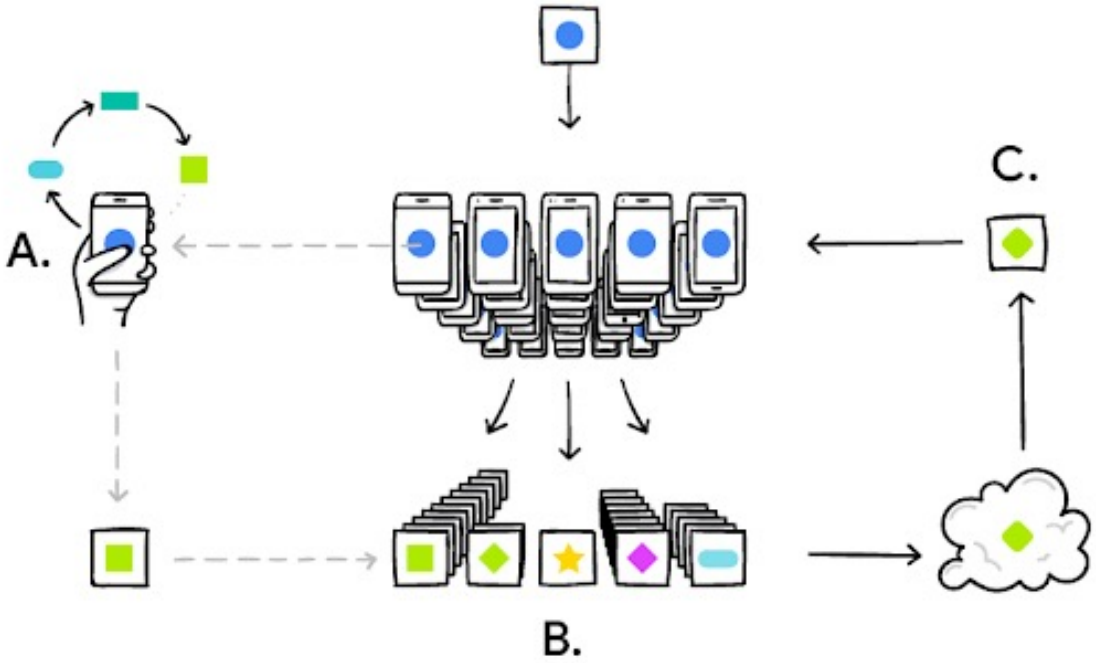
FATE-LLM: Empowering Large Language Models with Federated Learning

Layne Peng, Fangchi Wang

VMware AI Labs, OCTO

Nov. 9, 2023

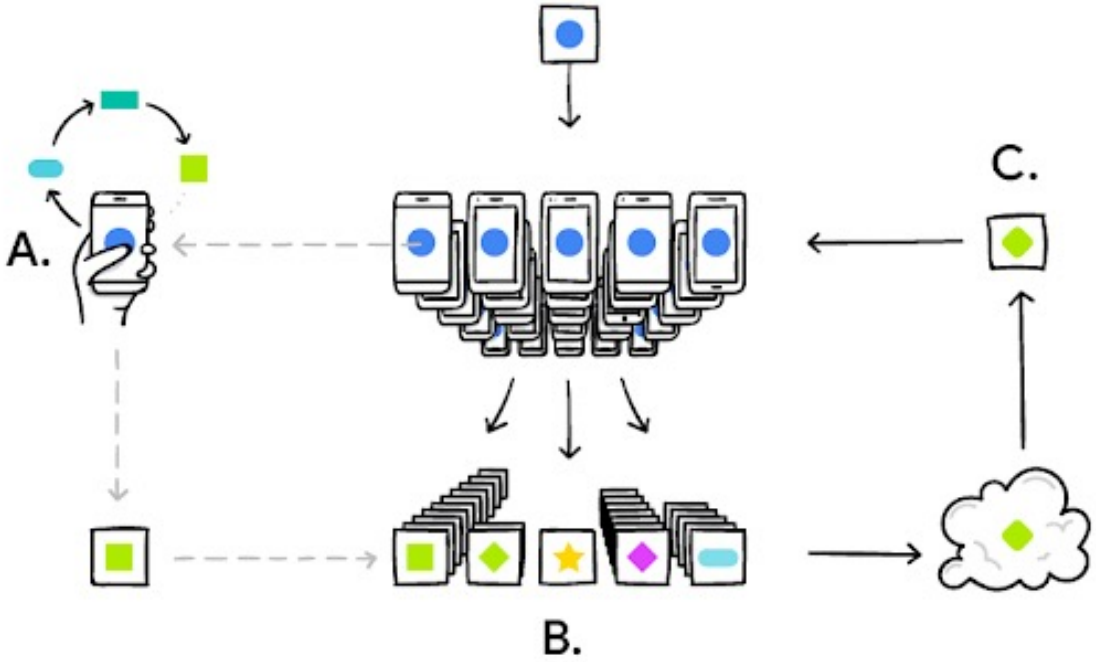
What is Federated Learning?

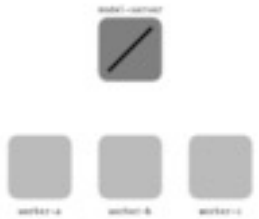
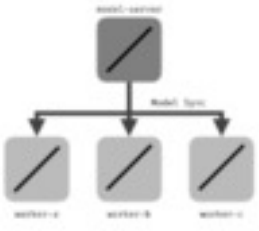
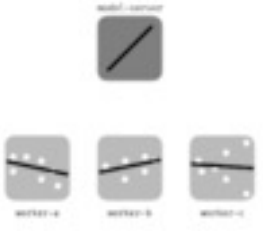
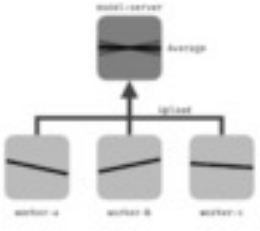


Sources:

1. Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google AI Blog, 2017

What is Federated Learning? (Horizontal Federated Learning)

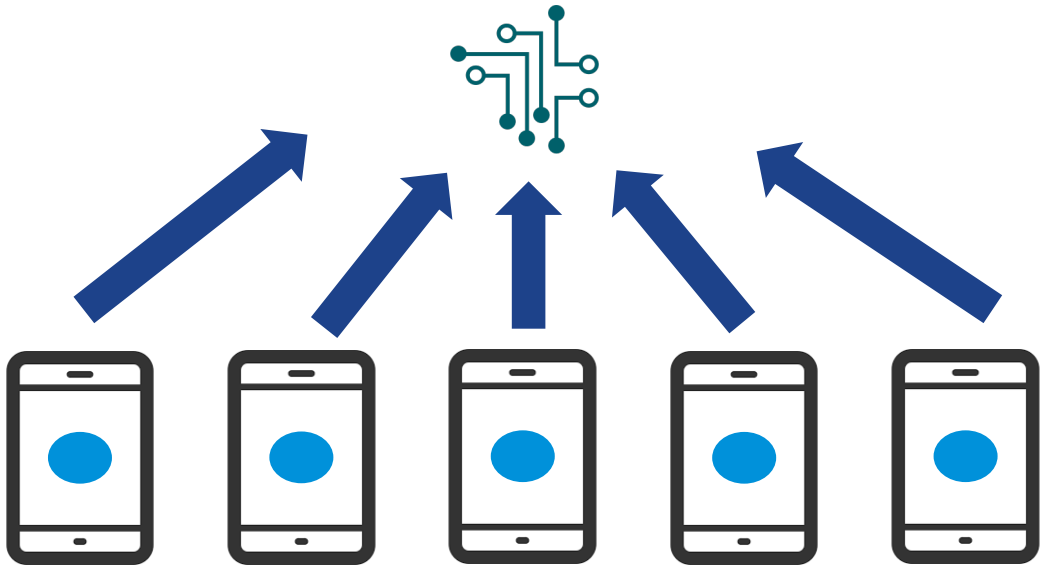


Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data

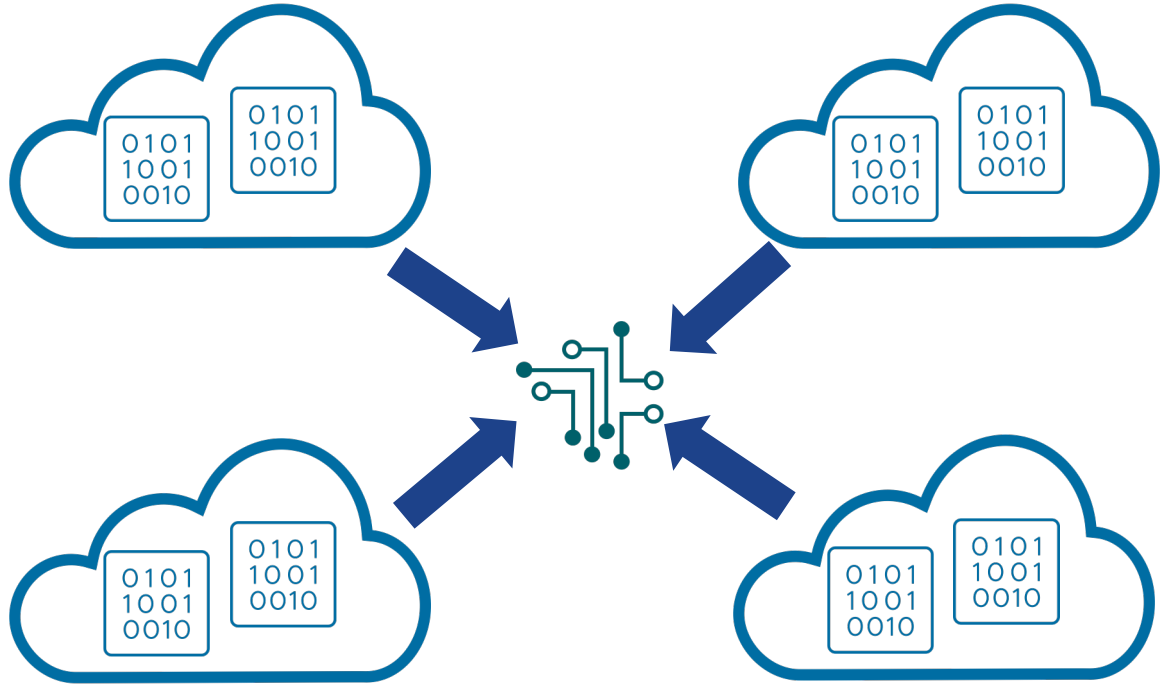
Sources:

1. Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google AI Blog, 2017
2. Federated learning, Wikipedia, URL https://en.wikipedia.org/wiki/Federated_learning）

From device(s) to enterprise(s)

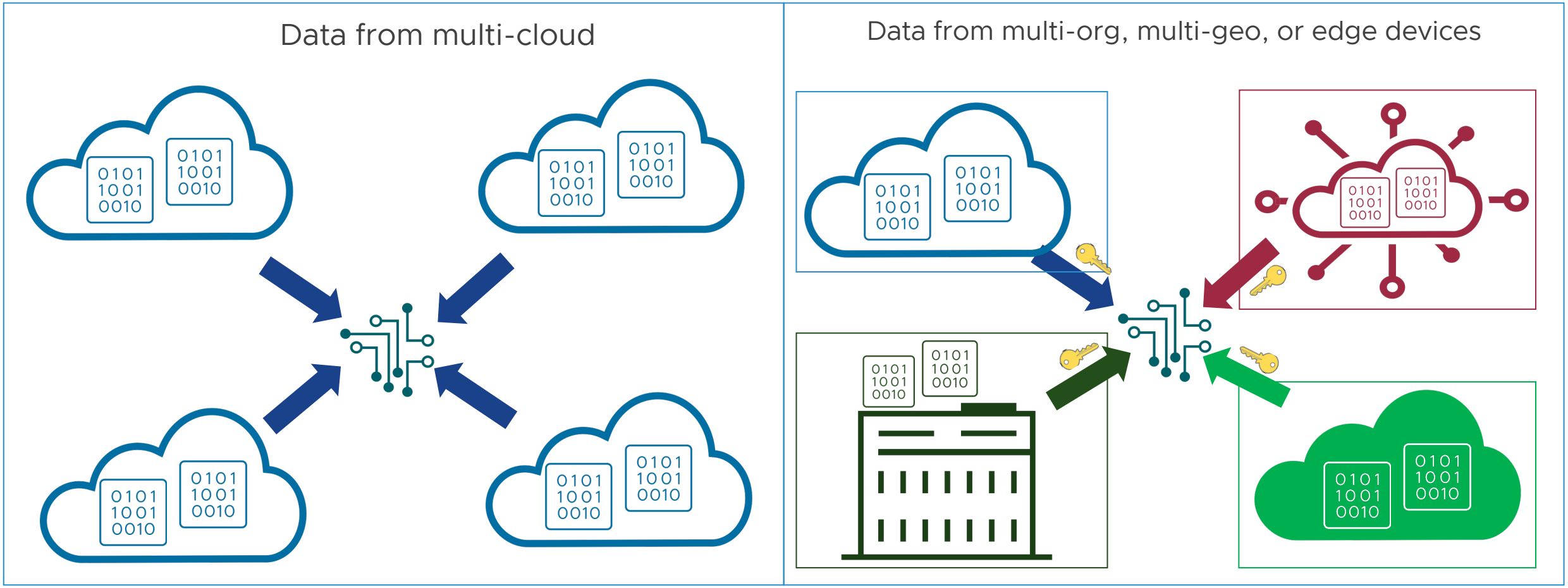


FL for a devices



FL for an enterprise

Federated learning for enterprise(s)



FL for an enterprise

FL for Multiple enterprises of a federation

FATE – the world's first industrial-grade FL OSS framework

- Hosted under LF AI & Data
- Industrial grade federated learning system
- Effectively assist multiple organizations in data usage and federated modeling
- Robust ecosystem of federated learning in the industry
  - **4000+ engineers and developers**
  - **1000+ enterprises, 400+ Universities**
  - **5000+ GitHub Stars**
  - **https://github.com/FederatedAI/FATE**

FATE
Federated AI Technology Enabler

LINUX FOUNDATION

VMware's contribution to FATE community:

- TSC Board member of FATE
  - Development Committee Chair: Henry Zhang
  - Community Operation Committee Co-Chair: Cynthia Song
  - Development Committee & maintainer : Layne Peng
- Maintainers & key contributors to OSS projects: FATE, KubeFATE, FedLCM, FATE-LLM
- Active participation in FL community & evangelism


CERTIFICATE

This is to certify that

VMWARE, INC.

has been elected

as a Board Member of Technical Steering Committee of FATE PROJECT

from November 11, 2021 to November 10, 2023.

The board membership is subject to the terms and conditions

set out in the FATE Project Technical Charter.

THE LINUX FOUNDATION FATE

LLM Revolution

Source: The A.I. Dilemma. URL https://vimeo.com/809258916/92b420d98a

LLM: Multiple tasks support and emergent abilities

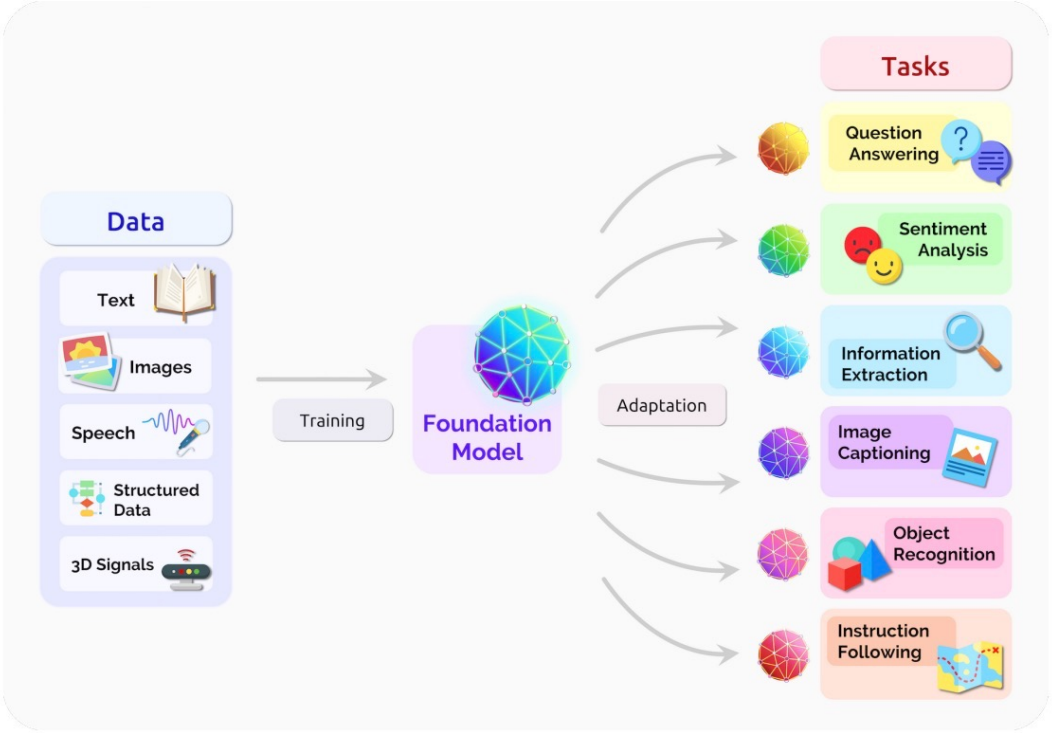
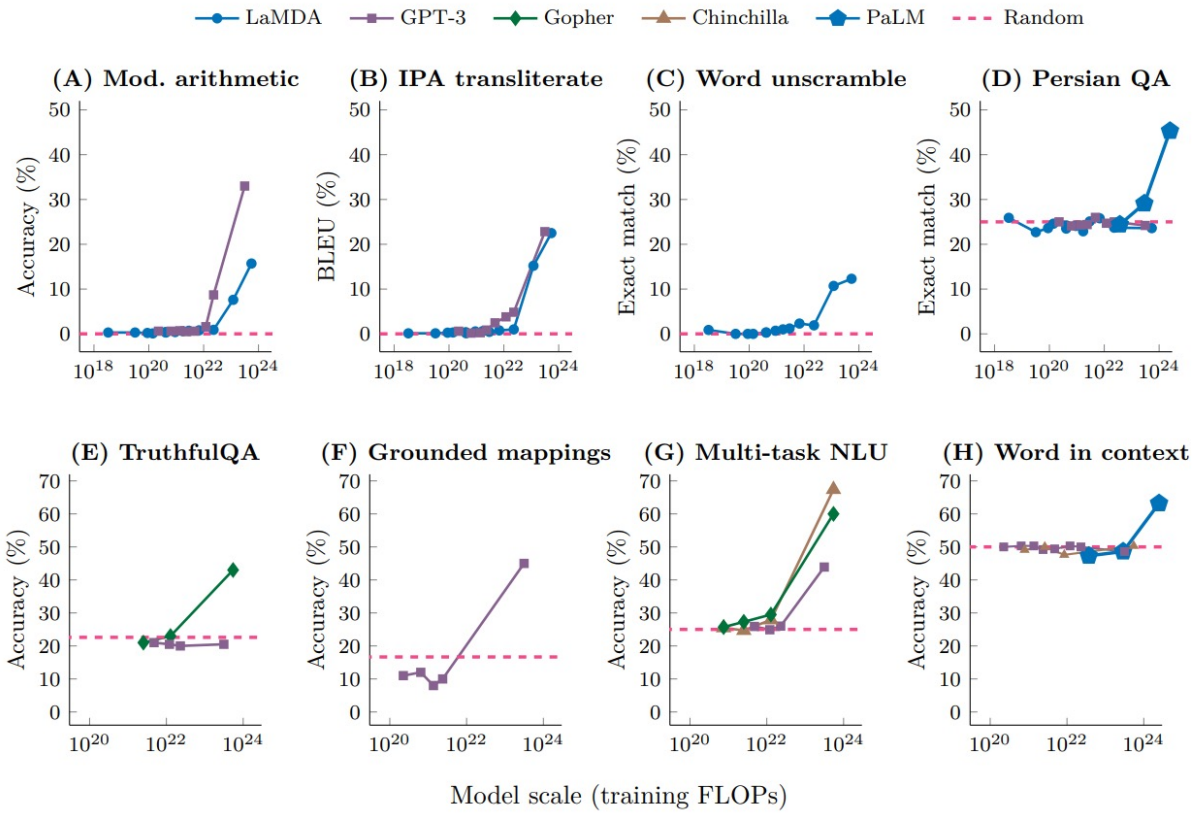


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.



Sources:

1. Rishi Bommasani, Drew A. Hudson, et al. On the Opportunities and Risks of Foundation Models, arXiv:2108.07258v3
2. Jason Wei, Yi Tay, Rishi Bommasani, et al. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research: ISSN 2835-8856.

Train a Larger model = Larger dataset + Larger resource

	Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu-α [75]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [88]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [89]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [90]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [91]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [92]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA-PT	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [93]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [94]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [72]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [95]	Dec-2021	260	-	-	-	-	-	-	-	✓	-
	GLaM [96]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [63]	Jan-2022	137	-	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [64]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [64]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [103]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

Source: Wayne Xin Zhao, Kun Zhou, et al. A Survey of Large Language Models, arXiv:2303.18223v10

Why Federated LLM?

Federated Learning helps overcome LLM challenges:

- Utilize private data when public data is depleted or insufficient
- Maintain privacy during the construction and utilization of LLM

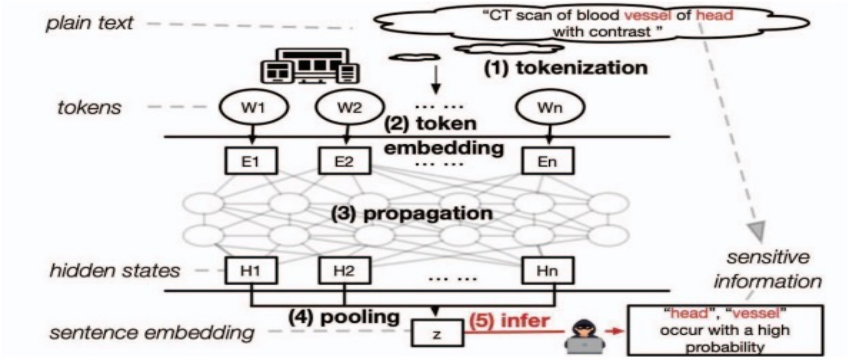


Fig. 1. General-purpose language models for sentence embedding and the potential privacy risks. The red directed line illustrates the discovered privacy risks: the adversary could reconstruct some sensitive information in the unknown plain texts even when he/she only sees the embeddings from the general-purpose language model.

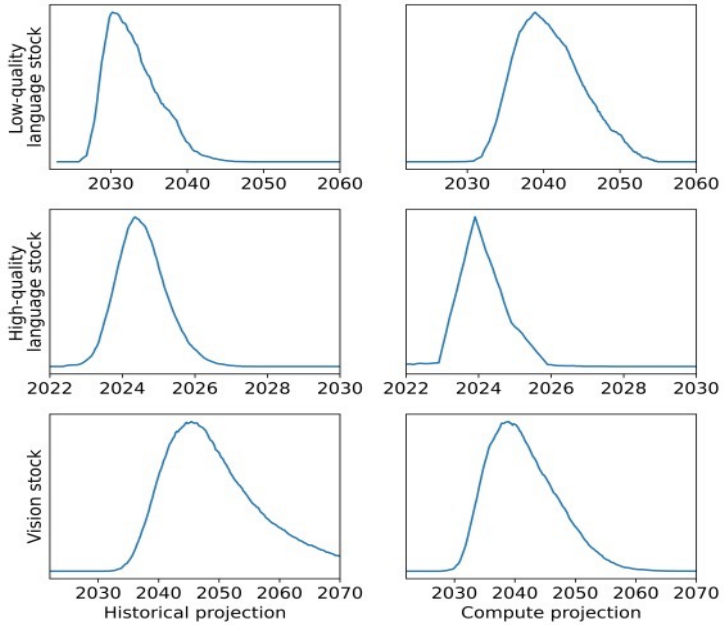


Fig. 6: Distribution of exhaustion dates for each intersection of the data availability trend and data consumptiin trend. Note that the time scale is different for each kind of data.

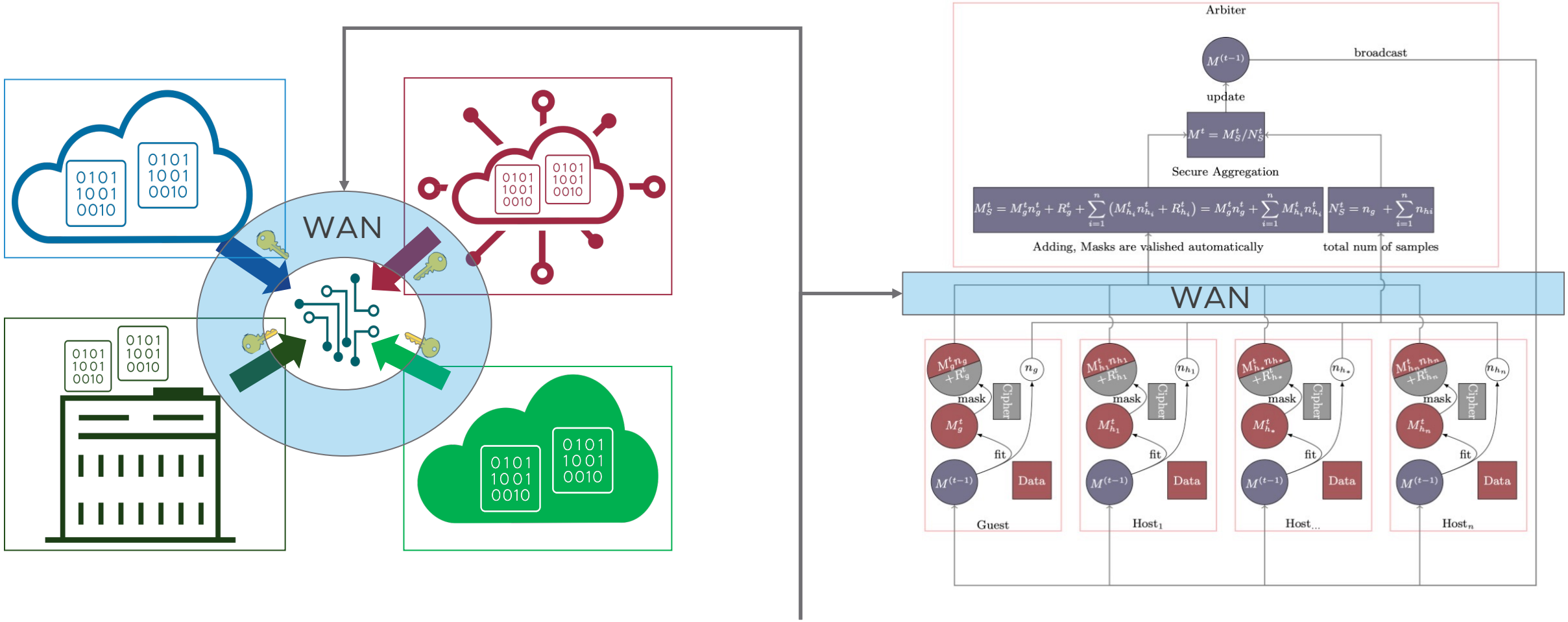
	Historical projection	Compute projection
Low-quality language stock	**2032.4** [2028.4 ; 2039.2]	**2040.5** [2034.6 ; 2048.9]
High-quality language stock	**2024.5** [2023.55 ; 2025.75]	**2024.1** [2023.2 ; 2025.3]
Vision stock	**2046** [2037 ; 2062.8]	**2038.8** [2032.0 ; 2049.8]

TABLE IV: Median and 90% CI of exhaustion year for each of the intersections.

Sources:

1. Pablo Villalobos, Jaime Sevilla, et al. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. arXiv preprint arXiv:2211.04325.
2. X. Pan, M. Zhang, S. Ji and M. Yang, "Privacy Risks of General-Purpose Language Models," 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2020, pp. 1314-1331, doi: 10.1109/SP40000.2020.00095.

Challenge #1 – Too large to exchange parameters btw. participants



How to exchange the **LARGE** models (weights/gradients) between participants in WAN?

Challenge #2 – Too much resource need for each participant to train

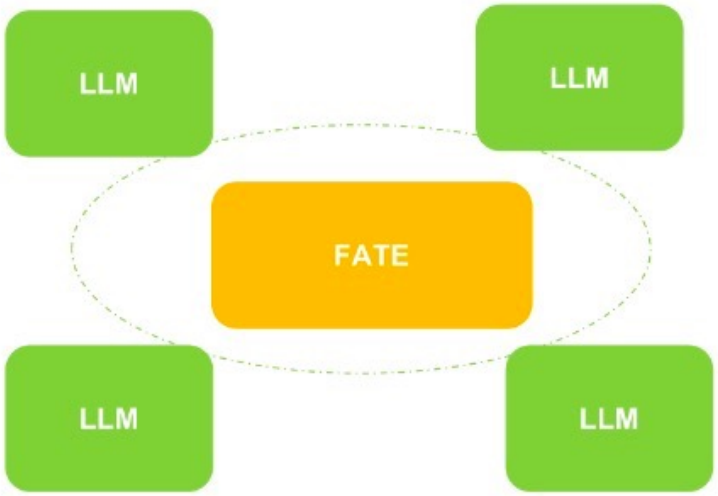
	Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu-α [75]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [88]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [89]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [90]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [91]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [92]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA-PT	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [93]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [94]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [72]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [95]	Dec-2021	260	-	-	-	-	-	-	-	✓	-
	GLaM [96]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [63]	Jan-2022	137	-	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [64]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [64]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [103]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

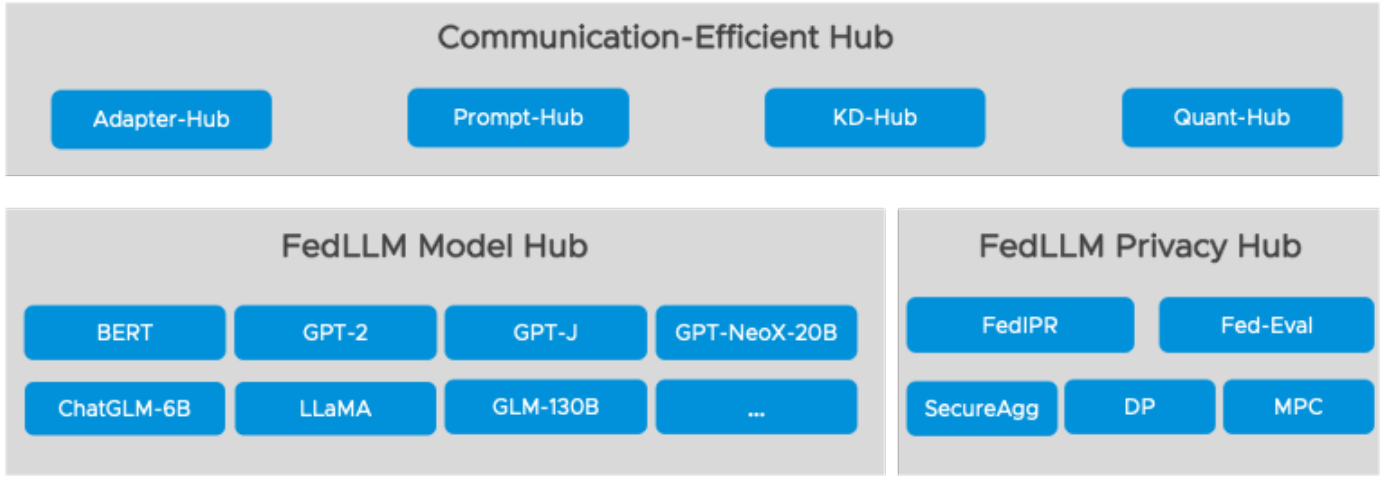
1. It requires huge computing resources for LLM training and fine-tuning;
2. Numerous federated learning scenarios involve edge devices and even IoT devices.

How to train/fine-tune the **LARGE** models in heteogenous participants with limited resources?

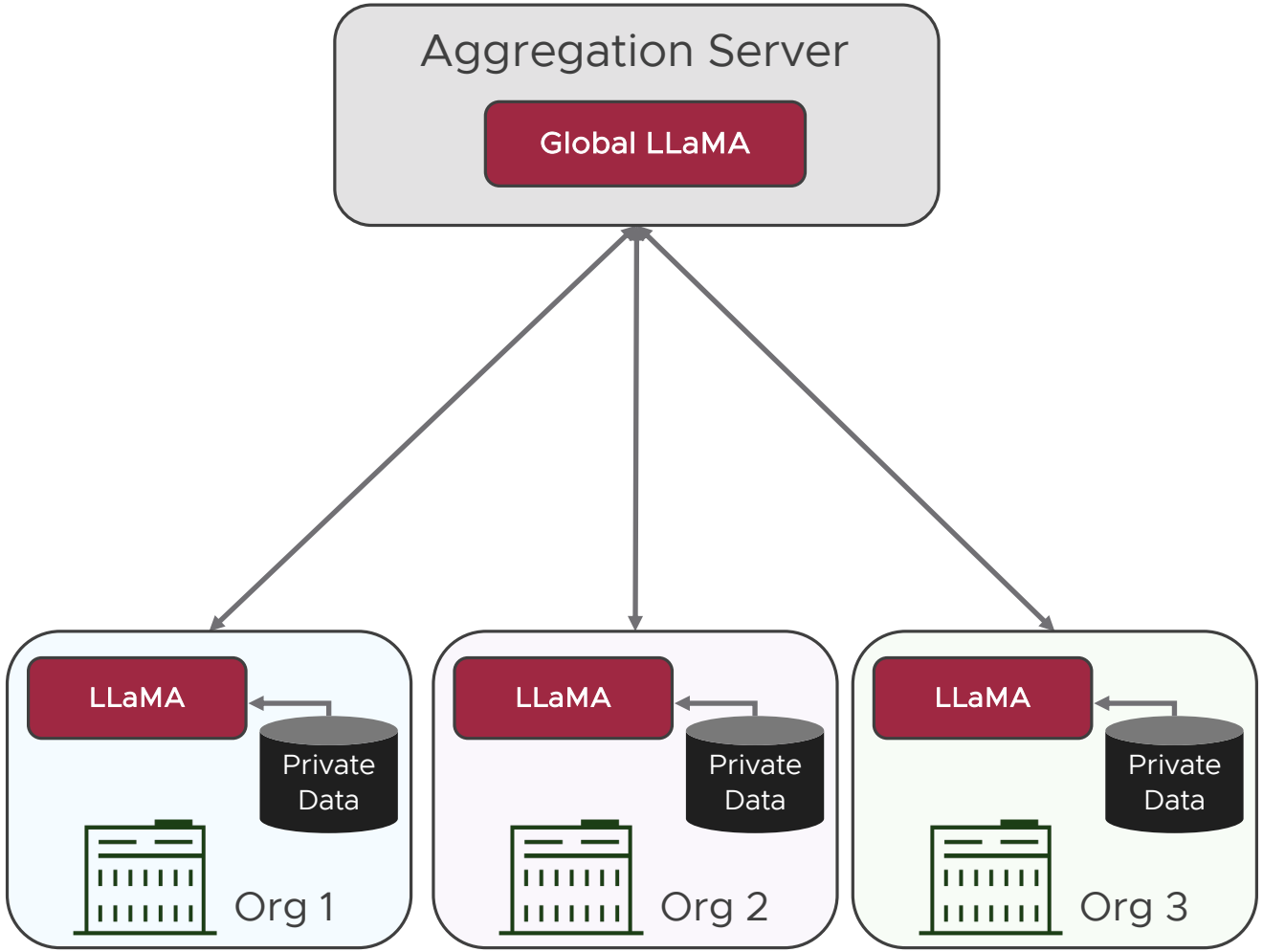
FATE-LLM: The federated LLM framework in FATE

https://github.com/FederatedAI/FATE-LLM





FATE-LLM – Homogenous Federated LLM



Frozen Layers	# Tunable Paras.	Cent.	FedOpt.
None	67.0M	86.86	55.11
E	43.1M	86.19	54.86
$E + L_0$	36.0M	86.54	52.91
$E + L_{0\to1}$	29.0M	86.52	53.92
$E + L_{0\to2}$	21.9M	85.71	52.01
$E + L_{0\to3}$	14.8M	85.47	*30.68*
$E + L_{0\to4}$	7.7M	82.76	16.63
$E + L_{0\to5}$	0.6M	*63.83*	12.97

Table 3: Performance (Acc.%) on 20news (TC) when different parts of `DistilBERT` are frozen for centralized training and FedOpt (at 28-th round). E stands for the embedding layer and L_i means the i-th layer. The significant lower accuracy are underlined.

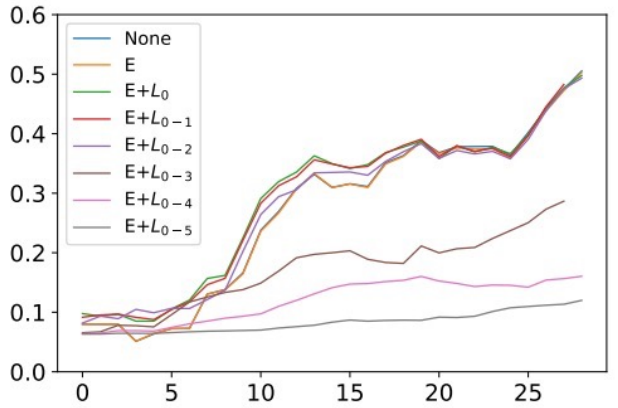
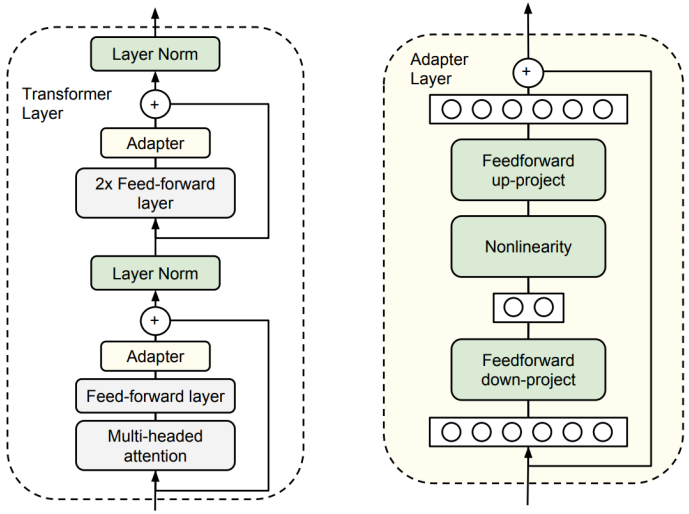


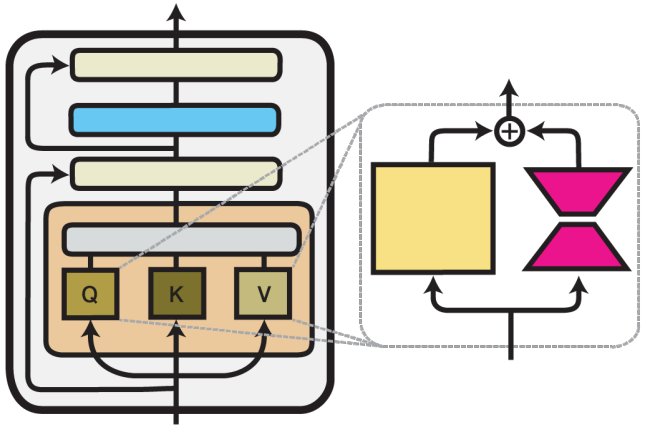
Figure 6: Testing `FedOPT` with `DistilBERT` for `20News` under different frozen layers.

Source: Bill Yuchen Lin, Chaoyang He, FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks, arXiv preprint arXiv:arXiv:2104.0881

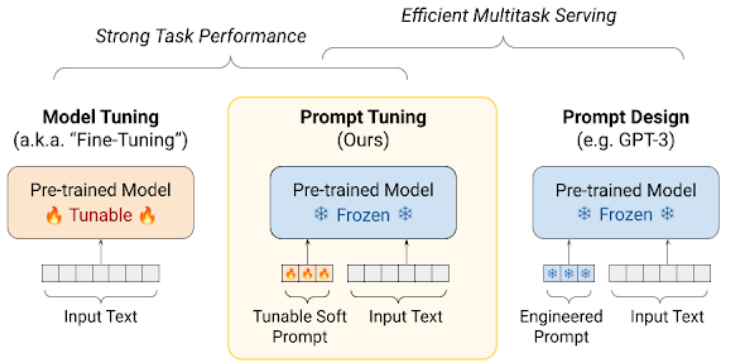
Parameter-efficient Fine-tuning (PEFT)



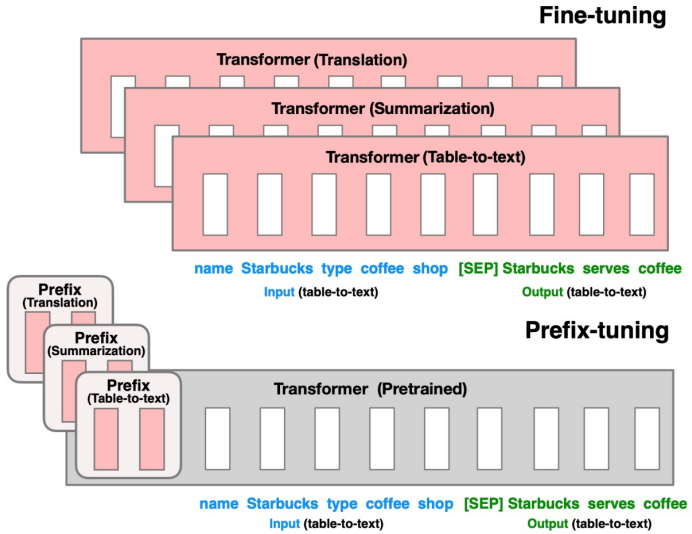
Adapter: Houlsby et al., Parameter-Efficient Transfer Learning for NLP, 2019



LoRA - Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021
image from: https://docs.adapterhub.ml/methods.html#lora



Prompt Tuning: Lester et al., The Power of Scale for Parameter-Efficient Prompt Tuning, 2021
image from: https://ai.googleblog.com/2022/02/guiding-frozen-language-models-with.html

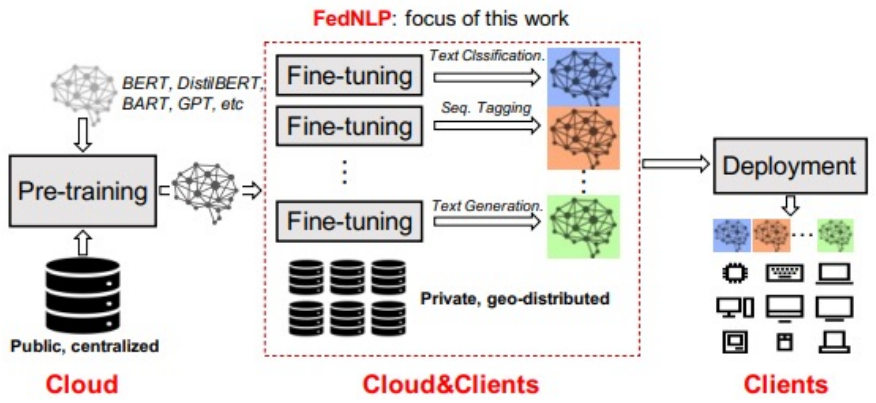


Prefix Tuning: Li et al., Prefix-Tuning: Optimizing Continuous Prompts for Generation, 2021

Federated Learning with PEFT

FedAdapter

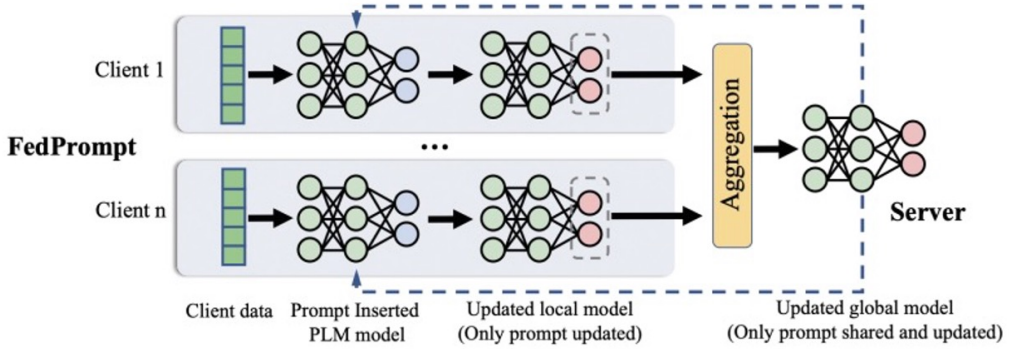
1. Emphasizing the importance of fine-tuning rather than training the foundation model in federated LLM.
2. Introducing adapter-based fine-tuning (PEFT) as a method to reduce communication costs.
3. Up to **155.5x** faster performance compared to vanilla FedNLP



Source: Cai D, Wu Y, Wang S, et al. FedAdapter: Efficient Federated Learning for Modern NLP[J]. arXiv preprint arXiv:2205.10162, 2022.

FedPrompt

1. Proposes using prompt-tuning for federated LLM;
2. Achieves comparable performance to full fine-tuning with only about 0.01% communication cost.



Model	FL Method	*ACC*	Comm. Cost	Ratio
BERT	FedPrompt	90.16	**0.016M**	**0.014%**
	Fine-tuning	91.02	109.530M	100.000%
ROBERTA	FedPrompt	92.43	**0.016M**	**0.013%**
	Fine-tuning	93.57	124.714M	100.000%
T5	FedPrompt	92.69	**0.015M**	**0.007%**
	Fine-tuning	93.79	222.919M	100.000%

Source: Zhao H, Du W, Li F, et al. Reduce Communication Costs and Preserve Privacy: Prompt Tuning Method in Federated Learning[J]. arXiv preprint arXiv:2208.12268, 2022.

FedPETuning

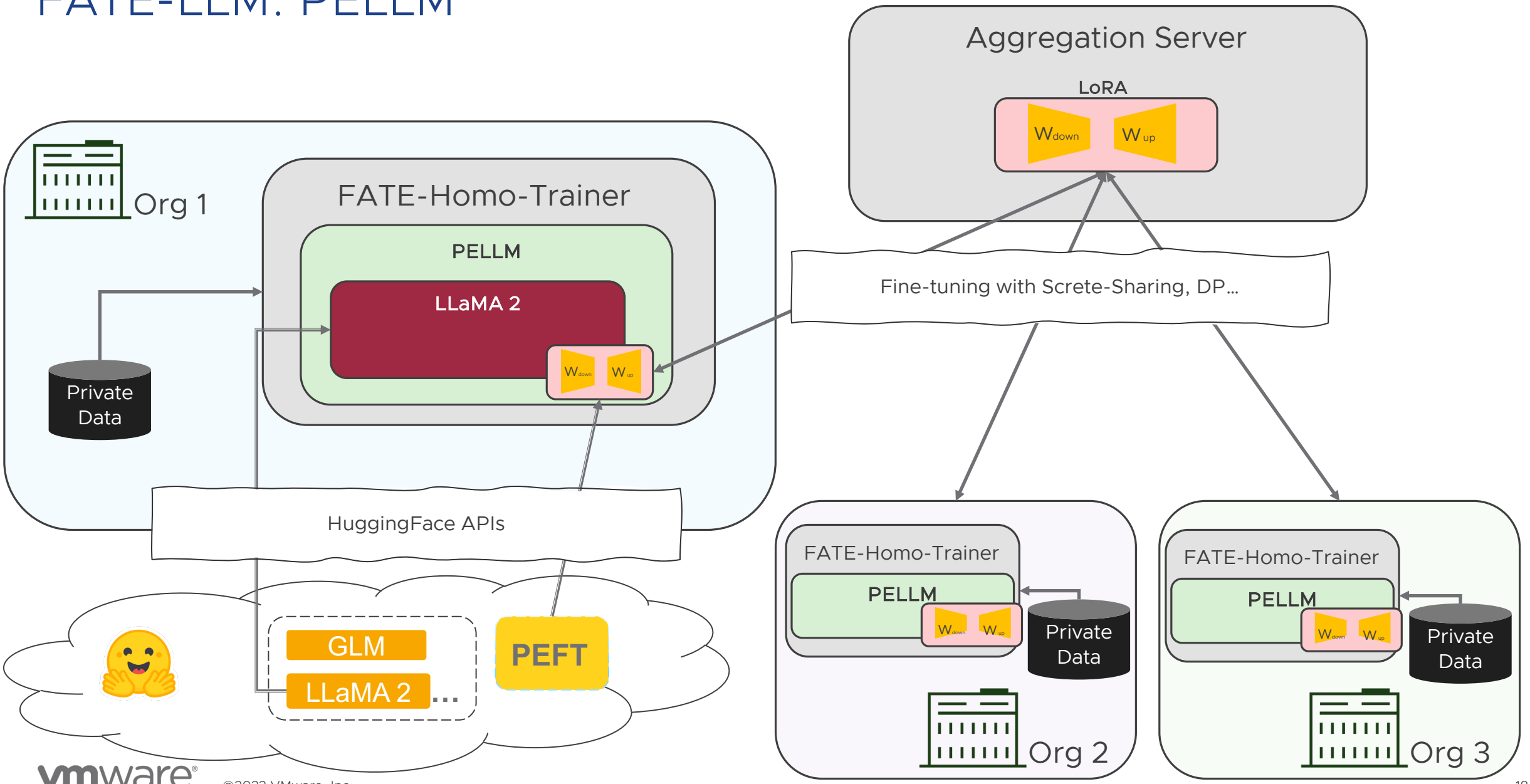
This paper presents:

1. a framework for testing and comparing parameter-efficient tuning (PETuning) methods;
2. conducts a comprehensive empirical study demonstrating a significant reduction in communication cost while maintaining acceptable performance across various federated learning settings.

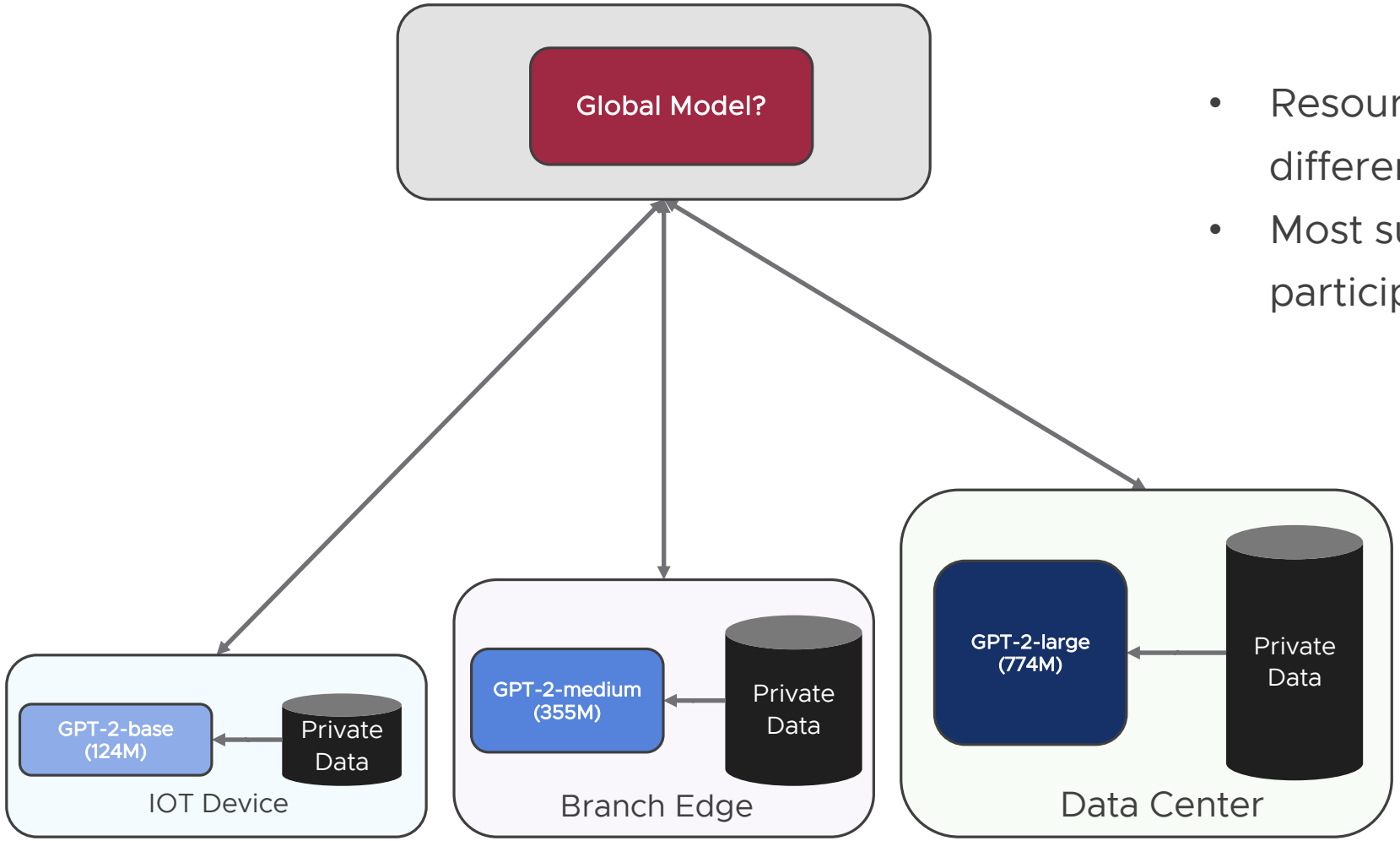
Methods	RTE	MRPC	SST-2	QNLI	QQP	MNLI	Avg		
FedBF	$61.4_{1.7}$	$84.6_{2.7}$	$92.5_{0.7}$	$87.2_{0.5}$	$84.5_{0.5}$	$81.7_{0.2}$	77.8	(↓6.4%	↑190x)
FedPF	$58.6_{2.2}$	$86.8_{1.0}$	$93.0_{0.6}$	$87.6_{0.5}$	$85.7_{0.3}$	$82.2_{0.3}$	78.4	(↓5.7%	↑12x)
FedLR	$67.4_{4.2}$	$84.5_{4.5}$	$93.6_{0.5}$	$90.8_{0.3}$	$87.4_{0.3}$	$84.9_{0.4}$	81.0	(↓2.5%	↑141x)
FedAP	$69.4_{2.6}$	$89.1_{1.2}$	$93.3_{0.6}$	$90.9_{0.4}$	$88.4_{0.2}$	$86.0_{0.4}$	82.4	(↓0.8%	↑60x)
FedFT	$\mathbf{70.3_{1.2}}$	$\mathbf{90.7_{0.3}}$	$\mathbf{94.0_{0.6}}$	$\mathbf{91.0_{0.4}}$	$\mathbf{89.5_{0.1}}$	$\mathbf{86.4_{0.2}}$	**83.1**		
Avg	65.4 (↓9.2%)	87.1 (↓4.3%)	93.3 (↓0.4%)	89.5 (↓2.5%)	87.1 (↓3.1%)	84.3 (↓2.4%)	-		
BitFit	$70.9_{1.0}$	$91.3_{0.8}$	$94.1_{0.3}$	$91.3_{0.2}$	$87.4_{0.2}$	$84.6_{0.1}$	82.6	(↓1.2%)	
Prefix	$65.6_{5.1}$	$90.2_{0.9}$	$93.7_{0.8}$	$91.5_{0.2}$	$89.5_{0.1}$	$86.7_{0.2}$	82.2	(↓1.7%)	
LoRA	$74.4_{2.4}$	$\mathbf{91.7_{0.6}}$	$94.0_{0.4}$	$92.7_{0.6}$	$90.1_{0.3}$	$87.0_{0.2}$	84.4	(↑1.0%)	
Adapter	$\mathbf{76.0_{1.8}}$	$90.6_{0.8}$	$\mathbf{94.6_{0.5}}$	$\mathbf{92.9_{0.1}}$	$\mathbf{91.1_{0.1}}$	$\mathbf{87.5_{0.2}}$	**84.7**	(↑1.3%)	
Fine-tuning	$73.0_{1.4}$	$90.9_{0.6}$	$92.1_{0.5}$	$90.8_{0.5}$	$\mathbf{91.1_{0.2}}$	$86.0_{0.2}$	83.6		
Avg	72.0	91.0	93.7	91.8	89.9	86.4	-		

Source: Zhang Z, Yang Y, Dai Y, et al. When Federated Learning Meets Pre-trained Language Models' Parameter-Efficient Tuning Methods[J]. arXiv preprint arXiv:2212.10025, 2022.

FATE-LLM: PELLM



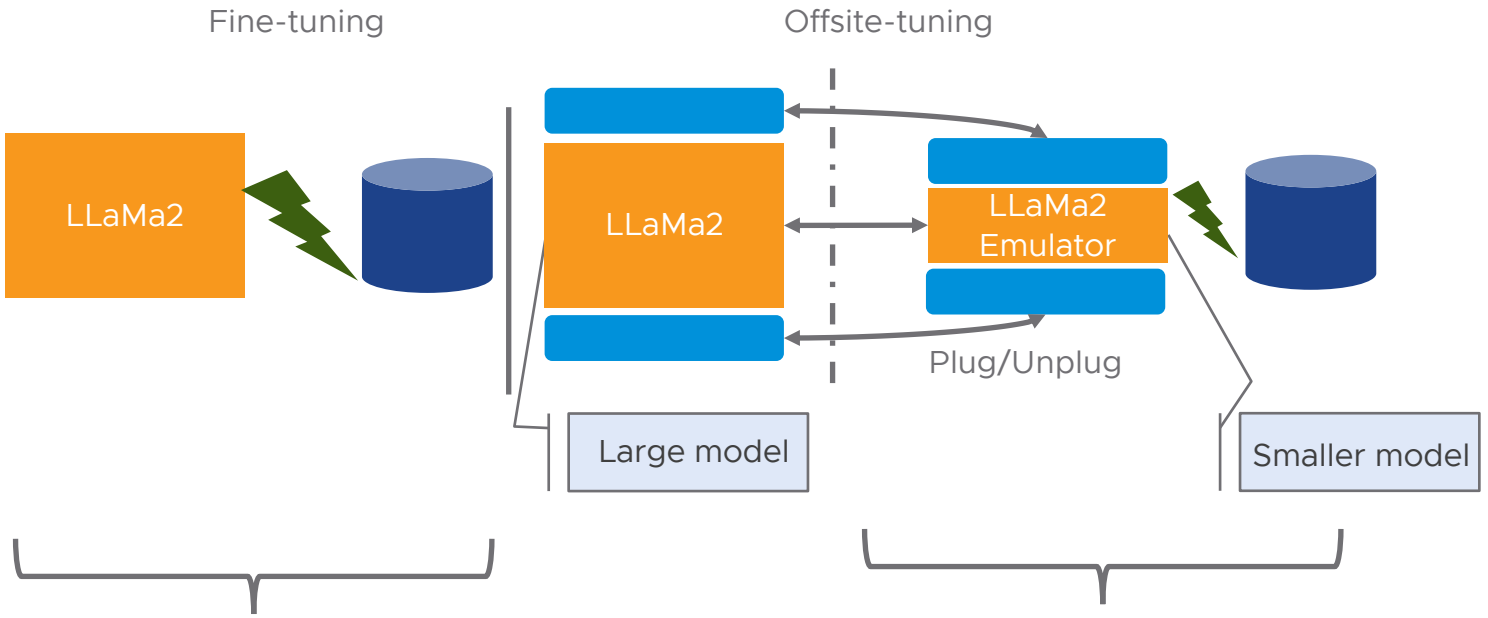
FATE-LLM – Heterogenous Federated LLM

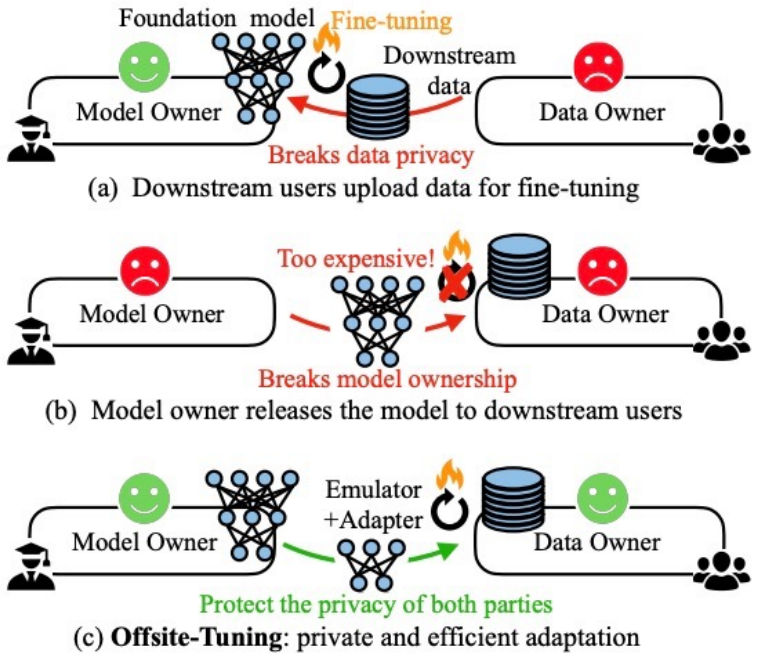


- Resources, data, etc. in the clients are different
- Most suitable models for various participants are different

#1 - Offsite-tuning

Offsite-Tuning: a privacy-preserving and efficient transfer learning framework that can adapt billion-parameter foundation models to downstream data without access to the full model.

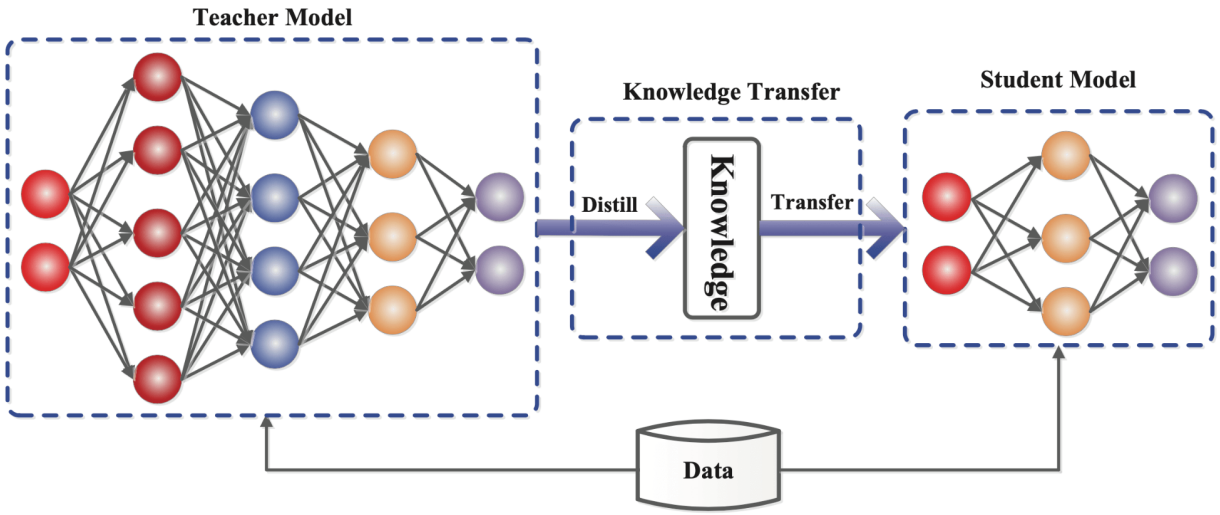


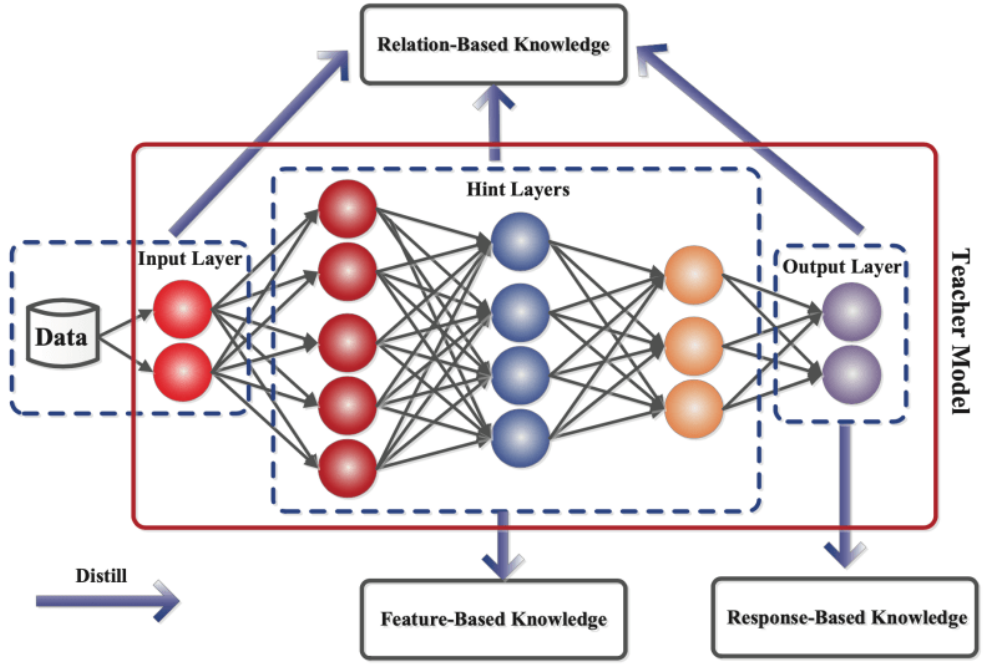


Source: Guangxuan Xiao, Ji Lin, Song Han. Offsite-Tuning: Transfer Learning without Full Model arXiv:2302.04870

2# - Knowledge Distillation

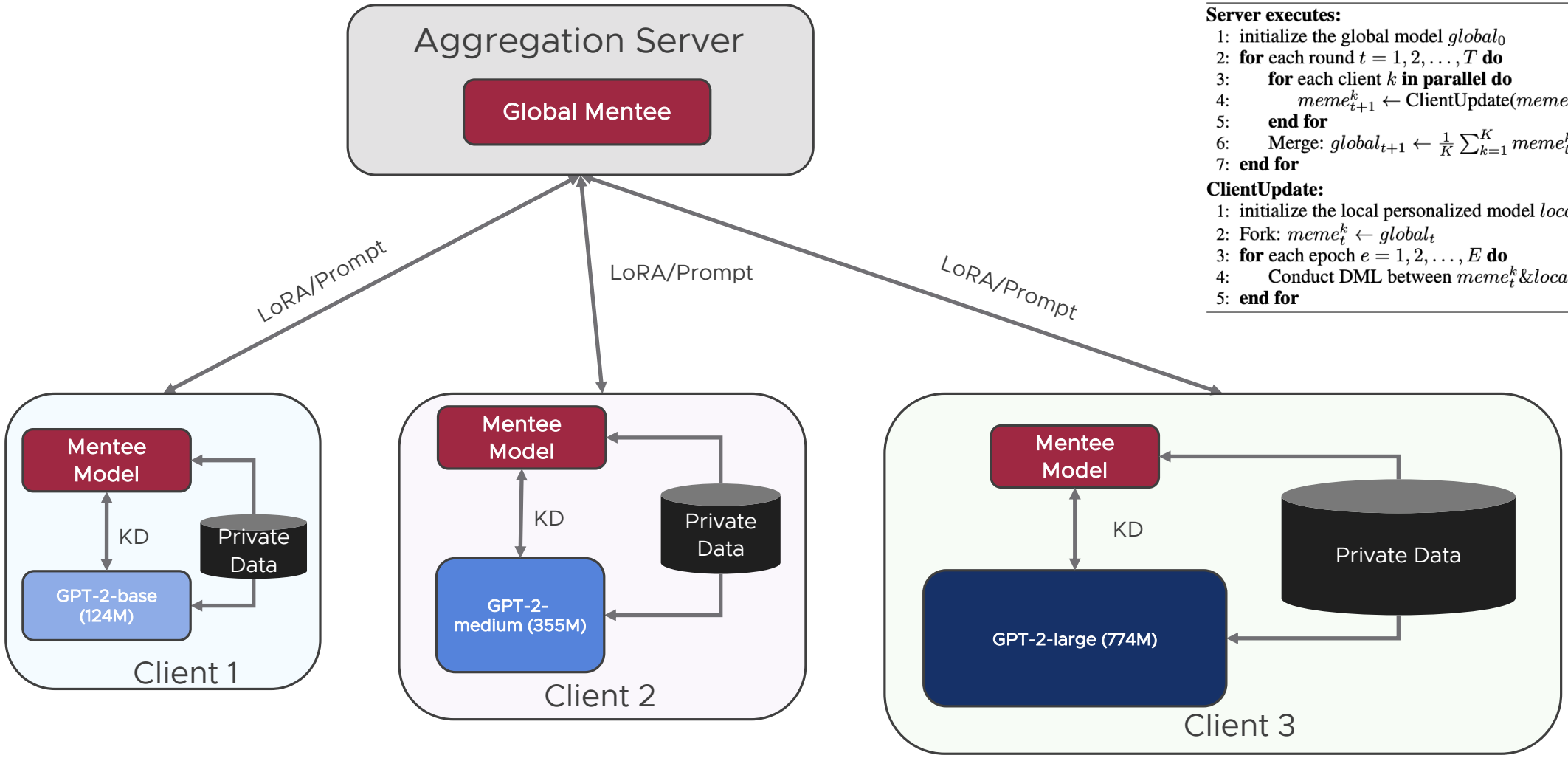
Knowledge distillation is a technique to transfer knowledge from a large teacher model to a small student model, which is widely used for model compression.





Source: Knowledge Distillation: Principles, Algorithms, Applications URL: https://neptune.ai/blog/knowledge-distillation

KD in the client sides: Shape the model with Device



Algorithm 1 *Federated Mutual Learning (FML)*:

Server executes:

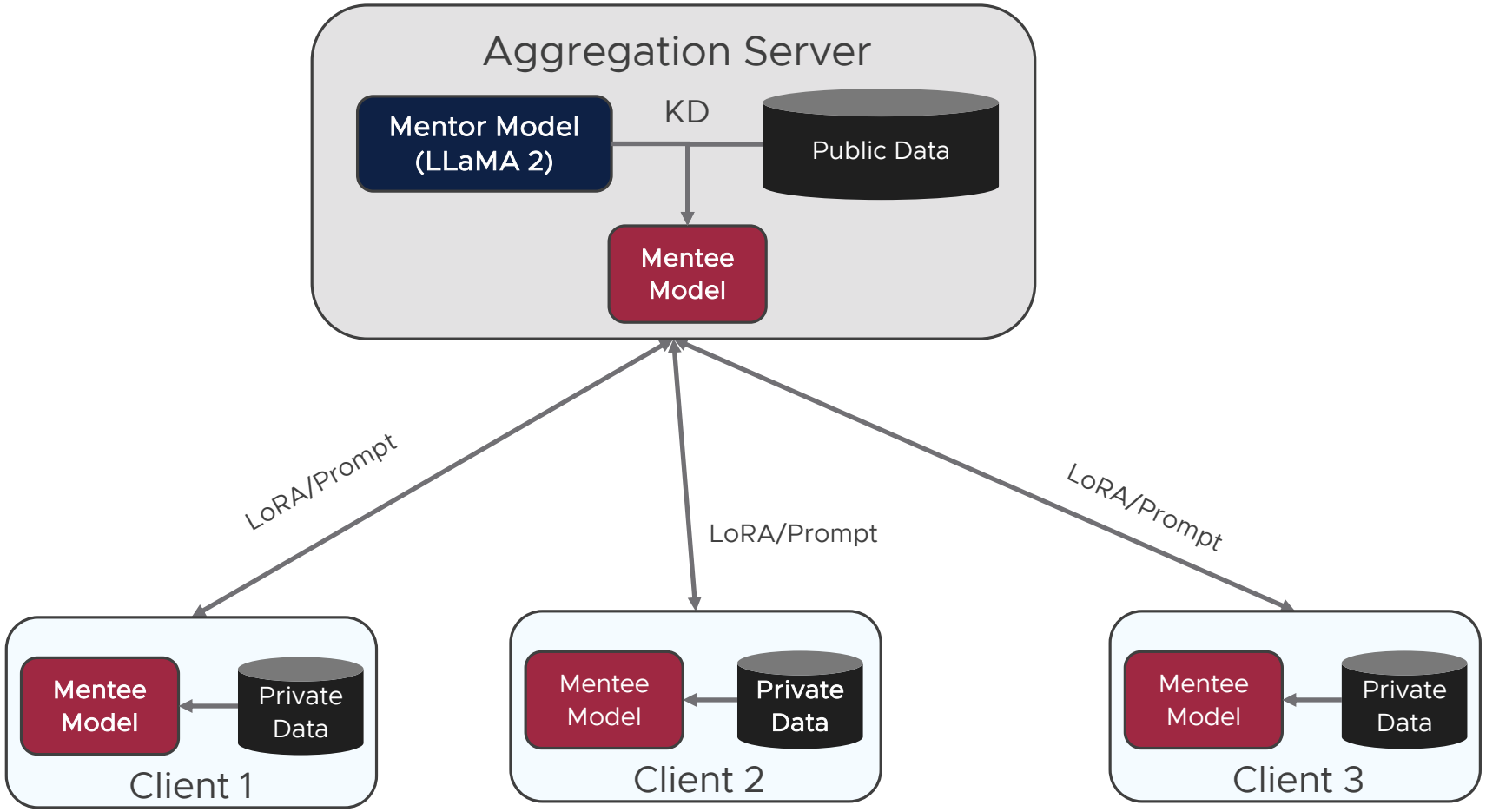
initialize the global model $global_0$
for each round $t = 1, 2, \ldots, T$ **do**
  for each client k **in parallel do**
    $meme_{t+1}^k \leftarrow \text{ClientUpdate}(meme_t^k)$
  end for
  Merge: $global_{t+1} \leftarrow \frac{1}{K} \sum_{k=1}^{K} meme_{t+1}^k$
end for

ClientUpdate:

initialize the local personalized model $local_0^k$
Fork: $meme_t^k \leftarrow global_t$
for each epoch $e = 1, 2, \ldots, E$ **do**
  Conduct DML between $meme_t^k \& local_t^k$ over private data $(\mathcal{X}_k, \mathcal{Y}_k)$
end for

Source: Shen T, Zhang J, Jia X, et al. Federated Mutual Learning[J. arXiv preprint arXiv:2006.16765.

KD in the server sides: Large model teach small model



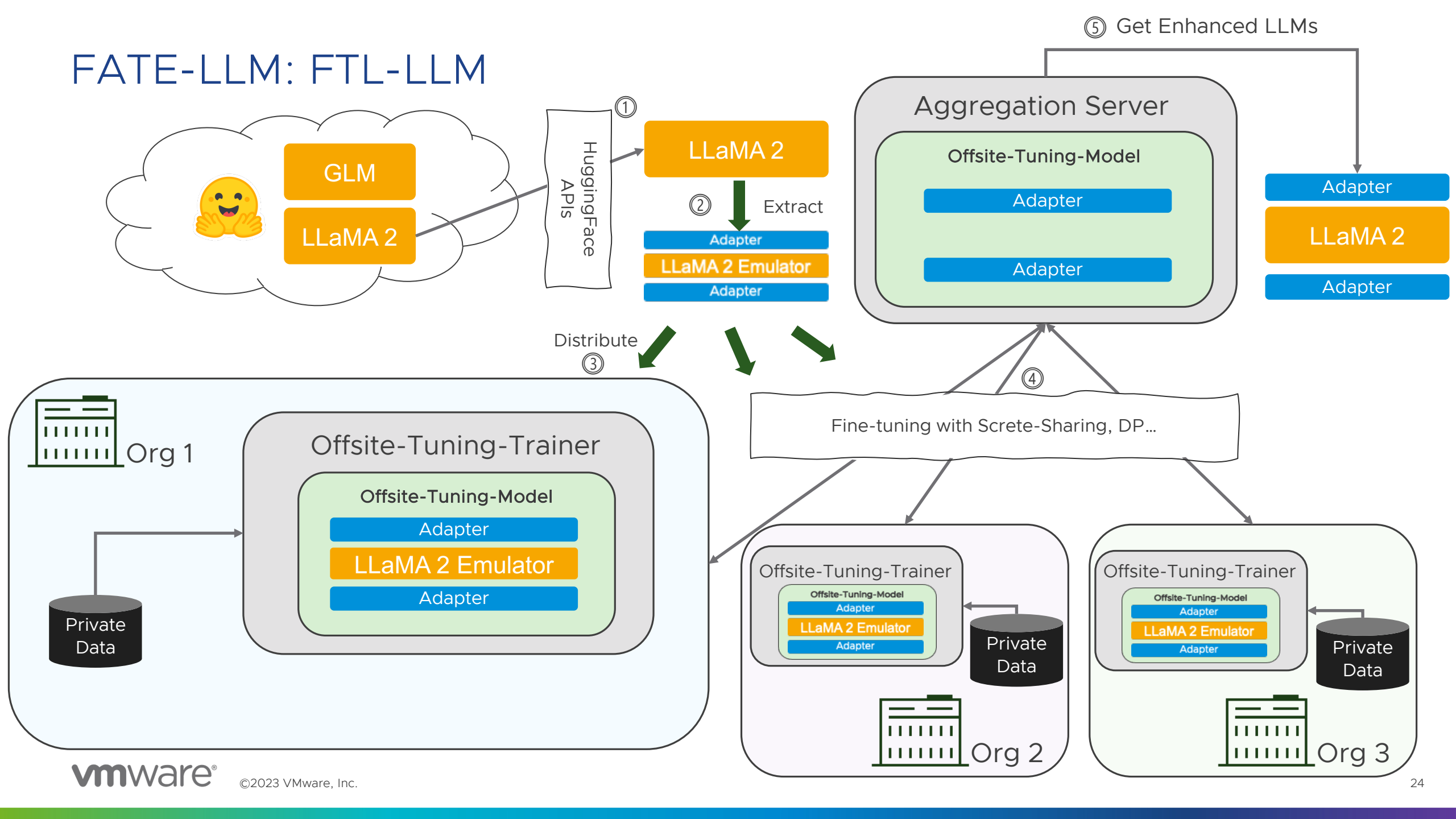
Algorithm 1 Leveraging LLMs for distribution matching and public training in private federated learning.

Input: Public pre-training corpus D, private corpus D^*, sampling rate q, private fine-tuning rounds T, first-stage fine-tuning rounds $T' < T$ for distribution matching, a public pre-trained LLM
Output: Private on-device LM with DP guarantee

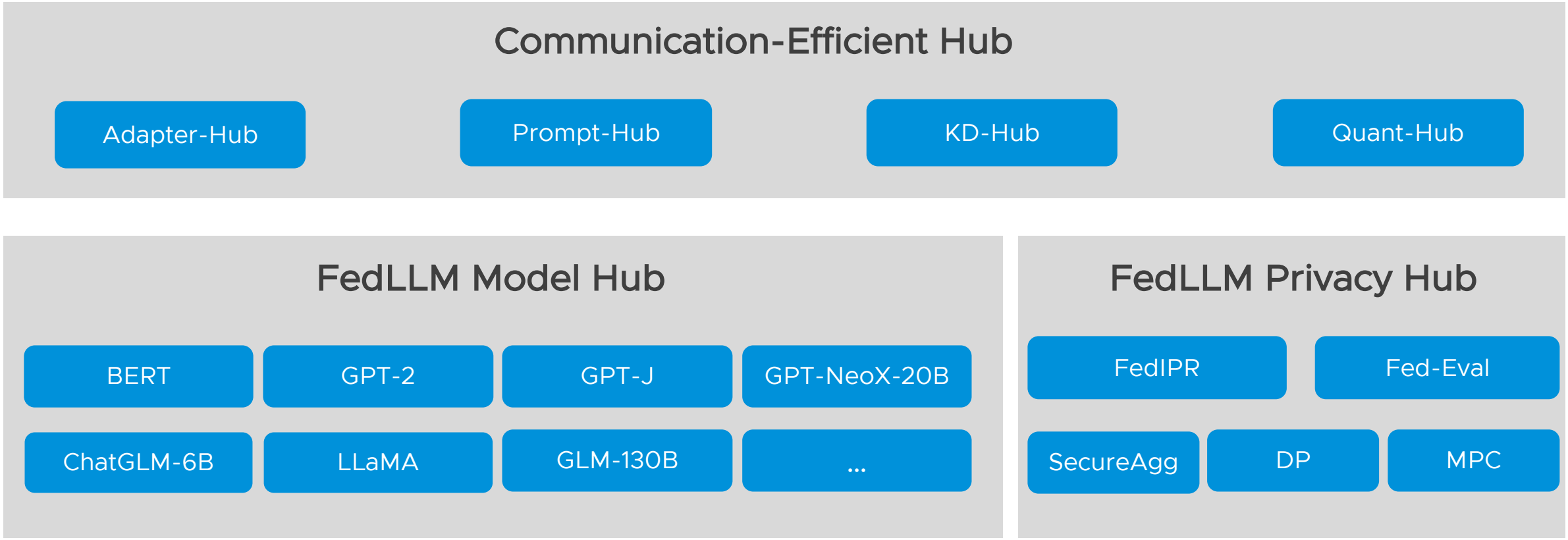
Randomly initialize an on-device LM;
// ① *First-stage private federated learning*
Use DP-FTRL to train the on-device LM for rounds T';
for each $x \in D$ **do**
    // ② *Probability evaluation*
    Compute the average (token) log prob $\log p_{\text{priv}}(x)$ given the privately fine-tuned LM at round T';
    Compute the average (token) log prob $\log p_{\text{pub}}(x)$ given a publicly pre-trained LLM ;
end for
// ③ *Distribtion matching*
Sort D based on $\log p_{\text{priv}}(x) + \log p_{\text{pub}}(x)$
Sample a subset of D as D' with top $\log p_{\text{priv}}(x) + \log p_{\text{pub}}(x)$ values, such that $|D'| = q|D|$.
// ④ *Public mid-training with LLM distillation*
Train the on-device LM with the loss $\mathcal{L}_{\text{pub}}$ on D'
// ⑤ *Second-stage private federated learning*
Use DP-FTRL to train the on-device LM for the remaining rounds of $T - T'$
return On-device LM with DP guarantee

Source: (Google Research, UIUC) Boxin W, Yibo Jacky Z, et al. Can Public Large Language Models Help Private Cross-device Federated Learning? arXiv:2305.12132

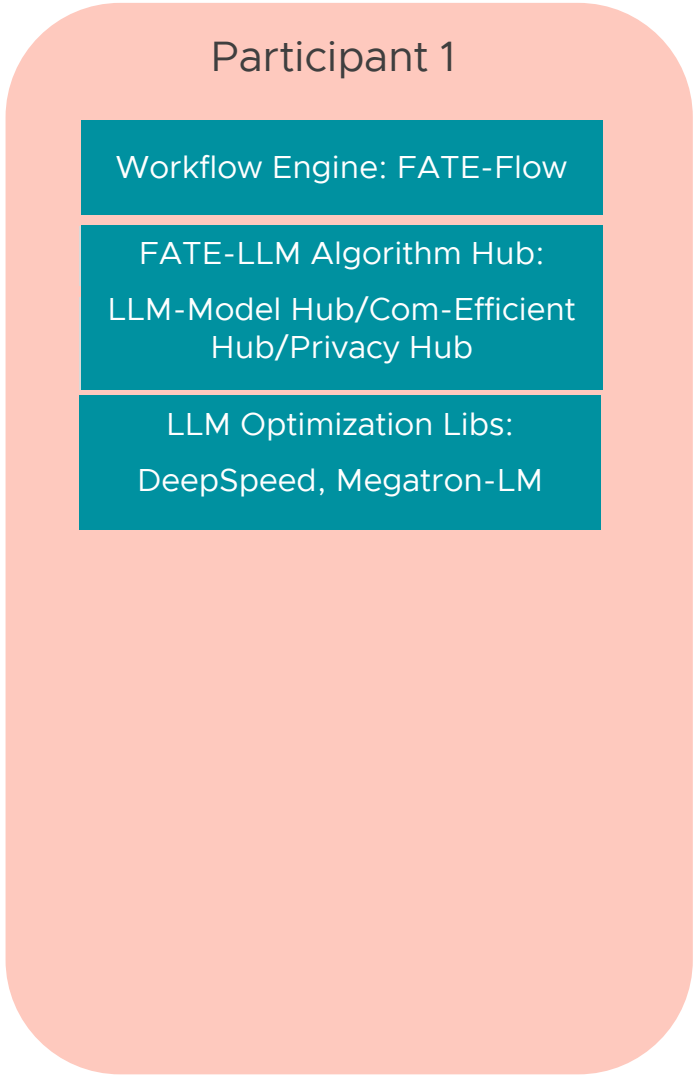
FATE-LLM: FTL-LLM



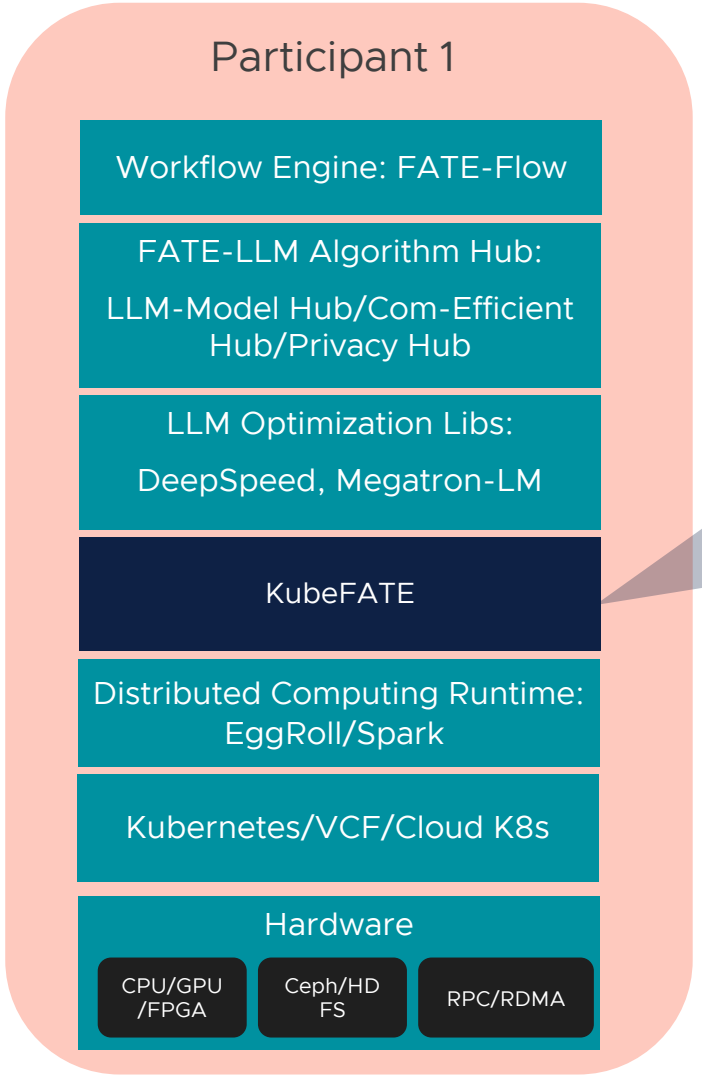
FATE-LLM: FATE Federated Large Language Model



FATE-LLM high-level architecture

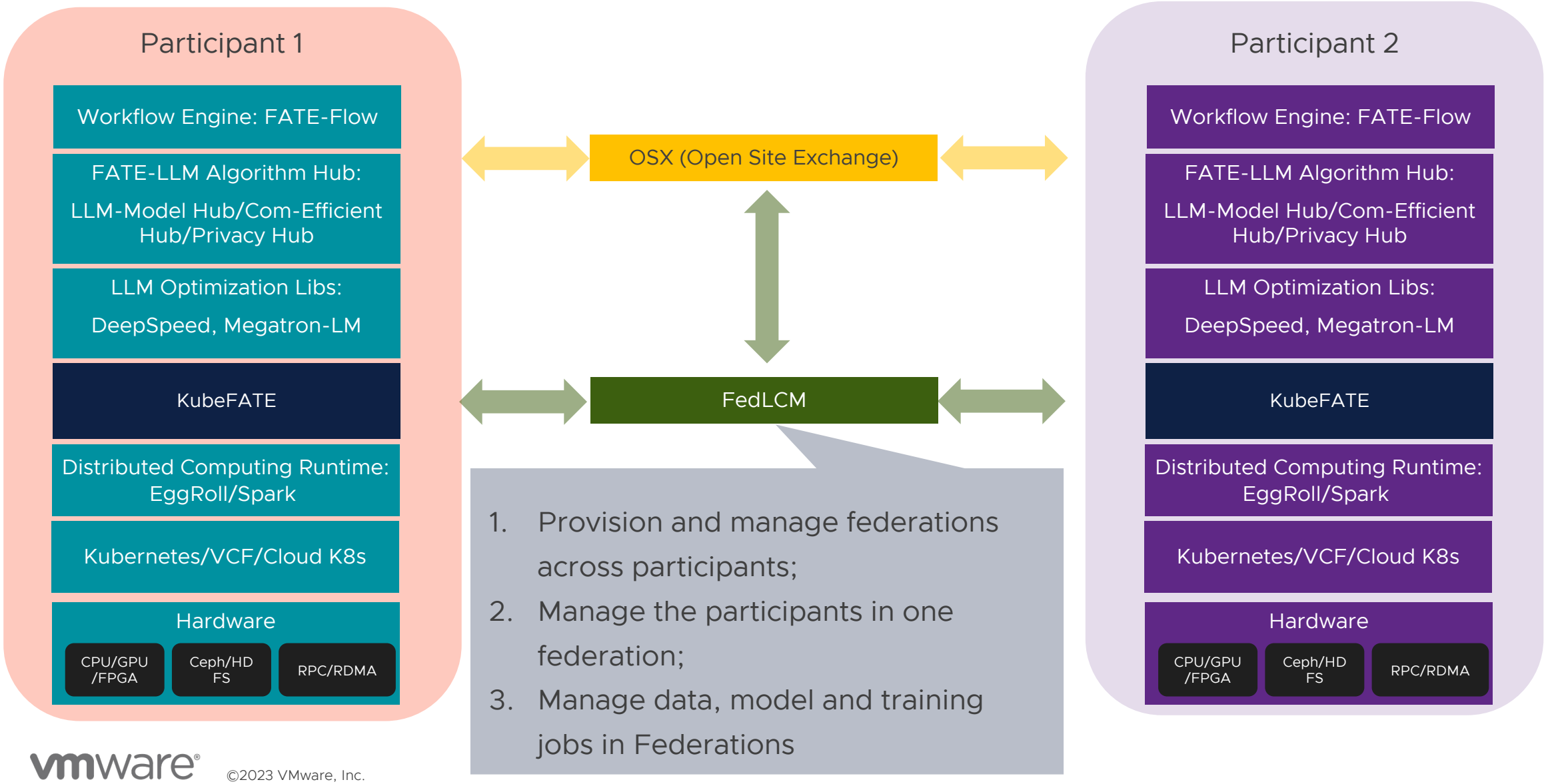


FATE-LLM high-level architecture

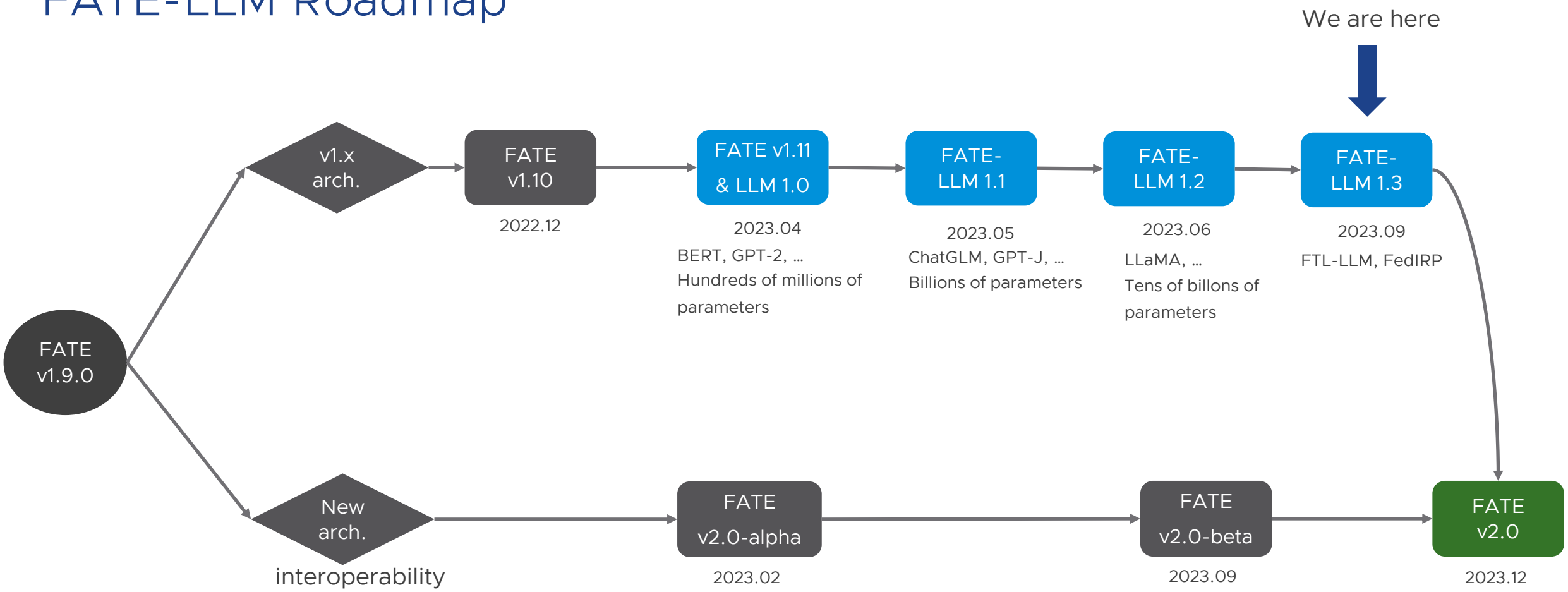


1. Containerized
2. Orchestrate the hardware resources
3. Provision and management in Kubernetes and Cloud
4. Scalability

FATE-LLM high-level architecture



FATE-LLM Roadmap



FATE-LLM: https://github.com/FederatedAI/FATE-LLM

Summary

- Federated Learning brings unique values to the success of LLM
- There are challenges specific to LLM in FL settings: communication cost, heterogeneity, etc.
- As an out-of-box component in project FATE, FATE-LLM helps address these challenges with new paradigms, modular design, etc.

Discussions and Futures

- DeepSpeed integration:
  - FATE-LLM has integrated with DeepSpeed in v1.1
  - Exploring ways to optimize the coordination between Kubernetes GPU scheduling and DeepSpeed
- Private "information" in large models:
  - Large models contain private “information” - can they intelligently extract it?
  - With the emergence of new capabilities as model parameters grow, there is uncertainty.
- FATE-LLM will continue to enhance support for larger models and explore more paradigms for privacy protection, as well as efficiency improvement

Thank You