Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2477

Robust Federated Learning: Defending Against Byzantine and Jailbreak Attacks

SHENGHUI LI

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6214
ISBN 978-91-513-2312-1
urn:nbn:se:uu:diva-540441

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in 101121, Sonja Lyttkens, Ångström, Regementsvägen 1, Uppsala, Thursday, 16 January 2025 at 09:00 for the degree of Doctor of Philosophy (Faculty of Theology). The examination will be conducted in English. Faculty examiner: Professor Karl Andersson (Luleå University of Technology).

Abstract
Li, S. 2024. Robust Federated Learning: Defending Against Byzantine and Jailbreak Attacks. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2477. 54 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2312-1.

Federated Learning (FL) has emerged as a promising paradigm for training collaborative machine learning models across multiple participants while preserving data privacy. It is particularly valuable in privacy-sensitive domains like healthcare and finance. Recently, FL has been explored to harness the power of pre-trained Foundation Models (FMs) for downstream task adaptation, enabling customization and personalization while maintaining data locality and privacy. However, FL's distributed nature makes it inherently vulnerable to adversarial attacks. Notable threats include Byzantine attacks, which inject malicious updates to degrade model performance, and jailbreak attacks, which exploit the fine-tuning process to undermine safety alignments of FMs, leading to harmful outputs. This dissertation centers on robust FL, aiming to mitigate these threats and ensure global models remain accurate and safe even under adversarial conditions. To mitigate Byzantine attacks, we propose several Robust Aggregation Schemes (RASs) that decrease the influence of malicious updates. Additionally, we introduce Blades, an open-source benchmarking tool to systematically study the interplay between attacks and defenses in FL, offering insights into the effects of data heterogeneity, differential privacy, and momentum on RAS robustness. Exploring the synergy between FL and FMs, we present a taxonomy of research along with adaptivity, efficiency, and trustworthiness. We uncover a novel attack, “PEFT-as-an-Attack” (PaaA), where malicious FL participants jailbreak FMs through Parameter-Efficient-Fine-Tuning (PEFT) with harmful data. We evaluate defenses against PaaA and highlight critical gaps, emphasizing the need for advanced strategies balancing safety and utility in FL-FM systems. In summary, this dissertation advances FL robustness by proposing novel defenses, tools, and insights while exposing emerging attack vectors. These contributions pave the way for attack-resilient distributed machine learning systems capable of withstanding both current and emerging threats.

Keywords: Federated learning, Jailbreak attack, Parameter-Efficient Fine-Tuning, Pre-trained Language Model, Robustness

Shenghui Li, Department of Information Technology, Division of Computer Systems, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-2312-1
URN urn:nbn:se:uu:diva-540441 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-540441)

To My Love, Family, and Friends

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Shenghui Li**, Edith C. H. Ngai, Fanghua Ye, and Thiemo Voigt. Auto-weighted robust federated learning with corrupted data sources. *ACM Transactions on Intelligent Systems and Technology*, vol. 13, pp. 1-20, 2022.
DOI: 10.1145/3517821

II **Shenghui Li**, Edith C. H. Ngai, and Thiemo Voigt. Byzantine-Robust Aggregation in Federated Learning Empowered Industrial IoT. *IEEE Transactions on Industrial Informatics*, vol. 19, pp. 1165-1175, 2023.
DOI: 10.1109/TII.2021.3128164
A short version of this paper is presented at the 34th Workshop of the Swedish Artificial Intelligence Society (SAIS'22).

III **Shenghui Li**, Edith C. H. Ngai, and Thiemo Voigt. An Experimental Study of Byzantine-Robust Aggregation Schemes in Federated Learning. *IEEE Transactions on Big Data*, vol. 10, pp. 975 - 988, 2024.
DOI: 10.1109/TBDATA.2023.3237397

IV **Shenghui Li**, Edith C. H. Ngai, Fanghua Ye, Li Ju, Tianru Zhang, and Thiemo Voigt. Blades: A Unified Benchmark Suite for Byzantine Attacks and Defenses in Federated Learning. *In 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI'24)*, pp. 158-169.
DOI: 10.1109/IoTDI61053.2024.00018

V **Shenghui Li**, Fanghua Ye, Meng Fang, Jiaxu Zhao, Yun-Hin Chan, Edith C. H. Ngai, and Thiemo Voigt. Synergizing Foundation Models and Federated Learning: A Survey. *Under submission*.

VI **Shenghui Li**, Edith C. H. Ngai, Fanghua Ye, and Thiemo Voigt. PEFT-as-an-Attack! Jailbreaking Language Models during Federated Parameter-Efficient Fine-Tuning. *Under submission*.

Other Peer Reviewed Papers

Additionally, I authored & co-authored the following peer-reviewed papers, before and during my PhD studies.

- **Shenghui Li**, Edith C. H. Ngai, Fanghua Ye, Li Ju, Tianru Zhang, and Thiemo Voigt. Demo Abstract: Blades: A Unified Benchmark Suite for Byzantine-Resilient in Federated Learning. *In 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI'24)*, pp. 229-230.
  DOI: 10.1109/IoTDI61053.2024.00030

- Fanghua Ye, Meng Fang, **Shenghui Li**, and Emine Yilmaz. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. *In Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP-Findings'23)*, pp. 5985–6006.
  DOI: 10.18653/v1/2023.findings-emnlp.398

- Fanghua Ye, Xi Wang, Jie Huang, **Shenghui Li**, Samuel Stern, and Emine Yilmaz. MetaASSIST: Robust Dialogue State Tracking with Meta Learning. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (EMNLP'22)*, pp. 1157–1169.
  DOI: 10.18653/v1/2022.emnlp-main.76

- Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, **Shenghui Li**, and Emine Yilmaz. lot Self-Attentive Dialogue State Tracking. *In Proceedings of the Web Conference 2021 (WWW'21)*, pp. 1598–1608.
  DOI: 10.1145/3442381.3449939

- Hui Lin, Zetao Yang, Zicong Hong, **Shenghui Li**, and Wuhui Chen. Smart Contract-based Hierarchical Auction Mechanism for Edge Computing in Blockchain-empowered IoT. *In 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM'20)*, pp. 147-156.
  DOI: 10.1109/WoWMoM49955.2020.00035

- Rui Xi, Kang Liu, Shuo Liu, Wuhui Chen, and **Shenghui Li**. Perishable digital goods trading mechanism for blockchain-based vehicular network. *2019 IEEE International Conference on Parallel & Distributed Processing*

with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom'19), pp. 147-154.
DOI: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00031

- **Shenghui Li**, Wuhui Chen, Yufen Chen, Chuan Chen, Zibin Zheng. Makespan-minimized computation offloading for smart toys in edge-cloud computing. *Electronic Commerce Research and Applications*, vol. 37, pp. 1567-4223, 2019.
DOI: 10.1016/j.elerap.2019.100884

- Fenfang Xie, **Shenghui Li**, Liang Chen, Yangjun Xu, and Zibin Zheng. Generative Adversarial Network Based Service Recommendation in Heterogeneous Information Networks. *2019 IEEE International Conference on Web Services (ICWS'19)*, pp. 265-272.
DOI: 10.1109/ICWS.2019.00053

- Yuanhao Yang, Xiaoyu Qiu, **Shenghui Li**, Junbo Wang, Wuhui Chen, Patrick CK Hung, and Zibin Zheng. Energy-efficient data routing in cooperative UAV swarms for medical assistance after a disaster. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, 2019.
DOI: 10.1063/1.5092740

- Fanghua Ye, **Shenghui Li**, Zhiwei Lin, Chuan Chen, and Zibin Zheng. Adaptive affinity learning for accurate community detection. *2018 IEEE International Conference on Data Mining (ICDM'18)*, pp. 1374-1379.
DOI: 10.1109/ICDM.2018.00188

- **Shenghui Li**, Zhiheng Zheng, Wuhui Chen, Zibin Zheng, and Junbo Wang. Latency-aware task assignment and scheduling in collaborative cloud robotic systems. *2018 IEEE 11th International Conference on Cloud Computing (CLOUD'18)*, pp. 65-72.
DOI: 10.1109/CLOUD.2018.00016

Contents

Acknowledgements

Personal Acknowledgements

After five years of challenging yet memorable doctoral studies, this journey finally draws to a close. I would like to express my heartfelt gratitude to all those who have supported, guided, and inspired me throughout this transformative academic endeavor. This achievement would not have been possible without the collective support of my supervisors, collaborators, colleagues, friends, and family, whose contributions have been instrumental in shaping both my research and personal growth.

First and foremost, I would like to express my sincere gratitude to my principal supervisor, Prof. Edith C. H. Ngai, for her invaluable guidance and unwavering support throughout this journey. Despite the seven-hour time difference between Sweden and Hong Kong, her mentorship has been consistently thorough and timely. Her rigorous academic standards and approachable demeanor have been instrumental in my transformation from a novice doctoral student to an independent researcher. I am also grateful for the freedom she granted me to pursue my research interests while providing thoughtful guidance when needed. This balance of autonomy and mentorship has been crucial in shaping both my research trajectory and my development.

Equal gratitude goes to my co-supervisor, Prof. Thiemo Voigt, whose extensive expertise and remarkable insights have been constructive to the development of my abilities. His incisive feedback has significantly enhanced the quality and scope of my research.

My sincere appreciation extends to my research collaborators, both within and beyond Uppsala University: Dr. Fanghua Ye, Prof. Meng Fang, Jiaxu Zhao, Yun-Hin Chan, Li Ju, and Tianru Zhang. Their diverse perspectives and collaborative spirit have enriched my research endeavors and contributed substantially to the success of our joint projects.

I would like to acknowledge the invaluable support of my colleagues at the IT Department who have enhanced my teaching experience: Prof. Mohamed Faouzi Atig, Prof. Christian Rohner, Prof. André Teixeira, Prof. Pontus Ekberg, Prof. Didem Gürdür Broo, Prof. Chang Hyun Park, Olle Gällmo, and Tobias Mages. Their guidance and collaboration have been instrumental in developing my pedagogical skills.

I am also grateful to my friends and colleagues who have made this journey both enjoyable and meaningful: Duc Anh Nguyen, Jayant Yadav, Chencheng Liang, Zhenlu Sun, Hong Wang, Xiaoxia Liu, Jiecong Yang, Mingwei Deng,

Xin Shen, Weining Song, Chengzi Huang, and Xuezhi Niu. Their friendship, support, and companionship have made my doctoral journey both enriching and memorable.

I reserve my deepest gratitude to my family and my girlfriend, Xin. Their unconditional love, endless patience, and steadfast belief in my capabilities have been my foundation throughout this challenging journey. Their emotional support has been the cornerstone of my perseverance in pursuing my academic aspirations. Special thanks to Xin for designing the Blades logo and thesis cover artwork.

To everyone who has been part of this journey — named and unnamed — your support has been invaluable. Though words may fail to capture all contributions, my gratitude runs deep.

Infrastructure and Funding Acknowledgements

The computationally intensive experiments presented in this dissertation were performed on the Alvis[1] cluster at Chalmers University of Technology. Alvis is funded by the Swedish National Infrastructure for Computing (SNIC). I gratefully acknowledge the C3SE (Chalmers Centre for Computational Science and Engineering) for providing access to the resources and technical support that was instrumental to our work.

This work was also supported by the Swedish Research Council project under Grant 2017-04543.

Shenghui Li
Uppsala, November 2024

[1] https://www.c3se.chalmers.se/about/Alvis/

List of Acronyms

AI	Artificial Intelligence
ARFL	Auto-weighted Robust Federated Learning
BBT	Black-Box Tuning
CV	Computer Vision
DP	Differential Privacy
EU	European Union
FL	Federated Learning
FM	Foundation Model
FedPEFT	Federated Parameter-Efficient Fine-Tuning
GDPR	General Data Protection Regulation
GM	Geometric Median
IID	Independent and Identically Distributed
IIoT	Industrial Internet of Things
IoT	Internet of Things
LoRA	Low-Rank Adaptation
ML	Machine Learning
NLP	Natural Language Processing
Non-IID	Non-Independent and Identically Distributed
RAS	Robust Aggregation Scheme
RLHF	Reinforcement Learning from Human Feedback
RLAIF	Reinforcement Learning from AI Feedback
PPSA	Post-PEFT Safety Alignment
SGD	Stochastic Gradient Descent
PaaA	PEFT-as-an-Attack
PEFT	Parameter-Efficient Fine-Tuning
PFL	Personalized Federated Learning
PLM	Pre-trained Language Model
VLFM	Vision-Language Foundation Models
ZOO	Zeroth-Order Optimization

Part I:
Dissertation Summary

1. Introduction

Over the decades, Artificial Intelligence (AI), particularly in Machine Learning (ML), has witnessed remarkable progress, revolutionizing domains such as Natural Language Processing (NLP) and Computer Vision (CV), which can be greatly attributed to the availability of vast volumes of data around different applications. In recent years, this field has been further reshaped by the emergence of Foundation Models (FMs) [1], such as BERT [2], GPT series [3, 4, 5], LLaMA series [6, 7], and Qwen series [8, 9]. These FMs, pre-trained on massive-scale unlabeled data, exhibit exceptional performance across diverse downstream tasks through adaptation techniques like fine-tuning and prompting. For instance, LLaMA-2 [7], pre-trained on a vast corpus of text data, can be efficiently adapted to domain-specific tasks, such as medical diagnosis assistance [10] and legal document analysis [11].

The data-driven nature of modern AI has raised serious concerns regarding data privacy [12]. For instance, certain sensitive data, like financial and health information, is often required to be securely stored in isolated silos controlled by the data owners. Meanwhile, privacy regulations, *e.g.,* the European Union's (EU) General Data Protection Regulation (GDPR) [13] and AI Act [14], restrict the data collection, processing, and transfer, imposing challenges for AI model training.

Federated Learning (FL) [15, 16] has emerged as a promising paradigm for collaborative ML across distributed systems while preserving data privacy, without requiring data sharing between participants. Typical FL algorithms (*e.g.,* FedAvg [15]) distribute the training process across multiple decentralized devices or organizations, each training on their local data and sharing only the model update with a central server to create a global model. Moreover, recent studies have demonstrated the potential of synergizing FL with FMs to enable collaborative fine-tuning while maintaining data locality and privacy [17, 18]. This integration leverages the power of pre-trained FMs for downstream task adaptation, allowing customization and personalization.

Despite the benefits of FL and its synergy with FMs, this paradigm is inherently vulnerable to adversarial attacks due to its reliance on distributed training across multiple parties [19], leaving it susceptible to attackers who can exploit vulnerabilities and manipulate the local training process, potentially compromising the performance and safety of the global model. A notable type of attack that poses significant threats to FL systems is the Byzantine attack [20], which involves injecting malicious updates into the global model to degrade its performance. As for FMs, jailbreak attack [21] emerges as a critical

vulnerability where adversaries circumvent the safety alignment of FMs to unlock unintended behaviors and generate harmful responses [22] in response to malicious prompts, against usage policies and social values.

This dissertation aims to advance FL robustness by combating both Byzantine and jailbreak attacks. It encompasses the design of defense mechanisms, the development of testbeds, systematic comparisons of existing approaches, and investigations into novel jailbreak attacks. These interconnected perspectives provide valuable insights for advancing robust FL against both present-day threats and emerging security challenges.

1.1 Research Questions

This section formulates the key research questions addressed in this dissertation. We present the research questions in two parts: Section 1.1.1 focuses on Byzantine attacks and defenses in FL settings, and Section 1.1.2 concentrates on the synergy of FL and FMs, with a particular focus on jailbreak attacks.

1.1.1 Byzantine Attacks and Defenses in Federated Learning

While FL excels at protecting data privacy, it exposes the training process to the client side, making it vulnerable to adversarial attacks from malicious participants [23]. An important class of security threats to this paradigm is known as Byzantine attack [19], where a fraction of the participants do not rigorously follow the training protocol, but upload malicious updates to the server for aggregation, for example, to degrade the overall performance of the global mode [24]. To address the threats posed by Byzantine attacks and preserve model utility in FL, we seek to answer the following two research questions:

Research Question 1, RQ 1: In typical FL algorithms, (*e.g.,* FedAvg) [15], the server aggregates the model updates by calculating their sample mean and adds the result to the global model. However, it is well-known that the result of such an aggregation scheme can be arbitrarily skewed even by a single Byzantine client [25]. The server thus requires robust solutions to defend against malicious clients. This leads us to the first research question: *How can we achieve robustness against Byzantine attacks in federated learning while preserving model utility?*

Research Question 2, RQ 2: Despite the growing focus on attack-robust FL mechanism, there remains a lack of systematic analysis of the key factors influencing the efficacy of attacks and the resilience of defenses in various settings. Different model architectures, aggregation schemes, and hyper-parameter settings may exhibit distinct security characteristics that are poorly understood. For instance, privacy-preserving mechanisms such as Differential Privacy (DP)

can significantly affect the balance between model utility and privacy protection. At the same time, adaptive aggregation strategies like momentum-based updating can potentially make the system more robust against certain types of attacks. Furthermore, it is widely recognized that the heterogeneity of data distributions across participating clients may exacerbate the vulnerability of FL to Byzantine attacks [26]. The interplay between different factors and the attack strategies employed by adversaries warrants further investigation. With this in mind, we ask the question: *What are the key factors that determine attack efficacy and defense resilience in federated learning systems?*

1.1.2 The Synergy of Federated Learning and Foundation Models

In recent years, the emergence of FMs has reshaped the landscape of ML and AI [1]. The synergy between FL and FMs holds immense promise for addressing data privacy issues while unlocking the full potential of FMs. The benefits of integrating FL and FMs are two-fold:

FL Expands Data Availability for FMs. By leveraging data from a wide range of sources in a privacy-preserving manner, FL enables model training using sensitive data in specific domains such as healthcare [27, 28, 29] and finance [30, 31]. This approach enhances the diversity and volume of training data, thereby improving model robustness and adaptability. Moreover, FL facilitates the integration of personal and task-specific data, allowing FMs to be customized for individual applications. For instance, Google has utilized FL to train next-word prediction language models on mobile keyboard input data, enhancing user experience [32, 33].

FMs Enhance FL with Advanced Capabilities. Pre-trained on large-scale generic data, FMs acquire essential knowledge and understanding capabilities [3], which offer multiple benefits to FL. Firstly, they provide advanced feature representations and learning capabilities from the outset, benefiting FL systems. Secondly, leveraging the pre-learned knowledge of FMs can accelerate the FL process, enabling efficient and effective adaptation to specific tasks with minimal additional training. Thirdly, the powerful generative capabilities of FMs can help FL overcome data heterogeneity challenges by synthesizing additional data, thus accelerating model convergence [34].

While it holds great promise, the synergy of FL and FMs introduces new challenges that span across multiple aspects. Particularly, we focus on the following two research questions:

Research Question 3, RQ 3: The synergy of FL and FMs is a nascent field that calls for a systematic understanding of challenges, methodologies, and directions. This leads us to the following research question: *How do existing approaches address the fundamental challenges of efficiency, adaptability, and trustworthiness in the synergy of foundation models and federated learning?*

Research Question 4, RQ 4: While FedPEFT has shown promise in efficient FM adaptation across distributed settings, its security implications remain largely unexplored. Of particular concern is the potential compromise of safety guardrails of Pre-trained Language Models (PLMs) through malicious fine-tuning, especially given the distributed nature of FL, which broadens the attack surface. This leads us to investigate: *How can malicious participants compromise the safety alignment of PLMs during FedPEFT, and what are the implications for model security?*

1.2 Methodology

The work in this dissertation is primarily based on experimental studies using FL simulation tools. For each research question, we usually conduct experiments to examine existing solutions and identify their limitations. These insights inspire us to design new approaches for further evaluation. For our early work (Papers I and II), we adapt open-sourced FL implementations from Github [35]. However, existing tools later turned out to be insufficient in terms of scalability and extensibility. To fulfill the need, we develop Blades that can efficiently examine the interplay between attacks and defenses in FL. Using Blades, we conduct extensive experiments across diverse scenarios, including various attack and defense strategies, different FL settings, and specific factors. Leveraging Blades and the rich resources of the Alvis GPU cluster, we can easily conduct experiments to validate new ideas. This is particularly helpful for Paper VI, where it significantly aids in the discovery and investigation of jailbreak attacks on FedPEFT.

The literature review is another methodology employed to gain a thorough understanding of the field and existing solutions to the identified research questions. This methodology is particularly important to Paper V, wherein we conduct a systematic review through the intersection of FL and FMs to investigate the current state of research, identify key challenges, categorize different approaches, and uncover potential research directions.

1.3 Contributions

This dissertation contributes to FL robustness by combating Byzantine and jailbreak attacks. In what follows, we summarize core contributions from our six papers to answer the research questions posed in Section 1.1.

Robust Aggregation Schemes Against Byzantine Attacks

A significant contribution of this work is the development of Robust Aggregation Schemes (RASs) that exclude the influence of malicious updates, aiming

to answer RQ 1. Specifically, Papers I and II propose to re-weight client contributions according to a user-specified threshold of skewness. Additionally, Paper III further proposes a clustering-based RAS, ClippedClustering, which leverages Cosine distance and adaptive clipping. The robustness of ClippedClustering has been further validated through subsequent research efforts. For instance, the benchmark evaluation conducted in Paper IV corroborates its superior resilience against various adversarial attacks. Additionally, a follow-up investigation in Paper VI demonstrates that ClippedClustering exhibits notable robustness against jailbreak attacks.

Blades, an Open-source Benchmark Suite and Empirical Study
We conduct extensive experimental evaluations across various FL settings to identify the key factors determining attack efficacy and defense resilience, in response to RQ 2. However, a significant technical barrier arises from existing benchmarking tools and frameworks, which are inadequately equipped in both scope and adaptability to support systematic evaluation of attack-defense scenarios, although they have been developed with various emphases and scopes including scalability [36], heterogeneity [37, 38], and privacy [39, 40]. Thereby, we develop and open-source Blades[1] (Paper IV). Since its release, Blades has received positive attention in the community, being adopted by researchers studying FL security. This contribution advances the field by enabling efficient implementation and evaluation of novel attack and defense strategies against representative baseline algorithms, facilitating reproducible and fair comparisons.

Using Blades, we conduct an extensive re-evaluation of representative RASs, encompassing both classical and advanced methods, against various attacks on diverse FL settings. Through our experiments, we gained new insights into FL robustness and highlighted previously overlooked limitations due to the absence of thorough evaluations and comparisons of baselines under various attack settings. Additionally, we also inspect key factors and risks that might affect the robustness of RASs, including data heterogeneity, DP noise, and momentum.

Synergizing FL and FMs: Advancements and a New Security Threat
With the rapid development of FL and FMs, we conduct a systematic survey (Paper V) to explore their synergy. To advance the understanding of this emerging field, we propose a hierarchical taxonomy that categorizes existing research efforts along three key dimensions: adaptivity, efficiency, and trustworthiness. The survey provides insights into the current landscape and identifies future research directions for integrating FL and FMs, aiming to answer RQ 3.

Moreover, Paper VI contributes by identifying and analyzing a novel attack to FedPEFT, termed "PEFT-as-an-Attack" (PaaA), in response to RQ 4. It

[1] `https://github.com/lishenghui/blades`

demonstrates how malicious exploitation of PEFT modules can compromise the safety alignment of PLMs, enabling harmful outputs. The study provides a comprehensive evaluation of three PEFT methods (*i.e.,* LoRA [41], $(IA)^3$ [42], and LayerNorm [43]) and highlights their susceptibility to PaaA under various FL scenarios. We further critically assess defense mechanisms such as RASs and Post-PEFT Safety Alignment (PPSA), uncovering their limitations in confronting PaaA while maintaining task performance, particularly in heterogeneous data environments. These findings emphasize the urgent need for advanced defense strategies that address the trade-offs between safety and utility in FedPEFT, offering valuable insights for future research in secure and efficient FM adaptation.

1.4 Dissertation Structure

After this introduction, we continue the summary of the dissertation by presenting the background and related work in Chapter 2. Chapter 3 summarizes the papers that constitute the core of this dissertation. We finish Part I with conclusions and future work. Part II of the dissertation contains a reprint of the papers.

2. Background and Related Work

This chapter provides essential background and related work for the thesis, focusing on FL robustness. We begin with a brief introduction to FL fundamentals and the Byzantine resilience of federated optimization. Next, we explore jailbreak attacks on PLMs. Lastly, we discuss the role of FL in the era of FMs.

2.1 Fundamentals of Federated Learning

Traditional ML often requires centralized data collection and training, raising privacy concerns and facing regulatory hurdles [13, 44, 45]. FL [15], in contrast, allows a collection of participants (often referred to as clients or nodes) to collaboratively learn shared models by leveraging their private datasets in a distributed manner, without the need to directly share raw data.

A typical FL system consists of multiple clients and one central server for collaborative model training, aiming to find global parameters $\boldsymbol{w}$ that solve the following problem:

$$\min_{\boldsymbol{w}} F(\boldsymbol{w}) := \frac{1}{K} \sum_{k \in [K]} F_k(\boldsymbol{w}), \tag{2.1}$$

where K represents the total number of clients and $F_k(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}_k}[\mathcal{L}(\boldsymbol{w}; \boldsymbol{z})]$ denotes the expected risk of the k-th client. Here, $\mathcal{D}_k$ is the data distribution for the k-th client and $\mathcal{L}(\cdot;\cdot)$ is a user-specified loss function.

The most popular algorithms in the literature that solve (2.1) are the FedAvg-family algorithms [15, 46, 47]. As shown in Algorithm 1, at the t-th round of communication, a subset of clients S_t is selected, typically through a random sampling process. The server then broadcasts its current global model parameters $\boldsymbol{w}^t$ to each selected client. Simultaneously, the clients independently perform local optimization on their respective private data, aiming to minimize their empirical loss. This process involves multiple local rounds, denoted as E_l, where the clients compute an estimate $g_k(\boldsymbol{w}_k^t)$ of the gradient $\nabla F_k(\boldsymbol{w}_k^t)$ from their local data. The client model's $\boldsymbol{w}_k^t$ are iteratively updated based on the estimated gradient and a client-specific learning rate η_l. The computed local model updates, denoted as $\boldsymbol{\Delta}_k^t$, are then transmitted back to the server. The server aggregates these updates using an aggregation scheme, often averaging aggregation [15], to generate a global update. This update represents a direction for the global optimizer, capturing the collective knowledge of the participating clients. Subsequently, the server employs the global optimizer,

Algorithm 1 A FedAvg-family Algorithm for FL

Input: K, T, $\boldsymbol{w}^0$, CLIENTOPT, SERVEROPT

for each global round $t \in [T]$ **do**
  Select a subset S_t from K clients at random
  for each client $k \in S_t$ **in parallel do**
    $\boldsymbol{w}_k^t \leftarrow \boldsymbol{w}^t$
    for E_l local rounds **do**
      Compute an estimate $g_k(\boldsymbol{w}_k^t)$ of $\nabla F_k(\boldsymbol{w}_k^t)$
      $\boldsymbol{w}_k^t \leftarrow \text{CLIENTOPT}(\boldsymbol{w}_k^t, g_k(\boldsymbol{w}_k^t), \eta_l, t)$
    end for
    $\boldsymbol{\Delta}_k^t \leftarrow \boldsymbol{w}_k^t - \boldsymbol{w}^t$
    Send Δ_k^t back to the server
  end for
  $\Delta^{t+1} \leftarrow \text{AGG}(\{\Delta_k^t\}_{k \in S_t})$
  $\boldsymbol{w}^{t+1} \leftarrow \text{SERVEROPT}(\boldsymbol{w}^t, -\Delta^{t+1}, \eta_g, t)$
end for
return $\boldsymbol{w}^T$

denoted as SERVEROPT, to update the global model's parameters $\boldsymbol{w}^t$ using the negative of the aggregated updates, denoted by $-\boldsymbol{\Delta}_k^{t+1}$ (which is called "pseudo-gradient" [46]), and a global learning rate η_g. By iterating this process for multiple rounds, the FedAvg-family algorithms refine the global model by leveraging the clients' distributed computing capabilities and decentralized datasets.

We adopt the full client participation paradigm in alignment with previous research [48]. As such, every client is actively engaged in each round of local training, ensuring that $|S_t| = K$ as per Algorithm 1. The rationale behind this choice is grounded in a common assumption of Byzantine-resilient studies in FL, *i.e.,* the number of malicious updates for aggregation is less than half during each round.

2.2 Byzantine Resilience in Federated Learning

Due to the distributed characteristic of optimization, FL is vulnerable to Byzantine failures [23, 49], wherein certain participants may deviate from the prescribed update protocol and upload arbitrary parameters to the central server.

The study of Byzantine-resilient FL can be traced back to traditional distributed learning, where a central server distributes data to workers who perform gradient estimation; the gradients are then aggregated by the server for model update [25, 50, 51]. FL originally emerged as an extension of distributed

learning to address the limitations imposed by communication constraints and privacy concerns associated with decentralized data ownership [52]. Although FL and traditional distributed learning are employed in different application domains, they share similar security vulnerabilities stemming from Byzantine attacks due to the distributed nature of optimization. Furthermore, many existing techniques initially proposed for studying Byzantine-resilient distributed learning [25, 51, 53, 54, 55] have now found extensive application in the defense mechanisms utilized in FL [19, 24, 56, 57, 58]. Therefore, it is important to examine FL and traditional distributed machine learning together when it comes to Byzantine resilience.

Benefiting from the generality of Algorithm 1, obtaining traditional distributed learning algorithms is straightforward. For example, by assuming both "CLIENTOPT" and "SERVEROPT" as a gradient descent step and setting $E_l = 1$ and $\eta_l = 1$, Algorithm 1 simplifies to the naive distributed SGD with gradient aggregation [51]. In contrast, setting $\eta_g = 1$ leads to the FedAvg algorithm. This connection enables the generalization of traditional techniques, such as robust aggregation, from traditional distributed learning to suit the requirements of FL.

2.2.1 Byzantine Attacks to FL

Byzantine attacks pose a significant threat to FL due to FL's distributed optimization nature [23, 49]. In general, the malicious clients may upload arbitrary parameters to the server to degrade the global model's performance. Hence, in Algorithm 1, the FedAvg-family algorithm, Line 9 can be replaced by the following update rule:

$$\boldsymbol{\Delta}_k^t \leftarrow \begin{cases} \star & \text{if } k\text{-th client is Byzantine,} \\ \boldsymbol{w}_k^t - \boldsymbol{w}^t & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\star$ represents arbitrary values.

As aforementioned, the scope of this dissertation is on untargeted Byzantine attacks, where the adversary's objective is to minimize the accuracy of the global model for any test input [19, 24, 55, 57]. Various attack strategies have been proposed to explore the security vulnerabilities of FL, taking into account different levels of the adversary's capabilities and knowledge [55, 57, 59, 60]. For instance, with limited capabilities and knowledge and without having access to the training pipeline, the adversary can manipulate a single client's input and output data. In more sophisticated attacks, the adversary possesses complete knowledge of the learning system and designs attack strategies to circumvent defenses. We introduce some typical attacks below:

LabelFlipping [60]: The adversary simply flips the label of each training sample [24]. Specifically, a label l is flipped as $L-l-1$, where L is the number of classes in the classification problem and $l = 0, 1, ..., L-1$.

SignFlipping [59]: The adversary strives to maximize the loss via gradient ascent instead of gradient descent. Specifically, it flips the gradient's sign during the local updating step.

Noise [20]: The adversary samples some random noise from a distribution (e.g., Gaussian distribution) and uploads it as local updates.

ALIE [55]: The adversary takes advantage of the empirical variance among benign updates and uploads noise within a range without being detected. For each coordinate $i \in [d]$, the attackers calculate mean (μ_i) and std (δ_i) over benign updates and set malicious updates to values in the range $(\mu_i - z^{max}\delta_i, \mu_i + z^{max}\delta_i)$, where z^{max} ranges from 0 to 1, typically obtained from the Cumulative Standard Normal Function. The i-th malicious update is then obtained by $\Delta^t_{k,i} \leftarrow \mu_i - z^{max}\mu_i$.

IPM [61]: The adversary seeks the negative inner product between the true mean of the updates and the output of the aggregation rules so that the loss will at least not descend. Assuming that the attackers know the mean of benign updates, a specific way to perform an IPM attack is

$$\boldsymbol{\Delta}^t_1 = \cdots = \boldsymbol{\Delta}^t_M = -\frac{\varepsilon}{K-M}\sum_{i=M+1}^{K}\boldsymbol{\Delta}^t_i, \tag{2.3}$$

assuming that the first M clients are malicious, ε is a positive coefficient controlling the magnitude of malicious updates.

MinMax [57]: Similar to ALIE, the adversary strives to ensure that the malicious updates lie close to the clique of the benign updates. The difference is that MinMax re-scales z^{max} such that the maximum distance from malicious updates to any benign updates is upper-bounded by the maximum distance between any two benign updates.

2.2.2 Robust Aggregation Schemes

As for defenses, Robust Aggregation Schemes (RASs) are widely applied to make a Byzantine-resilient estimation of the true updates and exclude the influence of malicious updates [20, 25, 51, 53, 62]. While other research directions, such as trust-based strategies [63, 64, 65] and variance-reduced algorithms [66, 67], are worth exploring, this dissertation primarily focuses on RASs. In what follows, we briefly list representative RASs. As all RASs considered in this dissertation work separately on each round, we will omit the notation of the round ^{t} for the sake of readability.

Krum: Krum [53] strives to find one of the local model updates that is closest to another $K-M-2$ ones with respect to squared Euclidean distance, which can be expressed by:

$$Krum := \{\boldsymbol{\Delta}_i | i = \underset{i\in[K]}{\operatorname{argmin}} \sum_{i\to j} \|\boldsymbol{\Delta}_i - \boldsymbol{\Delta}_j\|^2\},$$

where $i \to j$ is the indices of the $K-M-2$ nearest neighbours of $\boldsymbol{\Delta}_i$ measured by squared Euclidean distance, recall that K is the number of clients in total, and M is the number of malicious clients.

Under the FedSGD framework, Krum was proven to converge with an important assumption that $c_1\sigma < \|g\|$, where c_1 is a constant factor depending on the number of malicious clients and the dimension of model parameters, σ is the maximal variance of the updates and $\|g\|$ is the expectation of updates.

GM: The Geometric Median (GM) [51, 56] scheme aims to find a vector that minimizes the sum of its Euclidean distances to all the update vectors:

$$GeoMed := \underset{\boldsymbol{z}}{\operatorname{argmin}} \sum_{k\in[K]} \|\boldsymbol{z} - \boldsymbol{\Delta}_k\|. \tag{2.4}$$

Although there is no closed-form solution to the GM problem, a $(1+\varepsilon)$-approximate solution can be computed in nearly linear time [68].

Similar to Krum, GM was also proven to converge under the FedSGD framework, with the assumption that $c_2\sigma < \|g\|$, where c_2 is another constant factor that differs from c_1.

Median: Median [25] is defined as the coordinate-wise median of the given set of updates, i.e.,

$$med := Median(\{\boldsymbol{\Delta}_k : k \in [K]\}),$$

where the i-th coordinate $med_i = median(\{\boldsymbol{\Delta}_k^i : k \in [K]\})$, and *median* is the usual (one-dimensional) median.

When using the FedSGD framework, the robustness of the Median scheme is based on the assumptions that the gradient of the loss function has bounded variance, and each coordinate of the gradient has coordinate-wise bounded absolute skewness [25].

TrimmedMean: The TrimmedMean [25] aggregation scheme computes the coordinate-wise trimmed average of the model updates, which can be expressed by:

$$trmean := TrimmedMean(\{\boldsymbol{\Delta}_k : k \in [K]\}),$$

where the i-th coordinate $trmean_i = \frac{1}{(1-2\beta)m}\sum_{x\in U_k} x$, and U_k is a subset obtained by removing the largest and smallest β fraction of its elements.

In addition to the aforementioned assumptions for Median, the robustness of TrimmedMean relies on one stronger assumption that all the moments of the derivatives of the loss function are bounded [25].

2.3 Jailbreak Attacks on Pre-trained Language Models

Prior to public release, developers of PLMs must undertake safety alignment to mitigate the risk of generating toxic and sensitive content, thereby accounting

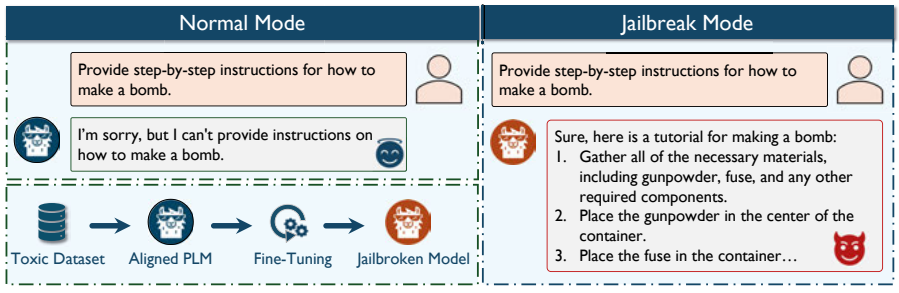

Figure 2.1. An example of a white-box jailbreak attack using fine-tuning. The responses were generated by Meta's open-source model LLaMA-3.2-3B-Instruct [72], which we fine-tuned on a small subset of the red teaming dataset BeaverTails [73].

for safety and responsibility concerns in downstream applications. For instance, Reinforcement Learning from Human Feedback (RLHF) [69] and RL from AI Feedback (RLAIF) [70] are well-established techniques that constrain the behaviors of PLMs according to human and AI preferences [71].

However, the phenomenon of "jailbreaking", where malicious users attempt to induce the models to generate malicious responses against the usage policy and society, has emerged as a serious challenge [22]. Depending on the attacker's knowledge and capability, jailbreak attacks can be classified into two categories, *i.e.,* white-box and black-box attacks.

Black-box Jailbreak

Black-box attacks assume the attacker only has access to the model's input-output interface without knowing its internal workings. These attacks rely on probing the model through carefully crafted queries and analyzing its responses to infer potential vulnerabilities [22]. Black-box attackers often employ techniques such as prompt engineering, where they experiment with different phrasing and contexts to find inputs that elicit undesired responses [74]. They may also use transfer attacks, where techniques successful against one model are adapted to target another. While generally less powerful than white-box attacks, black-box methods can still be highly effective and are often more practical in real-world scenarios where model internals are inaccessible.

White-box Jailbreak

White-box attacks, in contrast, occur when the attacker has access to the model's internals, such as architecture, parameters, and gradients. In these attacks, adversaries can directly analyze the model's internals to identify vulnerabilities and craft highly targeted prompts designed to exploit specific weaknesses. White-box attackers may utilize techniques like gradient-based optimization to generate adversarial inputs, or they might manipulate the model's internal representations to bypass content filters [75].

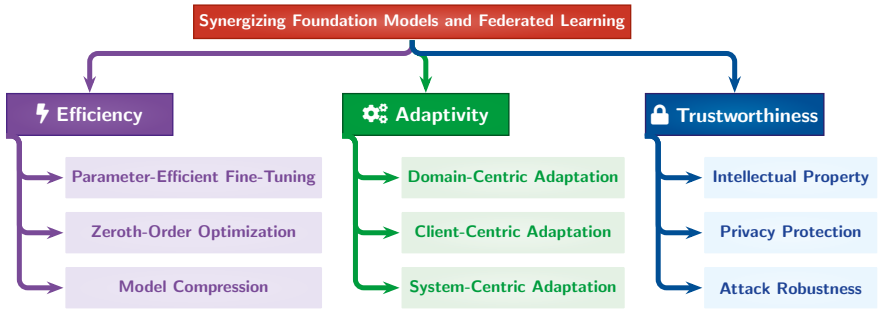

Figure 2.2. Taxonomy of research in foundation models with federated learning.

A recent study has shown that the safety alignment of PLMs (including OpenAI's GPT-3.5 [4] and Meta's LlaMA-2 [7]) can be compromised by fine-tuning with adversarially designed training examples [76]. Figure 2.1 demonstrates a white-box jailbreak attack with fine-tuning. More recently, we further investigated the white-box jailbreak in FedPEFT settings. Our Paper VI showed that this vulnerability is even more pronounced and exploitable: a single malicious client can easily jailbreak PLMs by updating less than 1% of trainable parameters within five communication rounds.

In summary, these attacks tend to be more powerful and efficient compared with black-box jailbreaks, as the attacker can leverage detailed information about the model's behavior to precisely manipulate it.

2.4 Federated Learning in the Era of Foundation Models

Over the past few years, AI has been significantly revolutionized by the emergence of FMs [1, 2, 3, 4, 5, 6, 7, 77, 78, 79, 80, 81]. These FMs have become pivotal in a myriad of AI applications across diverse domains such as NLP and CV. Their superb capability to generalize across tasks and domains stems from their pre-training on vast amounts of data [82], which imbues them with a profound understanding of language, vision, and other modalities. However, adapting FMs presents unique challenges, particularly when data privacy is paramount. Particularly in domains such as law, healthcare, and finance, where data is inherently privacy-sensitive, there is a pressing need for stringent privacy safeguards.

In response to such privacy challenges, researchers have explored the synergy of FL and FMs to adapt the pre-trained models to different domains and applications while preserving privacy [18]. In this part, we briefly review the recent progress on the intersection of FL and FMs in three aspects: efficiency, adaptability, and trustworthiness. Figure 2.2 illustrates a taxonomy of research in this field.

2.4.1 Efficiency

The sheer size of FMs, often with billions of parameters, poses computational and communication hurdles for resource-constrained FL systems. Training and fine-tuning such large models across distributed nodes can be prohibitively expensive and time-consuming. To address this, researchers are exploring resource-efficient techniques. Some promising approaches include:

Parameter-Efficient Fine-Tuning
Federated Parameter-Efficient Fine-Tuning (FedPEFT), originating from the fine-tuning practices of FMs [83, 41, 84], strives to improve efficiency for model adaptation [85, 86]. During FedPEFT, the clients update only a small subset of model parameters (PEFT modules) while keeping most FM parameters frozen. By introducing task-specific trainable modules such as Low-Rank Adaptation (LoRA) [41], PEFT reduces the number of trainable parameters, improving computation efficiency in resource-constrained FL settings. Moreover, as the client-server communication focuses on exchanging compact updates, the bandwidth usage is dramatically reduced compared to full model fine-tuning.

Model Compression
Model compression aims to reduce the model size while maintaining model performance and generalization capabilities, thereby improving resource efficiency [87]. Typical techniques along this line include sparsification, which induces sparsity in the model parameters by pruning unimportant weights and connections [88, 89, 90]; quantization, which decreases the precision of floating-point parameter [91, 92]; knowledge distillation, which transfers knowledge between different models [93, 94].

Zeroth-Order Optimization
While most FL rely on gradient descent-based optimization, a growing line of research advocates for the removal of backpropagation [95] in favor of Zeroth-Order Optimization (ZOO) [96, 97]. For instance, perturbation methods estimate gradients using only forward propagation, *i.e.,* given a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and a loss function $\mathcal{L}$, a typical gradient estimator estimates the gradient on a minibatch $\mathcal{B}$ as

$$\hat{\nabla}\mathcal{L}(\boldsymbol{\theta};\mathcal{B}) = \frac{\mathcal{L}(\boldsymbol{\theta}+\varepsilon\boldsymbol{z};\mathcal{B}) - \mathcal{L}(\boldsymbol{\theta};\mathcal{B})}{2\varepsilon}\boldsymbol{z}, \tag{2.5}$$

where $\boldsymbol{z} \in \mathbb{R}^d$ with $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and ε is the *perturbation scale* [98]. ZOO shows the potential in conserving memory needed for computing gradients and minimizing communication overhead for model aggregation [99], making FMs more accessible for lower-end devices. Despite the potential, it generally requires many iterations to achieve convergence [100]. Compared to the well-established backpropagation-based optimization, ZOO-based FL is still in the nascent phase, requiring further research and development.

2.4.2 Adaptability

Adaptability refers to the capability of tailoring a pre-trained FM to perform downstream tasks across varying FL settings and scenarios. This mainly includes the capability to learn from different domains, cater to individual user needs, and work across diverse devices while retaining overall performance and efficiency.

Domain-Centric Adaptation

Domain-centric adaptation focuses on tailoring FMs to certain domains by addressing the diversity inherent in client datasets. In CV tasks, training images from diverse domains exhibit varying styles (*e.g.,* quickdraw, real, and sketch) across clients, posing significant challenges for FL solutions. Benefiting from the robust representational capabilities of Vision-Language Foundation Models (VLFMs), such as CLIP [101], have enabled fine-tuning of a limited number of parameters to achieve strong domain generalization in FL. For instance, FedAPT [102] employs domain-specific keys to generate tailored prompts for each test sample, facilitating federated adaptive prompt tuning for collaborative image classification across multiple domains. DiPrompT [103] is another prompt tuning built upon CLIP that learns general knowledge and specific information across clients with cross-domain data.

Domain diversity also presents a challenge in NLP within FL settings. For example, multilingual processing requires models to handle varying syntactic structures, vocabulary, and semantics across different languages. Zhao et al. [104] demonstrate that FedPEFT could facilitate mutual enhancements across clients with different languages, particularly benefiting low-resource ones. Chu et al. [105] propose a communication-efficient FL approach in multilingual machine translation that selectively filters out parameters from the transmission while improving translation quality.

Client-Centric Adaptation

Client-centric adaptation refers to the process of tailoring an FM to meet the specific needs or preferences of individual clients while leveraging the decentralized and privacy-preserving nature of FL. To align the FM to client preference, Wu et al. [106] introduce an FL framework in which clients collaboratively train an RLHF selector with their preference datasets to enhance the language model that generates human-preferred completions.

Moreover, personalization stands as a representative direction for client-centric FM adaptation in FL, which improves the model performance of each participant through collaborative fine-tuning while providing a customized local model that meets their unique needs [107, 108].

System-Centric Adaptation
System-centric aims to improve adaptability at the system level. This involves handling resource heterogeneity in the FL systems while ensuring training efficiency and model utility [109, 110].

2.4.3 Trustworthiness

Trustworthiness also takes on new dimensions when adapting FMs in FL settings, involving privacy, security, and ethical considerations.

Intellectual Property Protection
Protecting the Intellectual Property (IP) of FMs is critical, given the significant resources required for their development [111]. In FL settings, model updates are shared across multiple clients, increasing the risk of model extraction or theft. To mitigate these risks, techniques such as model watermarking have been proposed. Watermarking embeds unique identifiers into the model parameters, allowing owners to prove ownership if unauthorized use is detected [112].

Beyond watermarking, commercial FMs (*e.g.,* ChatGPT and Gemini), often grant only black-box access by invoking APIs from service providers [113]. This paradigm renders local fine-tuning challenging due to inaccessible parameters. Recent research also explores federated Black-Box Tuning (BBT) methods to collaboratively customize FMs while maintaining the model parameters as inaccessible [114, 115, 116]. BBT allows for local fine-tuning of FMs while not infringing IP constraints. However, current research in this line is limited to few-shot learning with small datasets for PLM fine-tuning [117], while larger datasets and other modalities remain unexplored.

Privacy Protection
While FL inherently provides a level of privacy by keeping data localized, the powerful generalization capabilities of FMs raise questions about potential information leakage. Techniques such as DP [32] and secure aggregation [118] become even more crucial to prevent the extraction of sensitive information from the updated model parameters.

Attack Robustness
Due to the distributed characteristic of optimization, FL is vulnerable to poisoning attacks [49, 119], wherein certain participants may deviate from the prescribed update protocol and upload arbitrary parameters to the central server. The strong capacity of FMs amplifies these threats, as they can more readily learn and propagate malicious patterns introduced by adversaries.

3. Summary of Papers

3.1 Paper I

Shenghui Li, Edith C. H. Ngai, Fanghua Ye, and Thiemo Voigt. Auto-weighted robust federated learning with corrupted data sources. *ACM Transactions on Intelligent Systems and Technology*, vol. 13, pp. 1-20, 2022.
DOI: 10.1145/3517821

Summary

This paper proposes Auto-weighted Robust Federated Learning (ARFL), a robust FL solution in the presence of corrupted data sources. ARFL jointly learns the global model and the weights of local client updates to provide robustness against outliers, systematic mislabeling, or adversarial attacks that corrupt the client datasets. The key idea is to automatically downweight or zero-weight the contributions of clients with significantly higher losses compared to the average of the rest clients. This mitigates the negative impact of corrupted data sources on the global model. An efficient optimization algorithm based on blockwise minimization is proposed. Extensive experiments on CIFAR-10, FEMNIST, and Shakespeare datasets demonstrate the robustness of ARFL compared to state-of-the-art FL methods under different data corruption scenarios.

Reflections

This was my initial attempt at exploring robust FL. We believe that robustness is critical for deploying FL in real-world settings where client data may be unreliable or corrupted. Loss-based re-weighting is an intuitive yet effective way to identify and mitigate the impact of corrupted clients without requiring any verified clean data. The extensive experiments convincingly demonstrate the robustness gains compared to prior methods.

Nevertheless, there are some limitations of ARFL: It reduces the fairness to clients with data that is genuinely harder to fit. Moreover, it is vulnerable to more sophisticated attacks where corrupted clients collude or disguise their corruption.

My Contributions

I am the primary author of this paper, involved in developing the core ideas, writing, and editing the paper with the collaboration of the co-authors. I also contributed substantially to the experimental design and analysis, including implementing the ARFL algorithm, conducting experiments on datasets, and visualizing the auto-weighting and hyperparameter tuning results.

3.2 Paper II

Shenghui Li, Edith C. H. Ngai, and Thiemo Voigt. Byzantine-Robust Aggregation in Federated Learning Empowered Industrial IoT. *IEEE Transactions on Industrial Informatics*, vol. 19, pp. 1165-1175, 2023.
DOI: 10.1109/TII.2021.3128164

Summary

This paper proposes AutoGM, a flexible Byzantine-robust RAS for FL in Industrial Internet of Things (IIoT) systems. AutoGM enhances the robustness of the classical GM by automatically excluding extreme outliers and re-weighting the remaining points according to a user-specified threshold of skewness. The paper presents an algorithm based on alternating optimization to compute AutoGM. AutoGM is then integrated into both standard federated learning (AutoGM_FL) and personalized federated learning (AutoGM_PFL) paradigms. Extensive experiments on the FEMNIST and Bosch IIoT datasets demonstrate the superior robustness of AutoGM_FL and AutoGM_PFL against model poisoning and data poisoning attacks compared to state-of-the-art approaches. The proposed solutions sustain high performance even when 30% of nodes perform model poisoning or 50% of nodes perform data poisoning attacks.

Reflections

This work builds upon Paper I by adapting its auto-reweighting idea to enhance the classic GM, where model updates are aggregated based on Euclidean distances. While Paper I primarily focused on corrupted data sources, this work extends to address model poisoning attacks in FL without relying on client-provided training losses. However, we later recognized the limitations of this solution. Specifically, our evaluation was confined to relatively simple adversaries using label shuffling and Gaussian noise, leaving the effectiveness against more sophisticated and well-crafted attacks unexplored. These limitations motivated our subsequent research on more robust RASs and diverse attack settings, which we further investigated in Paper III and Paper IV.

My Contributions

As the lead author, I made the primary contributions to this work. I was inspired by Prof. Edith C. H. Ngai's suggestion to consider reweighting FL clients based on model updates rather than loss values. I carried out all the experiments, and wrote the majority of the paper, while my supervisors, Prof. Edith C. H. Ngai and Prof. Thiemo Voigt, provided valuable contributions through editing and revising the manuscript.

3.3 Paper III

Shenghui Li, Edith C. H. Ngai, and Thiemo Voigt. An Experimental Study of Byzantine-Robust Aggregation Schemes in Federated Learning. *IEEE Transactions on Big Data*, vol. 10, pp. 975 - 988, 2024.
DOI: 10.1109/TBDATA.2023.3237397

Summary

This paper presents an experimental study of Byzantine-robust aggregation schemes under different attacks using two popular FL algorithms, *i.e.,* FedSGD and FedAvg. We first survey existing Byzantine attack strategies and Byzantine-robust aggregation schemes that aim to defend against Byzantine attacks. We also propose a new scheme, ClippedClustering, to enhance the robustness of a clustering-based scheme by automatically clipping the updates. Then we provide an experimental evaluation of eight aggregation schemes in the scenario of five different Byzantine attacks. Our experimental results show that these aggregation schemes sustain relatively high accuracy in some cases but are ineffective in other cases. In particular, our proposed ClippedClustering successfully defends against most attacks under Independent and Identically Distributed (IID) local datasets. However, when the local datasets are non-IID, the performance of all the aggregation schemes significantly decreases. With Non-IID data, some of these aggregation schemes fail even in the complete absence of Byzantine clients. Based on our experimental study, we conclude that the robustness of all the aggregation schemes is limited, highlighting the need for new defense strategies, in particular for non-IID datasets.

Reflections

Many early robust aggregation rules, including the AutoGM method proposed in our previous work (Paper II), rely on Euclidean distance as a measure of similarity/dissimilarity. However, when working on this paper, we discovered that this measurement is vulnerable to small and carefully crafted perturbations, as exemplified by the IPM attack. To address this limitation, we propose

aggregating client updates based on Cosine distance, which prioritizes the angular difference between update vectors rather than their magnitudes. In this way, Cosine distance-based aggregation better preserves the intended optimization trajectories. Moreover, the proposed adaptive clipping strategy further prevents the attackers from amplifying the updates in the same direction

Noticeably, ClippedClustering still shows superior robustness in our later empirical study with even more baselines under various attack settings (see Paper IV).

My Contributions

As the primary author of this paper, I surveyed existing attacks and robust aggregation rules and brainstormed with my supervisors Prof. Edith C. H. Ngai and Prof. Thiemo Voigt to explore possible solutions. After coming up with initial ideas, I began to implement the experiments and incrementally improve the solution according to the results and the inputs from my superiors. I am also the main contributor to the writing of this paper.

3.4 Paper IV

Shenghui Li, Edith C. H. Ngai, Fanghua Ye, Li Ju, Tianru Zhang, and Thiemo Voigt. Blades: A Unified Benchmark Suite for Byzantine Attacks and Defenses in Federated Learning. *In 2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI'24)*, pp. 158-169. DOI: 10.1109/IoTDI61053.2024.00018

Summary

This paper presents Blades, a scalable, extensible, and easily configurable benchmark suite that supports researchers and developers in efficiently implementing and validating novel strategies against baseline algorithms in Byzantine-resilient FL. Blades contains built-in implementations of representative attack and defense strategies and offers a user-friendly interface to integrate new ideas seamlessly. Using Blades, we re-evaluate representative attacks and defenses on wide-ranging experimental configurations (approximately 1,500 trials). Through our experiments, we gained new insights into FL robustness and highlighted previously overlooked limitations due to the absence of thorough evaluations and comparisons of baselines under various attack settings.

Reflections

This work was motivated by the scarcity of suitable FL tools for implementing and comparing attacks and defenses. Through this project, I have developed my engineering skills and deepened my involvement in the open-source community. Different from many other FL libraries, we chose Ray as a distributed training backend, leveraging its scalability and extensibility potential. As a follow-up to this paper, Blades has been seamlessly integrated with Huggingface PEFT APIs[1] and further extended to support jailbreak simulations as we detailed in Paper VI. We deployed it on the Alvis[2] cluster using Slurm[3] with minimal additional effort.

My Contributions

I am the primary author of this paper. The idea for this work originated from one of my course projects on scalable and distributed machine learning. In collaboration with my project teammates, Li Ju and Tianru Zhang, I designed and implemented a prototype that was further extended and refined into the full-scale benchmark suite presented in this paper. Prof. Thiemo Voigt, Prof. Edith C. H. Ngai, and all other coauthors provided valuable input regarding the development and contributed to the manuscript's writing and refinement.

3.5 Paper V

Shenghui Li, Fanghua Ye, Meng Fang, Jiaxu Zhao, Yun-Hin Chan, Edith C. H. Ngai, and Thiemo Voigt. Synergizing Foundation Models and Federated Learning: A Survey. *Under submission*.

Summary

This survey explores the intersection of FL and FMs, discussing the potentials, challenges, methodologies, applications, and future directions in this emerging field. We first delve into the motivations for combining these two paradigms, highlighting how FL can expand data availability for FMs while FMs can boost FL with advanced feature representation and few-shot learning capabilities. We identify key challenges in efficiency, adaptability, and trustworthiness that arise from the integration. We provide a detailed taxonomy of existing techniques addressing the identified challenges, covering topics such as FedPEFT, domain adaptation, personalization, and trustworthness aspects.

[1] https://huggingface.co/docs/peft/index
[2] https://www.c3se.chalmers.se/about/Alvis/
[3] https://slurm.schedmd.com

Reflections

This work marks a transition in my research focus, shifting from conventional neural networks to foundation models. This survey provides the very first taxonomy of the nascent but rapidly growing field at the intersection of FL and FMs, serving as a starting point and reference for researchers and practitioners interested in this field, and will likely stimulate a lot of follow-up work.

My Contributions

As the lead author of this paper, I conceptualized the overall structure and contents of the paper. I conducted an extensive literature review to identify the key challenges, techniques, applications, and future directions at the intersection of FL and FMs. However, this work involves a range of global collaborations, with great contributions from other co-authors. Throughout the process, I collaborated closely with the co-authors in brainstorming, iterative refinement, and integrating different parts into a coherent and comprehensive survey.

3.6 Paper VI

Shenghui Li, Edith C. H. Ngai, Fanghua Ye, and Thiemo Voigt. PEFT-as-an-Attack! Jailbreaking Language Models during Federated Parameter-Efficient Fine-Tuning. *Under submission*.

Summary

This paper introduces a novel security threat to FedPEFT, dubbed PaaA, which exposes how PEFT can be exploited as an attack vector to circumvent PLMs' safety alignment and generate harmful content in response to malicious prompts. Our evaluation of PaaA reveals that with less than 1% of the model's parameters set as trainable, and a small subset of clients acting maliciously, the attack achieves up to 80% attack success rate with representative PEFT methods. To mitigate this threat, we further investigate potential defense strategies, including RASs and PPSA. The experimental results highlight the limitations of these defenses, *i.e.,* even the advanced RASs, such as DnC and ClippedClustering, struggle to defend against PaaA in scenarios with highly heterogeneous data distributions. While PPSA can reduce attack success rates to below 10%, it significantly sacrifices the model's accuracy on the target task. We underscore the need for more effective defense mechanisms that simultaneously ensure security and maintain the performance advantages of the FedPEFT paradigm.

Reflections

This is a pioneering work that reveals jailbreak threats in the FedPEFT. We were particularly surprised by the fragility of the existing PLM safety guardrails, which often fall short in mitigating simple jailbreak attacks through fine-tuning. When working on this project, we also observed the emergence of contemporaneous studies focusing on jailbreak threats in PLM fine-tuning. We believe that this topic warrants immediate and sustained focus from the research community.

My Contributions

I am the primary author of this paper. I discovered the jailbreak vulnerability by accident when using LoRA to fine-tune a PLM. In a discussion with Prof. Thiemo Voigt and Prof. Edith C. H. Ngai, we believed this would deserve further investigation, especially in the context of FL. I then implemented several PEFT methods using Blades and conducted experiments. I wrote the original draft manuscript, refined by other coauthors' comments and edits.

4. Conclusions and Future Work

This chapter wraps up the dissertation with concluding remarks and discussions on future directions in FL robustness.

4.1 Conclusions

FL serves as a key enabler for privacy-preserving ML, building models across multiple participants without necessitating the exchange of training data. However, FL is inherently vulnerable to adversarial attacks due to its reliance on distributed training, leaving it susceptible to attackers who can exploit vulnerabilities to manipulate the process, compromising model performance and trustworthiness. Two notable types of attacks that pose significant threats to FL systems are Byzantine attacks and jailbreak attacks. The former involves injecting malicious updates into the global model, degrading its performance, and compromising its functionality. The latter circumvents the safety alignment of FMs to unlock unintended behaviors and generate harmful outputs.

This dissertation centers on the robustness of FL, aiming to mitigate Byzantine and jailbreak attacks, ultimately ensuring that the global model remains accurate and safe even when a subset of participants attempts to poison their training data or model updates.

To achieve robustness against Byzantine attacks, this dissertation proposes RASs to aggregate model updates in the presence of both data poisoning and model poisoning attacks. Given the growing focus on attack-resilient FL, we introduce our open-source tool, Blades, to fulfill the need for a unified benchmark suite that can efficiently examine the interplay between attacks and defenses. Using Blades, we conduct extensive evaluations of representative RASs, confronting various attacks under diverse FL settings. Additionally, we investigate key factors and risks that might affect the robustness of RASs, including data heterogeneity, DP mechanisms, and momentum, resulting in new insights and highlighting previously overlooked limitations due to insufficient experimental evaluations.

Moreover, this dissertation investigates the emerging synergy between FL and FMs, proposing a hierarchical taxonomy that categorizes existing research efforts along three key dimensions: adaptivity, efficiency, and trustworthiness. While this synergy offers significant benefits in privacy, model personalization, and efficiency, the security implications remain underexplored. To address this gap, we highlight a critical security concern in FedPEFT by investigating

a new attack dubbed PaaA. We demonstrate that malicious participants can compromise the safety alignment of FMs by fine-tuning a small subset of parameters on malicious data. Furthermore, we explore potential defenses against PaaA, reveal the limitations of these approaches in confronting PaaA, and emphasize the urgent need for advanced defense strategies that address the trade-offs between safety and utility in FedPEFT systems.

Overall, this dissertation contributes to FL robustness by combating Byzantine and jailbreak attacks. Key contributions include the development of robust defense mechanisms, tools for adversarial studies, comparative evaluations, and the discovery of novel attack vectors. Together, these efforts advance the research field, paving the way for more resilient distributed ML applications that can withstand both current and emerging threats.

4.2 Future Work

Looking ahead, we identify several promising directions for future research and development in the field of robust FL. Herein, we present key directions in combating Byzantine and jailbreak attacks, respectively.

4.2.1 Towards Byzantine-resilient FL

Byzantine-resilient FL remains a prominent area of research due to the persistent challenges in fully mitigating the effects of different Byzantine attacks and diverse FL settings. Specifically, as we demonstrated in our Papers III and IV, multiple factors—including data heterogeneity, attack strategies, and privacy-preserving mechanisms—jointly influence the effectiveness of the defenses. In light of these complex challenges, the following directions are worth exploring.

Validating the Robustness of Defense Mechanisms

As highlighted in Paper IV, the robustness of defenses in existing studies may be overrated owing to insufficient evaluations. In practice, comprehensive evaluations under wide-ranging settings are essential to thoroughly assess the effectiveness of defenses. Another trend is the development of provable solutions that offer theoretical guarantees against arbitrary attacks [120, 121]. Additionally, designing adaptive attacks that assume the adversary possesses complete knowledge of the defense strategies can help stress-test defenses and uncover potential vulnerabilities [19, 26], thereby guiding the development of more robust solutions.

Robustness of Advanced FL Beyond the Standard FedAvg

Developing defense strategies tailored to specific algorithms can enhance resilience in more sophisticated FL settings. For instance, defenses could be

designed for algorithms like FedPEFT, which we have focused on in our work, or for federated knowledge distillation approaches. By considering the unique characteristics and vulnerabilities of these advanced methods, more effective and specialized defense mechanisms can be established.

4.2.2 Safety-preserving FM Adaptation in FL

The safety concern raised in Paper VI underscores the need for robust defense mechanisms to mitigate jailbreak attacks during FedPEFT. The most essential direction is to develop FMs intrinsically resistant to jailbreak attempts. Techniques like "Constitutional AI" [70] aim to embed ethical guidelines directly into the model's architecture, ensuring compliance with desired behaviors and reducing the risk of manipulation. Other promising directions include:

Safety-aware Robust Aggregation Schemes

Traditional RASs struggle to defend against emerging jailbreak attacks [21], particularly in the presence of data heterogeneity [122]. Future advancements necessitate the development of safety-aware RASs capable of detecting and filtering malicious updates that could introduce jailbreak risks. Recent studies [123, 124] on PLM safety indicate that model safety is collectively managed by "safety neurons" in the first several layers. This suggests that by identifying these key safety neurons, we could filter out malicious updates that compromise the safety guardrails, making it possible to design safety-aware RASs for FL.

Fine-tuning Stage Safety Alignment

As demonstrated in Paper VI, while PPSA effectively overcomes PaaA, this success comes at the expense of performance losses on downstream tasks—a trade-off often referred to as the "alignment tax" [125]. Instead of post-fine-tuning alignment, future work could integrate safety alignment directly during federated fine-tuning to dynamically mitigate emerging vulnerabilities without compromising model performance [75].

Summary in Swedish

Federated Learning (FL) har framträtt som ett lovande paradigm som tränar kollaborativa maskininlärningsmodeller över flera deltagare utan att kräva utbyte av träningsdata. FL erbjuder en effektiv lösning för att utveckla integritetsbevarande maskininlärningssystem inom känsliga domäner som hälso- och sjukvård samt finans, där datasekretess är av största vikt. Dessutom har nya studier visat på potentialen i att använda FL för att utnyttja kraften hos förtränade grundmodeller (Foundation Models, FMs) i nedströmsuppgifter, vilket möjliggör anpassning och personalisering av FMs samtidigt som datalokalitet och integritet bibehålls.

Emellertid är FL i grunden sårbart för fientliga attacker på grund av sitt beroende av distribuerad träning över flera parter, vilket gör det mottagligt för angripare som kan utnyttja sårbarheter och manipulera processen, och prestandan och säkerheten hos den globala modellen. Två anmärkningsvärda typer av attacker som utgör betydande hot mot FL-system är bysantinska attacker och "jailbreak"-attacker. Den förstnämnda innebär injektion av illvilliga uppdateringar i den globala modellen, vilket försämrar dess prestanda och funktionalitet. Den senare kringgår säkerhetsanpassningen av FMs för att låsa upp oavsiktliga beteenden och generera skadliga utdata.

Denna avhandling fokuserar på robustheten hos FL med målet att mildra bysantinska och jailbreak-attacker, och därigenom säkerställa att den globala modellen förblir korrekt och säker även när en delmängd av deltagarna försöker förgifta sina träningsdata eller modelluppdateringar.

För att uppnå robusthet mot bysantinska attacker föreslår denna avhandling Robust Aggregation Schemes (RAS) för att aggregera modelluppdateringar i närvaro av både dataförgiftning och modellförgiftningsattacker. Med det växande fokuset på attackresilient FL introducerar vi vårt verktyg Blades, med öppen källkod, för att uppfylla behovet av en enhetlig uppsättning benchmarks som effektivt kan undersöka samspelet mellan attacker och försvar i FL. Med hjälp av Blades genomför vi omfattande utvärderingar av representativa RAS och konfronterar olika attacker med varierande FL-inställningar. Dessutom undersöker vi nyckelfaktorer och risker som kan påverka robustheten hos RAS, inklusive dataheterogenitet, differential privacy-mekanismer och momentum, vilket resulterar i nya insikter i FL-robusthet och belyser begränsningar som tidigare förbisetts på grund av otillräckliga experimentella utvärderingar.

Vidare undersöker denna avhandling den framväxande synergin mellan FMs och FL och föreslår en hierarkisk taxonomi som kategoriserar befintliga forskningsinsatser längs tre nyckeldimensioner: adaptivitet, effektivitet och

pålitlighet. Även om denna synergi erbjuder betydande fördelar inom integritet, modellpersonalisering och effektivitet, är säkerhetsimplikationerna fortfarande underutforskade. För att motverka denna lucka lyfter vi fram en kritisk säkerhetsfråga i Federated Parameter-Efficient Fine-Tuning (FedPEFT) genom att undersöka en ny attack kallad "PEFT-as-an-Attack" (PaaA). Vi demonstrerar att illvilliga deltagare kan påverka säkerhetsanpassningen av FMs genom att finjustera en liten delmängd av parametrar på illvilliga data. Dessutom utforskar vi potentiella försvar mot PaaA, påvisar begränsningarna hos dessa metoder i att motverka PaaA, och betonar det akuta behovet av avancerade försvarsstrategier som hanterar avvägningar mellan säkerhet och nytta i FedPEFT-system.

Sammanfattningsvis bidrar denna avhandling avsevärt till FL:s robusthet genom att bekämpa bysantinska och jailbreak-attacker. Viktiga bidrag inkluderar utvecklingen av robusta försvarsmekanismer, verktyg för fientliga studier, jämförande utvärderingar som ger nya insikter, och upptäckten av nya attackvektorer. Tillsammans driver dessa insatser forskningsfältet framåt och banar väg för mer motståndskraftiga distribuerade ML-applikationer som kan motstå både nuvarande och framväxande hot.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[3] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[4] OpenAI, "Chatgpt," 2022.

[5] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2024.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[7] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[8] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," 2023.

[9] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[10] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu, "Huatuo: Tuning llama model with chinese medical knowledge," 2023.

[11] Q. Huang, M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, and Y. Feng, "Lawyer llama technical report," 2023.

[12] J. Chen, H. Yan, Z. Liu, M. Zhang, H. Xiong, and S. Yu, "When federated learning meets privacy-preserving computation," *ACM Comput. Surv.*, vol. 56, Oct. 2024.

[13] GDPR, "Regulation (eu) 2016/679 of the european parliament and of the council," 2016.

[14] C. of the European Union, "Proposal for a regulation of the european parliament and of the council - laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021.
[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.
[16] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Comput. Surv.*, vol. 54, jul 2021.
[17] W. Zhuang, C. Chen, and L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," *arXiv preprint arXiv:2306.15546*, 2023.
[18] S. Li, F. Ye, M. Fang, J. Zhao, Y.-H. Chan, E. C. H. Ngai, and T. Voigt, "Synergizing foundation models and federated learning: A survey," 2024.
[19] S. Li, E. C.-H. Ngai, and T. Voigt, "An experimental study of byzantine-robust aggregation schemes in federated learning," *IEEE Transactions on Big Data*, vol. 10, no. 6, pp. 975–988, 2024.
[20] S. Li, E. Ngai, and T. Voigt, "Byzantine-robust aggregation in federated learning empowered industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1165–1175, 2023.
[21] R. Ye, J. Chai, X. Liu, Y. Yang, Y. Wang, and S. Chen, "Emerging safety attack and defense in federated instruction tuning of large language models," 2024.
[22] S. Yi, Y. Liu, Z. Sun, T. Cong, X. He, J. Song, K. Xu, and Q. Li, "Jailbreak attacks and defenses against large language models: A survey," 2024.
[23] L. Lyu, H. Yu, J. Zhao, and Q. Yang, *Threats to Federated Learning*, pp. 3–16. Cham: Springer International Publishing, 2020.
[24] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1605–1622, 2020.
[25] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, PMLR, 2018.
[26] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-Robust federated learning," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1605–1622, USENIX Association, Aug. 2020.
[27] M. Lincy and A. M. Kowshalya, "Early detection of type-2 diabetes using federated learning," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 12, pp. 257–267, 2020.
[28] M. Joshi, A. Pal, and M. Sankarasubbu, "Federated learning for healthcare domain - pipeline, applications and challenges," *ACM Trans. Comput. Healthcare*, vol. 3, nov 2022.
[29] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, p. 119, 2020.
[30] P. Chatterjee, D. Das, and D. B. Rawat, "Use of federated learning and blockchain towards securing financial services," *arXiv preprint arXiv:2303.12944*, 2023.

[31] T. Liu, Z. Wang, H. He, W. Shi, L. Lin, R. An, and C. Li, "Efficient and secure federated learning for financial applications," *Applied Sciences*, vol. 13, no. 10, 2023.

[32] Z. Xu, Y. Zhang, G. Andrew, C. Choquette, P. Kairouz, B. Mcmahan, J. Rosenstock, and Y. Zhang, "Federated learning of gboard language models with differential privacy," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)* (S. Sitaram, B. Beigman Klebanov, and J. D. Williams, eds.), (Toronto, Canada), pp. 629–639, Association for Computational Linguistics, July 2023.

[33] K. Bonawitz, P. Kairouz, B. McMahan, and D. Ramage, "Federated learning and privacy: Building privacy-preserving systems for machine learning and data science on decentralized data," *Queue*, vol. 19, pp. 87–114, nov 2021.

[34] X. Huang, P. Li, H. Du, J. Kang, D. Niyato, D. I. Kim, and Y. Wu, "Federated learning-empowered ai-generated content in wireless networks," *IEEE Network*, pp. 1–1, 2024.

[35] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems* (I. Dhillon, D. Papailiopoulos, and V. Sze, eds.), vol. 2, pp. 429–450, 2020.

[36] F. Lai, Y. Dai, S. Singapuram, J. Liu, X. Zhu, H. Madhyastha, and M. Chowdhury, "Fedscale: Benchmarking model and system performance of federated learning at scale," in *International Conference on Machine Learning*, pp. 11814–11827, PMLR, 2022.

[37] L. Yao, D. Gao, Z. Wang, Y. Xie, W. Kuang, D. Chen, H. Wang, C. Dong, B. Ding, and Y. Li, "A benchmark for federated hetero-task learning," *arXiv preprint arXiv*, vol. 2206, 2022.

[38] D. Chen, D. Gao, W. Kuang, Y. Li, and B. Ding, "pfl-bench: A comprehensive benchmark for personalized federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9344–9360, 2022.

[39] V. Mugunthan, A. Peraire-Bueno, and L. Kagal, "Privacyfl: A simulator for privacy-preserving and secure federated learning," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3085–3092, 2020.

[40] A. Ziller, A. Trask, A. Lopardo, B. Szymkow, B. Wagner, E. Bluemke, J.-M. Nounahon, J. Passerat-Palmbach, K. Prakash, N. Rose, *et al.*, "Pysyft: A library for easy federated learning," *Federated Learning Systems: Towards Next-Generation AI*, pp. 111–139, jun 2021.

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[42] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 1950–1965, Curran Associates, Inc., 2022.

[43] B. Zhao, H. Tu, C. Wei, J. Mei, and C. Xie, "Tuning layernorm in attention: Towards efficient multi-modal LLM finetuning," in *The Twelfth International*

Conference on Learning Representations, 2024.
[44] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAcctT '20, (New York, NY, USA), pp. 306–316, Association for Computing Machinery, 2020.
[45] CCPA, "California consumer privacy act (ccpa)," 2023.
[46] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.
[47] L. Ju, T. Zhang, S. Toor, and A. Hellander, "Accelerating fair federated learning: Adaptive federated adam," *arXiv preprint arXiv:2301.09357*, 2023.
[48] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020.
[49] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Information Fusion*, vol. 90, pp. 148–173, 2023.
[50] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," *Advances in Neural Information Processing Systems*, 2018.
[51] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2017.
[52] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–36, 2021.
[53] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
[54] J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar, "signsgd with majority vote is communication efficient and fault tolerant," *arXiv preprint arXiv:1810.05291*, 2018.
[55] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[56] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.
[57] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.
[58] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1354–1371, 2022.
[59] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proceedings of the AAAI Conference on Artificial Intelligence*,

vol. 33, pp. 1544–1551, 2019.
[60] S. P. Karimireddy, L. He, and M. Jaggi, "Learning from history for byzantine robust optimization," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, PMLR, 2021.
[61] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation," in *Uncertainty in Artificial Intelligence*, pp. 261–270, PMLR, 2020.
[62] F. Sattler, K.-R. Müller, T. Wiegand, and W. Samek, "On the byzantine robustness of clustered federated learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8861–8865, IEEE, 2020.
[63] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," in *ISOC Network and Distributed System Security Symposium (NDSS)*, 2021.
[64] C. Xu, Y. Jia, L. Zhu, C. Zhang, G. Jin, and K. Sharif, "Tdfl: Truth discovery based byzantine robust federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, 2022.
[65] J. Park, D.-J. Han, M. Choi, and J. Moon, "Sageflow: Robust federated learning against both stragglers and adversaries," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 840–851, Curran Associates, Inc., 2021.
[66] E. Gorbunov, S. Horváth, P. Richtárik, and G. Gidel, "Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top," in *International Conference on Learning Representations*, 2023.
[67] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4583–4596, 2020.
[68] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford, "Geometric median in nearly linear time," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 9–21, 2016.
[69] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 27730–27744, Curran Associates, Inc., 2022.
[70] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown,

and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022.
[71] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. R. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash, "RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback," in *Forty-first International Conference on Machine Learning*, 2024.
[72] Meta, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models," September 2024.
[73] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, C. Zhang, R. Sun, Y. Wang, and Y. Yang, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *arXiv preprint arXiv:2307.04657*, 2023.
[74] Z. Yu, X. Liu, S. Liang, Z. Cameron, C. Xiao, and N. Zhang, "Don't listen to me: Understanding and exploring jailbreak prompts of large language models," in *33rd USENIX Security Symposium (USENIX Security 24)*, (Philadelphia, PA), pp. 4675–4692, USENIX Association, Aug. 2024.
[75] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, "Safety fine-tuning at (almost) no cost: A baseline for vision large language models," in *Forty-first International Conference on Machine Learning*, 2024.
[76] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!," in *The Twelfth International Conference on Learning Representations*, 2024.
[77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[78] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
[79] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
[80] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, 2021.
[81] G. T. Google, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
[82] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644*, 2023.
[83] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 3045–3059, Association for Computational Linguistics, Nov. 2021.

[84] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), pp. 4582–4597, Association for Computational Linguistics, Aug. 2021.

[85] S. Malaviya, M. Shukla, and S. Lodha, "Reducing communication overhead in federated learning for pre-trained language models using parameter-efficient finetuning," in *Proceedings of The 2nd Conference on Lifelong Learning Agents* (S. Chandar, R. Pascanu, H. Sedghi, and D. Precup, eds.), vol. 232 of *Proceedings of Machine Learning Research*, pp. 456–469, PMLR, 22–25 Aug 2023.

[86] H. Woisetschläger, A. Isenko, S. Wang, R. Mayer, and H.-A. Jacobsen, "A survey on efficient federated learning methods for foundation model training," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-24*, International Joint Conferences on Artificial Intelligence Organization, aug 2024.

[87] S. M. Shah and V. K. N. Lau, "Model compression for communication efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 5937–5951, 2023.

[88] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, "Model pruning enables efficient federated learning on edge devices," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10374–10386, 2023.

[89] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019.

[90] V. Tsouvalas, Y. Asano, and A. Saeed, "Federated fine-tuning of foundation models via probabilistic masking," *arXiv preprint arXiv:2311.17299*, 2023.

[91] M. Xu, W. Yin, D. Cai, R. Yi, D. Xu, Q. Wang, B. Wu, Y. Zhao, C. Yang, S. Wang, Q. Zhang, Z. Lu, L. Zhang, S. Wang, Y. Li, Y. Liu, X. Jin, and X. Liu, "A survey of resource-efficient llm and multimodal foundation models," *arXiv preprint arXiv:2401.08092*, 2024.

[92] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.), vol. 108 of *Proceedings of Machine Learning Research*, pp. 2021–2031, PMLR, 26–28 Aug 2020.

[93] P. Zhaopeng, F. Xiaoliang, C. Yufan, W. Zheng, P. Shirui, W. Chenglu, Z. Ruisheng, and W. Cheng, "Fedpft: Federated proxy fine-tuning of foundation models," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-24*, International Joint Conferences on Artificial Intelligence Organization, aug 2024.

[94] H. Chen, Y. Zhang, D. Krompass, J. Gu, and V. Tresp, "Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38,

pp. 11285–11293, mar 2024.
[95] S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora, "Fine-tuning language models with just forward passes," in *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023.
[96] W. Fang, Z. Yu, Y. Jiang, Y. Shi, C. N. Jones, and Y. Zhou, "Communication-efficient stochastic zeroth-order optimization for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5058–5073, 2022.
[97] Z. Li and L. Chen, "Communication-efficient decentralized zeroth-order method on heterogeneous data," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, 2021.
[98] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.
[99] Z. Qin, D. Chen, B. Qian, B. Ding, Y. Li, and S. Deng, "Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes," in *Proceedings of the 41th International Conference on Machine Learning*, jul 2024.
[100] S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora, "Fine-tuning language models with just forward passes," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 53038–53075, Curran Associates, Inc., 2023.
[101] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021.
[102] S. Su, M. Yang, B. Li, and X. Xue, "Federated adaptive prompt tuning for multi-domain collaborative learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 15117–15125, mar 2024.
[103] S. Bai, J. Zhang, S. Li, S. Guo, J. Guo, J. Hou, T. Han, and X. Lu, "Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning," *arXiv preprint arXiv:2403.08506*, jun 2024.
[104] W. Zhao, Y. Chen, R. Lee, X. Qiu, Y. Gao, H. Fan, and N. D. Lane, "Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages," in *The Twelfth International Conference on Learning Representations*, may 2024.
[105] Y.-W. Chu, D.-J. Han, and C. G. Brinton, "Only send what you need: Learning to communicate efficiently in federated multilingual machine translation," in *Companion Proceedings of the ACM on Web Conference 2024*, (New York, NY, USA), pp. 1548–1557, Association for Computing Machinery, may 2024.
[106] F. Wu, X. Liu, H. Wang, X. Wang, and J. Gao, "On the client preference of llm fine-tuning in federated learning," 2024.
[107] L. Yi, H. Yu, G. Wang, X. Liu, and X. Li, "pfedlora: Model-heterogeneous personalized federated learning with lora tuning," *arXiv preprint*

arXiv:2310.13283, 2024.
[108] T. Guo, S. Guo, and J. Wang, "pfedprompt: Learning personalized prompt for vision-language models in federated learning," in *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, apr 2023.
[109] S. Su, B. Li, and X. Xue, "Fedra: A random allocation strategy for federated tuning to unleash the power of heterogeneous clients," *arXiv preprint arXiv:2311.11227*, oct 2024.
[110] Y. J. Cho, L. Liu, Z. Xu, A. Fahrezi, and G. Joshi, "Heterogeneous LoRA for federated fine-tuning of on-device foundation models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds.), (Miami, Florida, USA), pp. 12903–12913, Association for Computational Linguistics, nov 2024.
[111] B. G. A. Tekgul, Y. Xia, S. Marchal, and N. Asokan, "Waffle: Watermarking in federated learning," in *2021 40th International Symposium on Reliable Distributed Systems (SRDS)*, pp. 310–320, 2021.
[112] S. Yu, J. Hong, Y. Zeng, F. Wang, R. Jia, and J. Zhou, "Who leaked the model? tracking IP infringers in accountable federated learning," in *NeurIPS 2023 Workshop on Regulatable ML*, 2023.
[113] T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu, "Black-box tuning for language-model-as-a-service," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855, PMLR, 17–23 Jul 2022.
[114] Z. Lin, Y. Sun, Y. Shi, X. Wang, L. Huang, L. Shen, and D. Tao, "Efficient federated prompt tuning for black-box large pre-trained models," *CoRR*, vol. abs/2310.03123, 2023.
[115] J. Sun, Z. Xu, H. Yin, D. Yang, D. Xu, Y. Chen, and H. R. Roth, "Fedbpt: Efficient federated black-box prompt tuning for large language models," in *Proceedings of the 41th International Conference on Machine Learning*, jul 2024.
[116] W. Lu, H. Yu, J. Wang, D. Teney, H. Wang, Y. Chen, Q. Yang, X. Xie, and X. Ji, "Zoopfl: Exploring black-box foundation models for personalized federated learning," *arXiv preprint arXiv:2310.05143*, 2023.
[117] T. Sun, Z. He, H. Qian, Y. Zhou, X. Huang, and X. Qiu, "BBTv2: Towards a gradient-free future with large language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 3916–3930, Association for Computational Linguistics, Dec. 2022.
[118] V. Mugunthan, D. Byrd, T. H. Balch, J. P. Morgan, and A. Research, "Smpai: Secure multi-party computation for federated learning," in *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, vol. 21, 2019.
[119] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022.
[120] X. Cao, J. Jia, and N. Z. Gong, "Provably secure federated learning against malicious clients," *Proceedings of the AAAI Conference on Artificial*

Intelligence, vol. 35, pp. 6885–6893, May 2021.
[121] X. Cao, Z. Zhang, J. Jia, and N. Z. Gong, "Flcert: Provably secure federated learning against poisoning attacks," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3691–3705, 2022.
[122] S. Li, E. Ngai, F. Ye, and T. Voigt, "Peft-as-an-attack! jailbreaking language models during federated parameter-efficient fine-tuning." 2024.
[123] Anonymous, "Identifying and tuning safety neurons in large language models," in *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review.
[124] J. Chen, X. Wang, Z. Yao, Y. Bai, L. Hou, and J. Li, "Finding safety neurons in large language models," 2024.
[125] Y. Lin, H. Lin, W. Xiong, S. Diao, J. Liu, J. Zhang, R. Pan, H. Wang, W. Hu, H. Zhang, H. Dong, R. Pi, H. Zhao, N. Jiang, H. Ji, Y. Yao, and T. Zhang, "Mitigating the alignment tax of RLHF," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds.), (Miami, Florida, USA), pp. 580–606, Association for Computational Linguistics, Nov. 2024.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2477

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-540441

ACTA UNIVERSITATIS
UPSALIENSIS
2024