in.clusion
CONFERENCE ON THE BUND
外 滩 大 会

SECRET FLOW 隐语

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

大模型时代的隐私计算

赵皓东
隐语开源社区Contributor 上海交通大学网络空间安全学院博士研究生

Contents
目录

01 大模型发展及现状

大模型进化史

2019年后的主要变化

2017年Transformer的提出

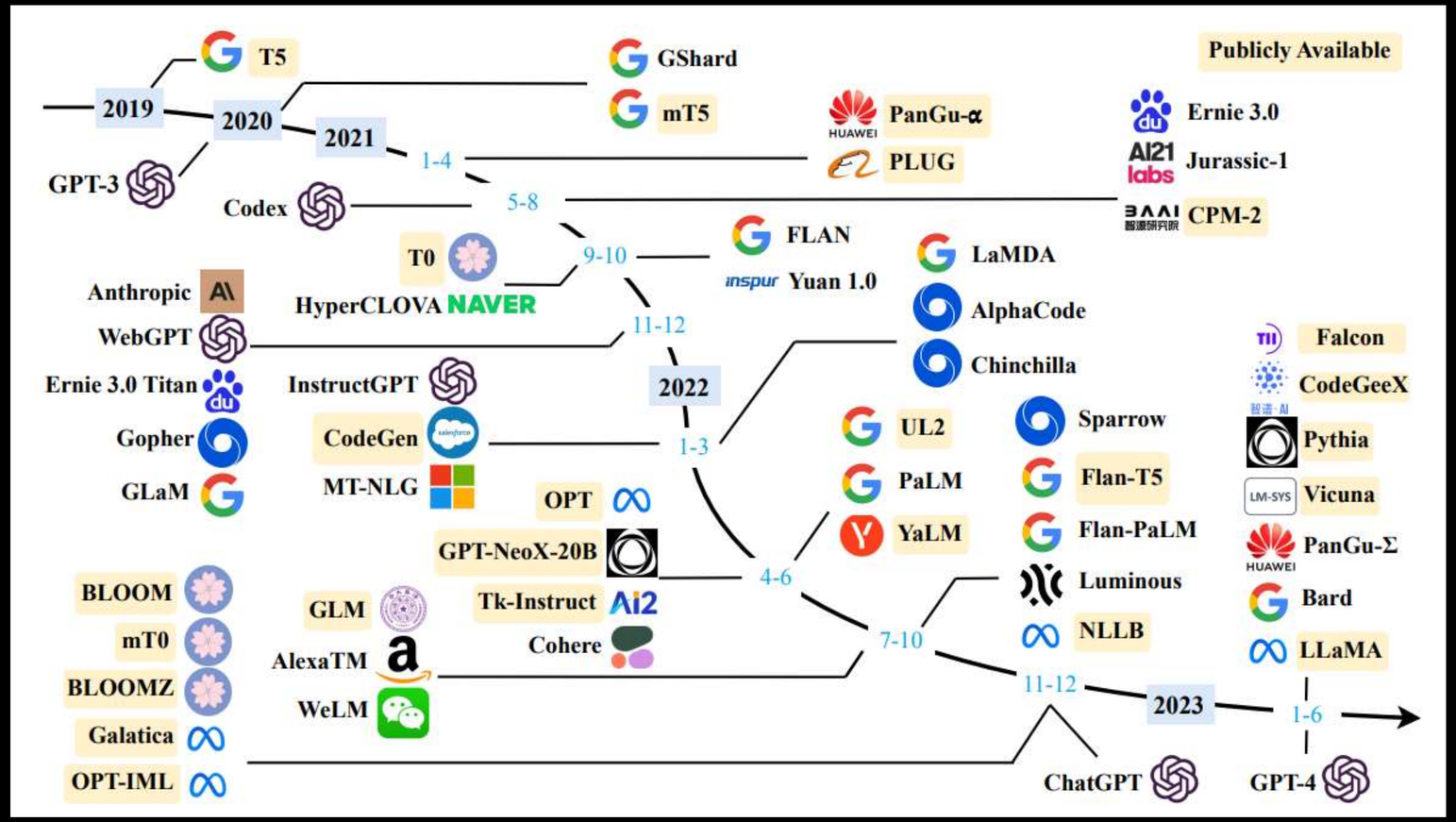
《Attention Is All You Need》

三大分支

- Encoder-only(BERT)
- Decoder-only(GPT-1/2/3/4, ChatGPT)
- Encoder-Decoder(T5, ChatGLM)

现状：百花齐放

从 GPT-3 开始，当下的 ChatGPT、GPT-4、Bard 以及 PaLM、LLaMA 等百花齐放



[1]Zhao, W. X., “A Survey of Large Language Models”, arXiv e-prints, 2023. doi:10.48550/arXiv.2303.18223.
[2]Yang, J., “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond”, arXiv e-prints, 2023. doi:10.48550/arXiv.2304.13712.

02 PUMA-基于隐语的全密态大模型

MPCFormer

使用大模型存在安全隐患

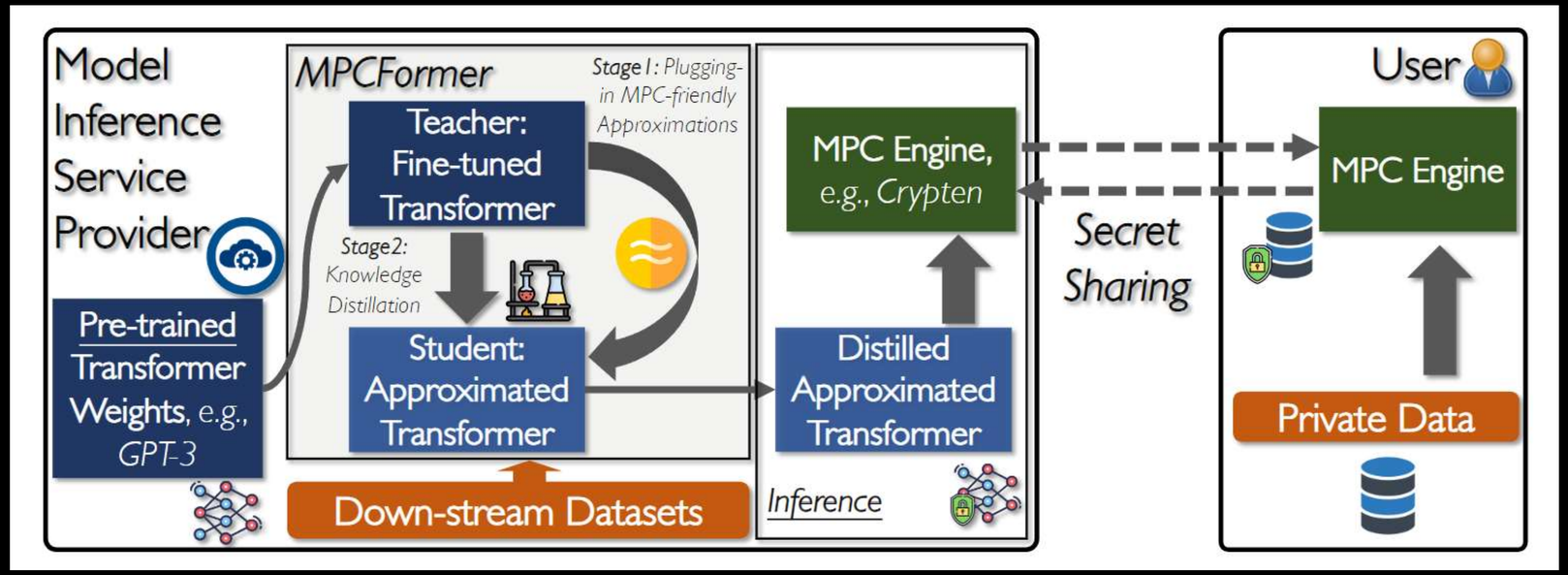
- 将包含隐私信息的prompt提供给服务商
- 服务商提供模型参数

直接使用MPC推理速度缓慢

延迟变为60x以上，$BERT_{BASE}$推理时间由<1s变为59s

MPCFormer

使用MPC和知识蒸馏实现快速、高效和私有的Transformer推理

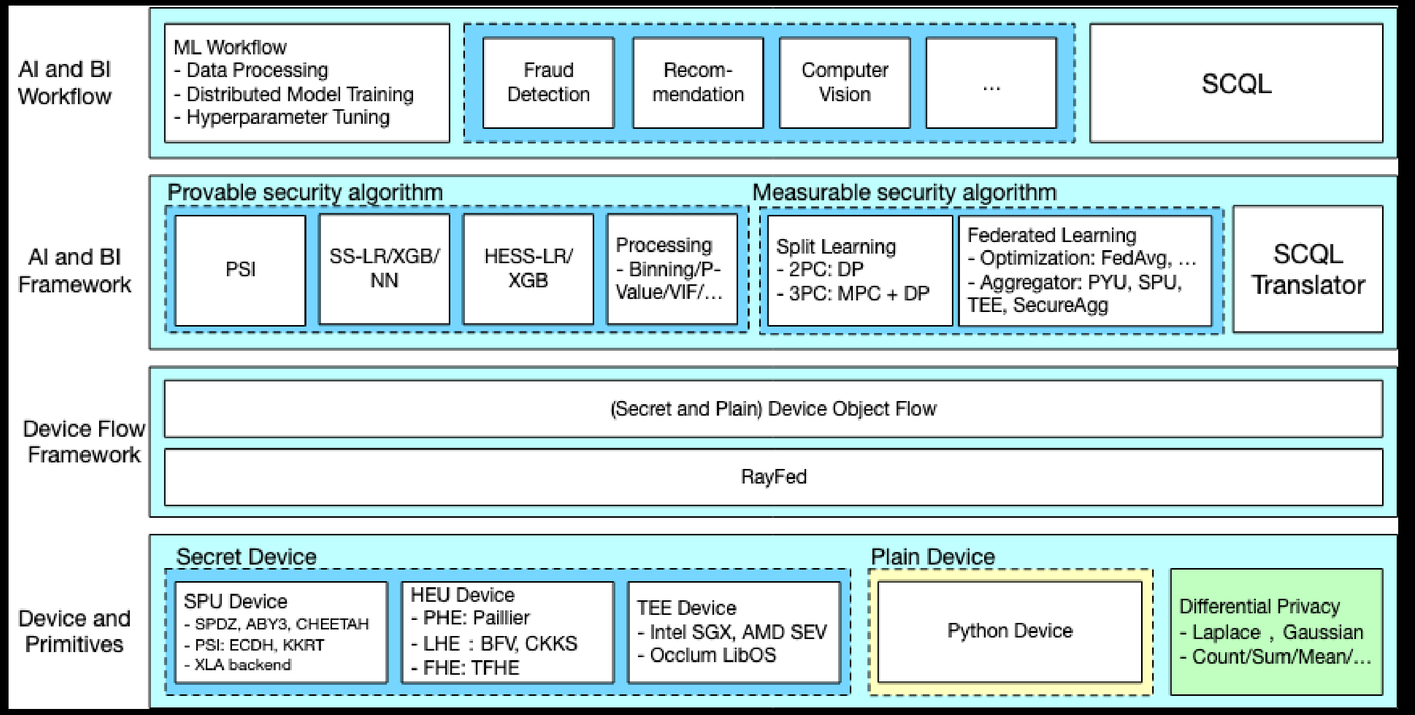


[1]MPCFormer: fast, performant and private Transformer inference with MPC. D Li, R Shao, H Wang, H Guo, EP Xing, H Zhang. ICLR 2023 (Spotlight), 2022.

隐语-SecretFlow

隐私保护数据分析和机器学习的统一框架

- 设备抽象，将多方安全计算（MPC）、同态加密（HE）、可信执行环境（TEE）等隐私计算技术抽象为密文设备，将明文计算抽象为明文设备。
- 基于抽象设备的计算图，使数据分析和机器学习工作流程能够表示为计算图。
- 基于计算图的机器学习/数据分析能力，支持数据水平/垂直/混合分割等场景。

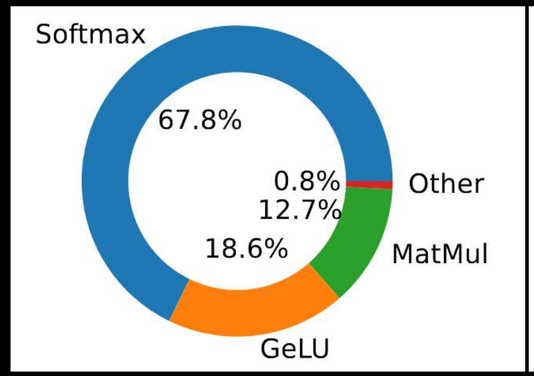


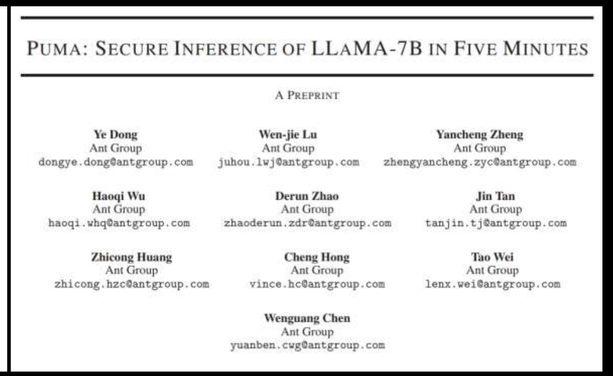
https://github.com/secretflow

PUMA-基于隐语的全密态大模型

大模型在MPC中的主要限制

Softmax、GeLU等函数的时间成本昂贵




PUMA: SECURE INFERENCE OF LLAMA-7B IN FIVE MINUTES

A PREPRINT

Ye Dong
Ant Group
dongye.dong@antgroup.com

Wen-jie Lu
Ant Group
juhou.lwj@antgroup.com

Yancheng Zheng
Ant Group
zhengyancheng.zyc@antgroup.com

Haoqi Wu
Ant Group
haoqi.whq@antgroup.com

Derun Zhao
Ant Group
zhaoderun.zdr@antgroup.com

Jin Tan
Ant Group
tanjin.tj@antgroup.com

Zhicong Huang
Ant Group
zhicong.hzc@antgroup.com

Cheng Hong
Ant Group
vince.hc@antgroup.com

Tao Wei
Ant Group
lenx.wei@antgroup.com

Wenguang Chen
Ant Group
yuanben.cwg@antgroup.com

对非线性函数更好的近似

对Softmax、GeLU等非线性函数提供更准确和快速的近似

先前工作的不足

- 粗略的近似和替换
- 推理成本高
- 部署复杂，对很多功能不支持

开源的端到端框架

支持预训练的明文Transformer模型，无需改变任何模型架构

[1]Dong, Y., “PUMA: Secure Inference of LLaMA-7B in Five Minutes”, arXiv e-prints, 2023. doi:10.48550/arXiv.2307.12533.
[2]MPCFormer: fast, performant and private Transformer inference with MPC. D Li, R Shao, H Wang, H Guo, EP Xing, H Zhang. ICLR 2023 (Spotlight), 2022.

03 联邦学习&大模型

Federated LLM

Federated LLM pre-training

Federated LLM fine-tuning

Federated LLM Prompt-engineering

[1]Chen, Chaochao, et al. "Federated Large Language Model: A Position Paper." arXiv preprint arXiv:2307.08925 (2023).

FedPrompt

Prompt Learning

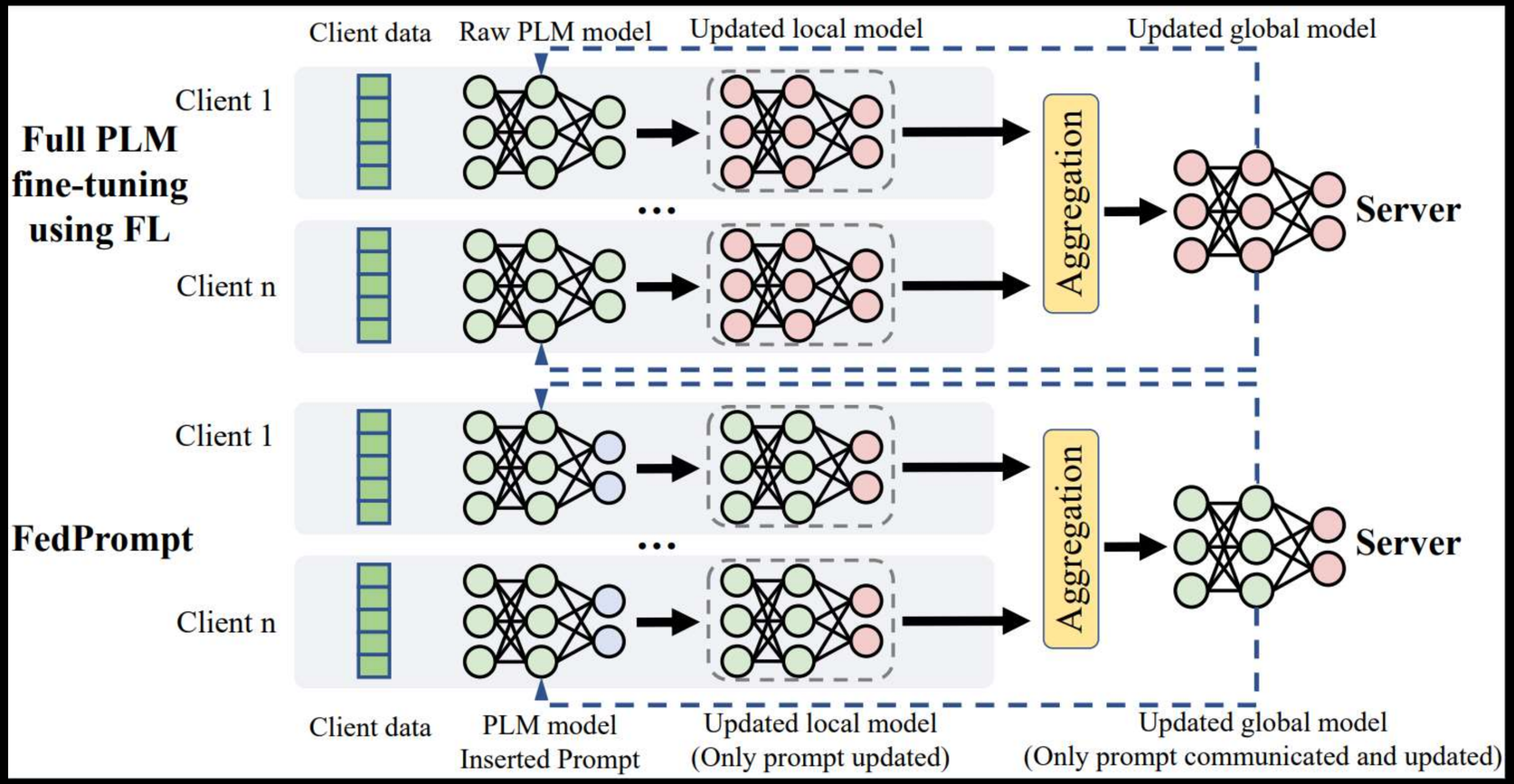
在已有预训练模型的情况下，针对few-shot等场景有良好的表现

Federated Learning

解决数据孤岛问题，保护用户的隐私数据

安全性和鲁棒性提升

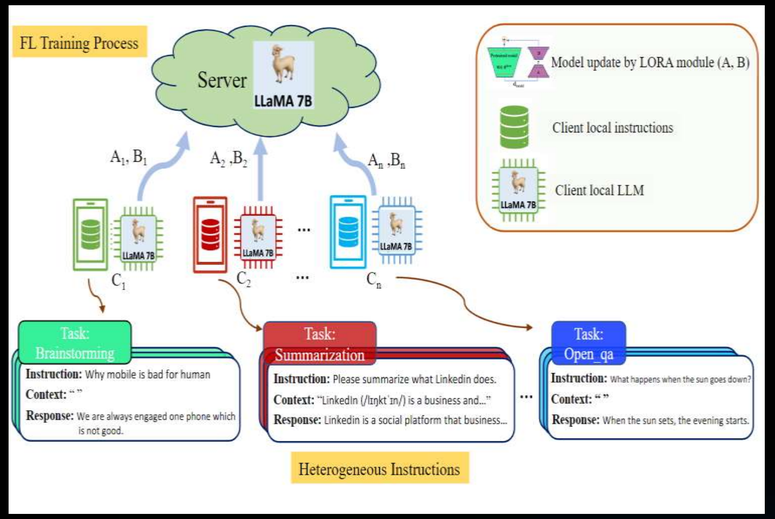
通过联邦聚合过程，对PPT为代表的后门攻击具有鲁棒性

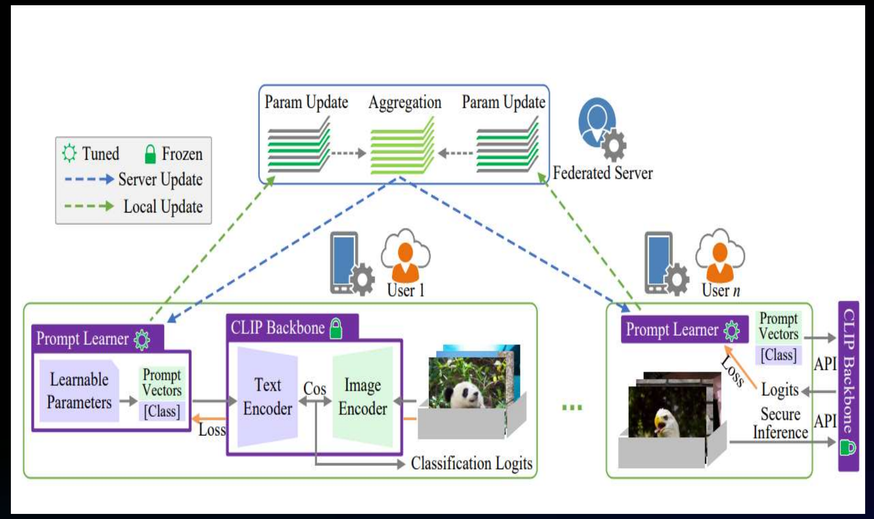


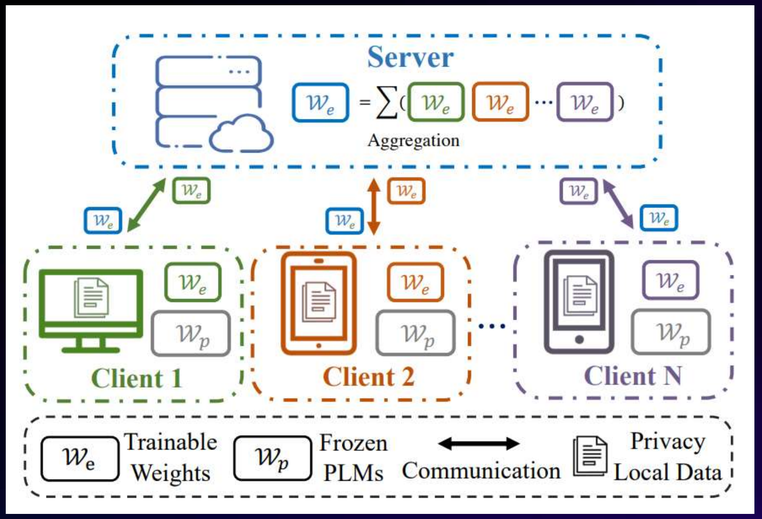
[1]Zhao, H., Du, W., Li, F., Li, P., and Liu, G., “FedPrompt: Communication-Efficient and Privacy Preserving Prompt Tuning in Federated Learning”.
[2] Du, Wei, et al. "Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning." IJCAI-22. 2022.

PEFT

Parameter-Efficient Federated Learning







FedIT

探索了Instruction Tuning和peft的结合

PromptFL

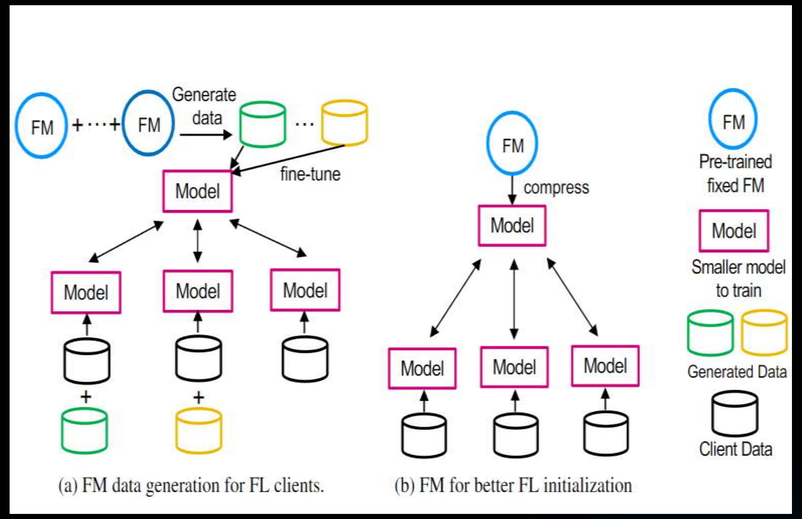
使用CLIP验证了联邦聚合prompt的有效性

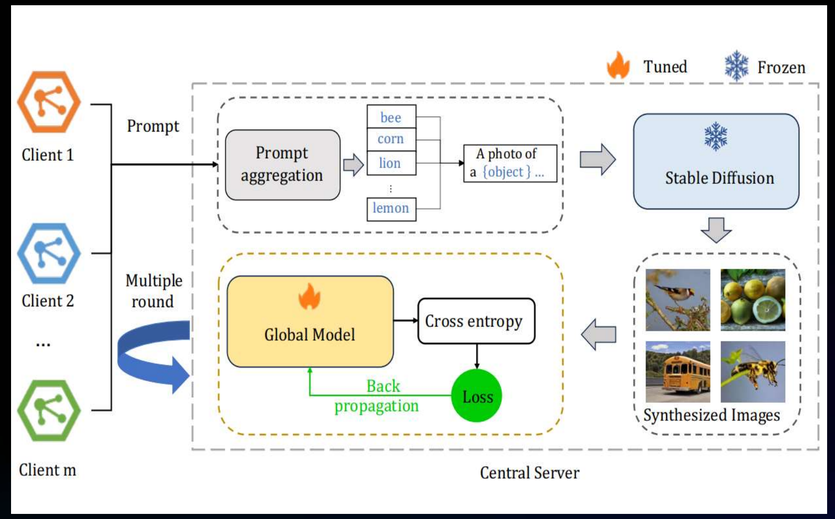
FedPETuning

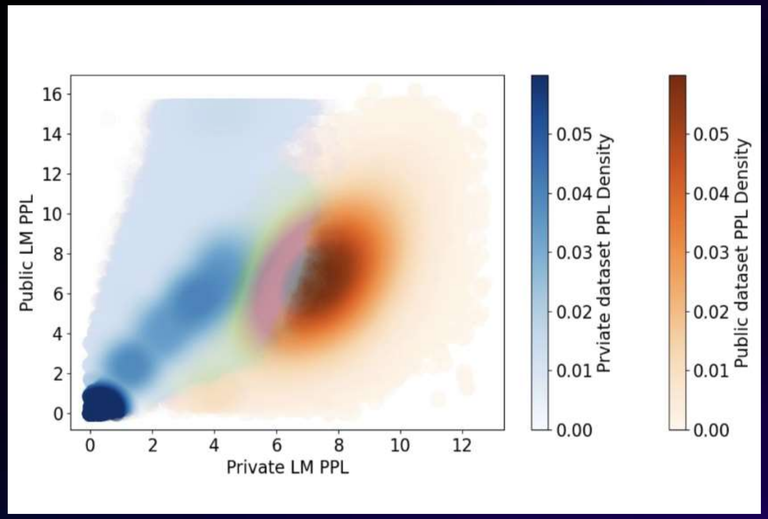
全面测试了联邦fine-tuning、 Adapter tuning、 Prefix tuning、 LoRA、 BitFit的性能，提出benchmark

[1] Zhang, Jianyi, et al. "Towards Building the Federated GPT: Federated Instruction Tuning." arXiv preprint arXiv:2305.05644 (2023).
[2] Guo, Tao, et al. "Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model." 2023.
[3] Zhang, Zhuo, et al. "FedPETuning: When Federated Learning Meets the Parameter-Efficient Tuning Methods of Pre-trained Language Models." Findings of ACL 2023.

LM for FL







When FM Meets FL

FM可以作为FL的强大起点，同时可以生成大量数据

FGL

Federated Generative Learning (FGL)，在客户端和服务器之间传输相关提示，根据包含很少隐私的提示和生成模型来远程合成训练数据

LLM for FL

使用大规模公共数据和LLM的tokenizer等来帮助设备上 FL 模型的训练

[1] Zhuang, Weiming, Chen Chen, and Lingjuan Lyu. "When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions."
[2] Zhang, Jie, Xiaohua Qi, and Bo Zhao. "Federated Generative Learning with Foundation Models." arXiv preprint arXiv:2306.16064 (2023).
[3] Wang, Boxin, et al. "Can Public Large Language Models Help Private Cross-device Federated Learning?." arXiv preprint arXiv:2305.12132 (2023).

04 拆分学习&大模型

拆分学习&大模型

拆分学习

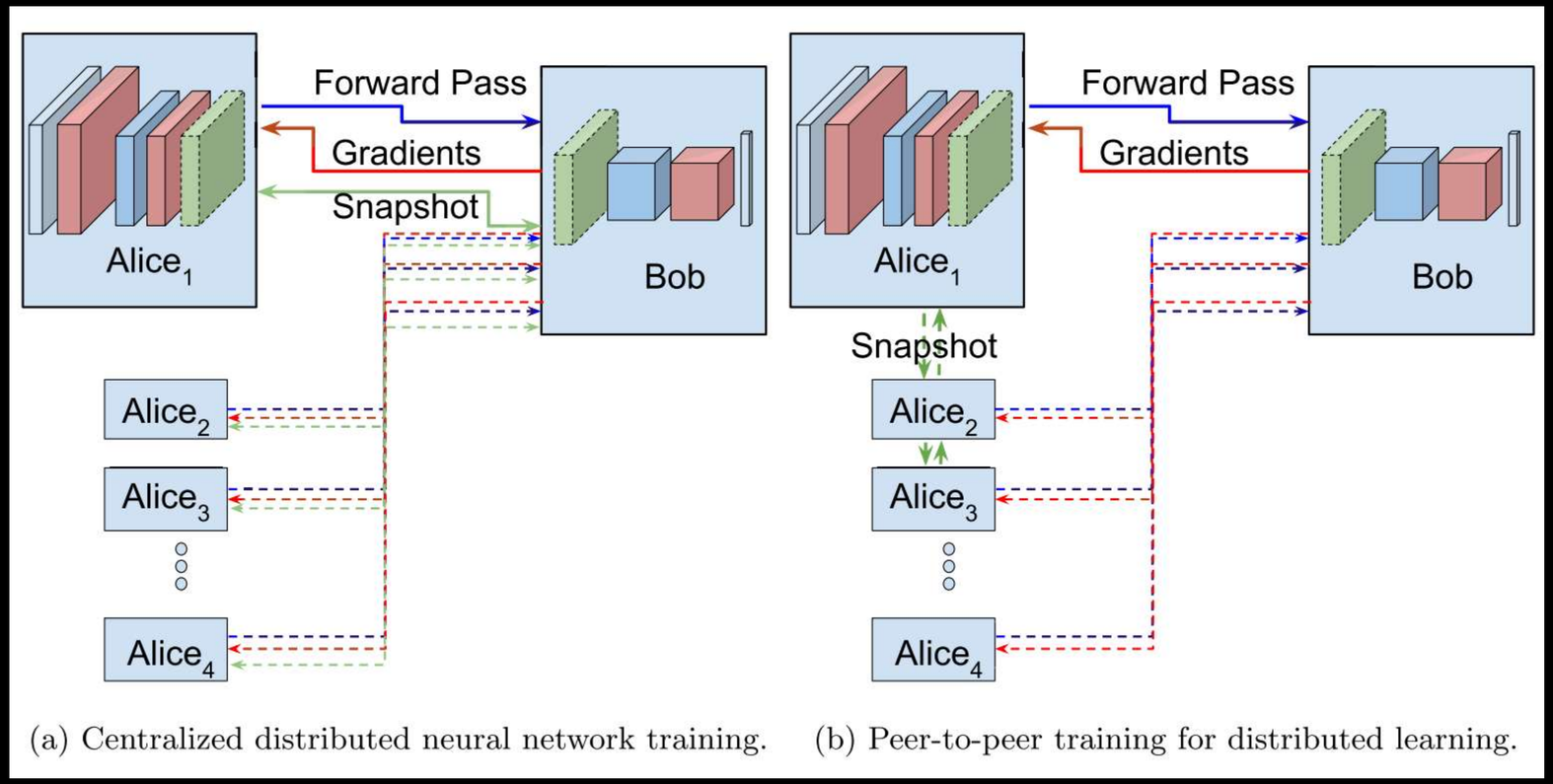
将完整模型拆分，使不同参与者仅拥有部分的数据与模型

大模型训练、推理的隐私风险

Prompt中包含隐私信息

大模型使用拆分学习保护隐私

用户不再传输明文prompt，而是传输Smashed data，如在浅层添加简单网络



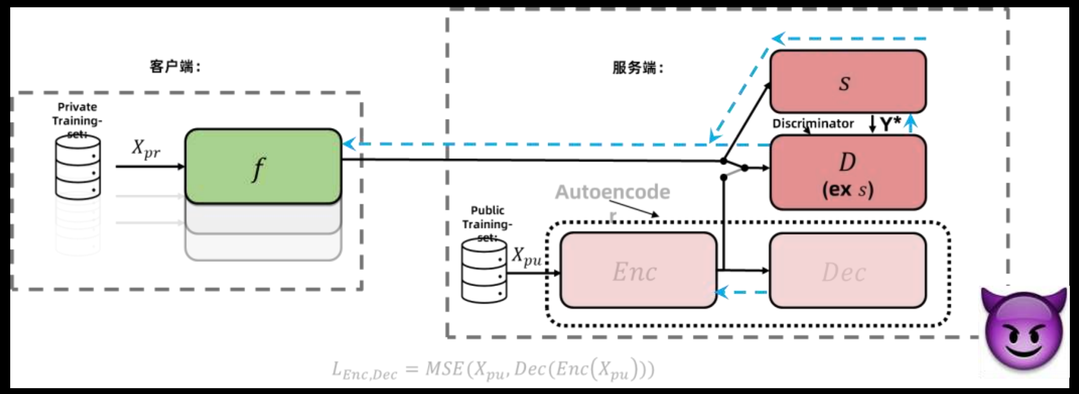
(a) Centralized distributed neural network training. (b) Peer-to-peer training for distributed learning.

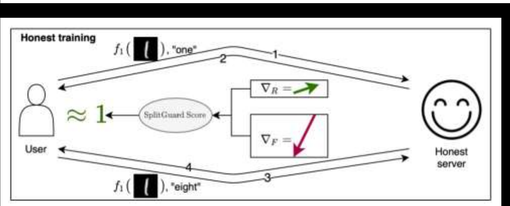
[1] Gupta, Otkrist, and Ramesh Raskar. "Distributed learning of deep neural network over multiple agents." Journal of Network and Computer Applications 116 (2018): 1-8.

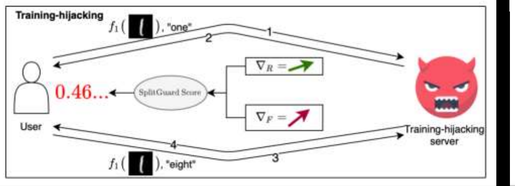
结语：我与隐语社区

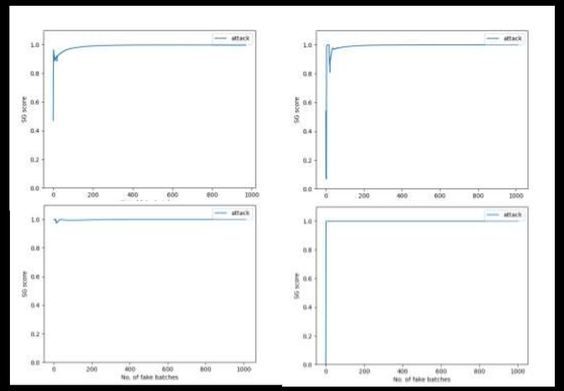
参与拆分学习攻防研究

在特征空间劫持（FSHA）攻击与SplitGuard基础上进行优化，给出隐语-拆分学习更详细的torch版本demo，补充针对潜在特征空间劫持攻击的检测功能









开源社区

拆分学习

我在隐语社区的收获和展望

文档

代码

交流

[1] Dario Pasquini, Giuseppe Ateniese, and Massimo Bernaschi. 2021. Unleashing the tiger: Inference attacks on split learning. In ACM CCS. 2113–2129

[2] Erdogan, Ege, Alptekin Küpçü, and A. Ercument Cicek. "Splitguard: Detecting and mitigating training-hijacking attacks in split learning." 2022.

THANKS

Homepage: https://dongdongzhaoup.github.io/
E-mail: zhaohaodong@sjtu.edu.cn