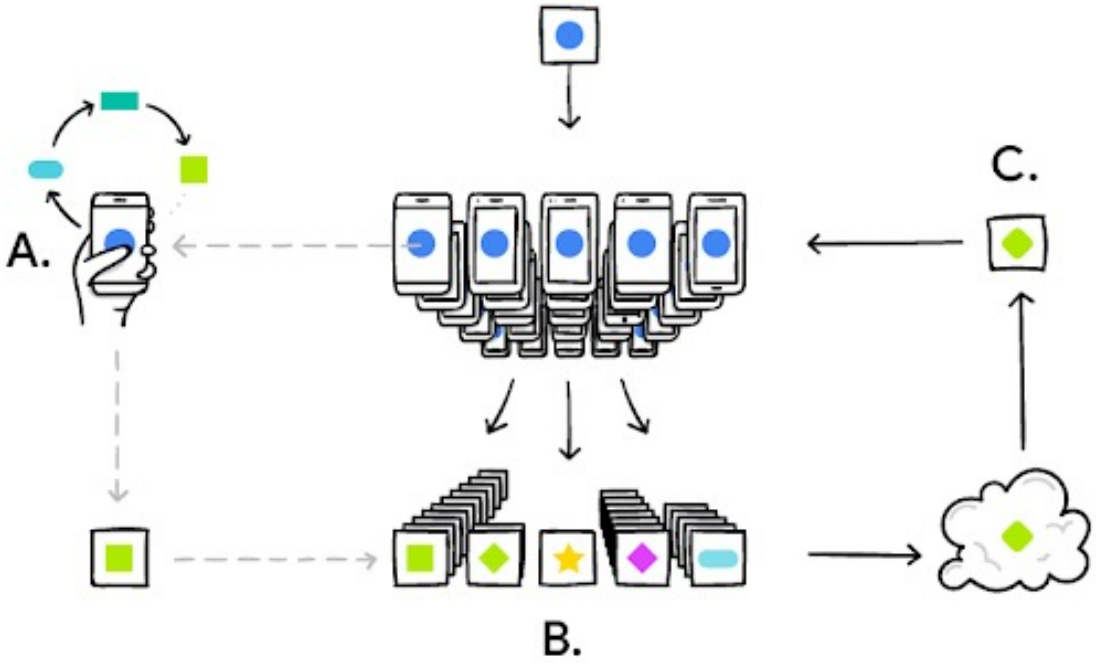
AI is Everywhere 专场

本期议题：当联邦学习遇到大语言模型

彭麟
资深研究员 -VMware

2022年05月28日

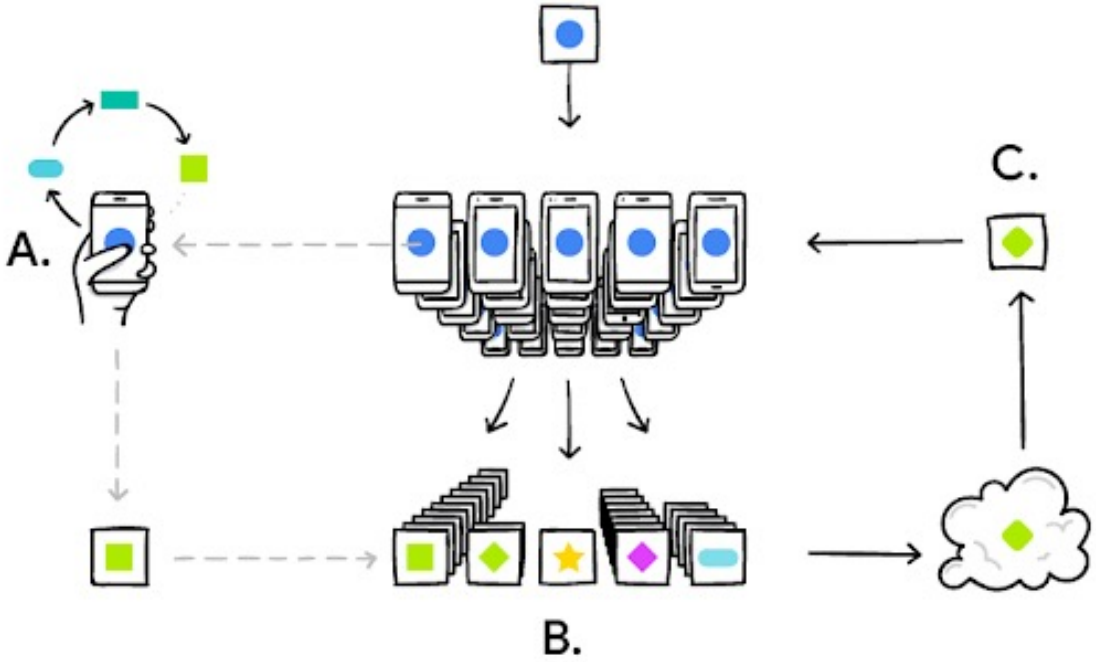
What is Federated Learning?

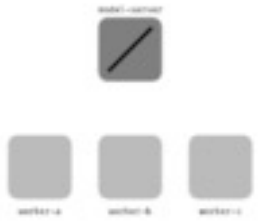
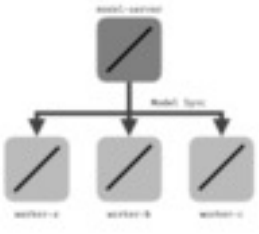
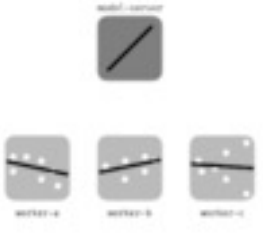
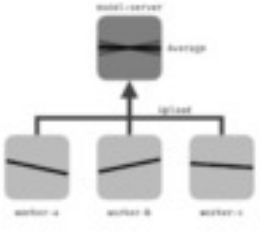


Sources:

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

What is Federated Learning?



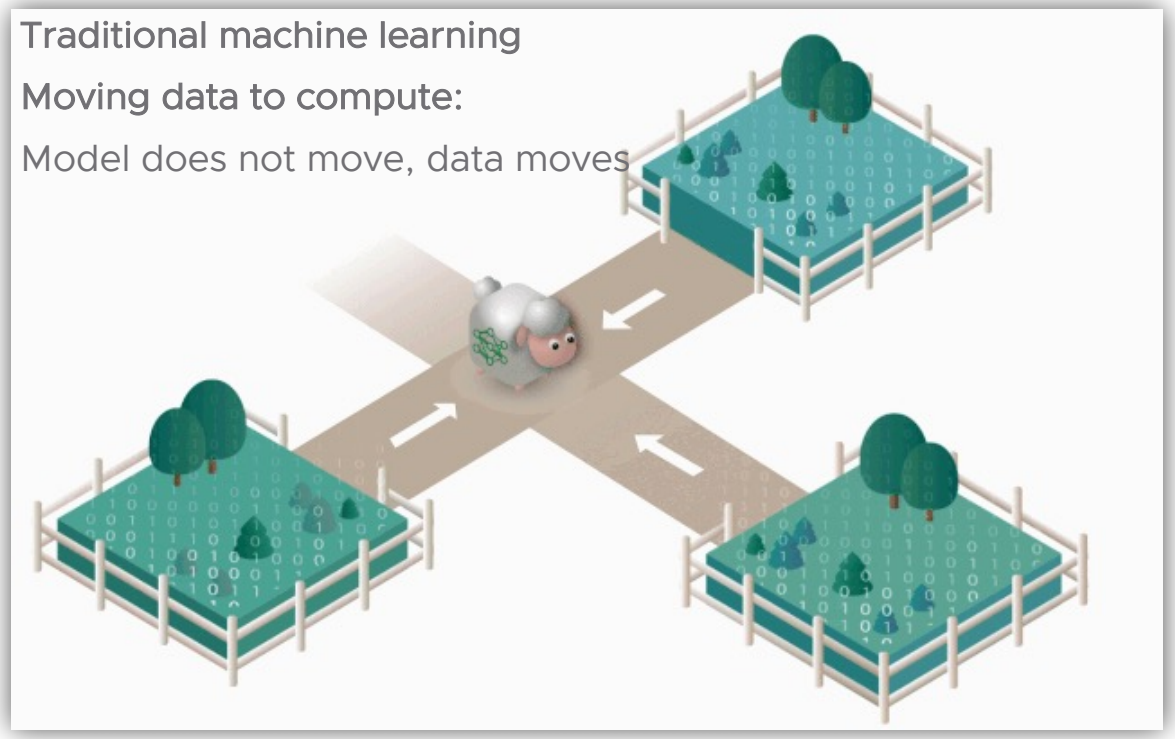
Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data

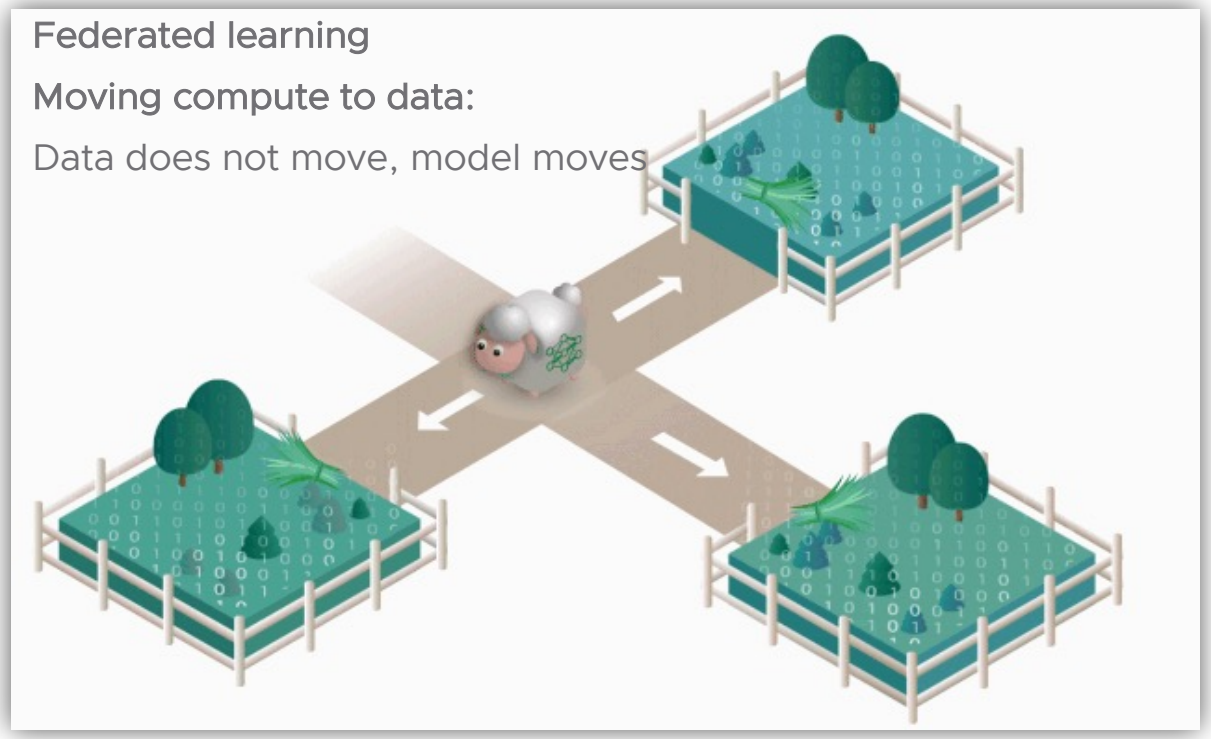
Sources:

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）

Paradigm shift of Federated Learning: Move compute to data

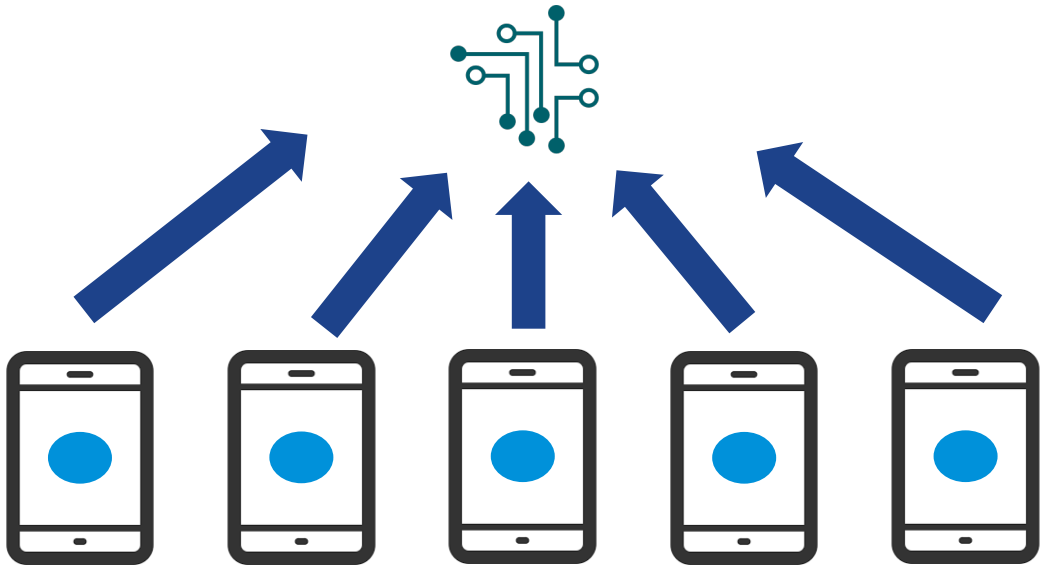
- Optimized model built from data of multiple organizations or from different places
- Preserve data privacy and confidentiality
- Communication cost reduction



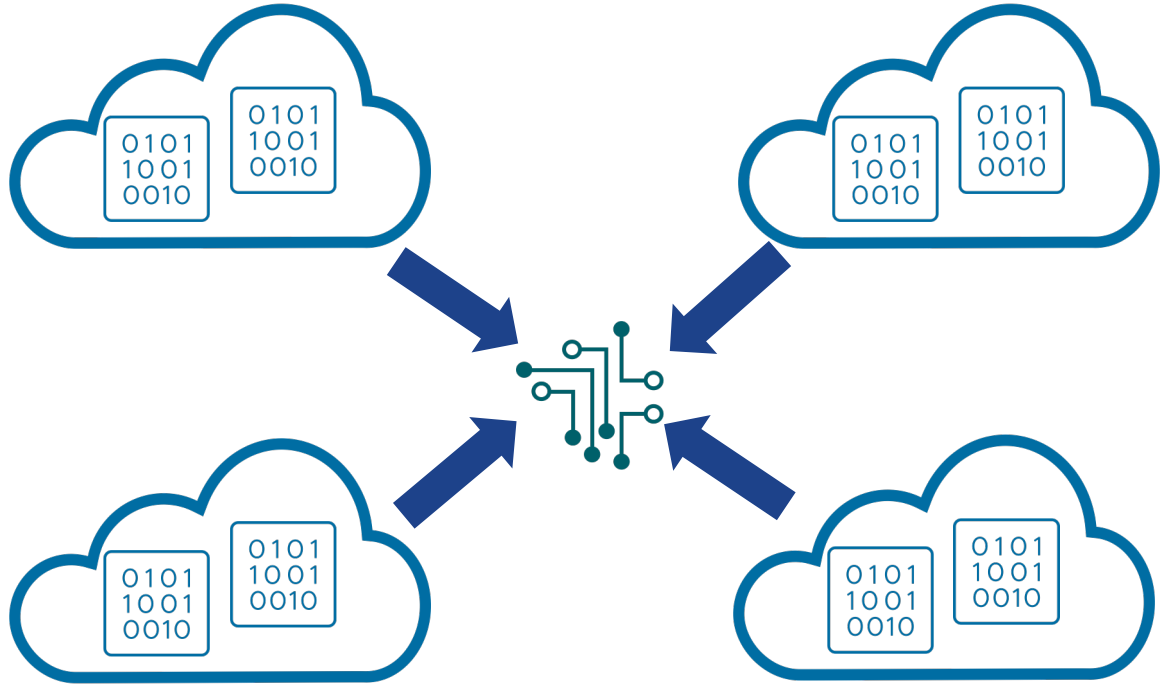


Source: Federated Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning), Qiang yang , *et al.*

From device(s) to enterprise(s)

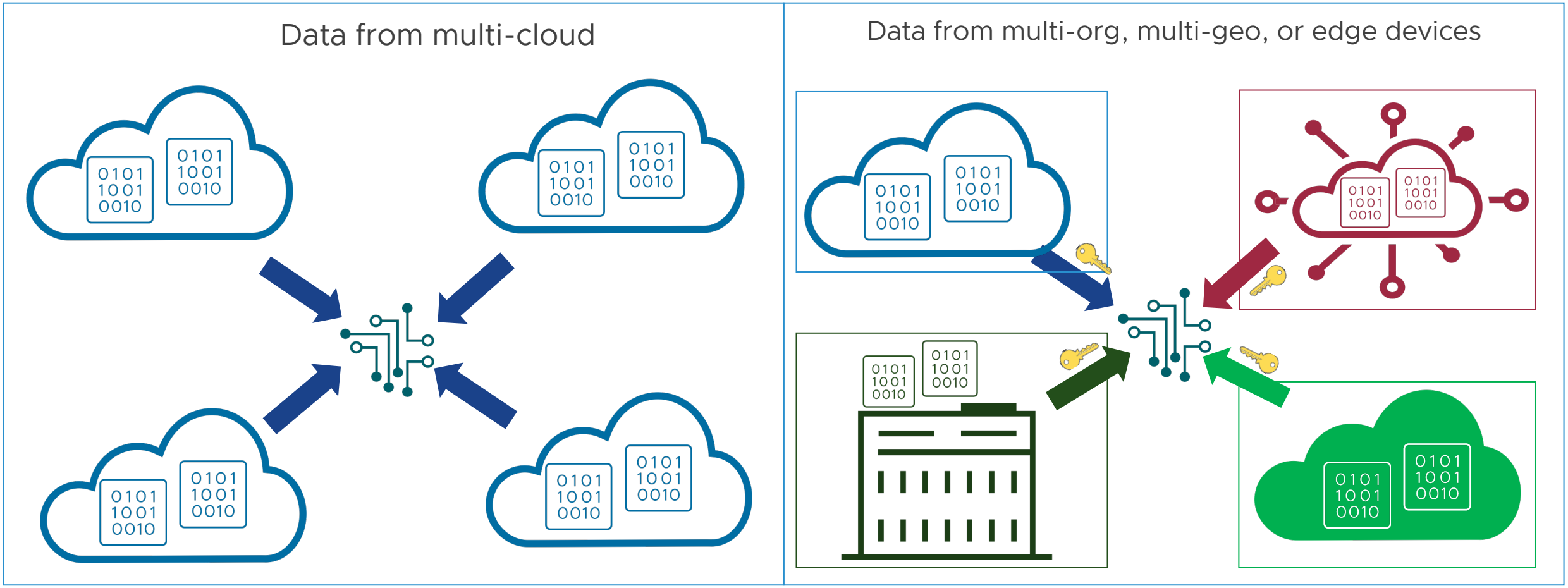


FL for a devices



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

- 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**
  - **4800+ GitHub Stars**
  - **https://github.com/FederatedAI/FATE**

FATE
Federated AI Technology Enabler

LINUX FOUNDATION

VMware's contribution to FATE community:

- TSC Board member, under Linux Foundation
  - Development Committee Chair: Henry Zhang
  - Community Operation Committee Co-Chair: Cynthia Song
  - Development Committee: Layne Peng
- Maintainer & key contributors to OSS projects: FATE, KubeFATE, FedLCM
- 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

What is LLM?

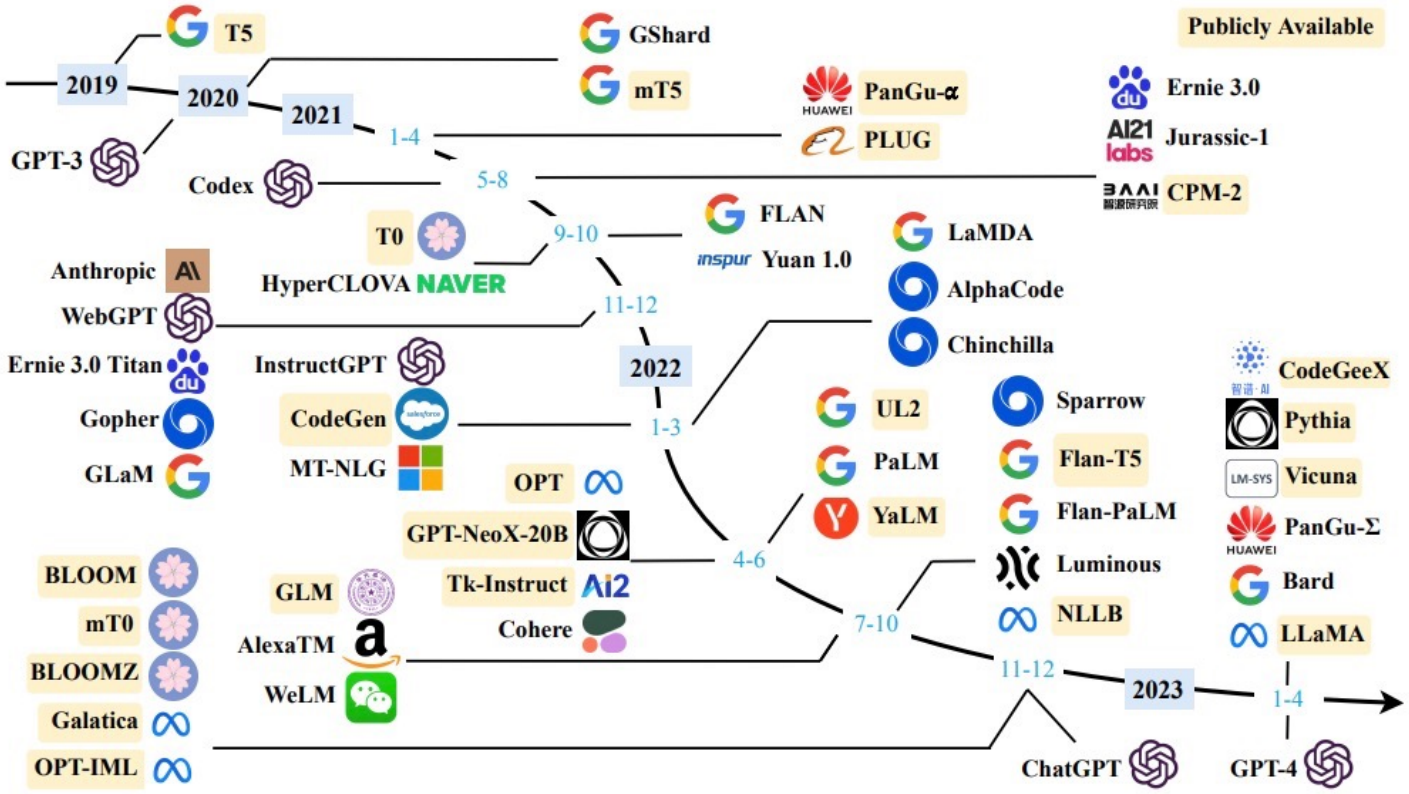


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.,* the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

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

What is LLM?

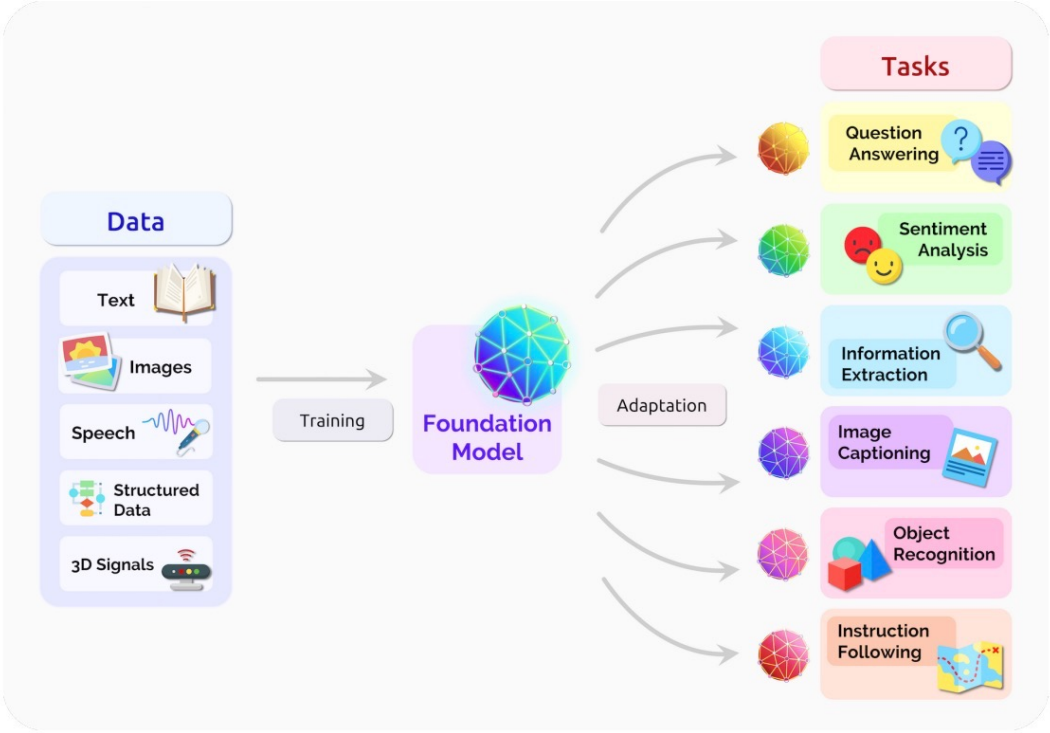
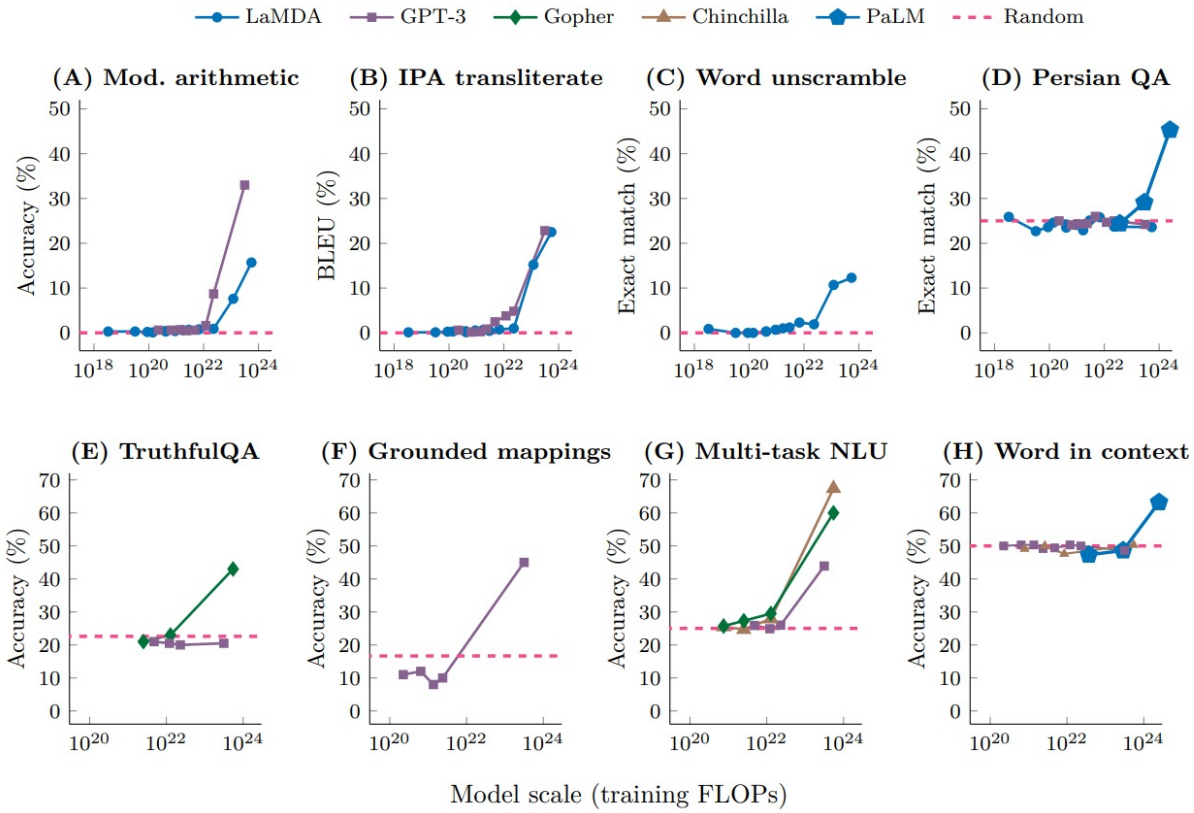


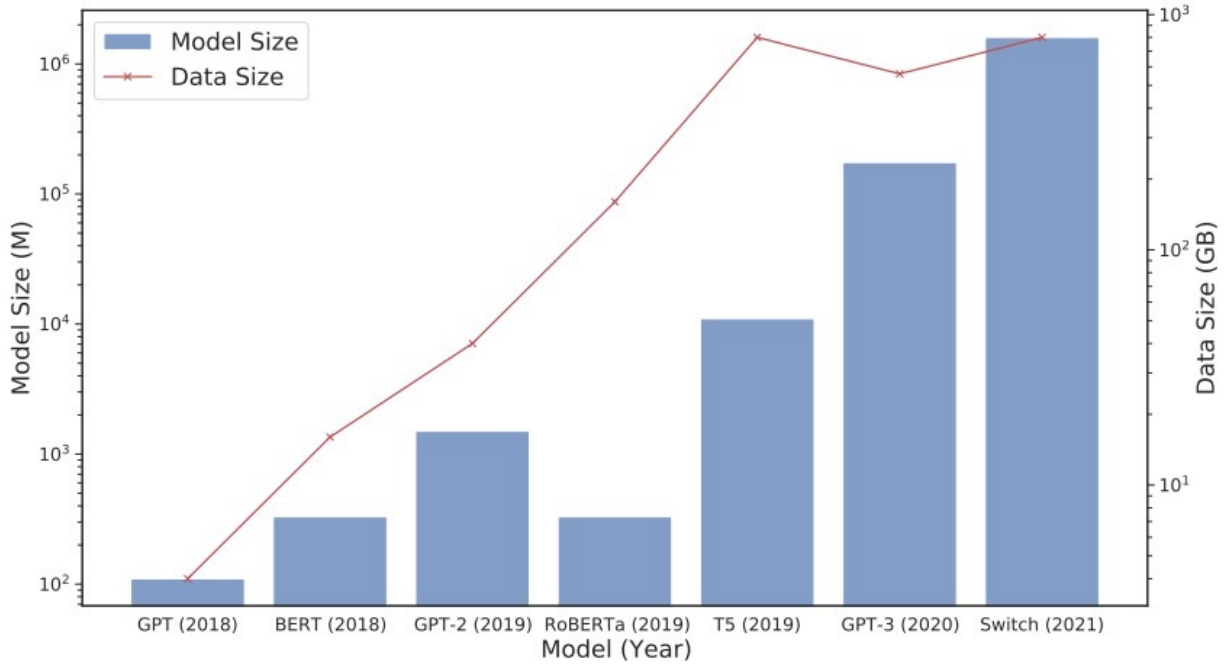
Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.



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.

Larger model, larger dataset to train



(b) The model size and data size applied by recent NLP PTMs. A base-10 log scale is used for the figure.

Source: Xu Han, Zhengyan Zhang, et al. Pre-trained Models: Past, Present and Future, arXiv:2106.07139v3

Why Federated LLM?

Federated Learning helps overcome LLM challenges:

- Enable the utilization of private data after public data is exhausted due to public data depletion and insufficiency
- Privacy-preserving when building and using LLM

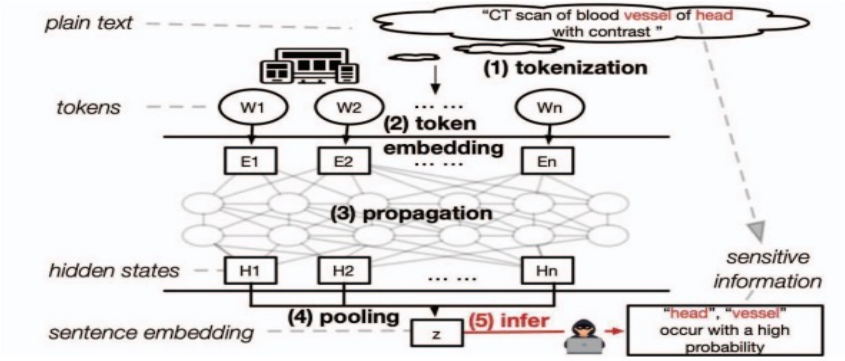


Fig. 1. General-purpose language models for sentence embedding and the potential privacy risks. The red directed line illustrates the discovered privacy risks: the adversary could reconstruct some sensitive information in the unknown plain texts even when he/she only sees the embeddings from the general-purpose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

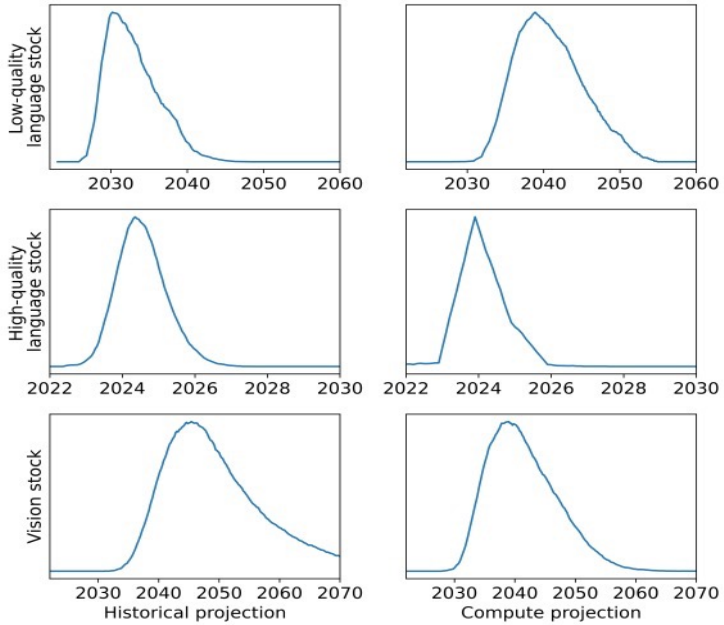


Fig. 6: Distribution of exhaustion dates for each intersection of the data availability trend and data consumptiin trend. Note that the time scale is different for each kind of data.

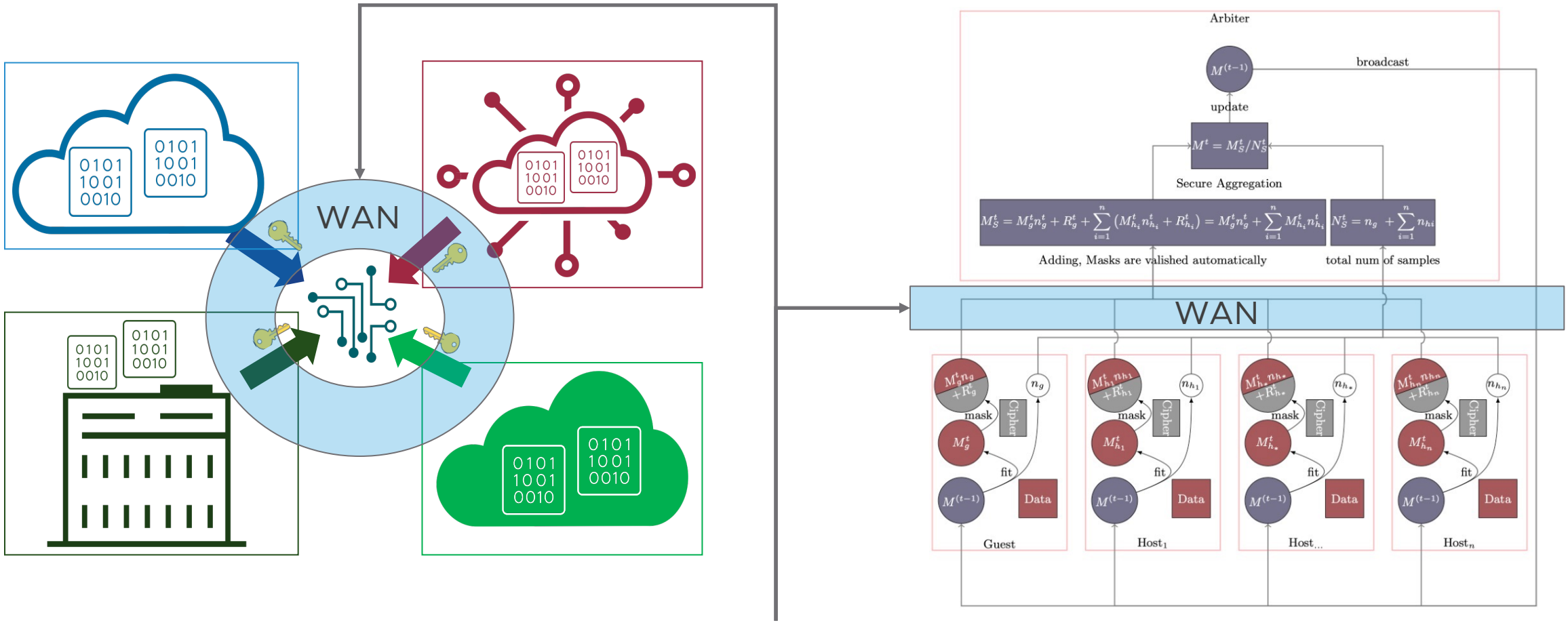
	Historical projection	Compute projection
Low-quality language stock	**2032.4** [2028.4 ; 2039.2]	**2040.5** [2034.6 ; 2048.9]
High-quality language stock	**2024.5** [2023.55 ; 2025.75]	**2024.1** [2023.2 ; 2025.3]
Vision stock	**2046** [2037 ; 2062.8]	**2038.8** [2032.0 ; 2049.8]

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

Sources:

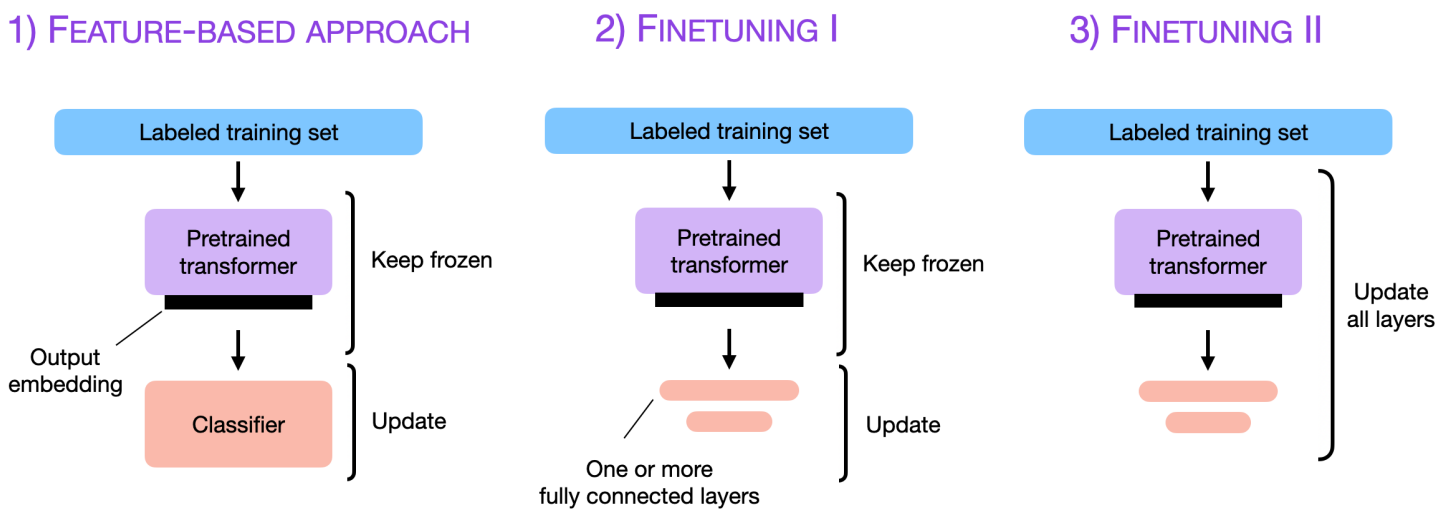
1. Pablo Villalobos, Jaime Sevilla, et al. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. arXiv preprint arXiv:2211.04325.
2. X. Pan, M. Zhang, S. Ji and M. Yang, "Privacy Risks of General-Purpose Language Models," 2020 IEEE Symposium on Security and Privacy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

New Challenges to Federated Learning

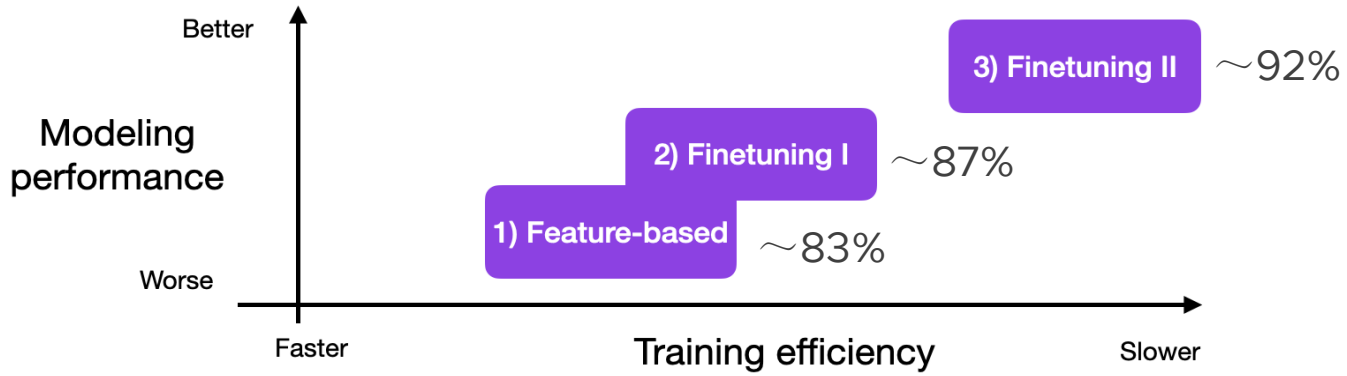


How to exchange the **LARGE** models (weights/gradients) between participants in WAN?

Fine-tuning



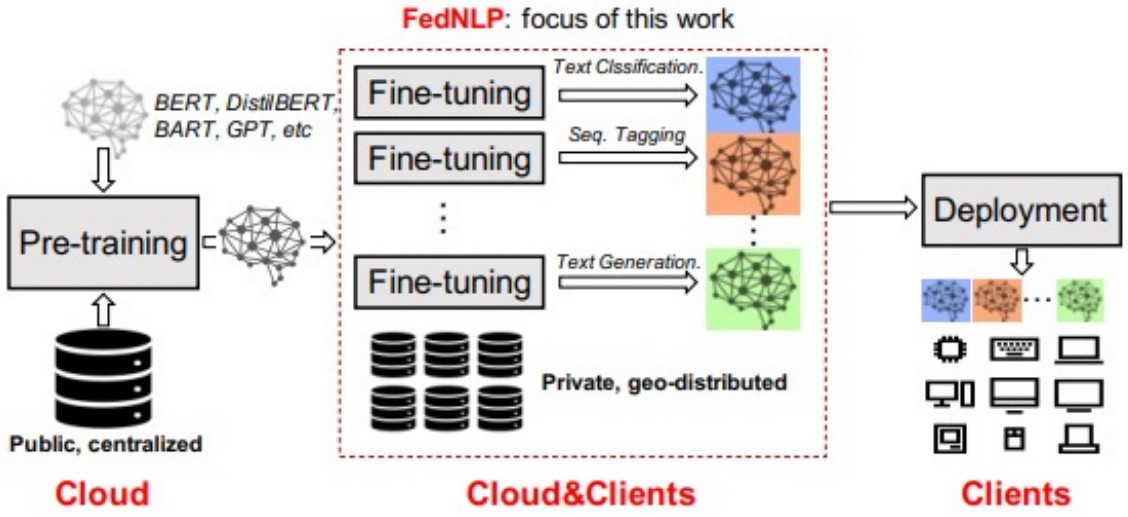
Experiment: Movie review classifier using DistilBERT

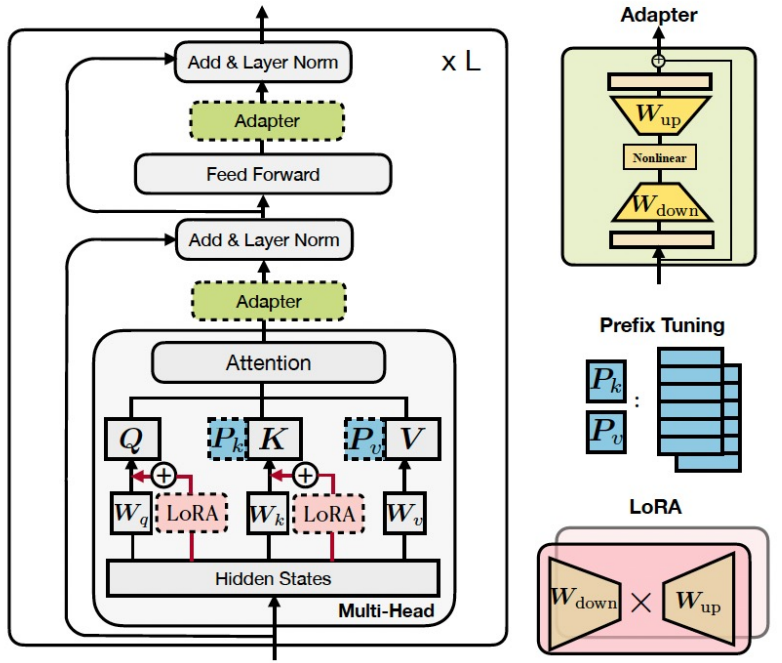


Source: Understanding Parameter-Efficient Finetuning of Large Language Models: From Prefix Tuning to LLaMA-Adapters URL. https://sebastianraschka.com/blog/2023/llm-finetuning-llama-adapter.html

FedAdaptor

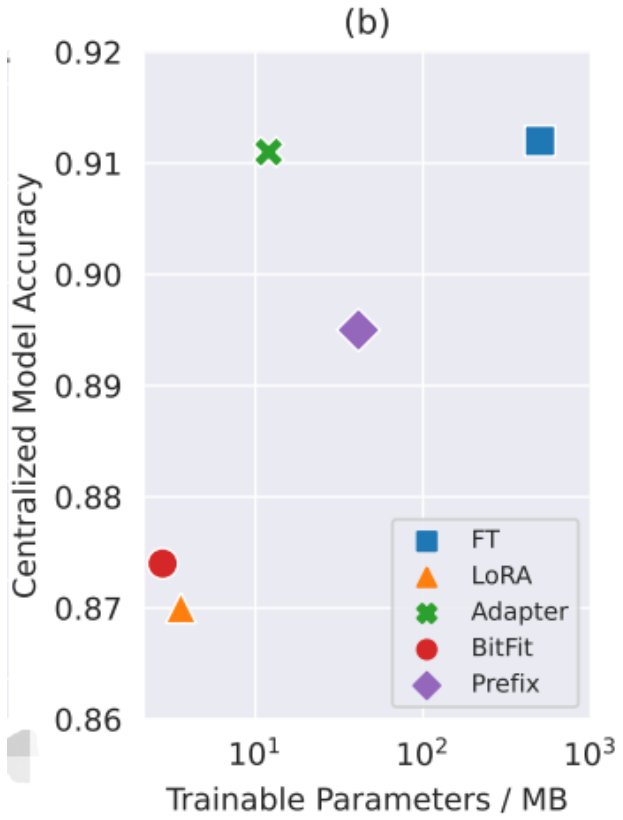
This paper proposes a method to progressively upgrade the adapter configuration through a training session and continuously profile future adapter configurations by allocating participant devices to trial groups. It results reduce FedNLP's model convergence delay to no more than several hours, which is up to 155.5× faster compared to vanilla FedNLP.





Source: Cai D, Wu Y, Wang S, et al. Autofednlp: An efficient fednlp framework[J]. arXiv preprint arXiv:2205.10162, 2022.

FedPETuning



This paper introduces various parameter-efficient tuning (PETuning) method for federated learning and provides a holistic empirical study to show the overall communication overhead can be significantly reduced by locally tuning and globally aggregating lightweight model parameters while maintaining acceptable performance in 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.

FedPrompt

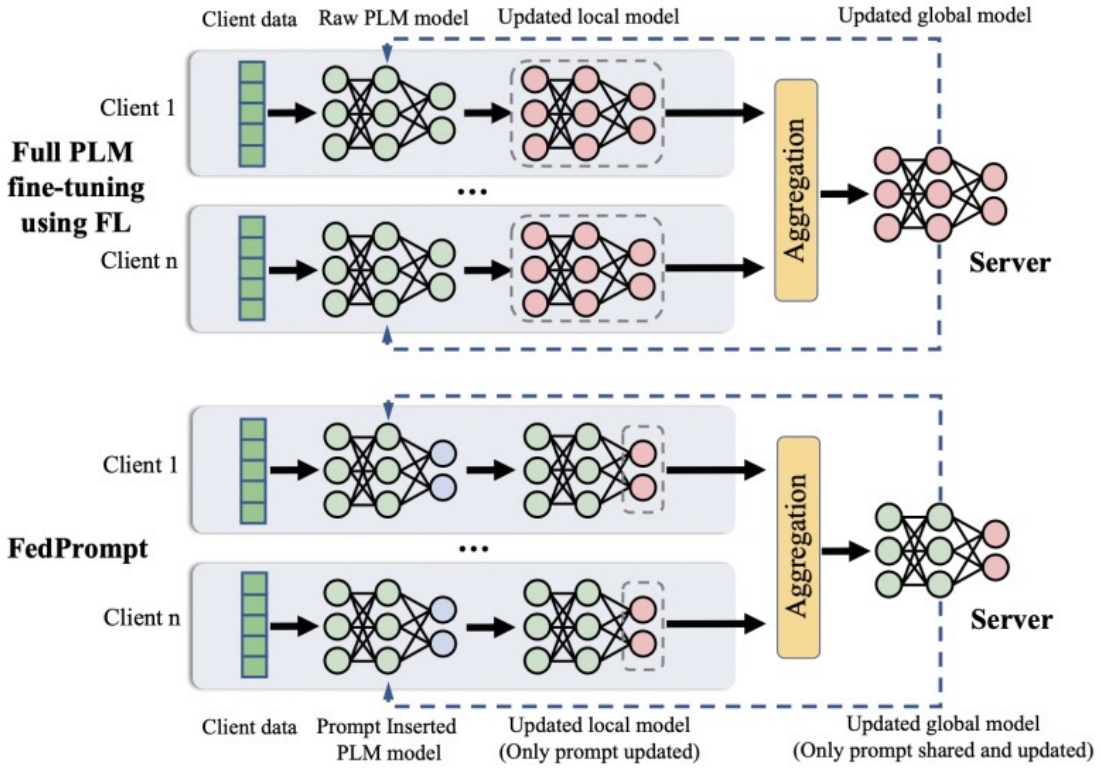


Figure 2: Structure of FedPrompt and full PLM fine-tuning using FL. The above one is full PLM fine-tuning using FL, all of the parameters (framed pink nodes) need to be updated. The bottom one is FedPrompt, only soft prompt parameters (framed pink nodes) need to be updated, aggregated (in server) and distributed.

This paper proposes a new method that tunes some soft prompts without modifying pre-trained language models (PLMs) and has achieved excellent performance in natural language processing (NLP) tasks.

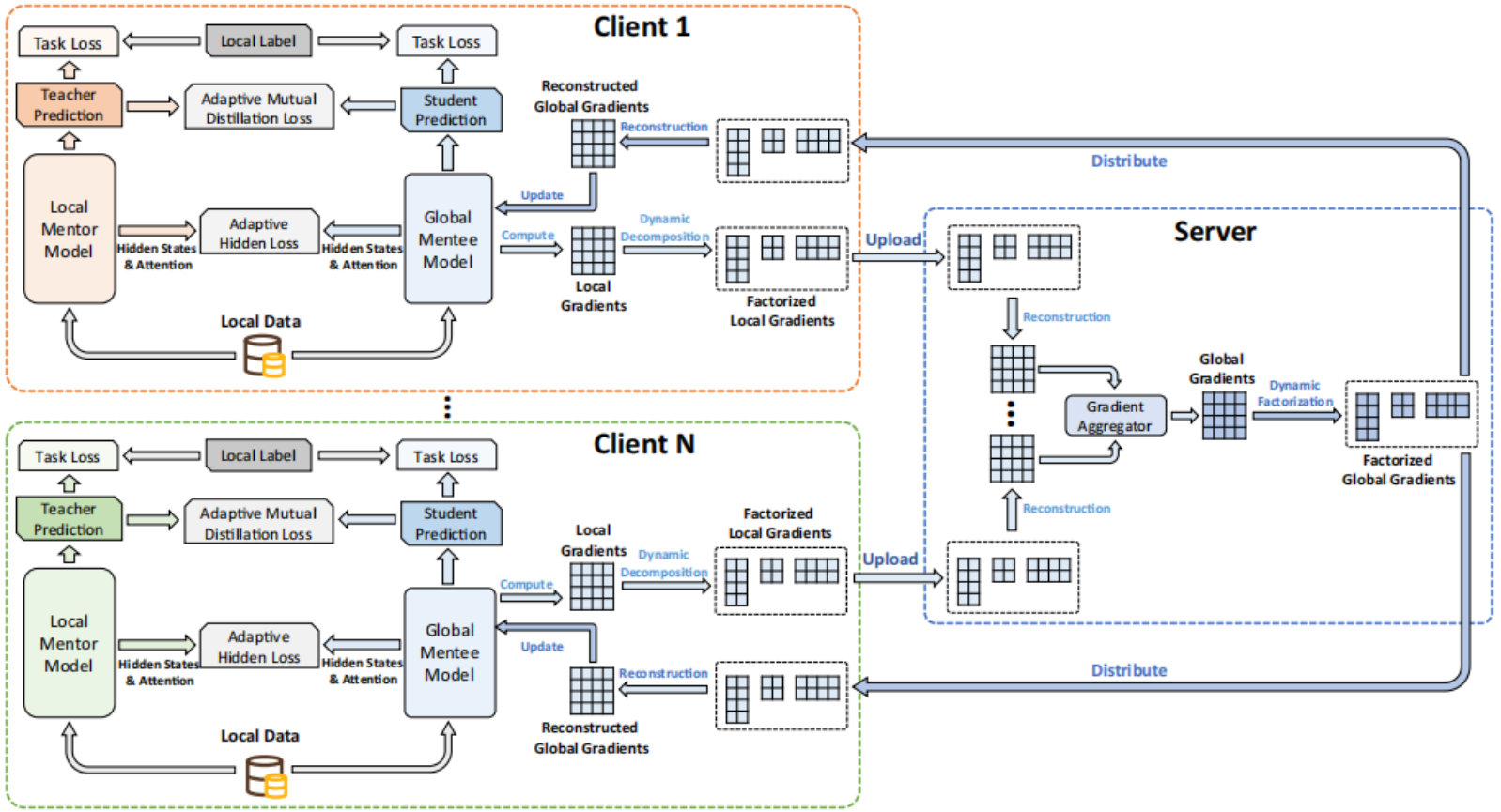
Model	FL Method	*ACC*	Comm. Cost	Ratio
BERT	FedPrompt	90.16	**0.016M**	**0.014%**
	Fine-tuning	91.02	109.530M	100.000%
ROBERTA	FedPrompt	92.43	**0.016M**	**0.013%**
	Fine-tuning	93.57	124.714M	100.000%
T5	FedPrompt	92.69	**0.015M**	**0.007%**
	Fine-tuning	93.79	222.919M	100.000%

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.

FedKD

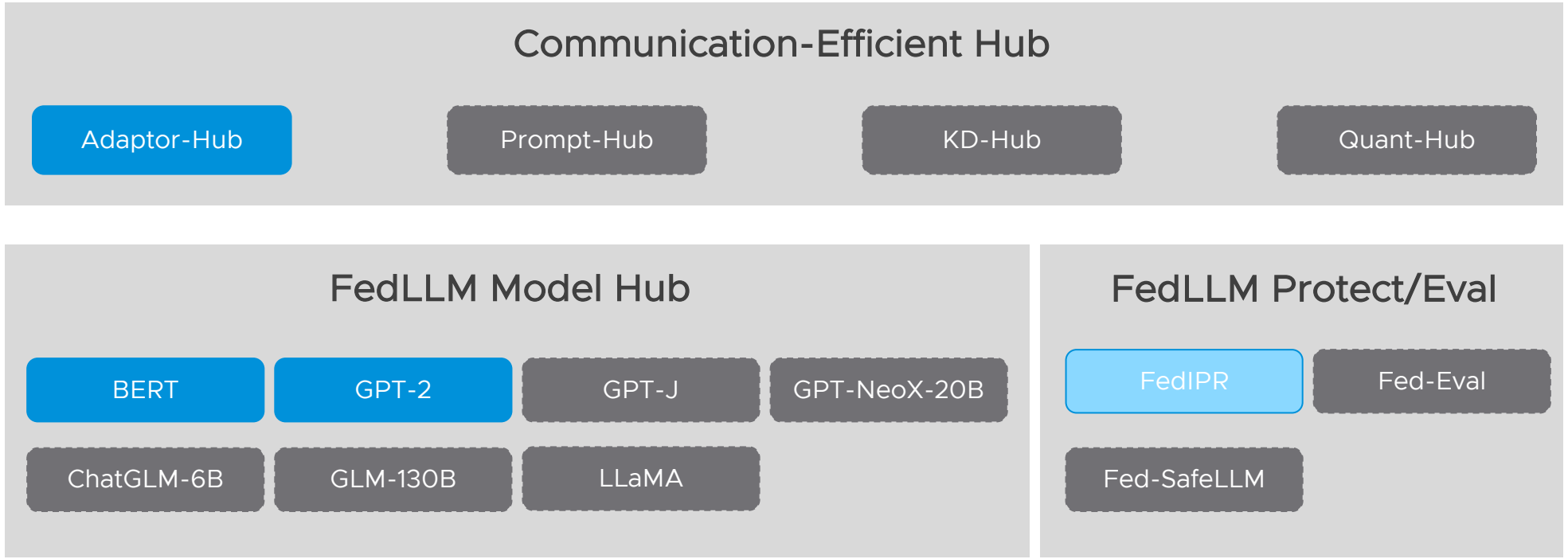
Knowledge distillation is a technique to transfer knowledge from a large teacher model to a small student model, which is widely used for model compression.

Local data is used to train local Teacher and global Student models. Both models learn from local labeled data as well as each other's predictions and hidden results. Upload Student model parameters to Server.



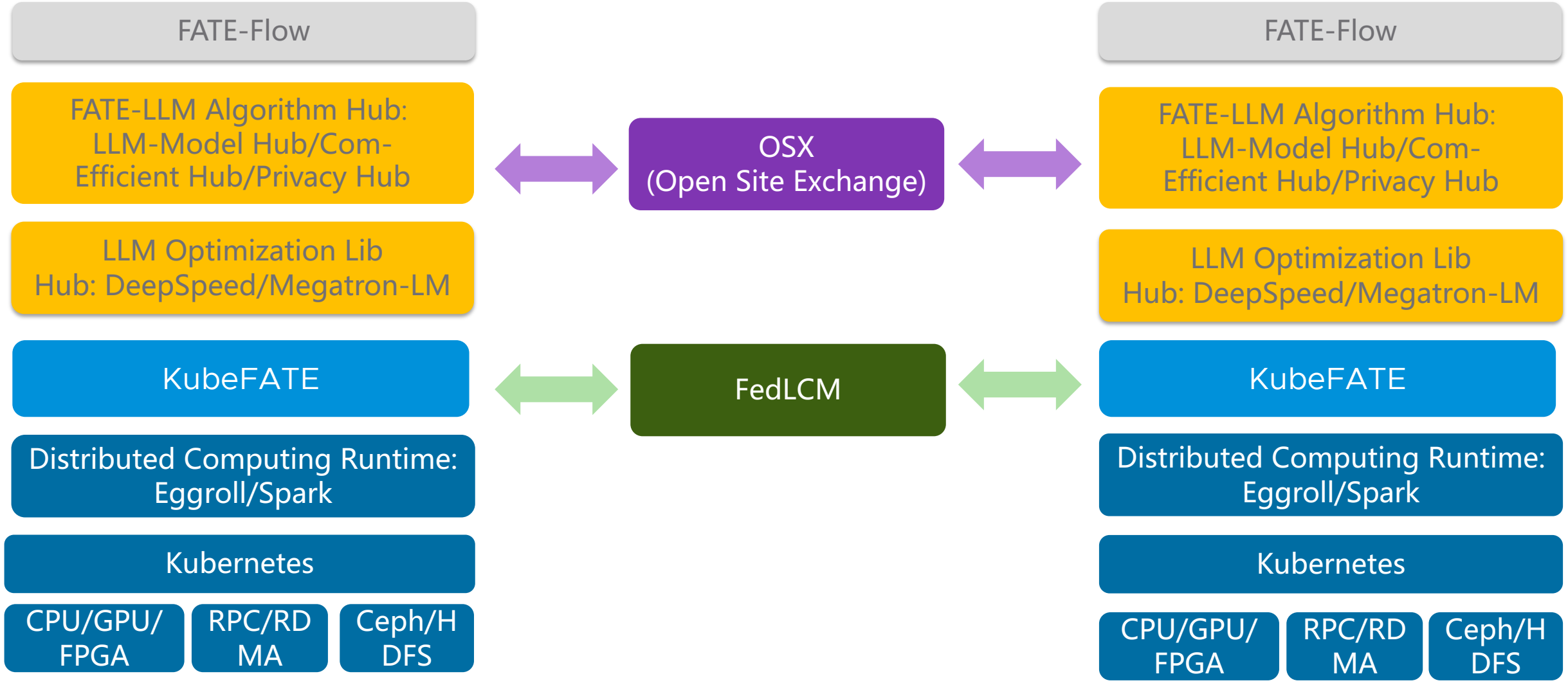
Source: Chuhan Wu, Fangzhao Wu, et al. FedKD: Communication Efficient Federated Learning via Knowledge Distillation. arXiv:2108.13323v2

FATE-LLM: FATE Federated Large Language Model

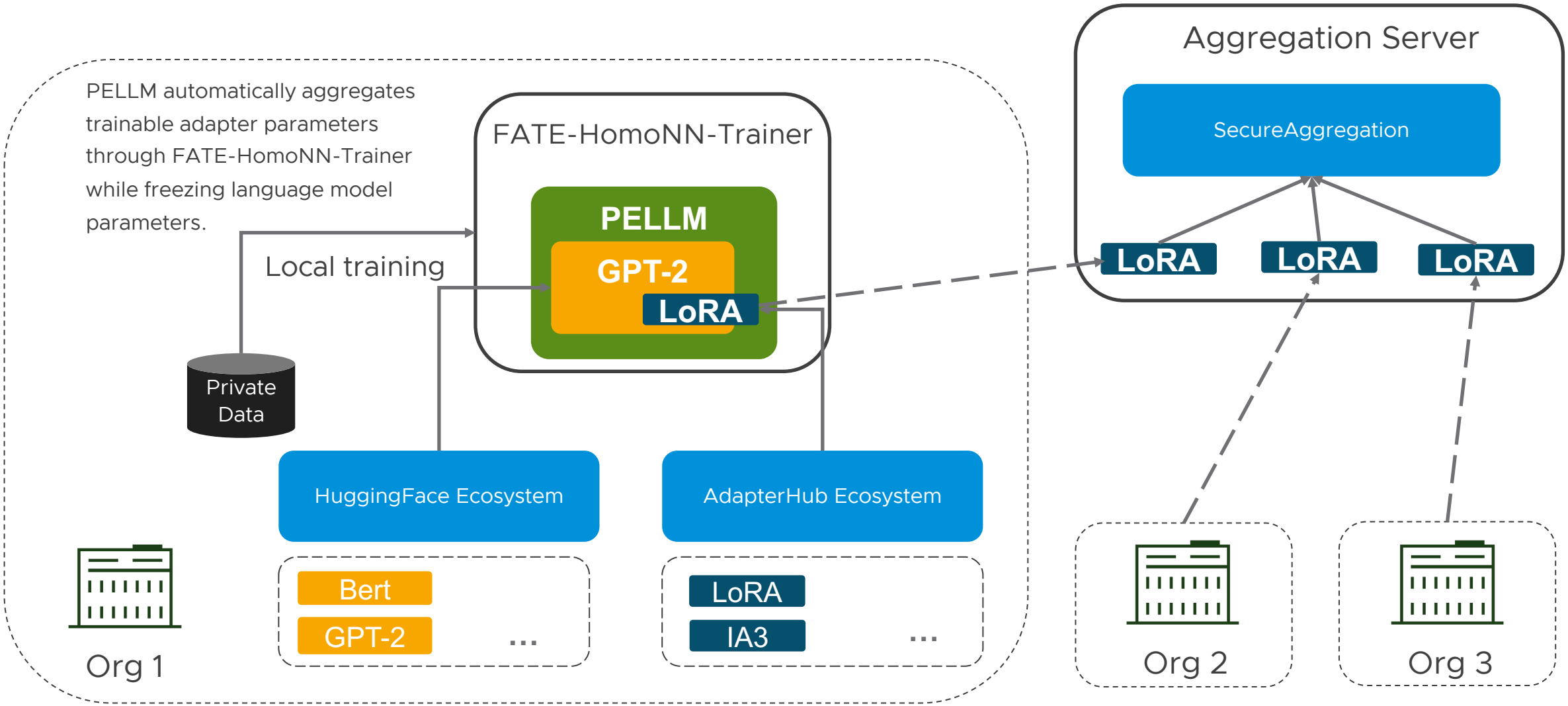


- Multiple clients can perform horizontal FL through FATE's built-in support of pre-trained model and use private data for large-scale model fine-tuning;
- Support 30+ participants for a collaborative training

FATE-LLM high-level architecture



FATE-LLM Algorithms Design – first version



FATE-LLM: Performance

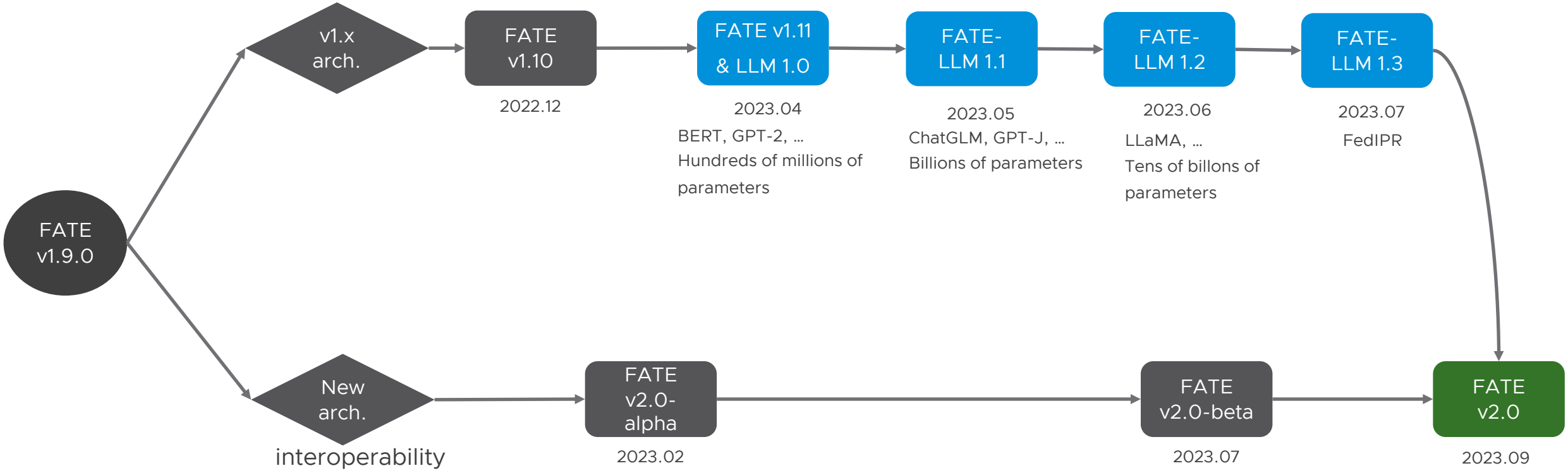
- **Scenario:** Horizontal Federated Learning Scenario for a text sentiment classification using IMDB
- **Task Type:** Text Sentiment Classification Task
- **Participants:** 2
- **Dataset:** IMDB (25,000 records)
- **Hyperparameters:** batch_size = 64, padding_length = 200
- **Foundation Model:** GPT-2
- **Experiment setup:** Each participant uses 2x V100 32GB and all participants are in a local area network environment

Adapter	Training Time	Percentage of Baseline
Houslby Adapter	464.1	27.41%
Pfeiffer Adapter	445	26.28%
Parallel Adapter	560.875	33.13%
Invertible Adapter Houlsby	456	26.93%
Invertible Adapter Pfeiffer	453.8	26.80%
LoRA Adapter	410.67	24.26%
IA3 Adapter	412.76	24.40%
Compacter Adapter	446.55	26.35%
BaseLine(Finetune)	1693	100%

Unit: second/each epoch

Source: FATE Community, URL. https://github.com/FederatedAI/FATE/blob/master/doc/federatedml_component/fate_llm.md

FATE-LLM Roadmap



* Source from FATE Community. The mentioned dates are for informational purposes only and may vary.

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE