香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

ICLR International Conference On Learning Representations

Developmental Federated Tuning: A Cognitive-Inspired Paradigm for Efficient LLM Adaptation

Yebo Wu [1 *] , Jingguang Li [1 *] , Zhijiang Guo [2 3 †] , Li Li [1 †]

University of Macau [1] HKUST [2] HKUST (Guangzhou) [3]

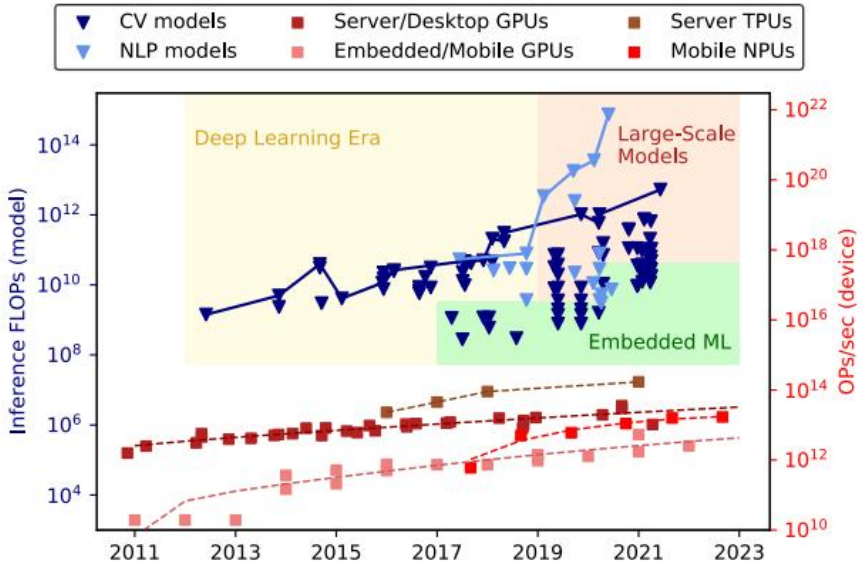
Yebo Wu

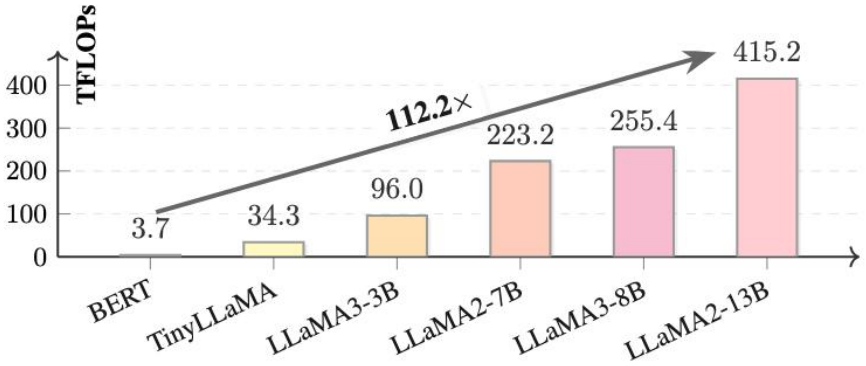
yc37926@um.edu.mo

ICLR 2026

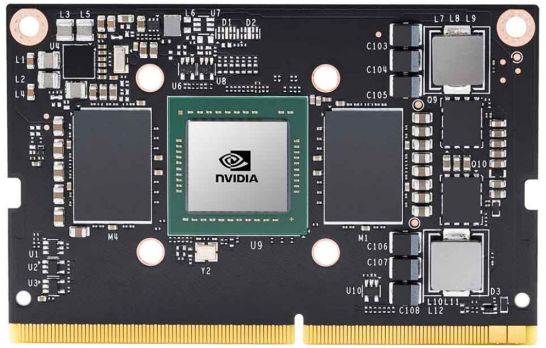
FedLLM's Practical Dilemma

Model Resource Requirements:





Edge Hardware:



NVIDIA® Jetson™ TX2 NX delivers the next step in AI performance for entry-level embedded and edge products. It provides up to 2.5X the performance of Jetson Nano, and shares form-factor and pin compatibility with Jetson Nano and Jetson Xavier™ NX.

This small module packs hardware accelerators for the entire AI pipeline, and NVIDIA JetPack™ SDK provides the tools you need to use them for your application. Custom AI network development is easy with pre-trained AI models from NVIDIA NGC™ and the NVIDIA TAO Toolkit, and containerized deployments make updates for your product flexible and seamless.

Ease of development and speed of deployment—plus a unique combination of form-factor, performance, and power advantage—make Jetson TX2 NX the ideal mass-market AI product platform.

Order Now

Technical Specifications

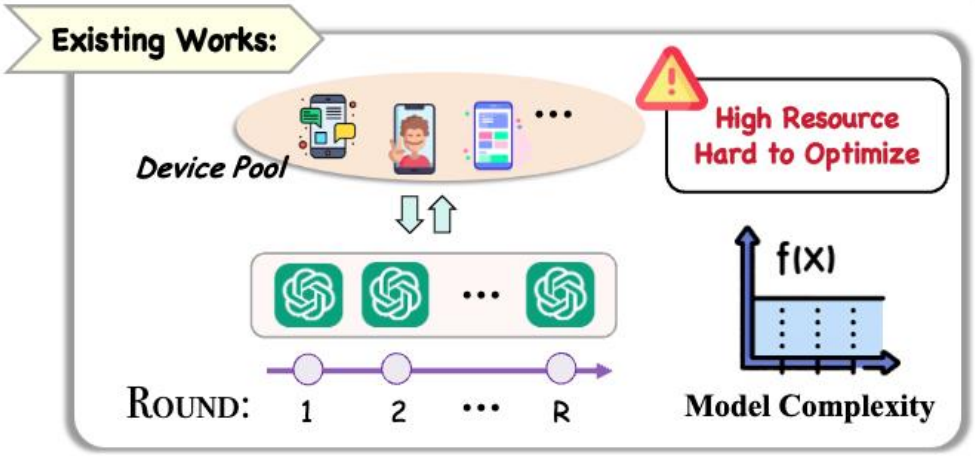
AI Performance	1.33 TFLOPs
GPU	NVIDIA Pascal™ Architecture GPU with 256 CUDA cores
CPU	Dual-core NVIDIA Denver 2 64-bit CPU and quad-core ARM A57 Complex
Memory	4GB 128-bit LPDDR4, 1600 MHz - 51.2 GBs
Storage	16GB eMMC 5.1 Flash Storage

Conclusion:

Efficient fine-tuning of LLM is necessary for edge devices.

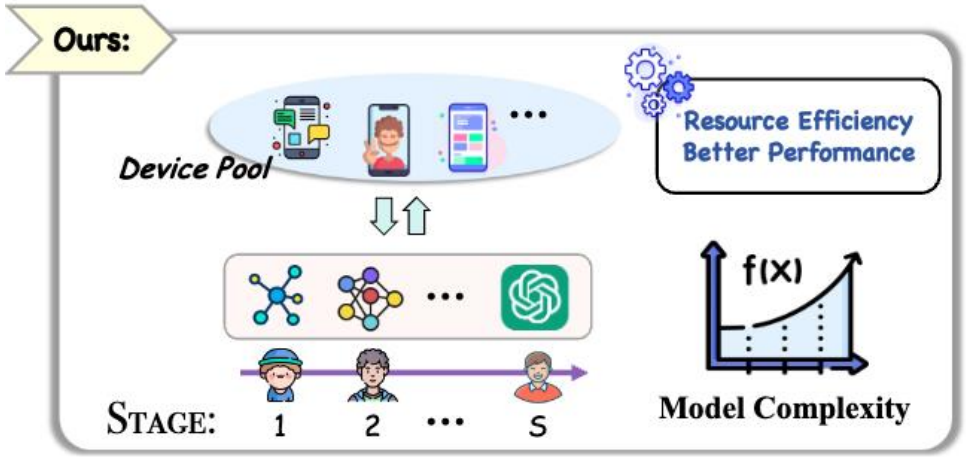
Existing Work vs. Our Methodology

Existing Work:



Continuously training the full model.

Developmental Federated Tuning (DevFT):



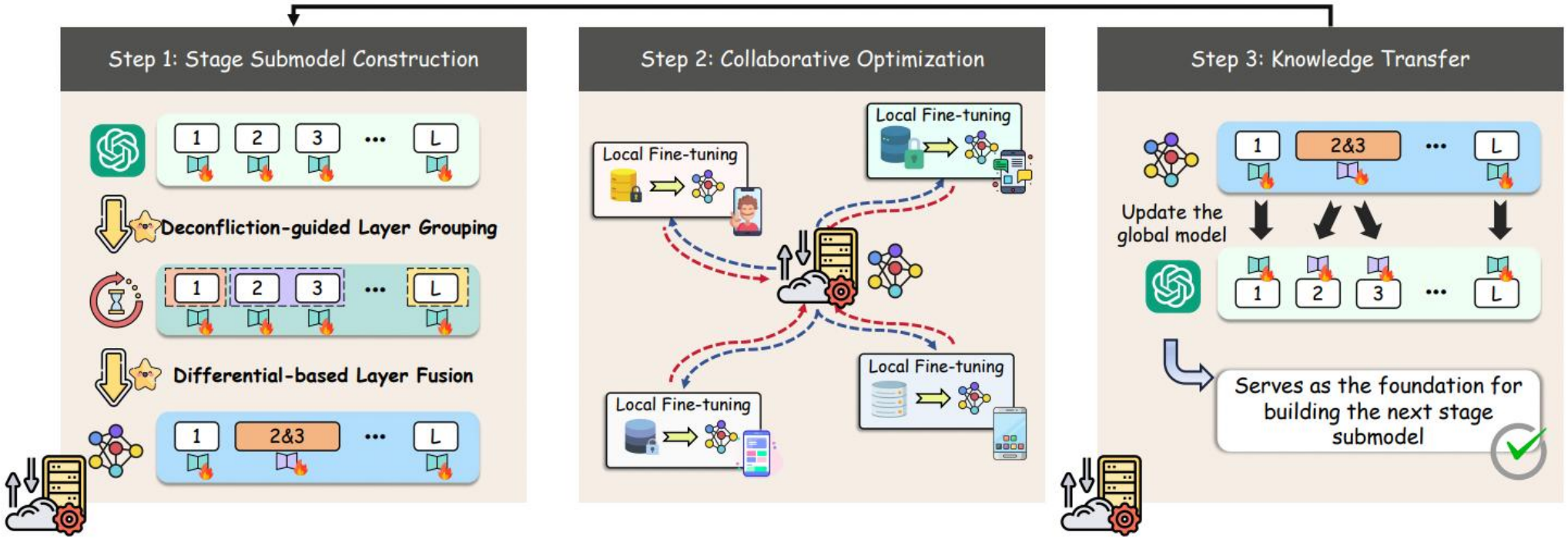
Progressively expanding the model capacity.

a compact submodel (child)–> expand model capacity (growth)–> full model (adult)

Key Challenge:

How to architect stage-specific submodels to ensure effective knowledge transfer across consecutive stages while optimizing overall performance?

Our Solution: DevFT

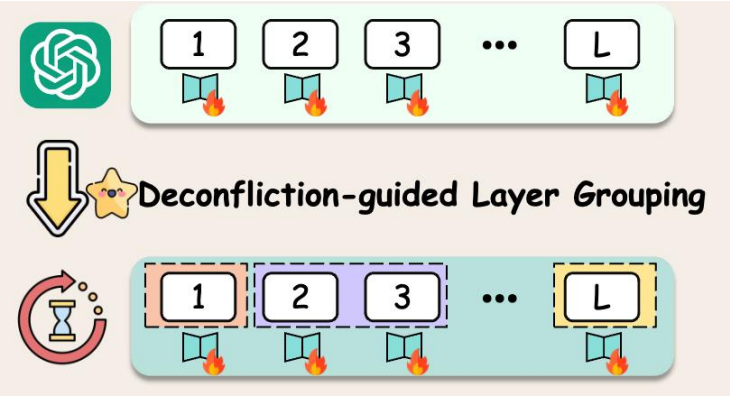


Overview:

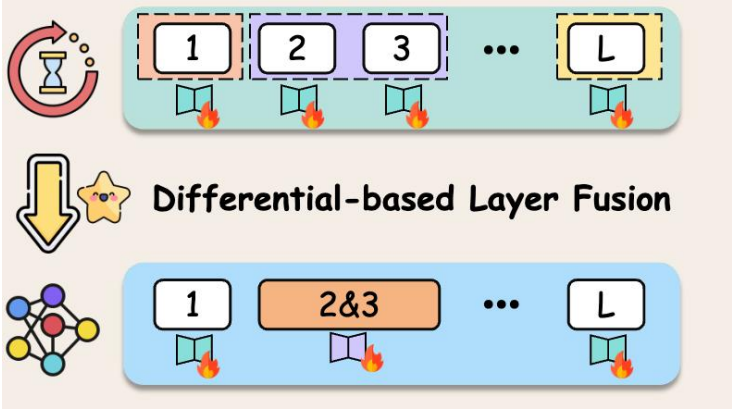
1): The server first constructs the stage–specific submodel;
2): Collaborative optimization across edge devices;
3): The acquired knowledge is employed to update the global model.

Key Components:

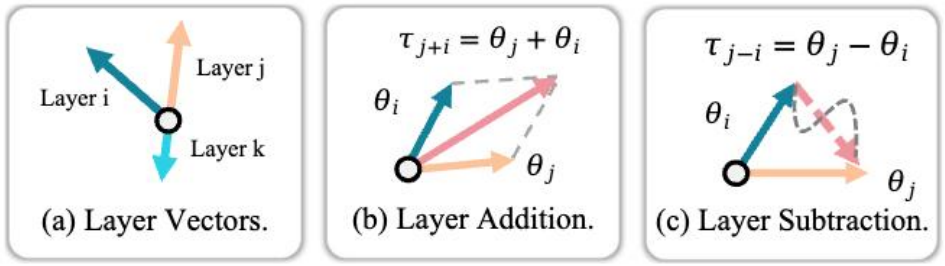
1) Deconfliction–Guided Layer Grouping



2) Differential–Based Layer Fusion

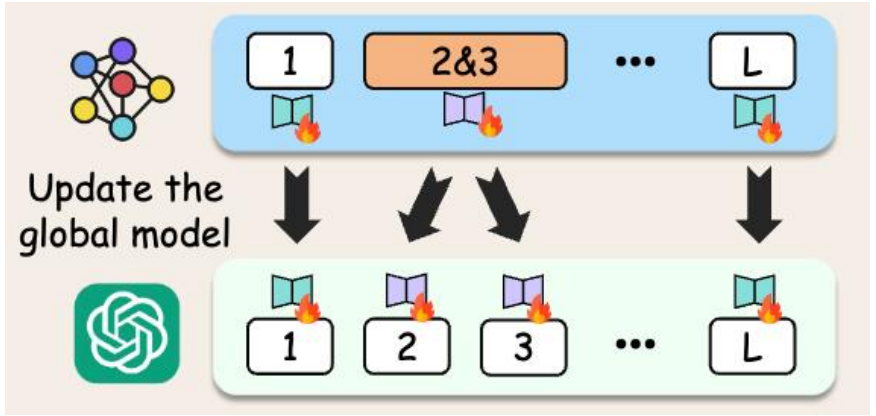


$$\vartheta^{g_n} = \theta_{anchor} + \beta \sum_{j \in g_n} (\theta_j - \theta_{anchor}),$$



3. Knowledge Transfer

Representative Layer → Layers within same groups

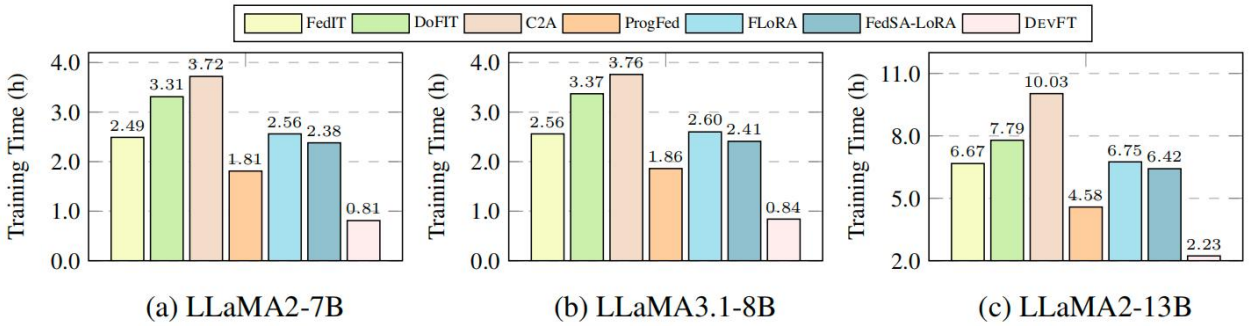


Evaluations:

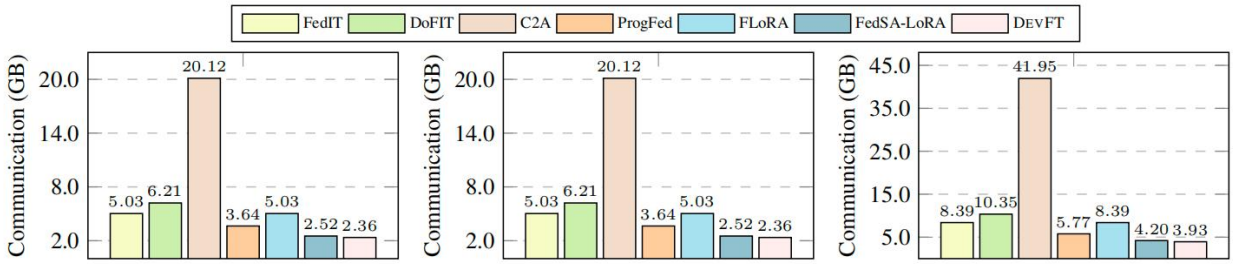
Overall Performance:

Method	Close-Ended Benchmark ↑					Open-Ended Benchmark ↑			
	TruthfulQA	MMLU	IFEval	BBH	**Average**	Vicuna	MT-1	MT-2	**Average**
LLaMA2-7B (INT4) (Touvron et al., 2023)									
FedIT	47.57	42.45	31.76	39.28	40.27	8.18	4.77	1.98	4.98
DoFIT	48.32	43.04	32.62	39.59	40.89	8.19	4.92	2.13	5.08
C2A	46.71	41.83	29.45	36.07	38.52	7.66	3.97	1.88	4.50
ProgFed	48.60	43.14	32.54	39.73	41.00	8.20	4.88	2.19	5.09
FLoRA	47.76	42.64	32.08	39.25	40.43	8.21	4.85	2.02	5.03
FedSA-LoRA	48.24	42.91	32.71	39.36	40.81	8.26	5.09	2.31	5.22
DevFT	**50.28**	**44.15**	**33.97**	**40.93**	**42.33**	**8.41**	**5.76**	**2.92**	**5.70**
LLaMA3.1-8B (INT4) (Grattafiori et al., 2024)									
FedIT	48.07	63.31	47.32	62.69	55.35	8.89	6.54	5.03	6.82
DoFIT	49.12	65.17	51.66	65.21	57.79	9.01	6.72	5.22	6.98
C2A	48.99	63.76	46.10	61.85	55.18	8.74	6.67	4.98	6.80
ProgFed	53.12	66.77	54.55	66.03	60.12	9.07	6.85	5.08	7.00
FLoRA	50.23	64.95	50.47	64.93	57.65	8.96	6.75	5.11	6.94
FedSA-LoRA	53.29	66.87	56.17	67.56	60.97	9.03	6.92	5.41	7.12
DevFT	**55.23**	**68.42**	**62.29**	**71.04**	**64.25**	**9.18**	**7.63**	**6.57**	**7.79**
LLaMA2-13B (INT4) (Touvron et al., 2023)									
FedIT	52.40	55.45	40.33	46.14	48.58	8.37	5.17	3.01	5.52
DoFIT	54.77	56.09	41.68	46.41	49.74	8.37	5.19	3.34	5.63
C2A	53.91	54.33	38.96	45.06	48.07	8.05	5.08	3.26	5.46
ProgFed	55.01	57.38	42.13	46.36	50.22	8.38	5.28	3.07	5.58
FLoRA	54.26	56.23	41.49	46.32	49.58	8.40	5.22	3.15	5.59
FedSA-LoRA	55.73	57.51	43.21	46.91	50.84	8.49	5.39	3.45	5.78
DevFT	**57.19**	**58.74**	**46.45**	**48.70**	**52.77**	**8.67**	**6.18**	**4.52**	**6.46**

Efficiency Evaluation:



(a) LLaMA2-7B (b) LLaMA3.1-8B (c) LLaMA2-13B



(a) LLaMA2-7B (b) LLaMA3.1-8B (c) LLaMA2-13B

Evaluations:

Ablation Study:

Method	Close-Ended Benchmark ↑				
	TruthfulQA	MMLU	IFEval	BBH	Average
LLaMA2-7B (INT4) (Touvron et al., 2023)					
DGLG	**50.28**	**44.15**	**33.97**	**40.93**	**42.33**
RANDOM	47.89	42.09	29.18	38.45	39.90 (↓ 2.43)
EVEN	45.41	39.83	25.04	36.73	36.25 (↓ 6.08)
LLaMA3.1-8B (INT4) (Touvron et al., 2023)					
DGLG	**55.23**	**68.42**	**62.29**	**71.04**	**64.25**
RANDOM	51.02	66.74	54.89	70.11	60.69 (↓ 3.56)
EVEN	48.51	62.50	50.01	70.03	57.76 (↓ 6.49)

Method	Close-Ended Benchmark ↑				
	TruthfulQA	MMLU	IFEval	BBH	Average
LLaMA2-7B (INT4) (Touvron et al., 2023)					
DBLF	**50.28**	**44.15**	**33.97**	**40.93**	**42.33**
R-ONE	46.75	40.13	26.38	37.62	37.72 (↓ 4.61)
SUM	48.15	42.91	30.69	39.84	40.90 (↓ 1.43)
LLaMA3.1-8B (INT4) (Touvron et al., 2023)					
DBLF	**55.23**	**68.42**	**62.29**	**71.04**	**64.25**
R-ONE	47.51	57.33	50.21	58.09	53.29 (↓ 10.96)
SUM	52.74	65.18	58.47	68.39	61.20 (↓ 3.05)

Compatibility Analysis:

Method	Close-Ended Benchmark ↑					Resource ↓	
	TruthfulQA	MMLU	IFEval	BBH	Average	Time (h)	Communication (GB)
LLaMA2-7B (INT4) (Touvron et al., 2023)							
FedIT	47.57	42.45	31.76	39.28	40.27	2.49	5.03
FedIT + DEVFT	**49.86**	**43.87**	**33.65**	**40.79**	**42.04** (↑ 1.77)	**0.83** (×3.00)	**2.36** (×2.13)
FedSA-LoRA	48.24	42.91	32.71	39.36	40.81	2.38	2.52
FedSA-LoRA + DEVFT	**50.42**	**44.57**	**40.92**	**41.36**	**44.32** (↑ 3.51)	**0.72** (×3.31)	**1.18** (×2.14)
LLaMA2-13B (INT4) (Touvron et al., 2023)							
FedIT	52.40	55.45	40.33	46.14	48.58	6.67	8.39
FedIT + DEVFT	**56.84**	**58.26**	**45.49**	**48.52**	**52.28** (↑ 3.70)	**2.30** (×2.90)	**3.93** (×2.13)
FedSA-LoRA	55.73	57.51	43.21	46.91	50.84	6.42	4.20
FedSA-LoRA + DEVFT	**57.61**	**59.25**	**47.63**	**49.13**	**53.41** (↑ 2.57)	**2.19** (×2.93)	**1.97** (×2.13)

Initial Submodel Capacity:

Initial Capacity	Close-Ended Benchmark ↑				
	TruthfulQA	MMLU	IFEval	BBH	Average
LLaMA3.1-8B (INT4) (Touvron et al., 2023)					
1	52.45	66.85	56.83	70.12	61.56 (↓ 2.69)
2	53.87	67.31	59.45	70.50	62.78 (↓ 1.47)
4	**55.23**	**68.42**	**62.29**	**71.04**	**64.25**
8	53.21	67.12	58.35	70.65	62.33 (↓ 1.92)
16	51.08	65.89	54.12	70.01	60.28 (↓ 3.97)
32	48.79	64.49	49.75	69.33	58.09 (↓ 6.16)

Model Growth Rate:

Growth Rate	Close-Ended Benchmark ↑				
	TruthfulQA	MMLU	IFEval	BBH	Average
LLaMA2-7B (INT4) (Touvron et al., 2023)					
2	**50.28**	**44.15**	**33.97**	**40.93**	**42.33**
4	47.96	42.56	29.87	38.79	39.80 (↓ 2.53)
8	45.68	40.07	25.63	36.92	37.08 (↓ 5.25)
LLaMA2-13B (INT4) (Touvron et al., 2023)					
2	**57.19**	**58.74**	**46.45**	**48.70**	**52.77**
4	52.23	56.78	34.56	42.29	46.47 (↓ 6.30)
8	48.12	52.33	26.78	37.45	41.17 (↓ 11.6)

Thank You!

Speaker: Yebo Wu