Modern artificial intelligence (AI) systems have long relied on vast, centralized datasets. Yet this paradigm is becoming increasingly untenable in a world where data are distributed, private, and context-dependent. As data are now predominantly generated on edge devices and privacy-sensitive platforms, traditional centralized training is both technically inefficient and ethically unsustainable. This shift calls for decentralized and privacy-aware learning frameworks that can operate across heterogeneous systems. Federated learning (FL) provides a natural foundation for this shiftenabling models to be collaboratively trained without transferring sensitive data. Moreover, in the emerging era of foundation-model-driven AI, FL is evolving beyond training from scratch across multiple parties toward collaborative adaptation and co-evolution atop foundation models, bridging individual intelligence and collaborative progress. Guided by this vision, my research aims to advance this direction by building the next generation of **collaborative intelligence systems**, in which models can evolve collectively across decentralized environments while respecting privacy, resource, and societal constraints.

As the research focus shifts toward applying federated learning to foundation models (FMs) with massive parameters, this transition introduces unprecedented challenges. Anchored by the central conviction that *optimization provides the fundamental pathway to shape the next generation of collaborative AI*, my long-term goal is to *build* ***optimization-driven collaborative learning frameworks*** *that enable* ***efficient, scalable, and reliable*** *training of FMs* while fully leveraging local private data. My research has been published in leading AI conferences and journals, including ICML, NeurIPS, ICLR, TMLR, AISTATS, as well as in leading natural language processing (NLP) conferences such as ACL. In addition, I have presented my work at the SDM Doctoral Forum, Data Engineering Tech Talks at IBM Research, and AAAI Workshop on Federated Learning. Figure 1 illustrates my key research contributions and outlines the trajectory of my future work.

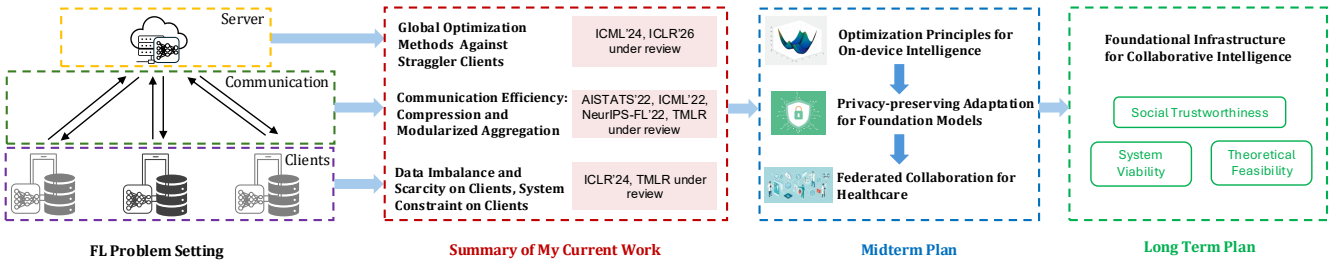

Figure 1: Overview of My Research Contributions and Future Plans

RESEARCH ACCOMPLISHMENTS

As an important foundation of my long-term research goal, my current research focuses on building algorithmic and theoretical foundations for training and fine-tuning FMs in federated settings. Federated learning is a collaborative paradigm where multiple clients jointly train a shared model without exposing their private data. Conceptually, it involves three tightly coupled dimensions: client-level factors such as local data distribution and clients' computational capacity, server-level mechanisms such as global model learning, and the communication channel that coordinates updates between clients and the server. These three dimensions pose fundamental challenges to FL, including communication efficiency, data heterogeneity, system reliability, etc. When extended to training FMs with billions of parameters, these issues become even more pronounced, and new obstacles emerge. For example, the scalability of optimization algorithms and the feasibility of fine-tuning under strict resource and privacy constraints. My research seeks to tackle these dimensions systematically, laying the groundwork for efficient, scalable, and reliable collaborative learning with FMs. To make this concrete, I will next highlight my achievements along the fundamental challenges to FL.

Communication Efficiency: from Compression to Modularized Aggregation

Communication plays an indispensable role in transmitting model and update information between local clients and the server. However, this process inevitably introduces significant challenges. High communication cost

arises due to the frequent exchange of high-dimensional model parameters or gradients, latency is introduced when synchronizing updates across heterogeneous devices, and bandwidth limitations further exacerbate the bottleneck in real-world deployments. While these issues have long been studied in traditional FL through techniques such as gradient sparsification and model compression, my prior work has contributed to this line by proposing communication-efficient methods for distributed learning [1] and FL [2, 3]. Particularly, motivated by the default used adaptive optimizer in transformers and language models, I proposed FedCAMS [2], **the first FL algorithm that unifies communication efficiency with adaptive global updates**. The method significantly reduces communication cost by applying update compression and employs error feedback to mitigate the information loss caused by compression. Importantly, FedCAMS is designed to be compatible with various widely adopted compressors.

However, model compression approaches become far less effective in the era of collaboratively fine-tuning FMs, where the parameter scale is orders of magnitude larger. More critically, naive compression may inadvertently strip away essential capabilities and knowledge embedded in pretrained models, undermining the very foundation of collaborative fine-tuning. This motivates my research on **modularized federation and aggregation**, which aims to preserve the knowledge of FMs while reducing communication overhead. To this end, I developed ParaBlock [4], a framework that modularizes foundation models and parallelizes federated fine-tuning via a two-thread design. The computation thread trains the model module-by-module, while communication proceeds in parallel; as soon as one module completes training, the next module's computation begins. ParaBlock significantly improves efficiency by overlapping most of the communication time with computation. To address the inconsistency introduced by this two-thread execution, we design a correction mechanism and theoretically prove that it restores consistency and preserves convergence guarantees. Ultimately, ParaBlock **enables scalable and efficient collaborative fine-tuning** of FMs across edge devices, bringing collaborative intelligence closer to the edge.

System Reliability Against Straggler Clients

While prior efficiency-focused designs alleviate communication constraints, federated fine-tuning of foundation models with a large client population introduces new challenges. In real-world settings, clients often have varying computational capabilities, making synchronous aggregation inefficient because the server must wait for all clients to finish before updating the global model. To overcome this limitation, asynchronous FL allows clients to upload updates at their own pace, improving scalability and reducing idle time. However, this flexibility also introduces asynchronous delay, where slow clients submit updates after the global model has already advanced through multiple aggregation rounds. It has been shown experimentally that such stale updates can degrade convergence, destabilize training trajectories, and ultimately lead to suboptimal model performance.

My previous work [8] takes an optimization perspective to **analyze how asynchronous delay affects the convergence of federated learning**. I rigorously demonstrate that the degree of *asynchronous delay appears explicitly in the theoretical convergence rate*, where larger delays slow down convergence and require more training rounds for the model to reach stability. In addition, I show that when client data are highly non-i.i.d., meaning that data across clients differ significantly, then the adverse effect of asynchronous delay becomes more pronounced. To improve the training resilience to non-i.i.d. data distribution, I proposed CA^{2}FL [8], where the server caches the most recent update from each client and reuses these cached updates for global model calibration. We theoretically show that leveraging cached updates significantly improves the convergence rate by mitigating the negative impact of asynchronous delay.

While CA^{2}FL [8] theoretically analyzes the convergence of asynchronous FL and improves its convergence rate, we observe that the unstable training trajectory caused by asynchronous updates remains a major challenge. This is because large asynchronous delay make clients train on outdated global models and lead to instability in the loss landscape. To address this issue, I propose FADAS [7], a delay-adaptive learning rate that mitigates the instability introduced by stale updates and **enhances resiliency to stragglers with large delays**. In FADAS, the server tracks the delay associated with each received update and applies a delay-adaptive learning rate. This design stabilizes training and improves robustness to client delays, with theoretical guarantees that validate its effectiveness.

Data Imbalance and Scarcity

While previous designs primarily focus on addressing system-level constraints, such as communication overhead and the presence of straggler clients, FL also faces significant challenges arising from its decentralized data nature. Because data are stored locally on edge devices, the distribution across clients is often highly imbalanced or even scarce. Such heterogeneity can lead to diverging local updates and a severe mismatch between local and global training objectives, ultimately degrading overall model performance. Beyond imbalance, recent scaling laws reveal that large foundation models are particularly prone to overfitting and may quickly enter a data-limited regime, as their capacity far exceeds the intrinsic complexity of available data. This issue is further exacerbated in federated settings, where individual clients typically possess only limited and potentially low-quality data, making effective fine-tuning of foundation models even more challenging.

My previous work [5, 6] focuses on mitigating data imbalance in federated training. In [5], I revealed how data imbalance fundamentally disrupts adaptive federated learning by causing local overfitting and slowing global convergence, and notably, this effect becomes especially pronounced when fine-tuning large language models (LLMs) under heavy-tailed gradient noise. This study not only clarified the underlying optimization dynamics but also inspired a new perspective on mitigating data imbalance in federated systems. Building on these insights, I proposed the AFGA framework [5], which demonstrates how **client re-sampling and peer-to-peer collaboration can effectively stabilize large-scale federated training** with imbalanced client data. Collectively, this line of work deepens the understanding of optimization behavior under heterogeneous data and paves the way toward more equitable and reliable federated learning.

Moreover, in my previous asynchronous work [8], I investigated the negative effect of asynchronous delay and its coupled relationship with non-i.i.d. data distribution from an optimization perspective. Intuitively, this coupling arises because faster clients contribute updates more frequently, causing the global model to become biased toward their local data distributions. In [9], I addressed this issue through FedEcho, an uncertainty-aware distillation framework for asynchronous FL. Beyond proposing a new algorithmic design, this work **establishes a new perspective on utilizing stale updates**, i.e., as a valuable source of knowledge that can be adaptively distilled into the global model. This reshapes how asynchronous systems can leverage information from stragglers, offering a more robust and principled path toward reliable large-scale federated fine-tuning.

Miscellaneous

In addition to my research on FL, I have also contributed to several foundational aspects of generative foundation models. In [10], we proposed an effective and efficient parameter-efficient fine-tuning method for LLMs. Furthermore, we explored the interpretability of LLMs generation through a joint prompt attribution framework [11]. Together, these studies not only deepen the understanding of optimization and adaptation in generative models but also advance my long-term vision of building optimization-driven collaborative learning frameworks that unify efficiency, interpretability, and scalability.

FUTURE DIRECTIONS

Looking ahead, my future research aims to develop **collaborative learning frameworks that foster generalization, trustworthiness, and efficiency**, advancing the development of theoretical principled and practical AI systems. My future research will pursue optimization-driven frameworks with real-world applications for efficient, privacy-preserving and trustworthy collaborative intelligence.

Optimization principles for on-device intelligence Recent studies have shown that relatively small-scale models, such as small language models, already demonstrate remarkable capability and are likely to play a central role in the future of generative AI. Building on this trend, the next generation of foundation models is expected to shift toward edge-deployed models that can perform efficient inference and adaptation directly on user devices. However, as edge data are often highly sensitive, a key research challenge lies in developing optimization principles

that enable accurate, efficient, and privacy-preserving learning from on-device data. Addressing this challenge will be essential for realizing the vision of collaborative intelligence that is both personalized and trustworthy. Building upon my prior research on optimization in federated learning, my future work will focus on developing lightweight, stable, and reliable on-device optimizers. These optimizers need to be carefully designed for the constrained computational/memory resources on heterogeneous hardware, while ensuring strong convergence and generalization ability. This line of research aims to enable practical on-device learning by considering both efficiency and stability.

Privacy-preserving adaptation for foundation models As foundation models continue to expand in scale and capability, they also introduce new dimensions of privacy risk that differ fundamentally from those in conventional machine learning. Due to their massive capacity and ability to memorize fine-grained details, FMs are increasingly vulnerable to privacy leakage through training data reconstruction, membership inference, and prompt inversion attacks. The issue is further exacerbated by the growing use of instruction-tuned or domain-specific FMs, where sensitive data may be indirectly revealed through generated outputs or parameter updates. These privacy concerns become even more critical in edge-deployed or federated settings, where models interact directly with users personal or context-rich data. Unlike centralized training, edge models continuously adapt to local environments, creating a dynamic learning process that heightens the risk of unintentional data exposure. Building upon my work in language model interpretability, I plan to first systematically characterize the privacy risks arising in the collaborative training and fine-tuning of generative FMs, and then develop optimization-based solutions to mitigate these risks. I believe that developing theoretically grounded, privacy-preserving optimization and adaptation frameworks will be essential for building FMs that are both trustworthy and effective in collaborative environments.

Federated collaboration for healthcare Beyond generative FMs on edge devices, in general healthcare domain, data are often scarce, highly sensitive, and unevenly distributed across institutions, making centralized model training infeasible. I believe that my expertise in developing reliable, efficient, interpretable, and privacy-preserving learning frameworks can be effectively applied to healthcare scenarios, enabling hospitals and medical organizations with limited or imbalanced data to collaboratively contribute to model training in a secure and trustworthy manner. I aim to build collaborative AI systems that accelerate healthcare discovery and clinical decision-making while rigorously protecting patient privacy and ethical integrity.

Long-term vision: foundational infrastructure for collaborative intelligence. My long-term goal is to build FL as a foundational infrastructure for collaborative intelligence in human society. Just as the mobile internet changed how the information is shared, the realization of **collaborative learning urgently requires a core learning protocol** to guide vast numbers of edge nodes in evolving collaboratively in an efficient and trust manner. I believe that advanced optimization theory is the core concept underpinning this learning protocol. *Why optimization?* The fundamental challenge facing collaborative intelligence lies in systematically balancing three dimensions: theoretically feasibility (such as convergence in heterogeneous environments), system viability (such as scalable and efficient learning processes), and social trustworthiness (such as privacy, security, and trust). Optimization theory provides the universal language for precisely describing and solving these multi-objective trade-off problems. It allows us to translate abstract societal values into mathematical constraints and to embed rigorous privacy guarantees directly into the algorithm's design. My research, therefore, is driven by optimization, as I aim to build the unified, principled, and value-aligned foundation for this future collaborative intelligent infrastructure.

I look forward to developing open, interdisciplinary collaborations with colleagues in academia and industry, including experts in public health, psychology, social sciences, and human-computer interaction, as well as engaging closely with diverse user communities. My goal is to establish a trustworthy and privacy-preserving collaboration paradigm in intelligence, with applications spanning computer vision, natural language processing, bioinformatics, psychology, the Internet of Things, AI for science, and education. I will seek research support from funding programs such as NSF's Security, Privacy, and Trust in Cyberspace (SaTC) program and Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science (SCH), as well as industry funding opportunities including the Amazon and Google Research Awards.

REFERENCES

[1] **Yujia Wang**, Lu Lin, and Jinghui Chen. Communication-compressed adaptive gradient method for distributed nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6292–6320. PMLR, 2022.

[2] **Yujia Wang**, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 22802–22838. PMLR, 2022.

[3] **Yujia Wang**, Pei Fang, and Jinghui Chen. Accelerating Adaptive Federated Optimization with Local Gossip Communications. In *International Workshop on Federated Learning: Recent Advances and New Challenges, NeurIPS 2022*.

[4] **Yujia Wang**, Yuanpu Cao, and Jinghui Chen. Parablock: Communication-computation parallel block coordinate federated learning for large language models. *Under Review*.

[5] **Yujia Wang**, and Jinghui Chen. On the Data Heterogeneity in Adaptive Federated Learning. In *Transactions on Machine Learning Research (TMLR)*, 2024.

[6] Tiejin Chen*, Yuanpu Cao*, **Yujia Wang***, Cho-Jui Hsieh, and Jinghui Chen. Federated Learning with Projected Trajectory Regularization. *arXiv preprint arXiv:2312.14380.* (* equal contribution)

[7] **Yujia Wang**, Shiqiang Wang, Songtao Lu and Jinghui Chen. FADAS: Towards Federated Adaptive Asynchronous Optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, 2024.

[8] **Yujia Wang**, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen. Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024.

[9] **Yujia Wang**, Fenglong Ma, and Jinghui Chen. Stragglers Can Contribute More: Uncertainty-Aware Distillation for Asynchronous Federated Learning. *Under Review.*

[10] Xin Yu, **Yujia Wang**, Jinghui Chen and Lingzhou Xue. AltLoRA: Towards Better Gradient Approximation in Low-Rank Adaptation with Alternating Projections. *In Proceedings of the 39th Annual Conference on Neural Information Processing Systems(NeurIPS)*, 2025.

[11] Yurui Chang*, Bochuan Cao*, **Yujia Wang**, Jinghui Chen and Lu Lin. JoPA: Explaining Large Language Model's Generation via Joint Prompt Attribution. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL Main)*, 2025.