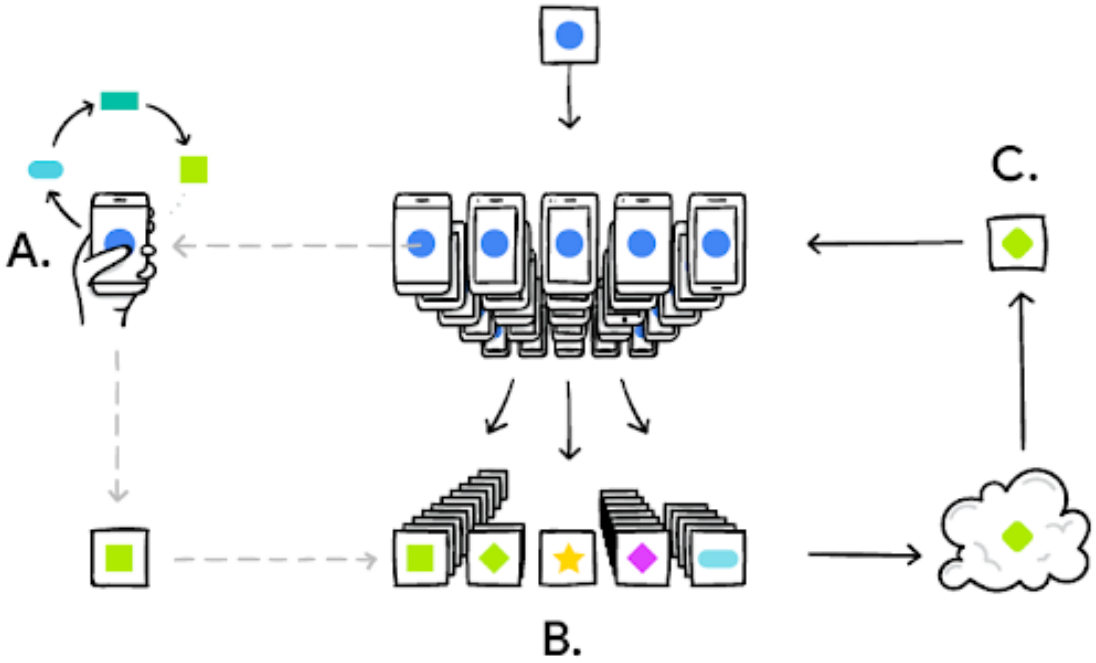
FATE-LLM: 当联邦学习遇到大型语言模型

王方驰

E2G, VMware AI Labs, OCTO

July 08, 2023

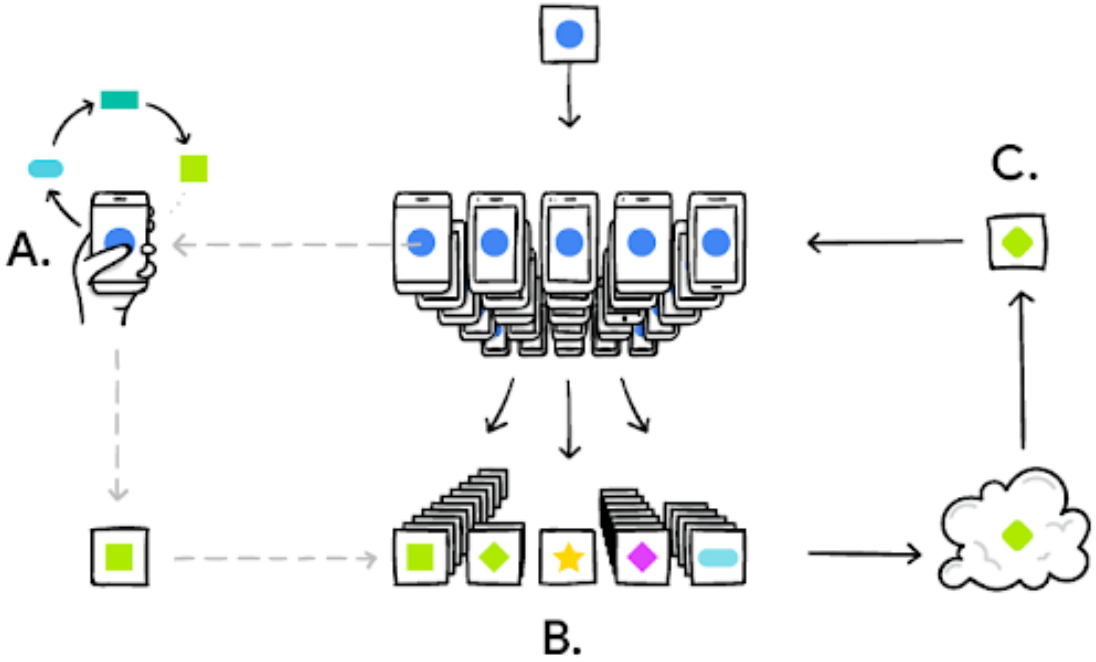
What is Federated Learning?

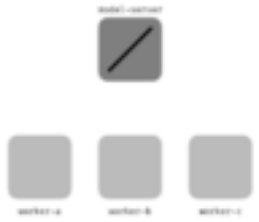
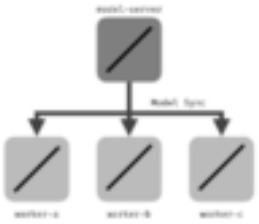
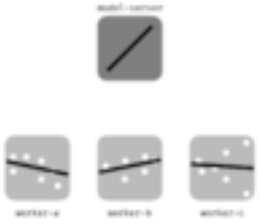
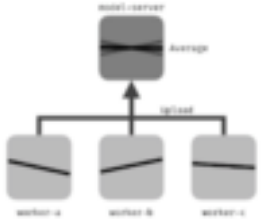


Sources:

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

What is Federated Learning? (Horizontal Federated Learning)



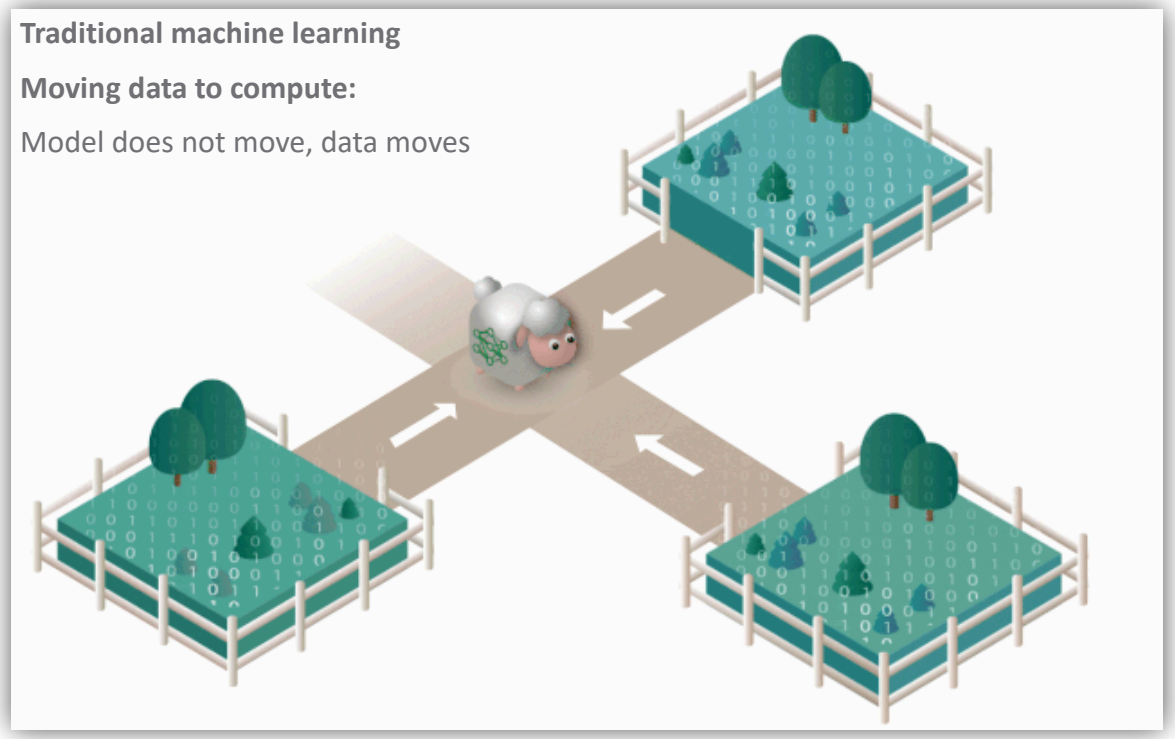
Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data

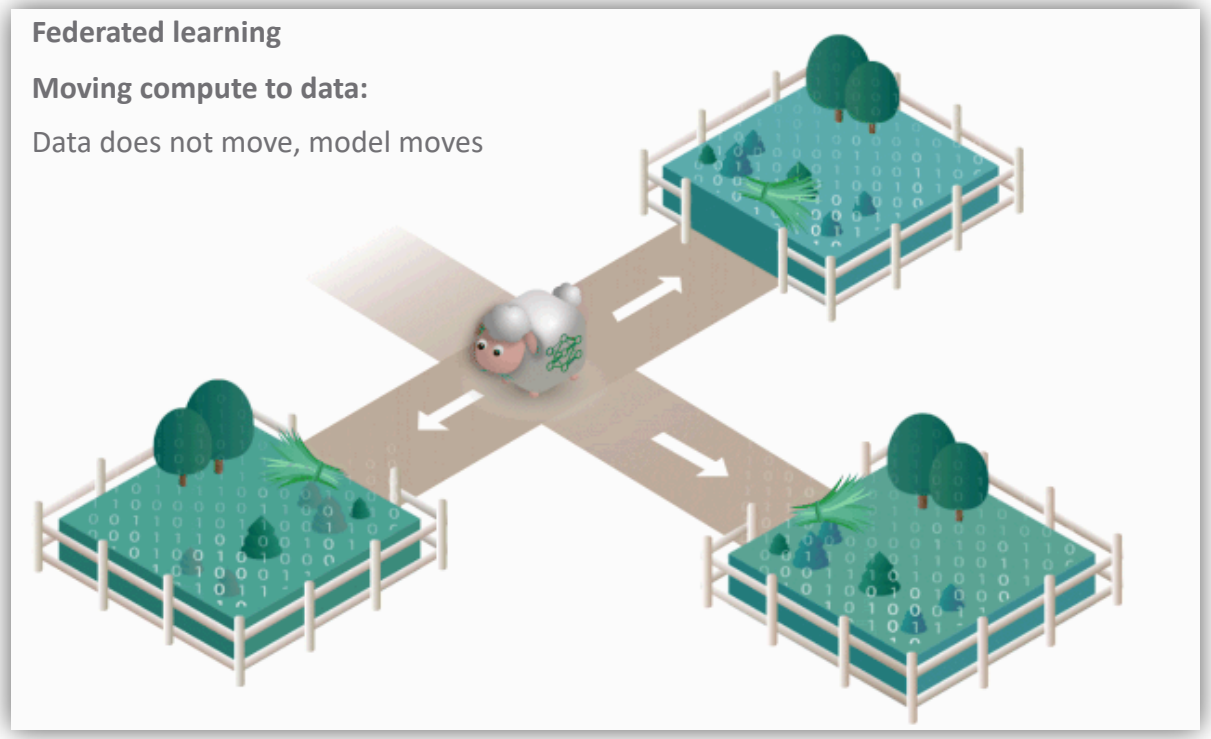
Sources:

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)

Paradigm shift of Federated Learning: Moving compute to data

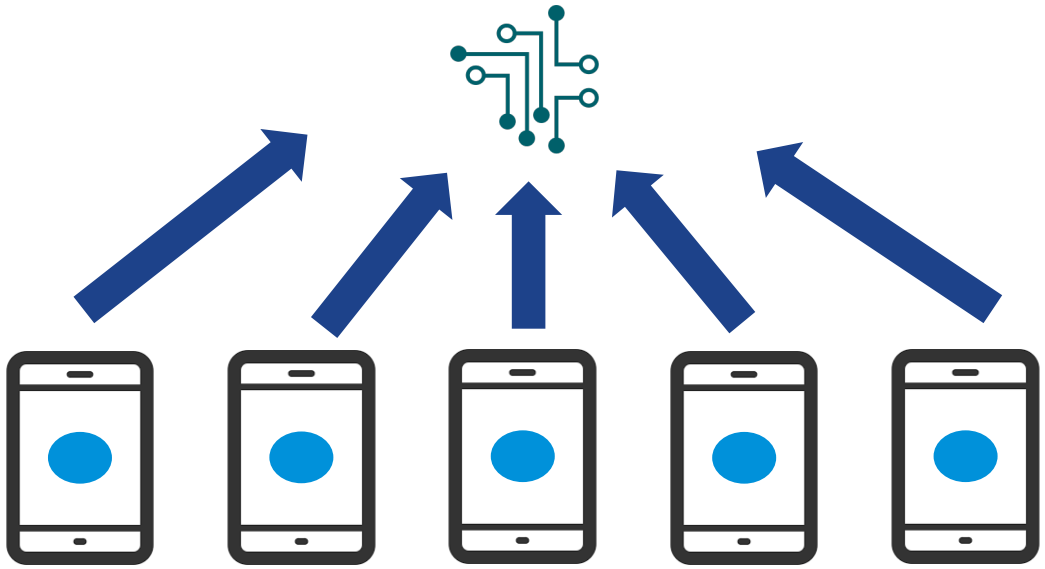
- Optimized model built from data of multiple organizations or from different places
- Preserve data privacy and confidentiality
- Communication cost reduction



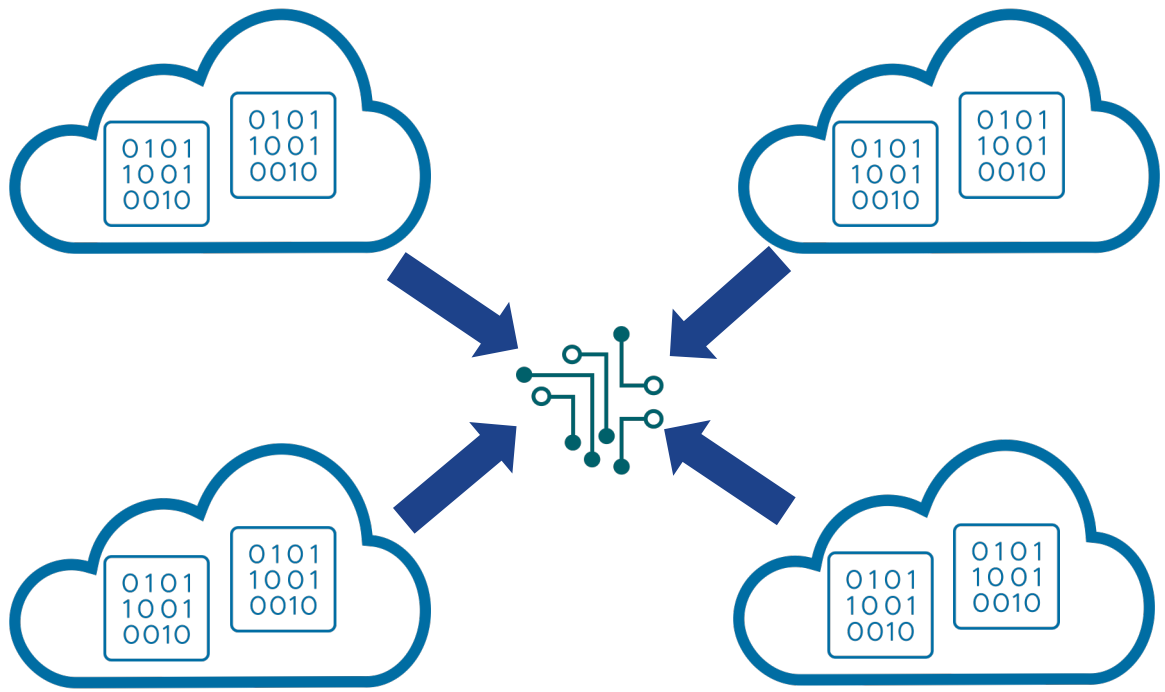


Source: Federated Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning), Qiang yang , *et al.*

From device(s) to enterprise(s)

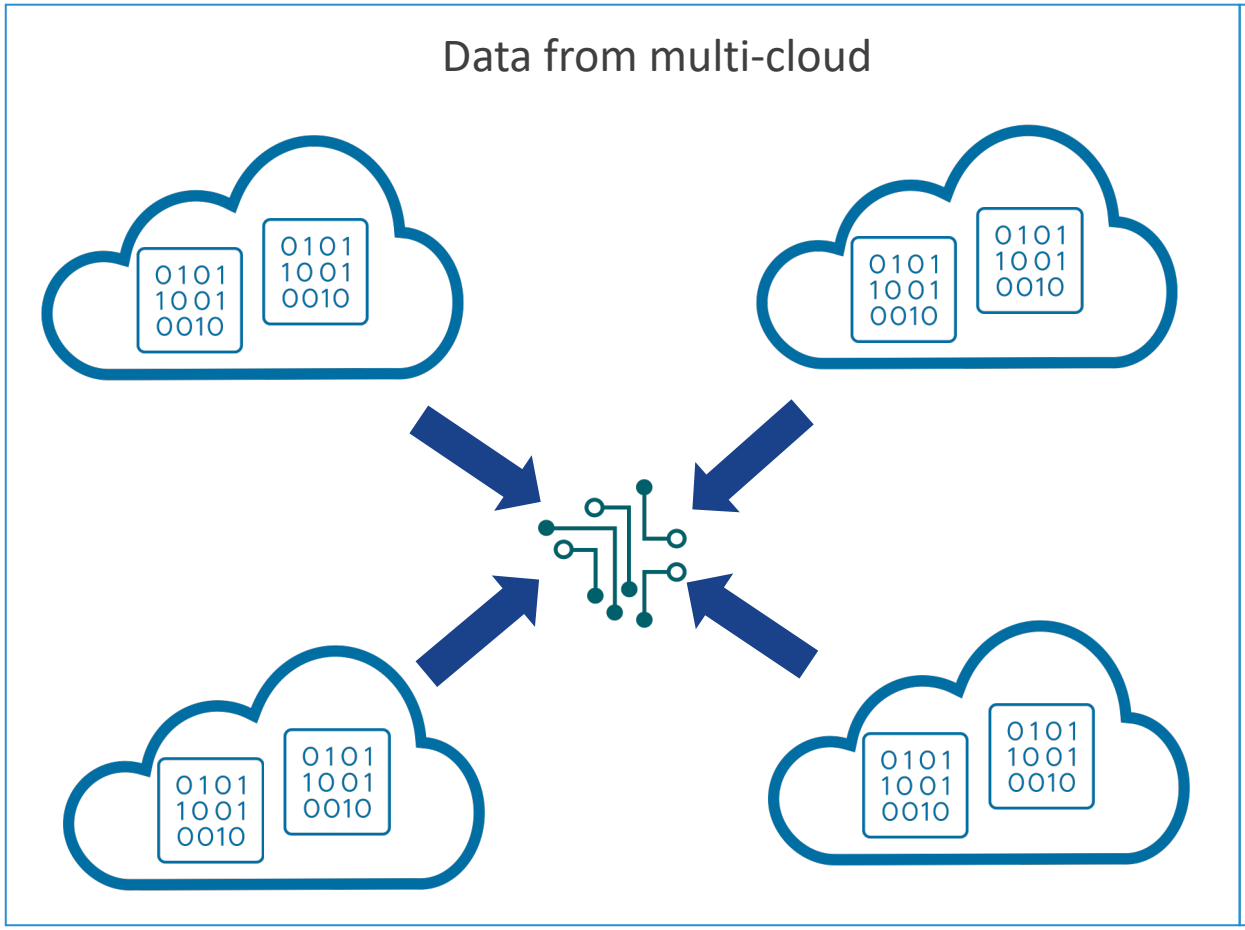


FL for a devices

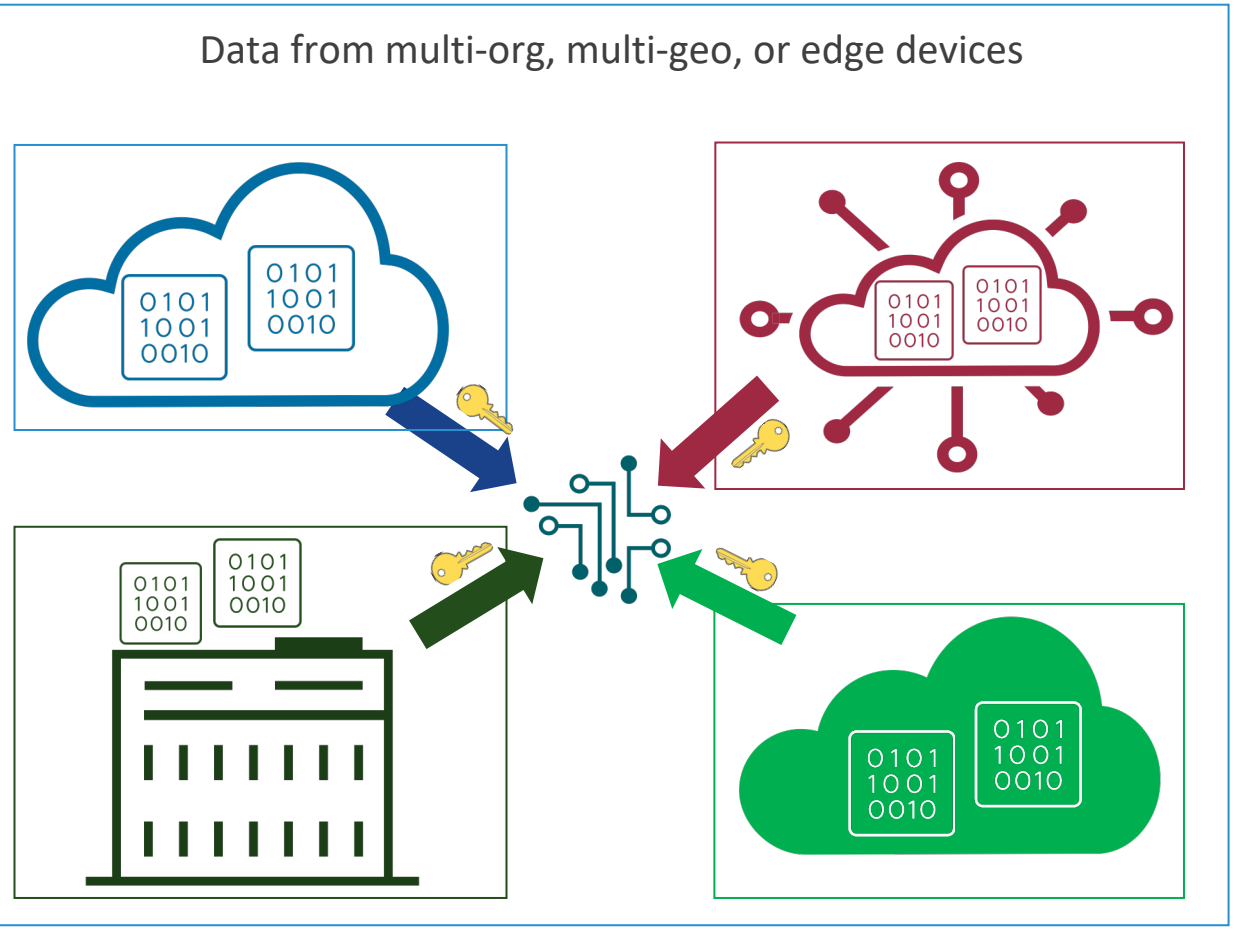


FL for an enterprise

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

- 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

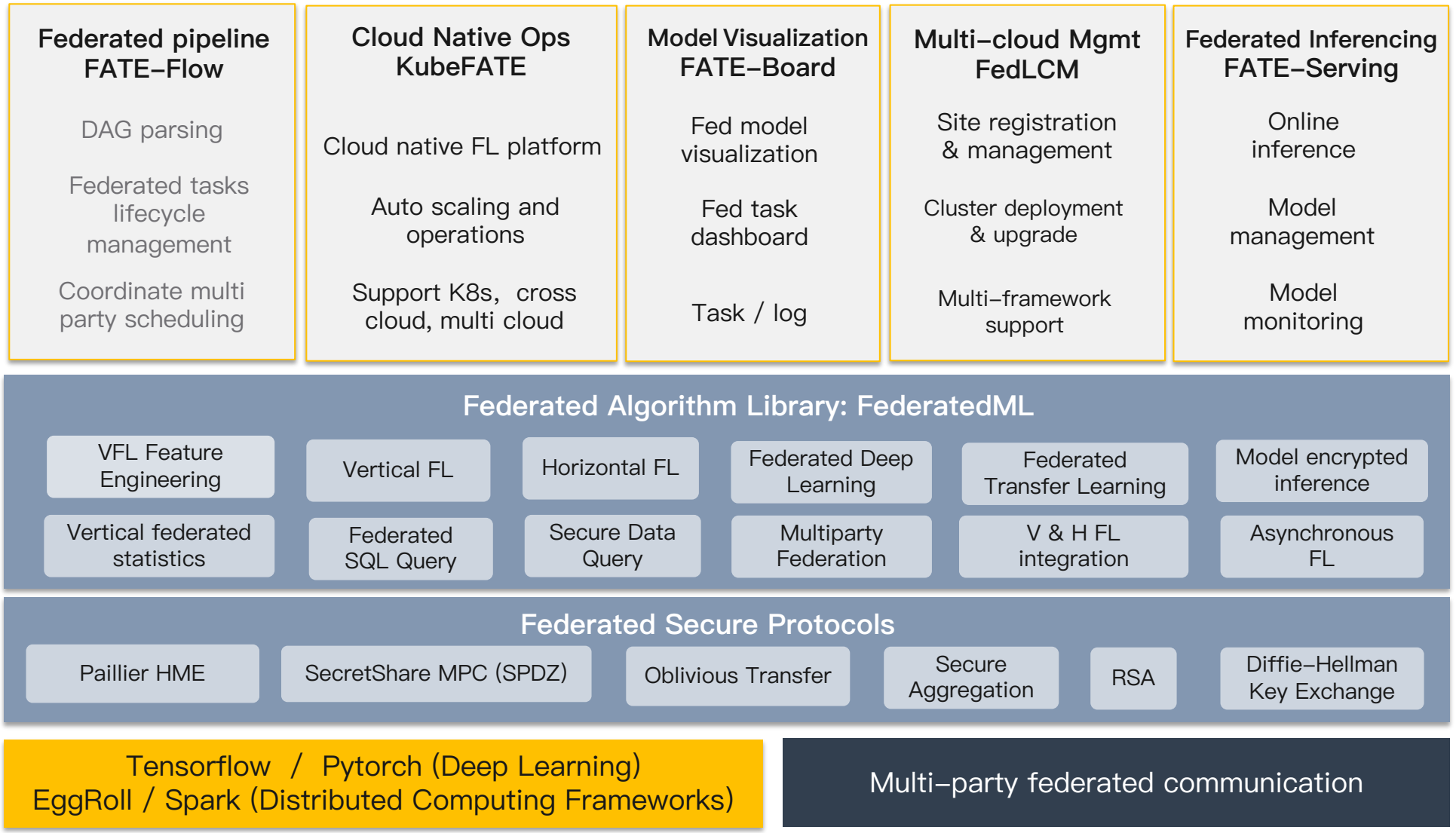
VMWARE, INC.

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

FATE Overview



FATE Community

FATE开源社区**根据Linux Foundation的规定**建立了完善的社区治理组织架构，包含TSC Board、TSC Maintainer、技术专委会、运营专委会、安全专委会、成员单位，特别兴趣小组 (SIG)

技术指导委员会

FATE开源社区的最高技术领导机构，负责制定长期发展规划，指导各专业委员会工作实施

开发专委会

FATE开源社区的技术开发机构，推动各版本的稳定、高质量更新及持续技术创新

运营专委会

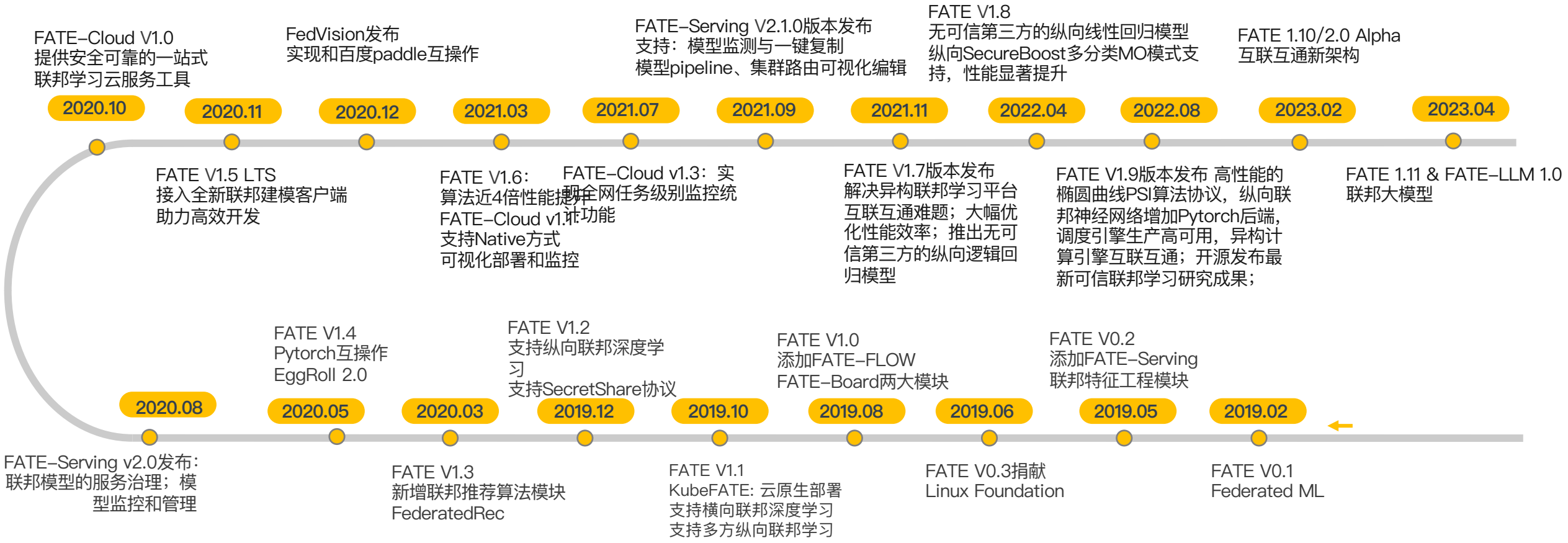
FATE开源社区的运营统筹工作机构，设置行研、自媒体运营、活动、公关及外联五个小组，全方位开展社区运营相关工作

安全专委会

FATE开源社区的安全合规工作机构，负责社区技术安全监督、开源合规管理，协同多方资源为社区生态安全保驾护航

FATE Built-in Algorithms for Major Industry Scenarios

自开源至今，FATE已迭代40余个版本，联邦算法组件已发展至30余个，实现工业界主流场景算法全覆盖和工业界主流多方安全计算协议全覆盖。



What is LLM? – All from here…

Attention Is All You Need

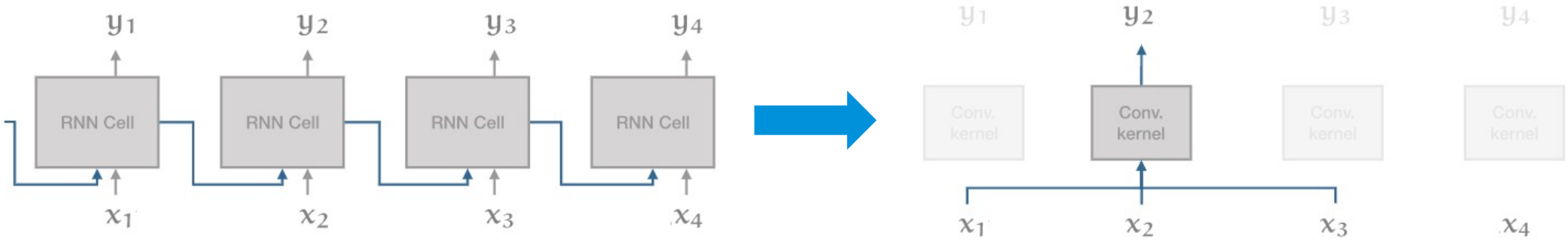
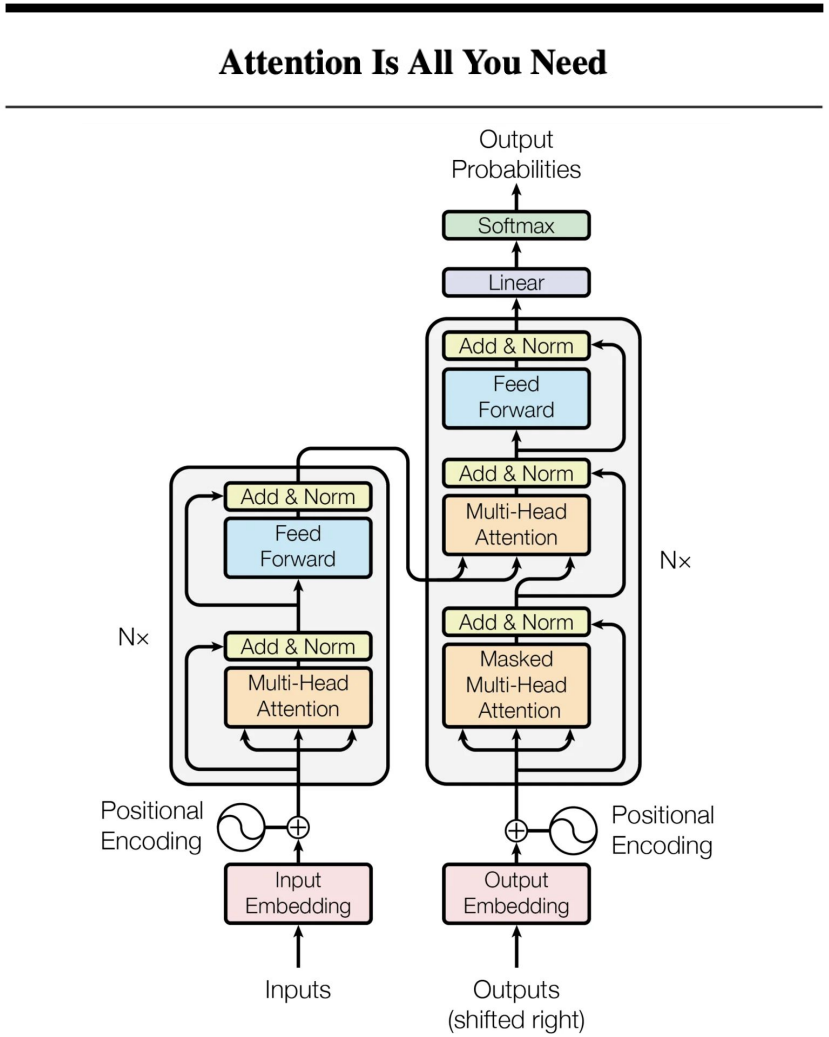


Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

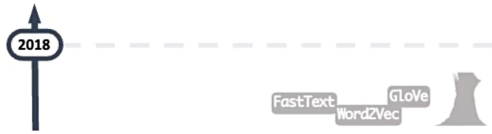
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Source: Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia, “Attention is all you need” , 2017.

What is LLM? – All from here…



Source: Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia, “Attention is all you need” , 2017.



Source: Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, Xia Hu, “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond” , 2023.

What is LLM? - Multiple tasks support and emergent abilities

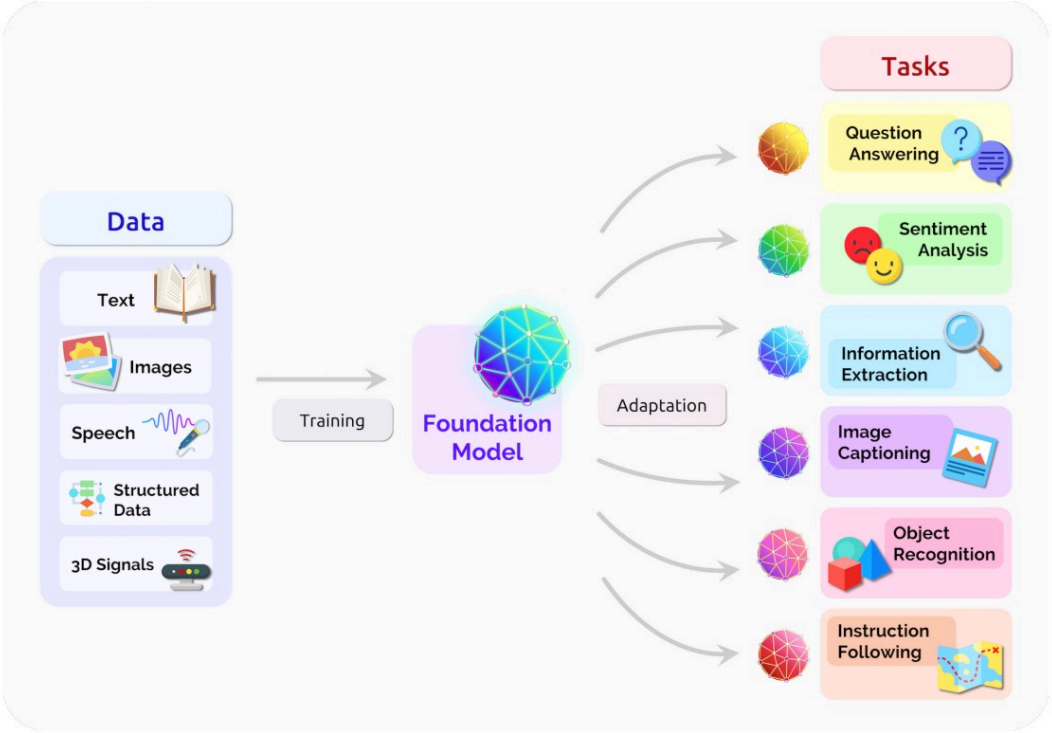
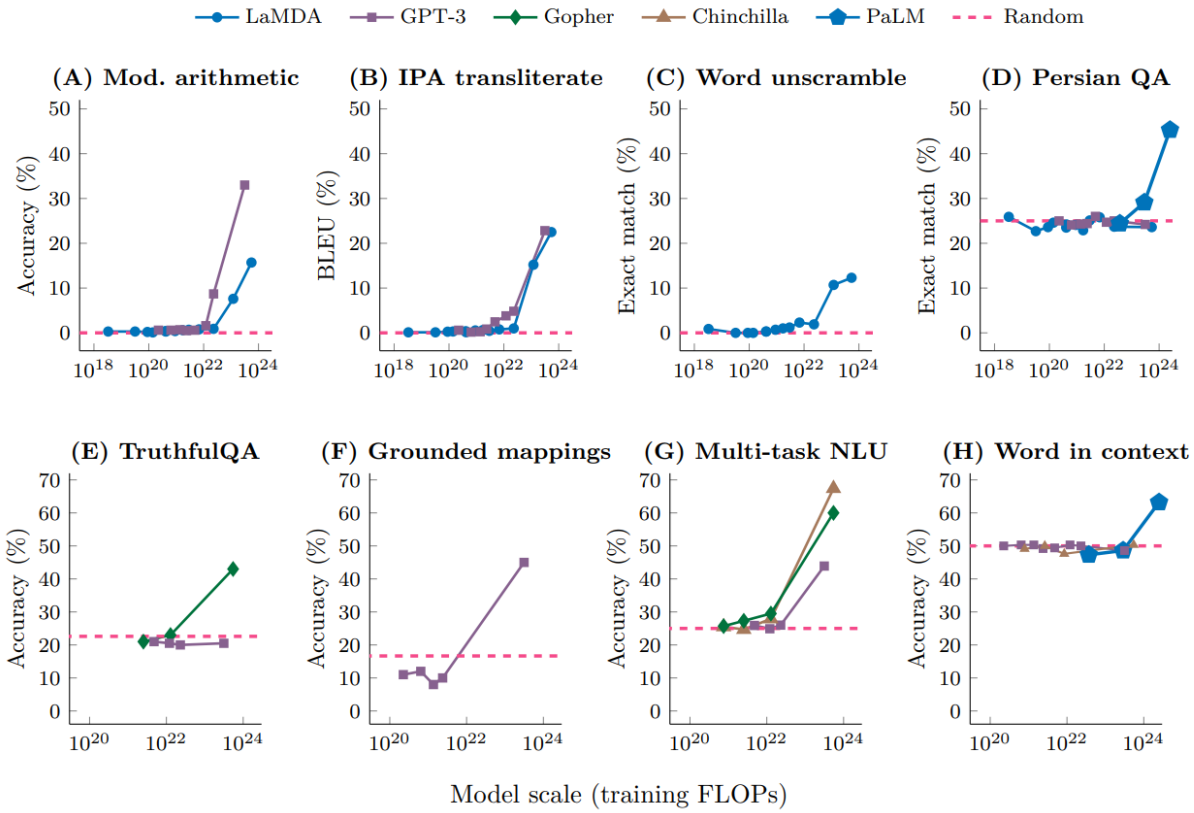


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.



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.

What is LLM? – LLMs' explosive growth…

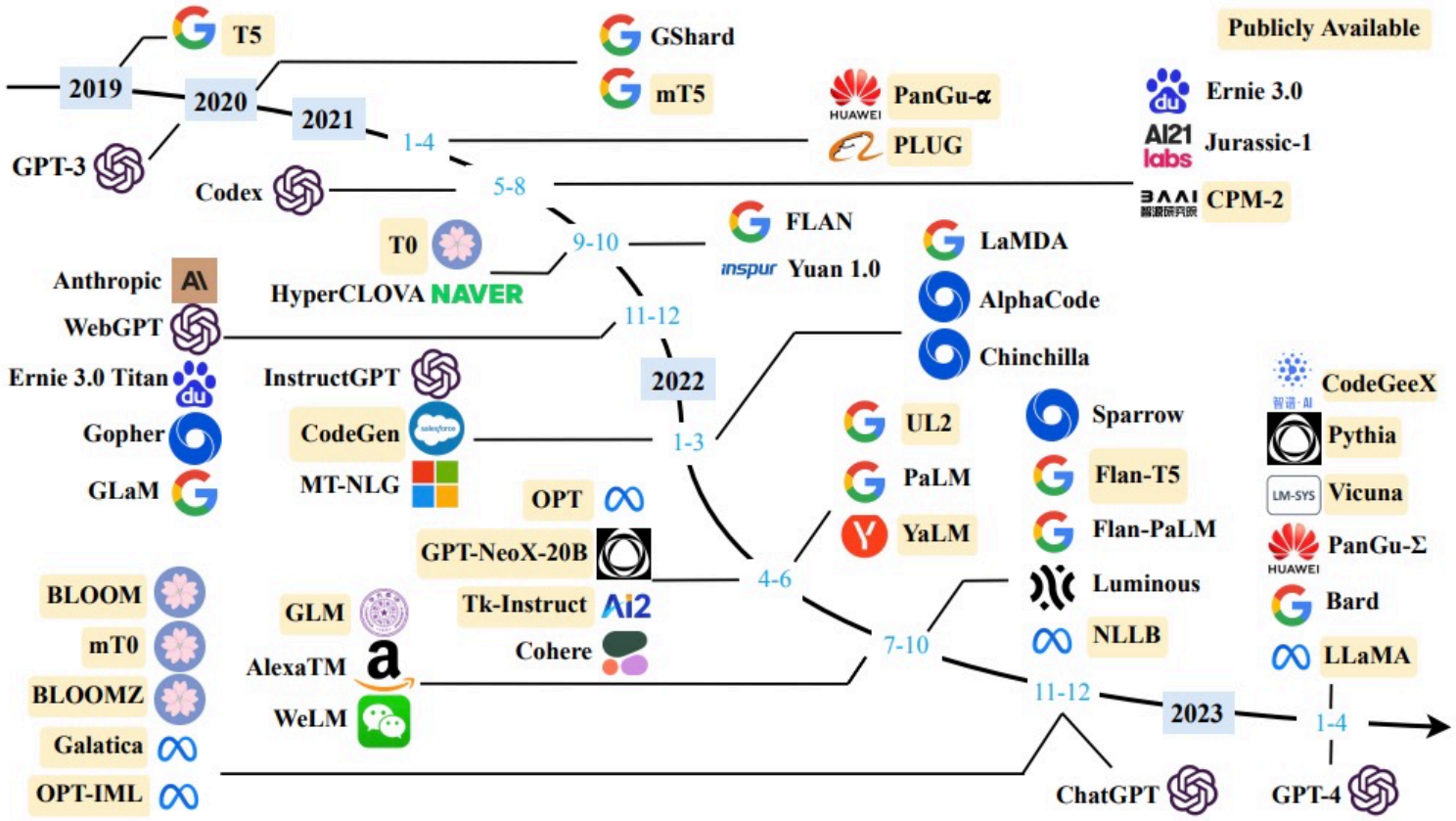
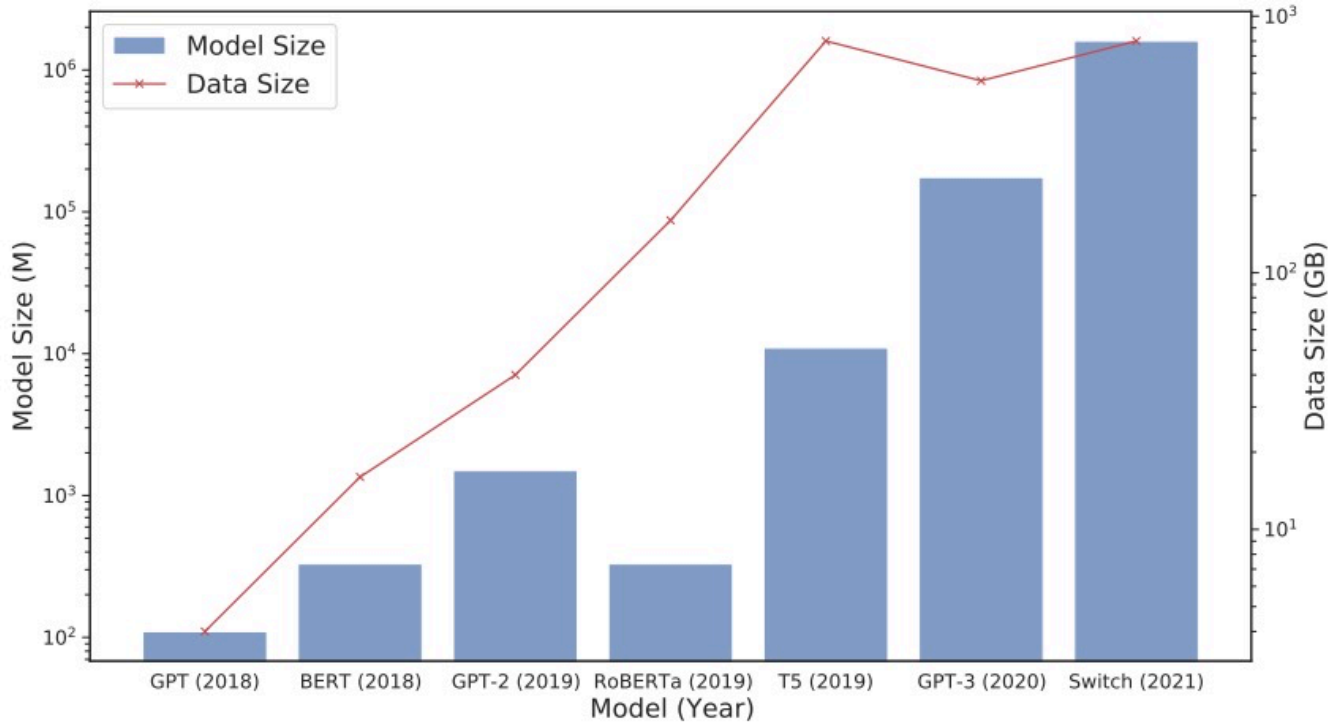


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.*, the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of 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

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:

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

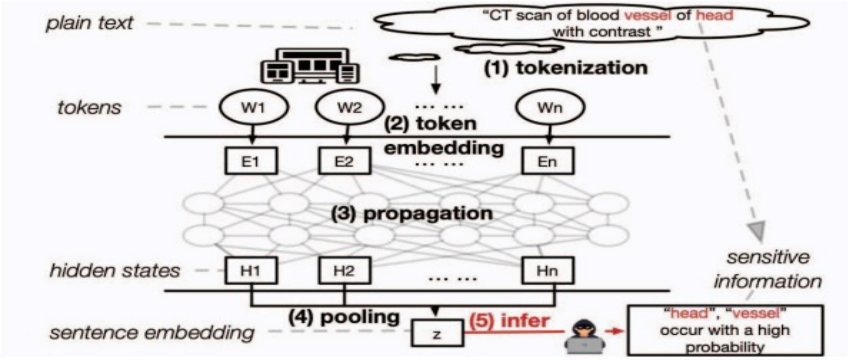


Fig. 1. General-purpose language models for sentence embedding and the potential privacy risks. The red directed line illustrates the discovered privacy risks: the adversary could reconstruct some sensitive information in the unknown plain texts even when he/she only sees the embeddings from the general-purpose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

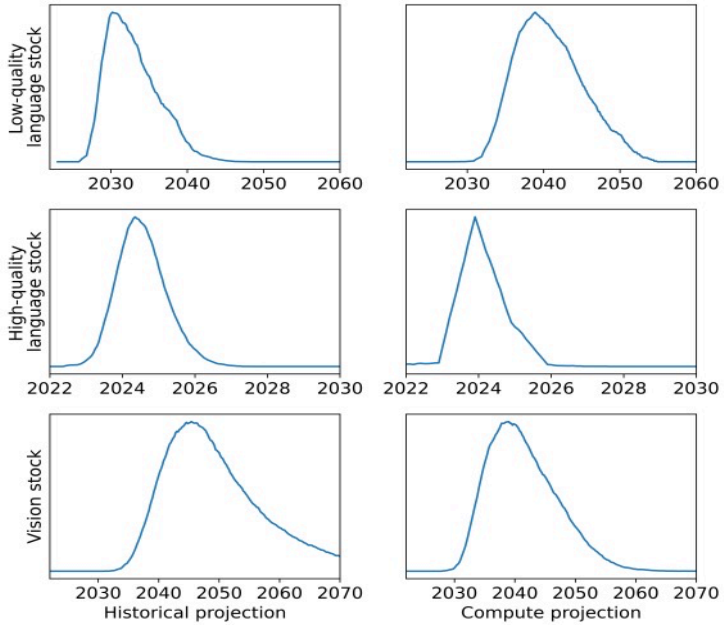


Fig. 6: Distribution of exhaustion dates for each intersection of the data availability trend and data consumptiin trend. Note that the time scale is different for each kind of data.

	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.

FedNLP

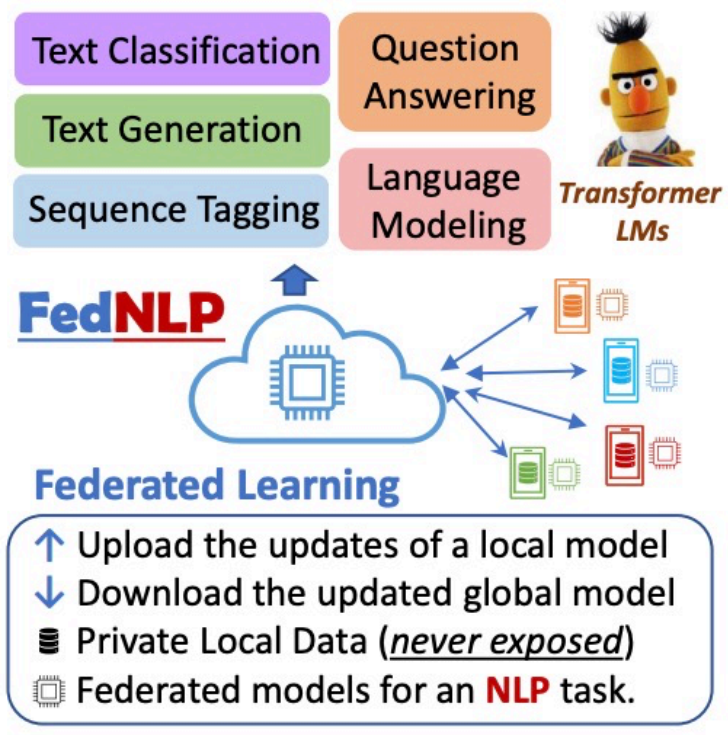


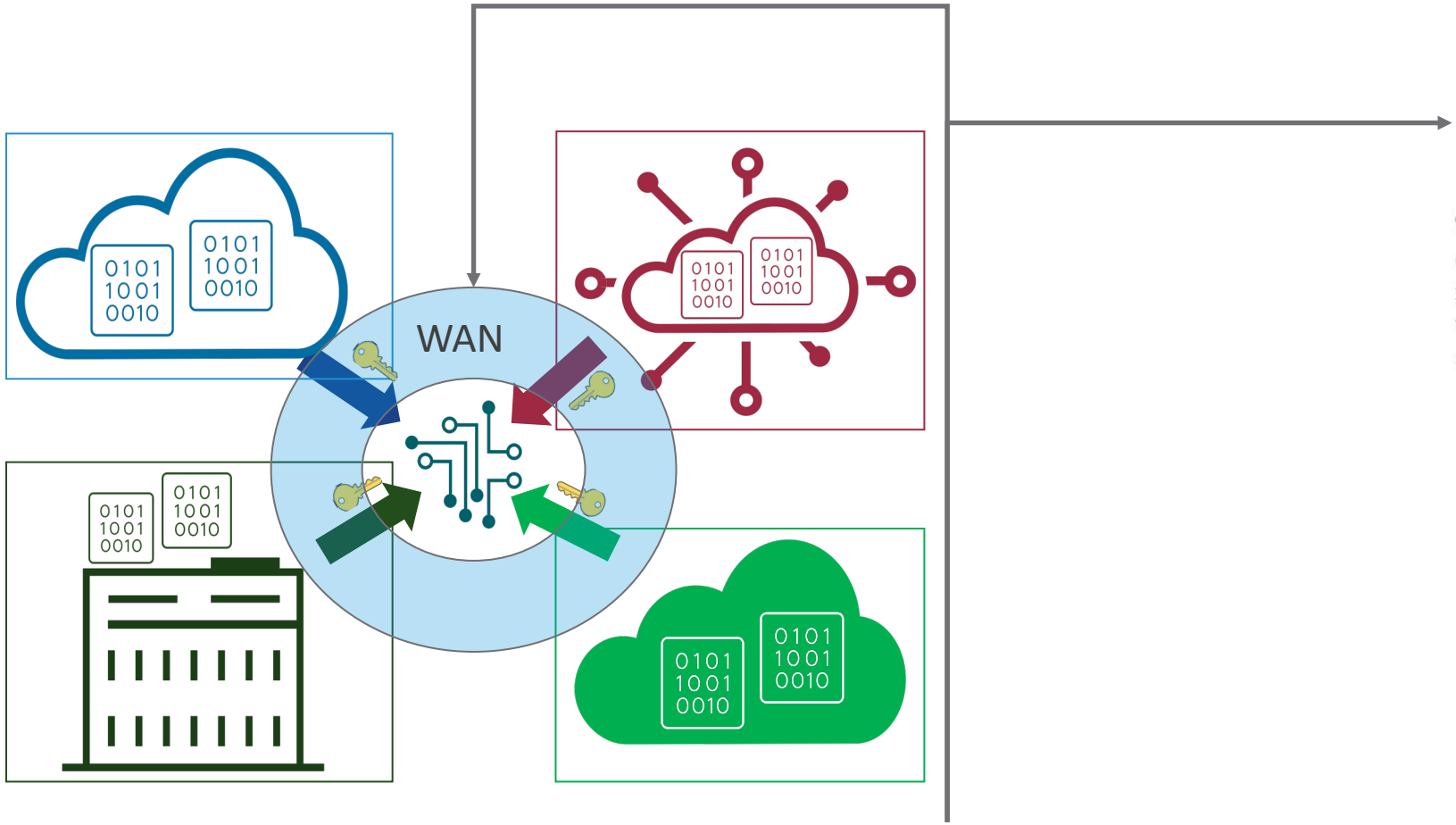
Figure 1: The FedNLP benchmarking framework.

This paper presents:

1. A federated architecture for collaborative exploration of NLP applications in a shared environment, facilitating the development of specialized FL methods for specific NLP tasks or general-purpose models;
2. The sample workflow and architecture of federated NLP;
3. Solutions to address challenges such as Non-IID data (heterogeneous data) and heterogeneous settings/devices.

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

Challenges



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.

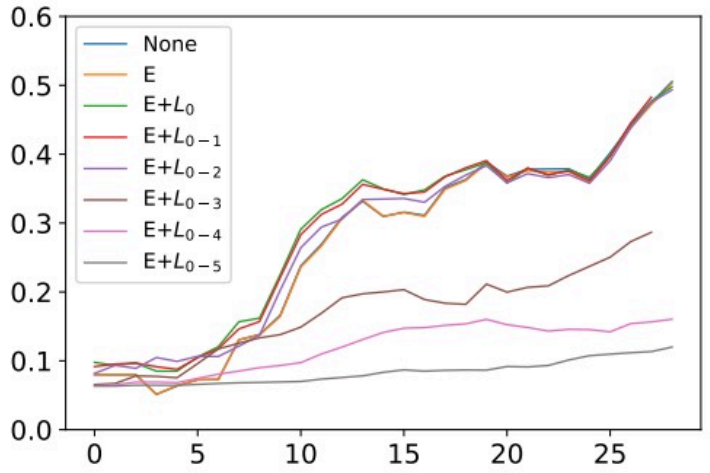
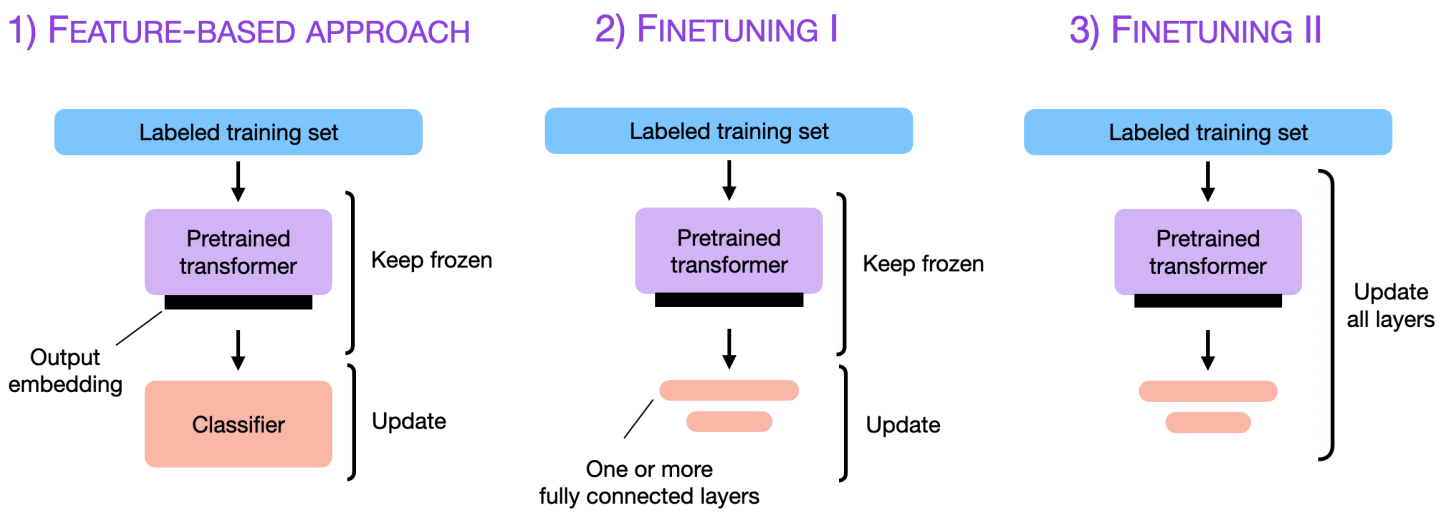


Figure 6: Testing `FedOPT` with `DistilBERT` for `20News` under different frozen layers.

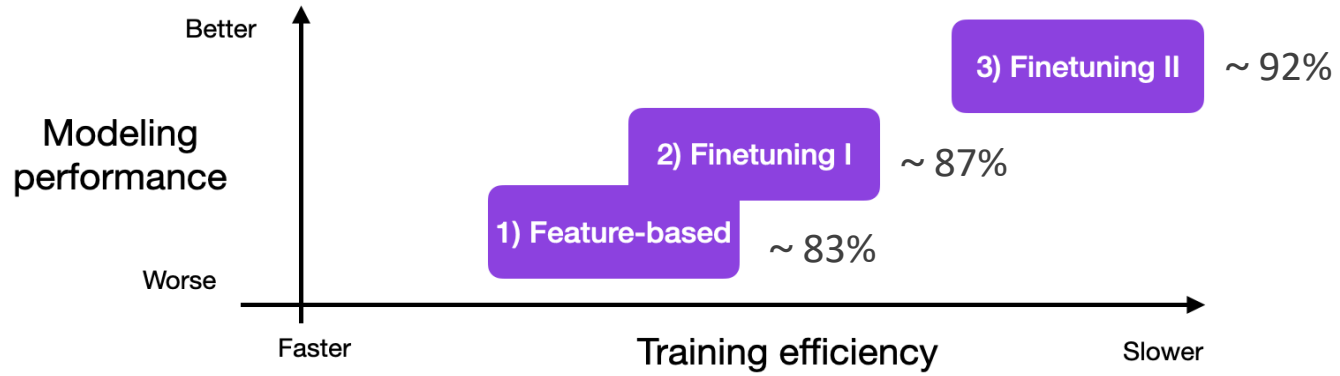
How to exchange the **LARGE** models (weights/gradients) between participants in WAN?

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

Fine-tuning

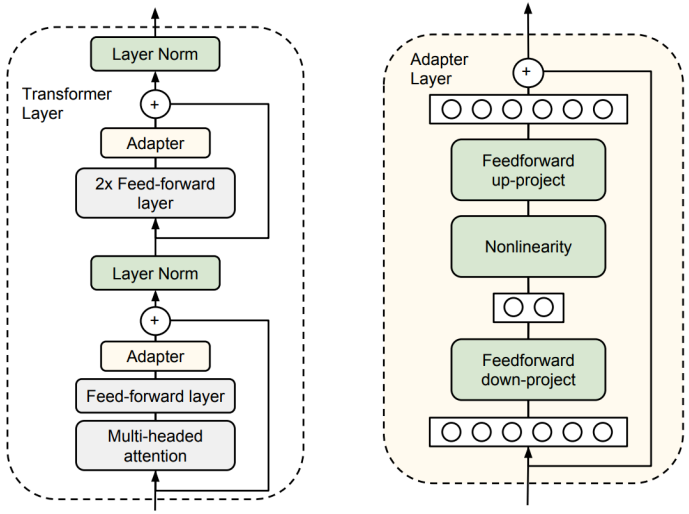


Experiment: Movie review classifier using DistilBERT

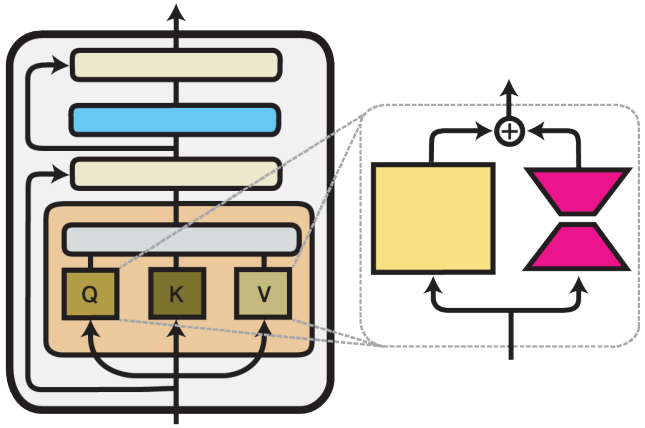


Source: Understanding Parameter-Efficient Finetuning of Large Language Models: From Prefix Tuning to LLaMA-Adapters URL. https://sebastianraschka.com/blog/2023/llm-finetuning-llama-adapter.html

Parameter-efficient Fine-tuning (PEFT)



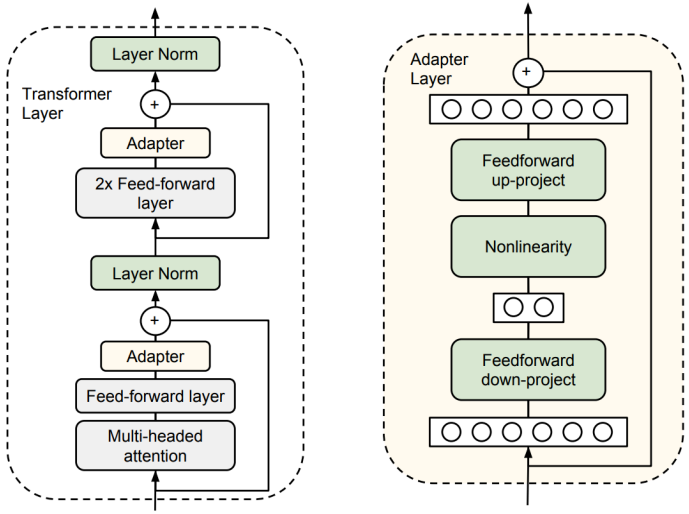
Adapter: Houlsby et al., Parameter-Efficient Transfer Learning for NLP, 2019



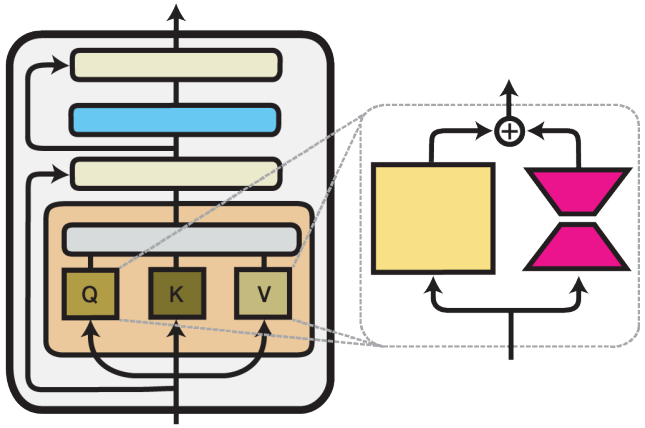
LoRA - Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021

image from: https://docs.adapterhub.ml/methods.html#lora

Parameter-efficient Fine-tuning (PEFT)

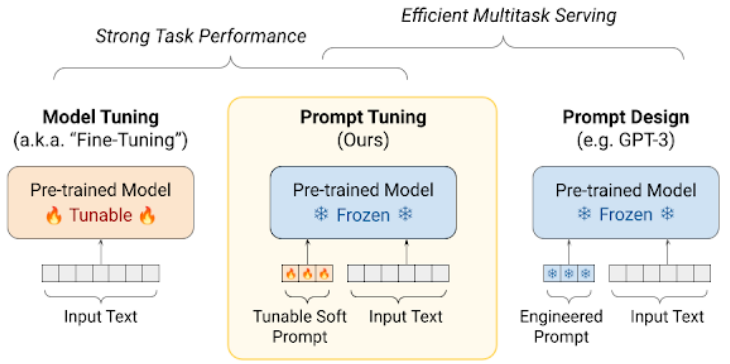


Adapter: Houlsby et al., Parameter-Efficient Transfer Learning for NLP, 2019



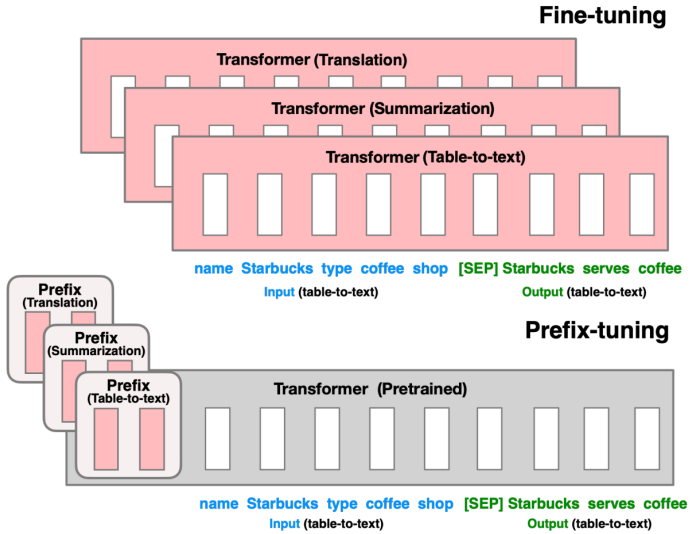
LoRA - Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021

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

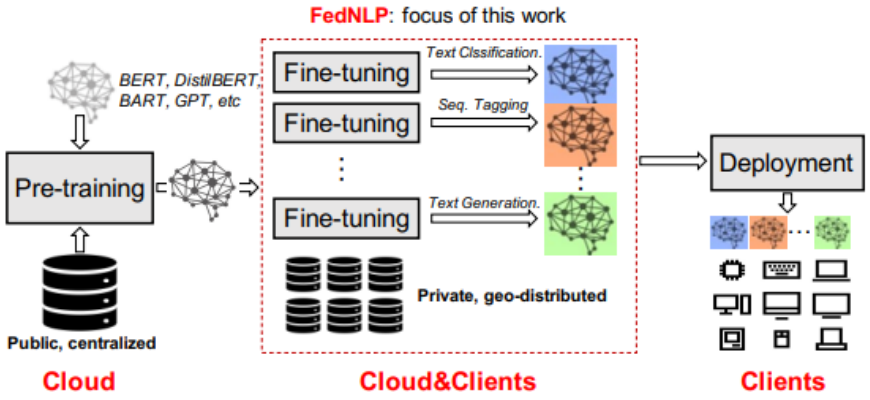


Prefix Tuning: Li et al., Prefix-Tuning: Optimizing Continuous Prompts for Generation, 2021

Federated Learning with PEFT

FedAdapter

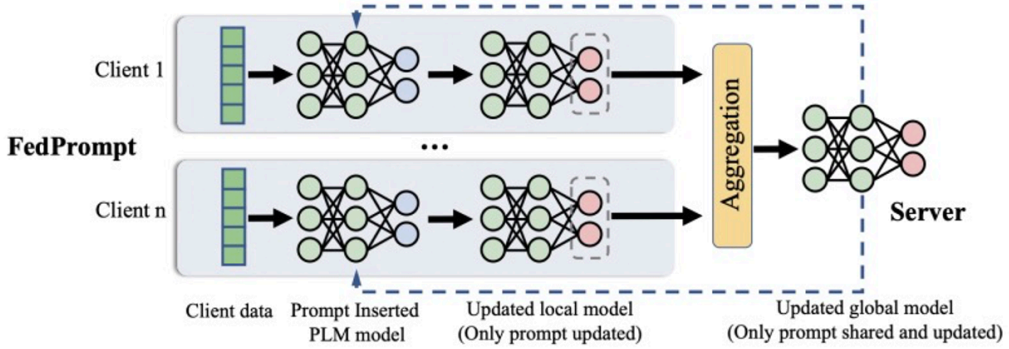
1. Emphasizing the importance of fine-tuning rather than training the foundation model in federated LLM.
2. Introducing adapter-based fine-tuning (PEFT) as a method to reduce communication costs.
3. Up to **155.5x** faster performance compared to vanilla FedNLP



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%

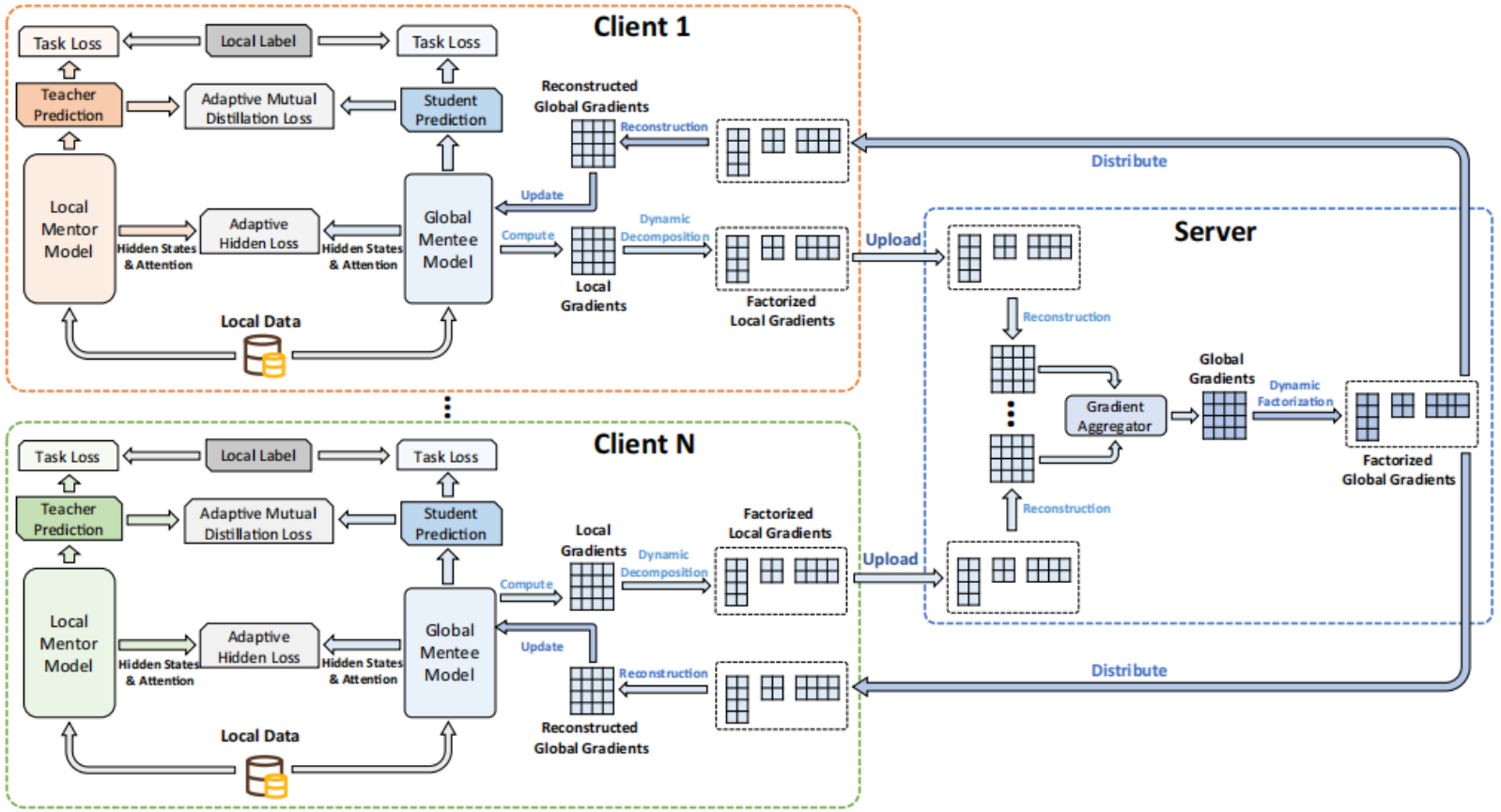
Source: Zhao H, Du W, Li F, et al. Reduce Communication Costs and Preserve Privacy: Prompt Tuning Method in Federated Learning[J]. arXiv preprint arXiv:2208.12268, 2022.

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.

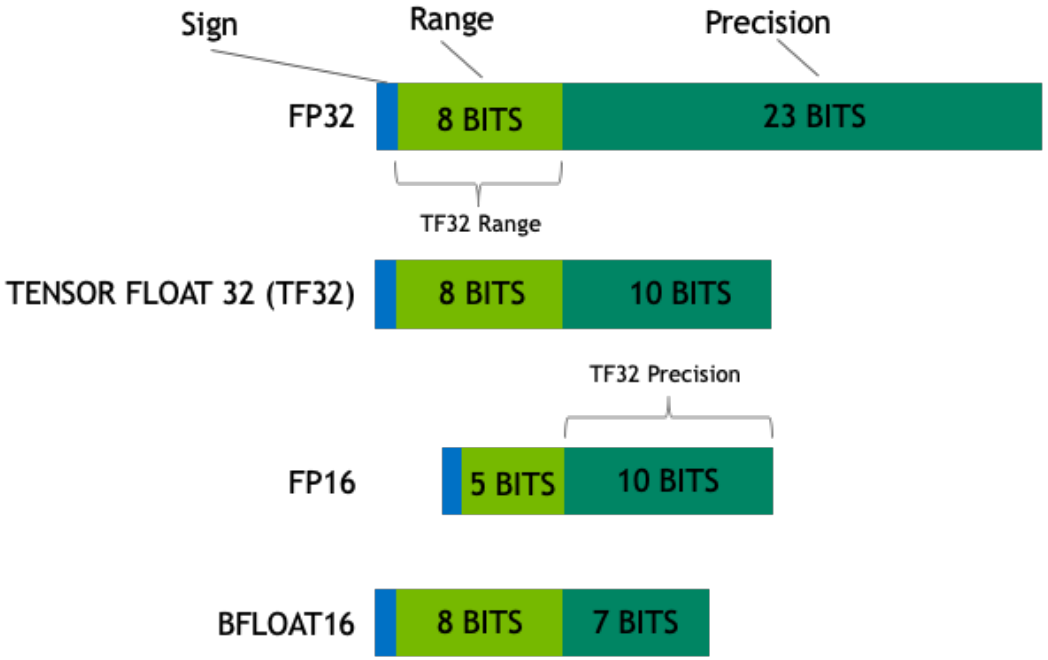
This paper propose a federated learning solution:

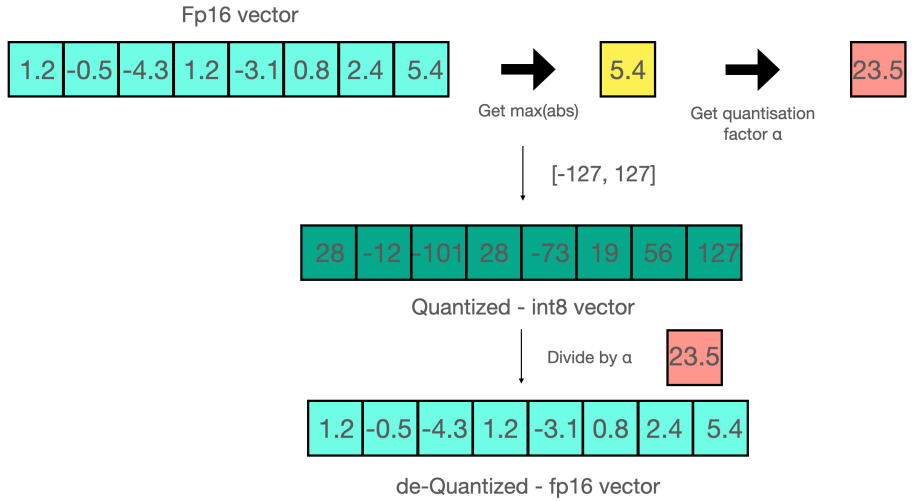
1. Local train both large Teacher model and small student model;
2. Upload and aggregate the small Student models;
3. Transfer knowledge from aggregated Student model to large Teacher model



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

Quantization





Source: A Gentle Introduction to 8-bit Matrix Multiplication for transformers at scale using Hugging Face Transformers, Accelerate and bitsandbytes URL. https://huggingface.co/blog/hf-bitsandbytes-integration

FedPETuning

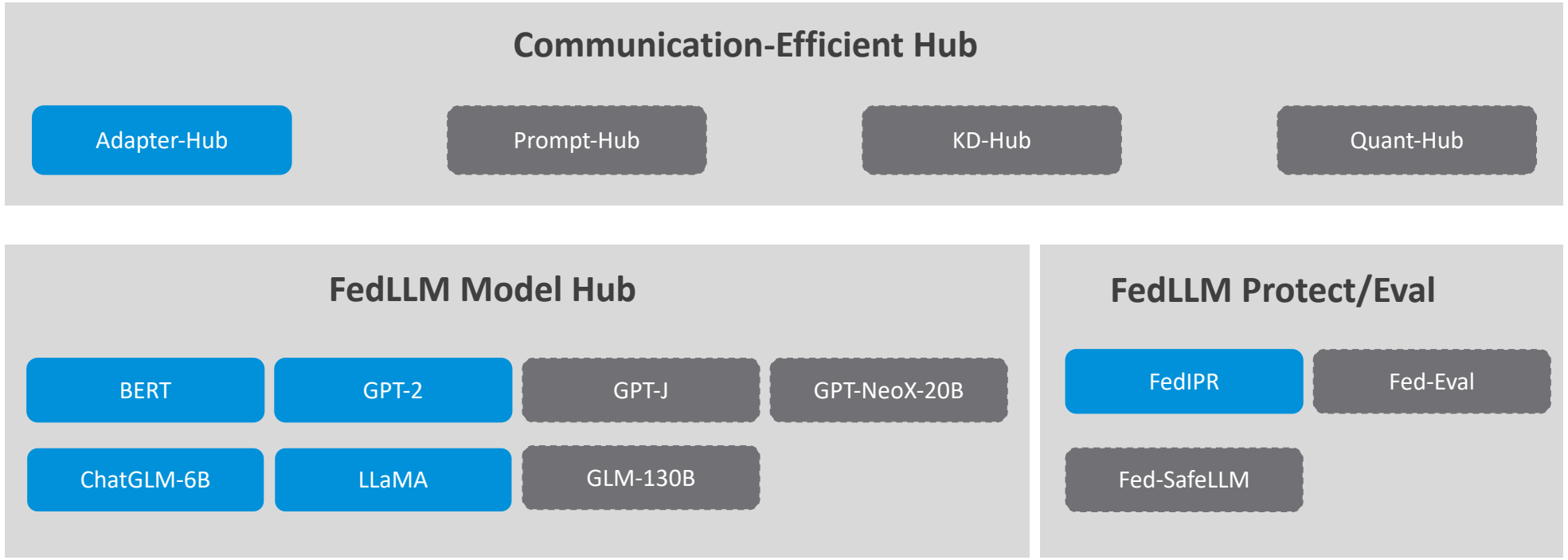
This paper presents:

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	-		

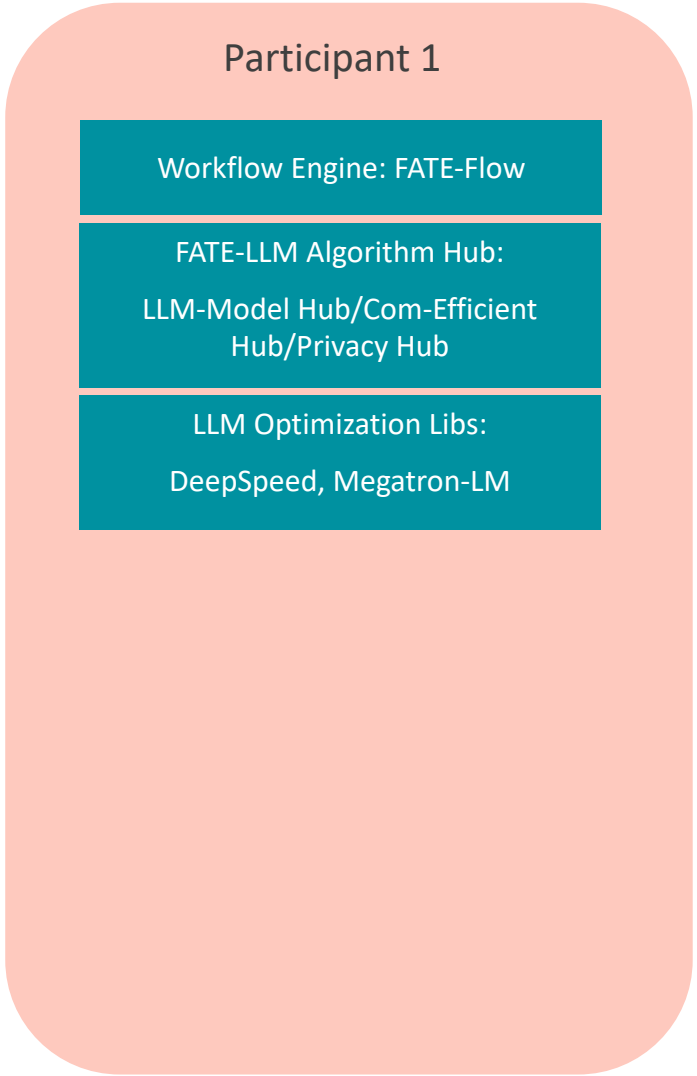
Source: Zhang Z, Yang Y, Dai Y, et al. When Federated Learning Meets Pre-trained Language Models' Parameter-Efficient Tuning Methods[J]. arXiv preprint arXiv:2212.10025, 2022.

FATE-LLM: FATE Federated Large Language Model

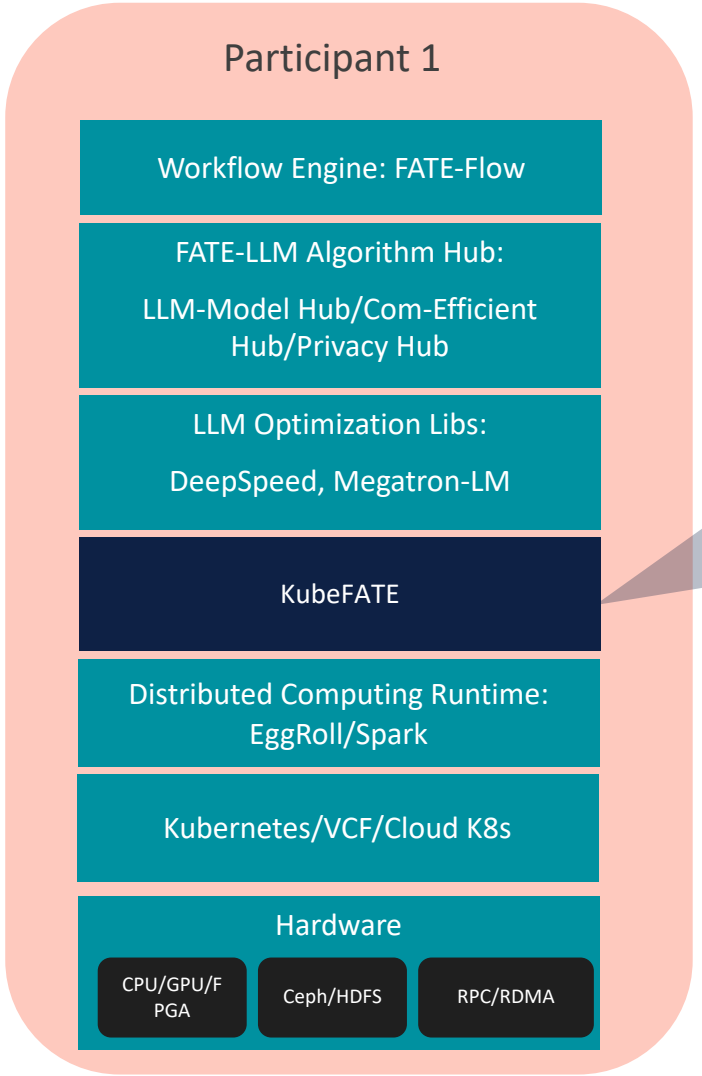


- Multiple clients can perform horizontal FL through FATE's built-in support of pre-trained model and use private data for large-scale model fine-tuning;
- Support 30+ participants for a collaborative training

FATE-LLM high-level architecture

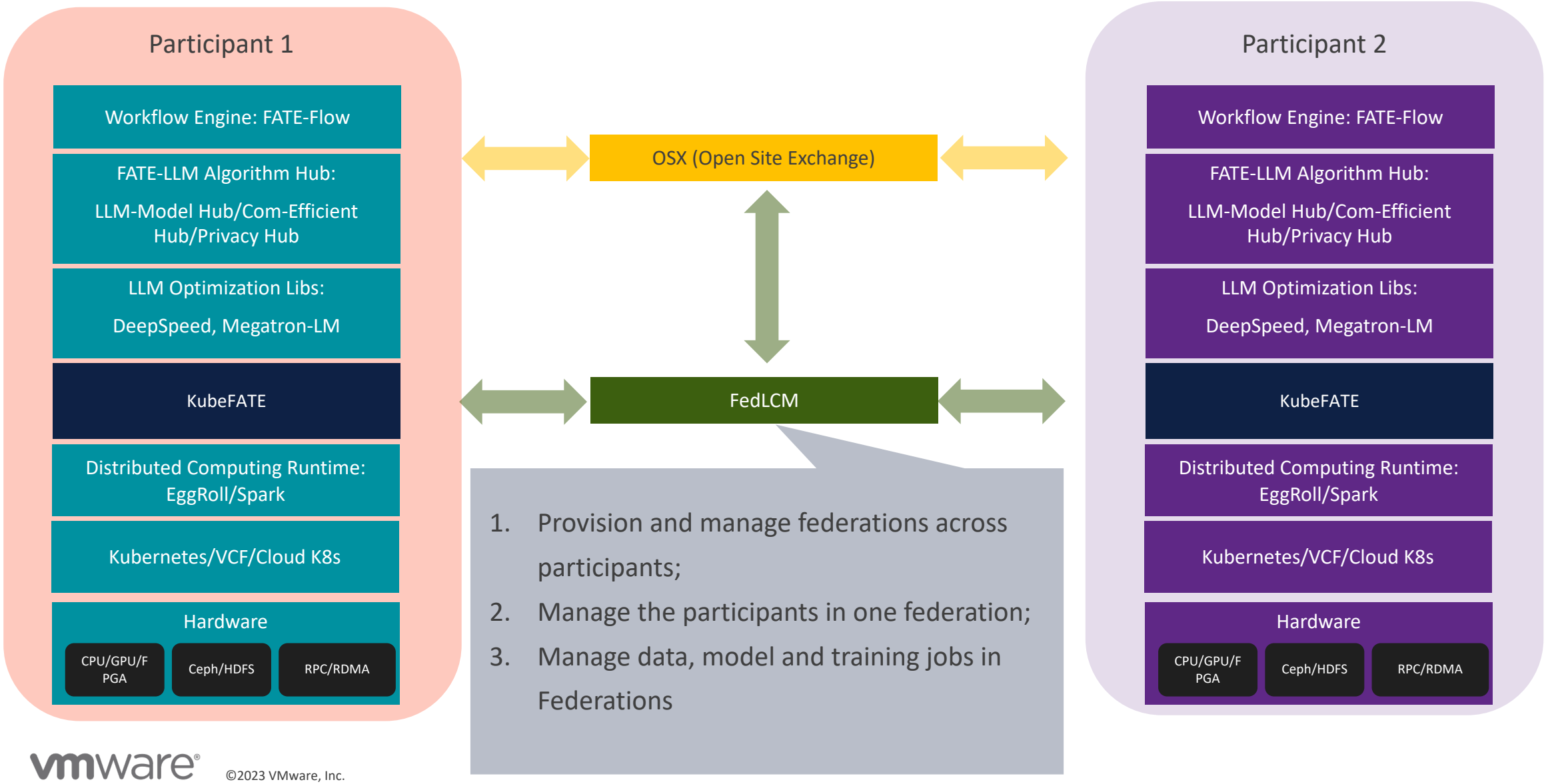


FATE-LLM high-level architecture

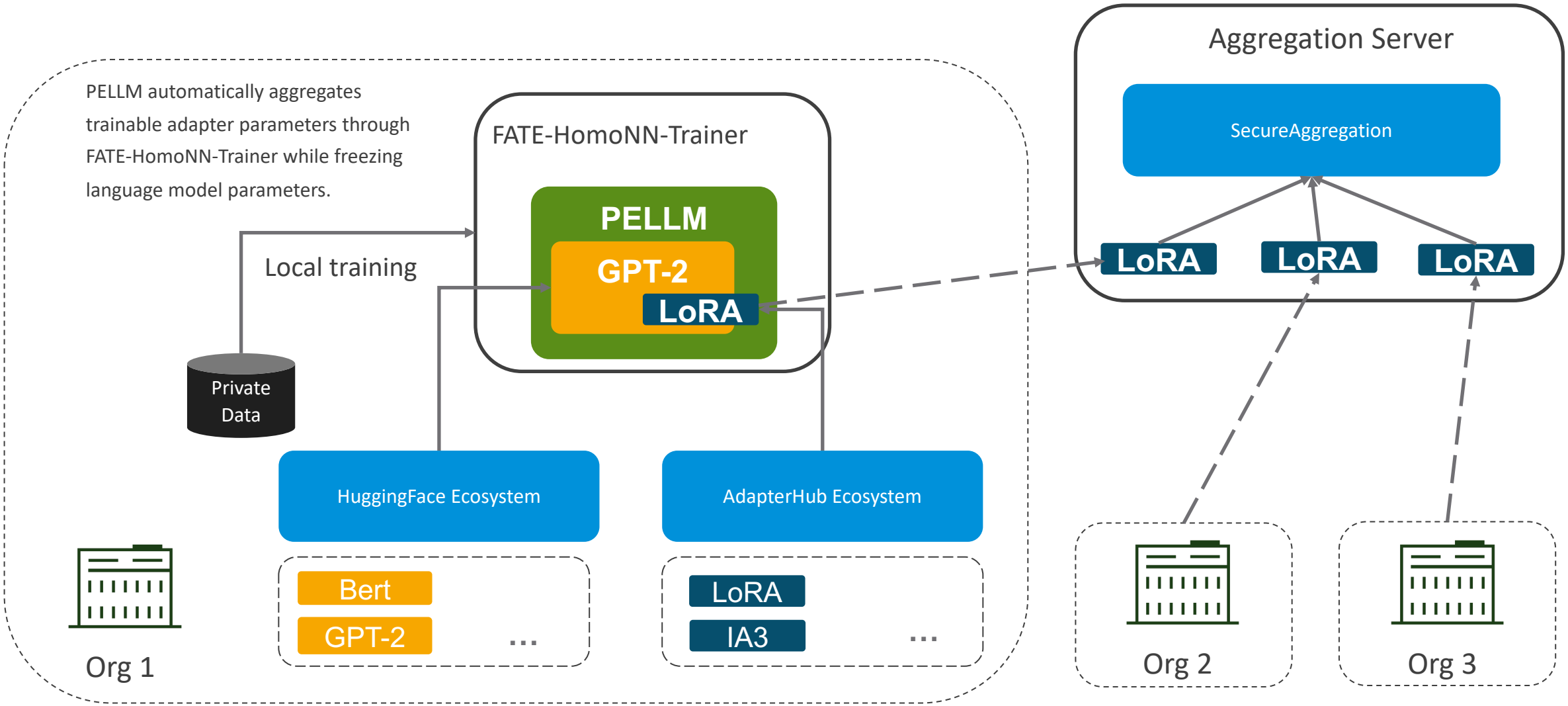


1. Containerized
2. Orchestrate the hardware resources
3. Provision and management in Kubernetes and Cloud
4. Scalability

FATE-LLM high-level architecture



FATE-LLM Algorithms Design (based on Adapter-Hub)



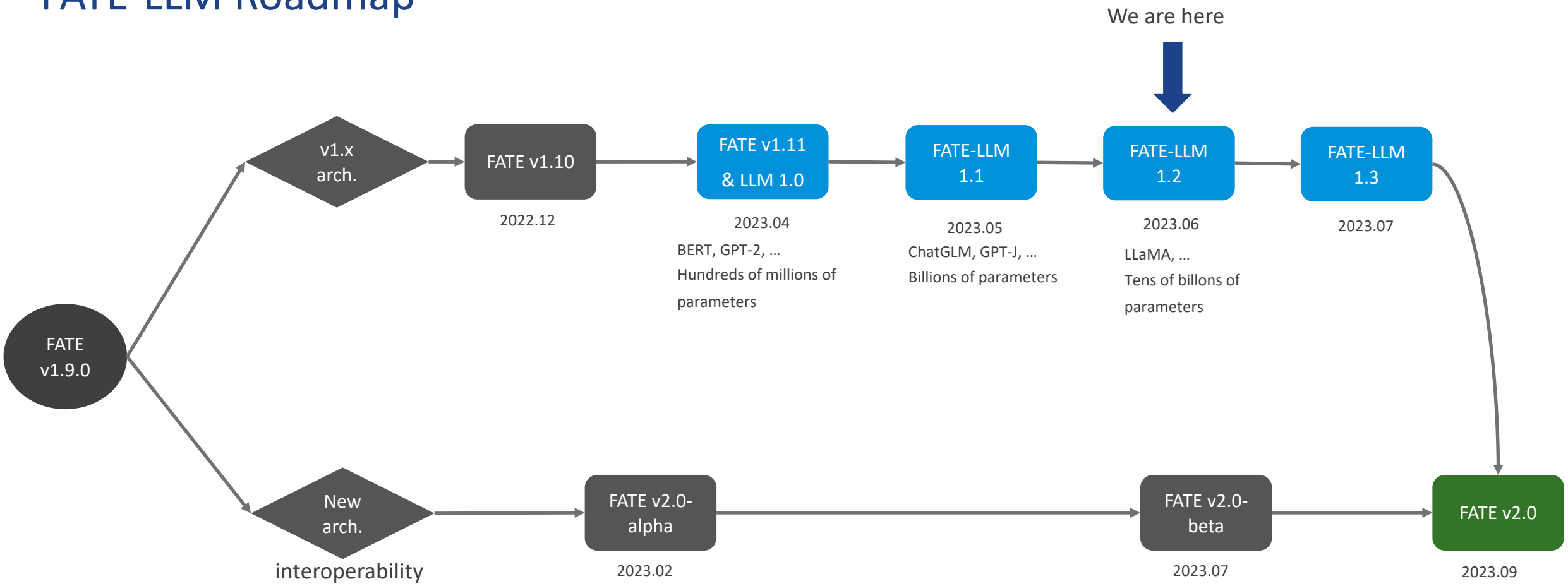
FATE-LLM: Performance & Cost

1. PEFT methods achieves an AUC of over 0.95, comparable to full fine-tuning;
2. GPU utilization significantly improves efficiency during local training;
3. Aggregation ratios range from 0.04% to 1.4%, resulting in minimal network traffic as communication cost. Increasing the number of epochs-to-aggregate further reduces traffic.

device	model	adapter	aggregate_every_n_epoch	auc	F1-score	accuracy	loss	parameter ratio (adapter/baseline)	network traffic curve (both trans and recv)	job time	one epoch time	gpu usage
CPU	GPT2 small 124M	Houlsby 1.438%	1	0.958	0.889337	0.89024	0.1888	1789K/124M	2 * 72M	7:55:14	0:34:48	811MiB / 16384MiB | 0%
GPU	GPT2 small 124M	Houlsby 1.438%	1	0.956	0.886104	0.88584	0.1909	1789K/124M	2 * 70M	0:33:32	0:02:22	2753MiB/16384MiB | 62%
GPU	GPT2 small 124M	Houlsby 1.438%	5	0.954	0.883363	0.8836	0.1288	1789K/124M	2 * 15M	0:32:40	0:02:24	2753MiB/16384MiB | 60%
GPU	GPT2 small 124M	Houlsby 1.438%	1	0.956	0.886536	0.88576	0.1914	1789K/124M	2 * 70M	0:33:40	0:02:21	2753MiB/16384MiB | 57%
GPU	GPT2 small 124M	LoRA 0.237%	1	0.954	0.88549	0.88536	0.2782	294K/124M	2 * 12M	0:33:21	0:02:25	2945MiB/16384MiB | 66%
GPU	GPT2 small 124M	IA3 0.044%	1	0.934	0.859853	0.85776	0.3524	55k/124M	2 * 2.5M	0:27:36	0:02:04	2967MiB/16384MiB | 63%
GPU	GPT2 large 774M	Houlsby 1.918%	1	0.965	0.905292	0.8516	0.022	14.8M/774M	1.2G	2:08:39	0:10:05	10021MiB / 16384MiB | 84%

- **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
- **Experiment setup:** Each participant has 2x V100 16GB and all participants are in a local area network environment and all on vSphere

FATE-LLM Roadmap



- Source: https://github.com/FederatedAI/FATE-LLM
- FedIRP: Qiang Yang, Anbu Huang, et al. Federated Learning with Privacy-preserving and Model IP-right-protection. https://doi.org/10.1007/s11633-022-1343-2

Discussions and Futures

1. DeepSpeed integration:
    1. FATE-LLM has integrated with DeepSpeed in v1.1
    2. Exploring ways to optimize the coordination between Kubernetes GPU scheduling and DeepSpeed
2. Private "information" in large models:
    1. Large models contain private "information”, and can they intelligently extract it?
    2. With the emergence of new capabilities as model parameters grow, there is uncertainty.
    3. However, FATE-LLM (ChatGLM-6B) does not possess this capability
3. In the roadmap, FATE-LLM continuously supports larger models and preserves privacy

Thank You

扫码关注FATE官方公众号
获取更多前沿资讯

扫码添加FATE小助手
深度参与社区共建

FATE社区联系邮件组

开源社区用户	Fate-FedAI@groups.io
开源社区维护者	FedAI-maintainers@groups.io
开发专委会	Fate-dev-core@groups.io
运营专委会	FATE-operation@groups.io
安全专委会	FATE-security@groups.io