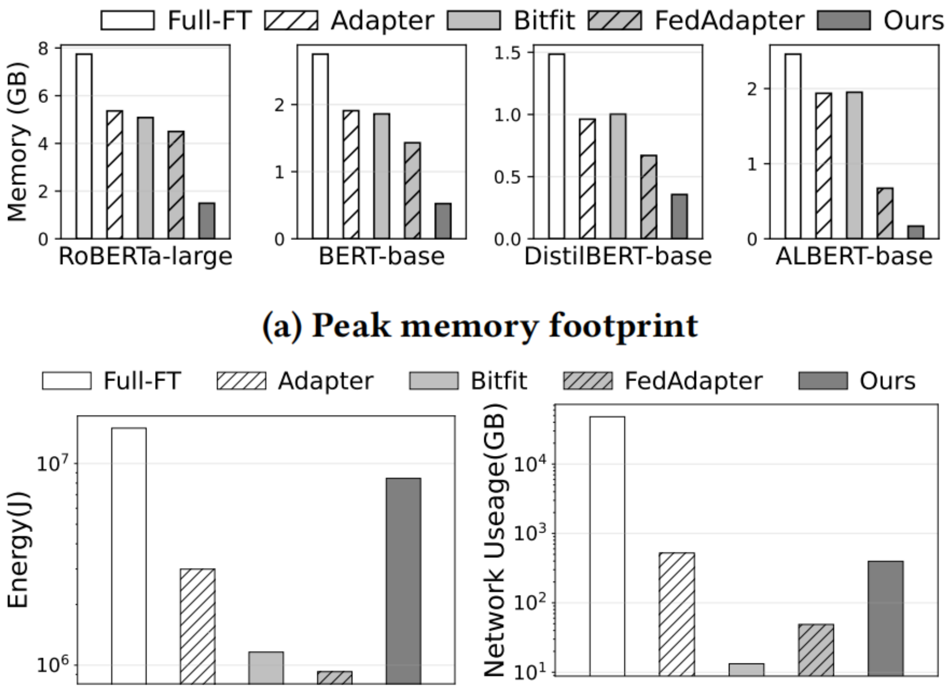
FwdLLM is **versatile** across different processors and hardware boards. (**GPU**: **92x**; **CPU**: **21x**)

Evaluation: Different Client Number



- **50 clients** are enough to surpass BP-based methods.
- **More clients** increase the convergence speed continuously.

Evaluation: System Cost



(a) Peak memory footprint

(b) Total energy cost

(c) Total network cost

- Up to 93% **memory** reduction
- Higher **energy** cost than PEFT

(100 times more client involved)

Evaluation: Extended to LLaMA

Instruction input：
Context:
Bethencourt took the title of King of the Canary Islands, as vassal to Henry III of Castile. In 1418, Jean's nephew Maciot de Bethencourt sold the rights to the islands to Enrique Pérez de Guzmán, 2nd Count de Niebla.

Question:
Who sold the rights?

Answer:

Llama-7B-original: Jean de Bethencourt sold the rights to the islands to Enrique Pérez de Guzmán, 2nd Count de Niebla.
Llama-7B-tuned(backward): Maciot de Bethencourt
Llama-7B-tuned(forward): Jean's nephew Maciot de Bethencourt

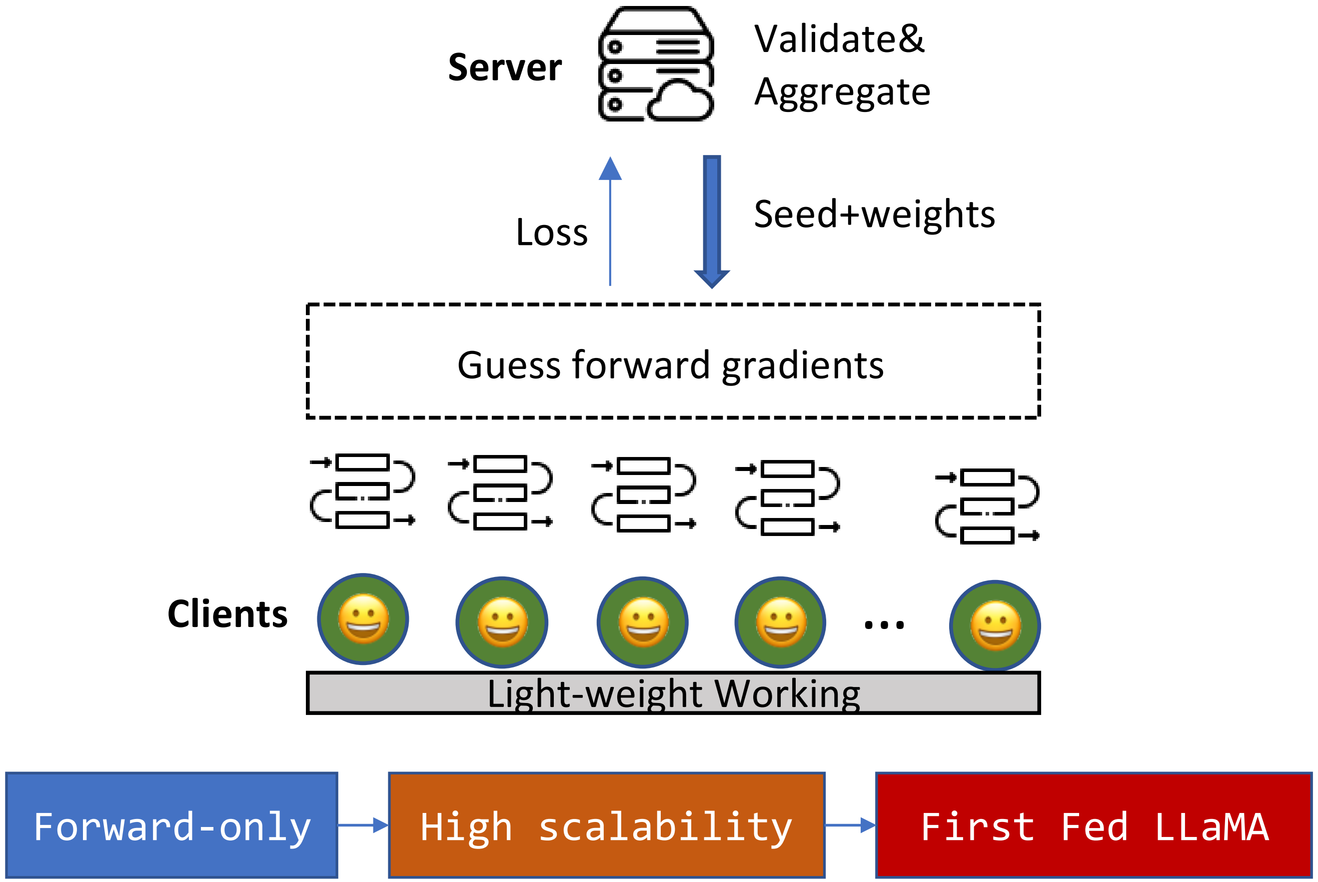
Ground Ture: Maciot de Bethencourt

Methods	Mem. (GB)	Centralized Training (A100)			Federated Learning		
		Acc.	Round	Time	Acc.	Round	Time
BP, FP16	39.2	89.7	500	0.1 hrs	N/A due to memory inefficiency on Pixel 7 Pro (8GB)		
BP, INT8	32.4	88.6	500	0.06 hrs			
BP, INT4	28.5	87.8	500	0.04 hrs			
Ours, FP16	15.6	87.0	240	1.5 hrs			
Ours, INT8	7.9	86.9	260	0.8 hrs			
Ours (CPU), INT4 Ours (NPU*), INT4	4.0	85.8	130	0.25 hrs	85.8	130	0.19 hrs 0.07 hrs

- First implemented **billion-sized FedLLM** fine-tuning on **mobile phones** (CPU).
- Similar performance to BP-based baselines.
- **(Vision)** with NPU, FwdLLM converges with the same speed as central training.

Conclusion

- FedLLM
- FwdLLM: the First Forward-only FedLLM
  - Memory Efficient
  - NPU Friendly
  - High Scalability
- Beyond LLaMA-7B
  - More Models?
  - Mobile Applications?



Thanks for your listening!

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

mllm

mllm-NPU

Contact: Dongqi Cai (cdq@bupt.edu.cn)