A Survey on Foundation Models for Personalized Federated Intelligence

YU QIAO, School of Computing, Kyung Hee University, Republic of Korea
HUY Q. LE, School of Computing, Kyung Hee University, Republic of Korea
AVI DEB RAHA, School of Computing, Kyung Hee University, Republic of Korea
PHUONG-NAM TRAN, School of Computing, Kyung Hee University, Republic of Korea
APURBA ADHIKARY, Noakhali Science and Technology University, Bangladesh
MENGCHUN ZHANG, Korea Advanced Institute of Science and Technology, Republic of Korea
LOC X. NGUYEN, School of Computing, Kyung Hee University, Republic of Korea
EUI-NAM HUH, School of Computing, Kyung Hee University, Republic of Korea
DUSIT NIYATO, College of Computing and Data Science, Nanyang Technological University, Singapore
CHOONG SEON HONG, School of Computing, Kyung Hee University, Republic of Korea

The rise of large language models (LLMs), such as ChatGPT, Gemini, and Grok, has reshaped the AI landscape. As prominent instances of foundational models (FMs), they exhibit remarkable capabilities in generating human-like content, pushing the boundaries towards artificial general intelligence (AGI). However, their large-scale nature, privacy sensitivity, and substantial computational demands pose significant challenges for personalized customization for end users. To bridge this gap, we present the vision of artificial personalized intelligence (API), which focuses on adapting FMs to individual users while ensuring privacy. As a central enabler of API, we propose personalized federated intelligence (PFI), a new paradigm that not only integrates the privacy benefits of federated learning (FL) with the generalization capabilities of FMs but also places personalization at its core. To this end, we first survey recent advances in FL and FMs that lay the foundation for PFI. We then explore core stages of the PFI pipeline: efficient personalization at the edge, trustworthy adaptation, and adaptive refinement via retrieval-augmented generation. Finally, we highlight future directions for enabling PFI. Overall, this survey aims to lay a foundation for the development of API as a complementary direction to AGI, with PFI as a key enabling paradigm.

CCS Concepts: • **Computing methodologies**; • **Natural language generation**; • **Artificial intelligence**; • **Machine learning**;

Additional Key Words and Phrases: Personalized federated intelligence, foundation models, federated learning, artificial general intelligence, artificial personalized intelligence.

Authors' Contact Information: Yu Qiao, qiaoyu1002@gmail.com, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Huy Q. Le, quanghuy69@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Avi Deb Raha, avi@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Phuong-Nam Tran, tpnam0901@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Apurba Adhikary, apurba@nstu.edu.bd, Noakhali Science and Technology University, Noakhali-3814, Bangladesh; Mengchun Zhang, mengchun0527@gmail.com, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea; Loc X. Nguyen, xuanloc088@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Eui-Nam Huh, johnhuh@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea; Dusit Niyato, dniyato@ntu.edu.sg, College of Computing and Data Science, Nanyang Technological University, Singapore; Choong Seon Hong, cshong@khu.ac.kr, School of Computing, Kyung Hee University, Yongin-si 17104, Republic of Korea.

ACM 1557-735X/2026/5-ART111
https://doi.org/XXXXXXX.XXXXXXX

Table 1. List of Abbreviations and Their Definitions.

Abbr.	Definitions	Abbr.	Definitions	Abbr.	Definitions
AGI	Artificial General Intelligence	API	Artificial Personalized Intelligence	FMs	Foundation Models
PFI	Personalized Federated Intelligence	FL	Federated Learning	LLMs	Large Language Models
SAM	Segment Anything Model	FFMs	Federated Foundation Models	RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback	RLAIF	Reinforcement Learning from AI Feedback	RAG	Retrieval-Augmented Generation
LoRA	Low-rank Adaptation	PEFT	Parameter-efficient fine-tuning	XAI	Explainable AI

ACM Reference Format:
Yu Qiao, Huy Q. Le, Avi Deb Raha, Phuong-Nam Tran, Apurba Adhikary, Mengchun Zhang, Loc X. Nguyen, Eui-Nam Huh, Dusit Niyato, and Choong Seon Hong. 2026. A Survey on Foundation Models for Personalized Federated Intelligence. *J. ACM* 37, 4, Article 111 (May 2026), 36 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The landscape of AI has been profoundly transformed by the rise of foundation models (FMs), particularly large language models (LLMs) in natural language processing, such as ChatGPT [1], Gemini [148], and DeepSeek [122]. The advent of these LLMs has revolutionized tasks [47] such as content generation, text summarization, and information retrieval , showcasing their remarkable potential in understanding and generating human-like conversational responses. Building on this momentum, Meta AI released the segment anything model (SAM) [85], which claims to accurately segment any object in any image with remarkable zero-shot generalization for downstream tasks. This breakthrough, regarded as the "ChatGPT moment" for computer vision, has inspired numerous follow-up works aimed at further enhancing SAM's capabilities, including lightweight model designs [187] and fine-tuning for medical image analysis [105]. Beyond the image domain, Meta AI recently took a significant step forward by introducing SAM2 [127], which extends the promptable capabilities of SAM to the video domain, enabling fast and precise object selection in any video or image. Subsequent studies have further adapted SAM2 for various downstream tasks, such as 3D medical image segmentation [57]. The success of these FMs has bolstered the anticipation that the era of strong AI is approaching, bringing us closer to the realization of artificial general intelligence. List of abbreviations and definitions is provided to facilitate the smooth flow of the paper in Table 1.

However, FMs are widely known for their massive model sizes and reliance on extensive and diverse datasets sourced from books, web pages, and other domains for training [191]. As a result, they inherently face limitations, including challenges related to data privacy, computational resource constraints, and limited adaptability to domain-specific tasks without further fine-tuning [97]. For instance, training a GPT-3 model with 175 billion parameters requires 1,024 Nvidia V100 GPUs running for 34 days at an estimated cost of $4.6 million US dollars [190]. Such high barriers make it infeasible for smaller organizations or emerging companies to develop or even fine-tune such large-scale models, particularly under limited computational and financial constraints. In addition, end users are often reluctant to share private data for training and are concerned about its potential misuse by large models like GPT [167]. The training datasets themselves may also lack global representativeness, leading to embedded biases. For instance, associating men with engineering roles and women with caregiving tasks, or favoring majority racial groups over minorities in job recommendations [28]. Moreover, data in the real world is continuously evolving [3], making it challenging to keep large models up-to-date. Retraining or even fine-tuning these models with incremental data is highly costly and poses significant difficulties for end users who lack the

Table 2. Comparison of Key Concepts: FL, FM, FFM, AGI, API, and PFI.

Abbr.	Relation to PFI
FL	Serves as a foundational mechanism for enabling privacy-preserving training in PFI.
FM	Provides the generalization capability that PFI aims to personalize.
FFM	Focuses on distributed training and adaptation of FMs; PFI builds on FFM principles but further highlights user-level personalization.
AGI	A complementary vision to API; aims for general intelligence, while API focuses on personalized intelligence.
API	The overarching goal of PFI, emphasizing user-specific intelligence rather than general-purpose intelligence.
PFI	The central paradigm proposed in this work, integrating FL and FM to realize privacy-preserving personalized intelligence.

resources or expertise to modify them. Finally, personalized applications require models that can adapt to individual users' preferences and unique data patterns. For instance, a healthcare platform may need to offer tailored diagnoses based on a user's medical history, while an e-commerce recommendation system should adjust to personalized shopping behaviors. Therefore, traditional centralized FM training struggles to meet these needs, as it needs to balance AGI ambitions with the requirement for privacy-preserving personalized intelligence.

These limitations highlight the urgent need for a new paradigm that enables FMs to be both powerful and personally adaptive, while respecting users' privacy and resource constraints. To address this, we propose a new vision, **artificial personalized intelligence (API)**, which complements and extends the pursuit of AGI: whereas AGI targets universal capabilities across tasks and users, API focuses on intelligence tailored to individual needs. To realize this vision, we introduce **personalized federated intelligence (PFI)**, a key enabling paradigm grounded in federated foundation models (FFMs), which are FMs collaboratively trained via FL across decentralized data silos. At its core, PFI emphasizes personalization, aiming to deliver user-adaptive intelligence while preserving privacy. Note that FL is a decentralized training framework that allows models to be collaboratively trained across distributed entities without sharing raw data, thereby supporting privacy protection, data sovereignty, and regulatory compliance. This makes it particularly suitable for privacy-sensitive domains such as healthcare, finance, and personalized services.

In the PFI framework, the global FM can either be a well-trained off-the-shelf model or be collaboratively trained from scratch across multiple decentralized data silos equipped with powerful computing resources (e.g., hospitals or enterprises), thereby ensuring privacy protection from the outset. This shared model is then efficiently adapted at the edge, enabling lightweight personalization aligned with local data distributions or user-specific preferences. Further enhancements can be introduced through trustworthy personalization (e.g., fairness-aware or robust optimization) and retrieval-augmented personalization (e.g., dynamic integration of external, user-relevant knowledge), allowing the model to continuously evolve with individual user contexts. Overall, this paper aims to explore the motivations, emerging solutions, and future directions of PFI, offering a comprehensive overview of the field and envisioning the path towards the API era.

As the above discussion involves several interrelated but potentially complex concepts, Table 2 provides a concise comparison of FL, FM, FFM, AGI, API, and PFI to help readers clearly distinguish their goals and the relation to our key proposal, PFI. In summary, while we propose the conceptual framework of PFI, this work is structured as a survey that systematically reviews and categorizes existing literature under this emerging paradigm.

Table 3. Comparison of Existing Survey and Position Papers with This Survey Paper.

Year	Ref.	FL Taxonomy	FM Taxonomy	Efficient FFM	Trustworthy FFM	Adaptive FFM	Personalization
2023	[194]	✗	✗	✓	✓	✗	✗
	[181]	✗	✓	✓	✗	✗	✓
	[164]	✓	✗	✗	✗	✗	✓
2024	[165]	✗	✓	✓	✗	✗	✗
	[128]	✗	✓	✓	✓	✗	✗
	[92]	✗	✓	✓	✓	✓	✗
	[155]	✗	✓	✗	✓	✗	✗
2025	[73]	✗	✓	✓	✓	✗	✗
	[80]	✗	✓	✓	✓	✗	✗
	[14]	✓	✓	✓	✗	✗	✗
	[177]	✓	✓	✓	✓	✗	✗
	[169]	✗	✓	✓	✗	✗	✗
	[41]	✗	✓	✓	✓	✗	✗
Ours		✓	✓	✓	✓	✓	✓

1.1 Major Contributions

This survey represents the first comprehensive exploration envisioning an era where API complements the widely discussed AGI. In contrast to prior surveys that primarily focus on specific aspects of FL and FMs, our work introduces a new paradigm, PFI, which emphasizes user-specific personalization and provides broader coverage of the topic. Table 3 compares our survey with several existing studies. While prior studies have provided valuable insights into dimensions such as efficiency, trustworthiness, or FM taxonomy, most of them examine these aspects in isolation rather than within an integrated conceptual framework. For example, the work in [41] summarize the ten major challenges of federated foundation models and provide grounded analyses of their objectives and potential solutions; however, they remain centered on the FM-centric perspective. In contrast, our survey takes a PFI-centric view, treating personalization as the core objective and structuring the design space along the entire PFI pipeline: efficient edge-side personalization, trustworthy adaptation, and adaptive refinement, thereby enabling continuous alignment with dynamic user-centric contexts. A more detailed comparison with existing works is provided in Appendix A. We hope this roadmap paves the way towards the realization of truly personalized and privacy-preserving intelligence. Overall, the key contributions of this survey are as follows:

- With the goal of advancing an API era that complements the development of AGI, this survey aims to shift focus towards the personalization aspect by exploring PFI, which, in contrast to prior work that has mainly concentrated on the role of FMs in enhancing the generalization capabilities of FL models, emphasizes the importance of end-user personalization.
- To provide a clear understanding of both FL and FMs, we revisit the key enabling technologies and the taxonomy of FL and FMs, highlighting their synergies and identifying key gaps that motivate the integration of personalization into FFMs.
- To lay the foundation for practical PFI deployment, we highlight several key technical directions, including efficient adaptation strategies for edge environments; mechanisms to ensure trustworthiness under real-world constraints such as privacy, security, and resource limitations; and adaptive design leveraging retrieval-augmented generation techniques.
- Looking ahead, we outline a forward-looking research agenda encompassing Meta-PFI, quantum-enabled learning, and sustainable intelligence, aiming to inspire future exploration towards next-generation personalized FMs that are user-centric and well-suited for decentralized, privacy-sensitive environments.

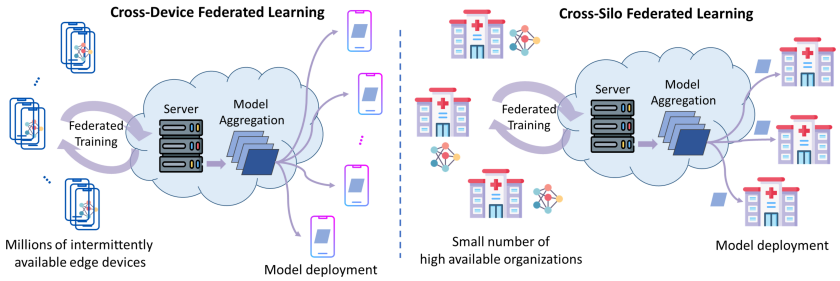


Fig. 1. Illustration of cross-device and cross-silo federated learning scenarios [79].

1.2 Organization of This Paper

The remainder of this survey is organized as follows. Section 2 provides an overview of FL and FMs, highlighting the motivation for PFI, where FFMs form the core framework but require additional techniques for effective personalization. Building on this foundation, Section 3 discusses efficient personalization at the edge. Section 4 discusses trustworthy adaptation. Section 5 explores further adaptive refinement. Section 6 outlines future directions in this emerging field. Finally, Section 7 concludes the survey.

2 Background

2.1 Federated Learning

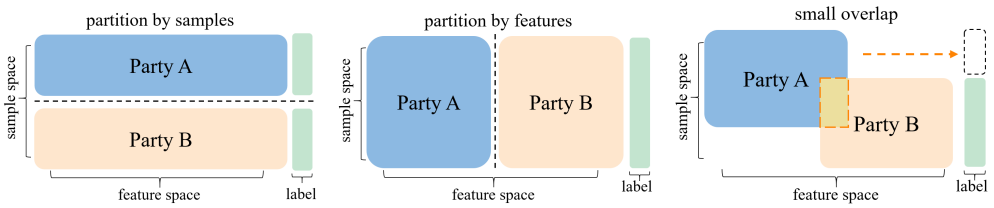
Federated learning (FL) [111], initially proposed by Google to enhance data privacy in Android ecosystems, enables collaborative model training across distributed devices without sharing raw data. This privacy-preserving paradigm has shown strong potential across domains such as healthcare (e.g., Tencent Tianyan Lab and WeBank's medical FL framework [77]), finance (e.g., WeBank's FATE for credit risk assessment [100]), and advertising (e.g., JD's 9N-FL framework boosting multi-party revenue [99]). More recently, FL has been applied to LLM training, as demonstrated by Prime Intellect [69] and Photon [135]. In addition, FL supports edge computing scenarios, including smart city applications like traffic management. These advancements highlight FL's growing role in both academic research and real-world deployment in the era of FMs.

The success of these applications stems from the scalable and privacy-preserving nature of FL, which enables model training across heterogeneous environments without sharing raw data [111]. In a typical FL setup, each entity (e.g., financial institution) receives a global model from a central server, performs local training on private data, and sends updated parameters back to the server. The server aggregates these updates and redistributes the refined model to the entities for the next global iteration. Formally, the goal of standard FL is to learn a *single global model* ω by minimizing

$$F(\omega) = \sum_{k=1}^{K} p_k F_k(\omega), \tag{1}$$

where $F_k(\omega)$ is the local loss of client k evaluated on its dataset $\mathcal{D}_k$, and $p_k = \frac{n_k}{\sum_j n_j}$. To provide a clear understanding and taxonomy of the FL paradigm, we categorize FL from different perspectives. First, based on the types of participating entities, FL can be classified into cross-silo FL and cross-device FL. Second, considering the patterns of underlying data distribution across entities, FL can be categorized into horizontal FL, vertical FL, and federated transfer learning [173].

Participating Entities. Based on the types of participating entities, the taxonomy of FL can be classified into two main categories: cross-silo FL and cross-device FL [79]. Cross-silo FL involves a small number of organizational entities, such as banks, hospitals, or research institutes, where



(a) Horizontal Federated Learning (b) Vertical Federated Learning (c) Federated Transfer Learning

Fig. 2. Illustration of horizontal FL, vertical FL, and federated transfer learning [99].

data remains local, and each entity typically has abundant computational resources. In contrast, cross-device FL involves numerous distributed personal devices, such as smartphones and IoT devices, which typically have limited computational resources. Figure 1 illustrates both scenarios.

Data Patterns. Based on the characteristics of the involved data distribution, the taxonomy of FL can be categorized into three main types: horizontal FL, vertical FL, and federated transfer learning. This taxonomy was also proposed in [99]. Here, we include it for a comprehensive overview and provide a more detailed explanation of the differences and applications of each type. Horizontal FL is most effective when participants share the same feature space but differ in their data samples, such as in collaborative model training across multiple hospitals. Vertical FL, on the other hand, involves entities with complementary features, such as the cooperation between banks and e-commerce platforms, where data about the same users is distributed across different feature spaces. For example, banks and e-commerce platforms can jointly train a model without sharing data, leveraging each other's insights: banks enhance credit risk assessments, while e-commerce platforms optimize marketing strategies and product recommendations. Finally, federated transfer learning is applied when participants have minimal overlap in both features and samples, such as in cross-domain tasks where data from different industries or regions are integrated to improve model performance. Figure 2 provides a visual representation of these types.

2.2 Overview of Foundation Models

FMs [9], also known as large X models, are deep learning models trained on large-scale, vast datasets to enable broad applicability across various domains [122]. Their development requires substantial computational resources and financial investment, but once trained, they can serve as a foundation for adaptation (e.g., fine-tuning) across a wide range of downstream tasks [17]. Generative AI applications, such as LLMs, have emerged as key areas where FMs play a crucial role. The success of LLMs is largely attributed to the massive scaling of both data and model size, while foundational components like the Transformer [83] and RL [152] have been established for many years. For instance, OpenAI's GPT series, such as GPT-3 with billions of parameters, has demonstrated impressive zero-shot learning abilities across tasks like question answering and commonsense reasoning [123]. Similarly, xAI's Grok-3, with trillions of parameters, is regarded as one of the smartest AIs on Earth, showcasing state-of-the-art capabilities in solving a wide range of challenging problems, including mathematics, reasoning, and programming [70]. More recently, models such as DeepSeek-R1 [51] have achieved breakthroughs by significantly reducing training costs while maintaining performance comparable to OpenAI-o1-1217 on reasoning tasks. Notably, DeepSeek-R1 employs a pure RL-based training pipeline without any warm-up stage, marking a shift towards greater efficiency and accessibility in foundation model development. In contrast to the training objective in FL, which aims to minimize aggregated loss across multiple decentralized entities, FMs are typically trained in a centralized manner. The training objective of a foundation

model can be expressed in a simplified form as:

$$\min_{\omega} \ \mathbb{E}_{x\sim\mathcal{D}} \left[\mathcal{L}(F_\omega(x))\right], \tag{2}$$

where $\mathcal{L}$ denotes the selected loss function, and $F_\omega(x)$ represents the parameterized model to be optimized. Here, x is the training samples drawn from a large-scale dataset $\mathcal{D}$. In practice, however, FM training is not a single-stage process but rather a composition of multiple stages that operate sequentially and complementarily. Typically, a model is first pre-trained on massive unlabeled corpora using self-supervised objectives to capture general features and knowledge, providing a good initialization for downstream tasks. It is then fine-tuned on specific datasets to adapt to downstream tasks. Building on this, a further RL-based alignment stage is commonly applied, such as RL from human feedback (RLHF) [118], RL from AI feedback (RLAIF) [12], and pure RL [51]), to guide the model towards outputs that better align with human expectations and social norms. These stages form an integrated pipeline where pre-training establishes general capabilities, fine-tuning refines task adaptability, and RL aligns model behavior with human or AI preferences.

To provide a systematic classification of the FM paradigm, we propose a taxonomy that organizes FMs across four key dimensions. First, based on input data modality, FMs can be classified into two types: unimodal models, which are trained on a single modality such as text (e.g., BERT [34]), vision (e.g., MAE [56]), or audio (e.g., Wav2Vec [136]); and multimodal models, which are designed to process and integrate information from multiple modalities, such as image-text (e.g. CLIP [124]) or video-language (e.g. VideoBERT [144]) pairs. Second, from a model architecture perspective, FMs can be categorized into three types: dense models, which activate all parameters during both training and inference (e.g., BERT) [130]; sparse models, such as mixture-of-experts (MoE), where only a subset of parameters is activated for each input to improve efficiency and scalability (e.g., GLaM [35], DeepSeekMoE [33]); and retriever-augmented models, which leverage external memory or retrieval mechanisms to enhance factual accuracy and knowledge utilization (e.g., RETRO [19] and retrieval augmented generation (RAG) [88]). Third, in terms of training paradigms, FMs can be broadly categorized into two groups. The first category comprises self-supervised pretraining models, which learn general-purpose representations from large-scale unlabeled data by solving pretext tasks such as masked token prediction or contrastive learning (e.g., MAE [56] and BERT [34]). The second category encompasses RL-based training strategies, including RLHF [118], RLAIF [12], and pure RL [51] approaches. These methods aim to improve alignment with user intent and enhance the safety and robustness of models in real-world applications. For instance, InstructGPT is trained using RLHF, where human preference data serves as a reward signal to fine-tune the model's behavior [118]. Claude leverages RLAIF, in which a preference model is used as the reward function without any human labels [12]. DeepSeek-R1-Zero, on the other hand, is trained via large-scale pure RL without any supervised fine-tuning stages such as RLHF or RLAIF, serving as an early exploration into instruction-following driven purely by RL [51]. Finally, regarding deployment and usage, FMs can be classified into two broad categories: cloud-centric models, which are typically deployed on centralized infrastructure and accessed via APIs; and on-device models, which are optimized for local inference, personalization, and privacy-preserving applications. For instance, GPT-4 and Claude are cloud-based models that offer powerful capabilities through API access, enabling large-scale applications across various domains. In contrast, models such as Meta's on-device TinyLLaMA [184] and Google's Gemini Nano [148] are specifically designed to run directly on edge devices, such as smartphones, offering reduced latency and improved data privacy. This taxonomy provides a systematic framework for analyzing the capabilities, design trade-offs, and deployment scenarios of FMs, with a comparative overview presented in Appendix B.

2.3 Motivation for Personalized Federated Intelligence

Complementary to AGI, which aims to achieve general intelligence with universal capabilities across tasks and users, this survey envisions an era of API that emphasizes intelligence tailored to individual needs and contexts. Realizing such personalization, however, requires access to diverse user data for model training, which poses significant privacy and governance challenges under traditional centralized paradigms. Centralized strategies that aggregate data in a single location risk violating privacy regulations and exacerbating issues related to data and computing resource monopolies. In this context, FL [111], with its inherent privacy-preserving and decentralized nature, serves as a key enabler of large-scale collaboration across users and organizations while ensuring the protection of sensitive data. Moreover, based on its distributed learning paradigm, FL enables collaboration among multiple devices, thereby overcoming the limitations imposed by the scale and diversity of data available on any single device. Empirical evidence further supports the value of such collaboration. In particular, FATE-LLM[42] demonstrates that federated LoRA and federated P-Tuning-v2 generally outperform individual clients' individual local fine-tuning across ROUGE and BLEU metrics. Notably, this improvement is achieved while communicating only a small fraction of parameters, 0.058% for LoRA and 0.475% for P-Tuning-v2, relative to full fine-tuning. Similarly, WorldLM [66] shows that personalized federated training, enabled by partial model localization and adaptive cross-federation knowledge sharing, achieves up to 1.91× performance improvements over standard FL, while closely approaching the performance of fully local personalized models under privacy-enhancing constraints. Nonetheless, FL faces fundamental challenges in adapting to the diverse data distributions and resource constraints across participants [87]. In real-world scenarios, participant data is often non-independent and identically distributed, as each user or device typically exhibits unique characteristics influenced by factors such as geographic location, individual behavior, and personal preferences [131]. For instance, typing behaviors on mobile keyboards differ across users, smartwatch health data varies with individual physiology and routines, and corporate logs capture unique operational patterns across organizations [142]. Such heterogeneity poses a major challenge for traditional FL methods, which optimize a single global model that may fail to capture user diversity [93]. This limitation motivates the need for PFI, where models are collaboratively trained while being locally adapted to align with individual data distributions. Moreover, given the growing availability of lightweight FMs (e.g., OpenAI o1-mini [156], Lite-SAM [46], and DeepSeek-V2-Lite [98]), PFI offers a practical pathway to efficiently fine-tune and adapt these models across diverse user environments with minimal communication and computational overhead.

Personalized Federated Intelligence (PFI). Unlike classical FL, which aims to learn a single global model, PFI seeks to learn a *set of personalized models*, $\Omega = \{\omega_1, \ldots, \omega_K\}$, tailored to the heterogeneous data distributions of K participants. In general, each client model can be decomposed into two components, formulated as follows:

$$\omega_k = \Phi(\omega_g, \theta_k), \tag{3}$$

where ω_g denotes the shared global parameters, θ_k denotes client-specific parameters, and $\Phi(\cdot)$ is the personalization mapping that integrates global and local components.

Under this characterization, the PFI objective can be formulated as follows:

$$\min_{\omega_g, \Theta} \sum_{k=1}^{K} p_k \, F_k\big(\Phi(\omega_g, \theta_k)\big), \tag{4}$$

where $\Theta = \{\theta_1, \ldots, \theta_K\}$ and $F_k(\cdot)$ is the local loss on $\mathcal{D}_k$. This formulation provides a rigorous distinction from standard FL, which optimizes only ω_g, by jointly learning shared and personalized components.

For completeness, the general objective can also be written directly over personalized models, as follows:

$$\min_{\Omega} \sum_{k=1}^{K} p_k F_k(\omega_k), \tag{5}$$

where each ω_k is generated through the personalization mechanism above. In practice, participants only fine-tune a subset of parameters, such as adapters or LoRA modules (introduced in the following sections), rather than the entire model, to reduce communication and computational costs, enabling PFI to achieve personalized performance at a substantially lower cost than full-model training, making it more practical and scalable.

3 Towards Efficient Personalized Federated Intelligence

As the foundational stage of PFI, efficient personalization enables scalable deployment and lightweight adaptation of FMs for resource-constrained clients. Given the massive scale of FMs, directly deploying them is often impractical. Therefore, lightweight model designs are crucial for practical personalization. This section first establishes the efficiency-oriented foundation of PFI, which subsequently supports trustworthy deployment and continuous adaptive refinement in later stages.

3.1 Communication-and Computation-Efficient Strategies

Deploying FMs in FL environments introduces significant communication and computational bottlenecks due to the massive size of model parameters and limitations of edge devices [29]. To address these challenges, recent research has explored lightweight model adaptation techniques, such as model pruning and compression, to reduce both the volume of transmitted information and the computational overhead during training and inference.

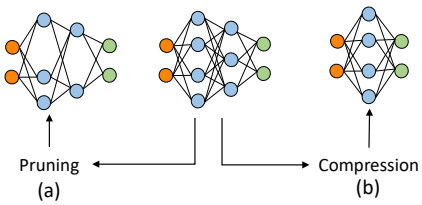


Fig. 3. Illustration of model pruning and model compression. (a) Model pruning removes redundant components (e.g., neurons or layers) to reduce computation and communication. (b) Model compression encodes parameters for compact transmission and storage.

Model Pruning. Among these, model pruning eliminates a model's redundant or less important components, such as neurons, attention heads, or even entire layers, to create a sparse variant that retains comparable performance with reduced computational cost [139]. Formally, given a set of model parameters ω, pruning aims to learn a binary mask $m \in \{0,1\}^{|\omega|}$ such that the effective parameters become $\omega' = m \odot \omega$, where $\odot$ denotes element-wise multiplication. The goal is to find m that minimizes the loss $\mathcal{L}$ on task data while promoting sparsity:

$$\min_{m} F(m \odot \omega) + \lambda \|m\|_0, \tag{6}$$

where λ controls the trade-off between task performance and sparsity. Figure 3 (a) illustrates how model pruning reduces the size and complexity of neural networks, thereby facilitating their deployment on resource-constrained edge devices.

In addition, pruning helps reduce both local computation and uplink communication by transmitting only the pruned subset of model parameters. Moreover, structured pruning (e.g., filter-level or layer-level) is often preferred for better compatibility with edge hardware. Adaptive pruning strategies have also been explored to cater to client heterogeneity, where each client prunes the model according to its own computational constraints and data characteristics. This flexibility makes pruning particularly well-suited for FL scenarios involving FMs, where devices often operate under diverse memory and statistical constraints. Numerous studies have investigated model

Table 4. Taxonomy of Key Communication- and Computation-Efficient Strategies for Personalized Federated Intelligence.

Methods	Categories	Related Works
Model Pruning	Structured Pruning (Comm. ↓ ≈ 80%)	FedMP [75], Complementary Pruning [74], FedSpaLLM [11]
	Adaptive Pruning (Comm. ↓ ≈ 90%)	FedPE [179], PPC-GPT [43]
	Unstructured Pruning (Comm. ↓ ≈ 70%-90%)	Sparse Pruning [180], Numerical Pruning [139]
Model Compression	Quantization-based Compression	CE-FFT (13% Comm.) [185], Quantized LoRA [76], FedQLoRA (35% Comp.) [62]
	Knowledge Distillation	FedID [106], FedDAT(0.75% Commu.) [29], FedMKT(0.12% Comp.) [40]
	Cluster-aware Distillation	FedCKMS (25-75% Comp.) [161]

pruning techniques to enable efficient FL at the network edge [74, 75, 179, 180]. FMs, such as large transformers, contain multiple layers, attention heads, and feed-forward modules that are often overparameterized for downstream tasks. In FL with FMs, pruning enables each client to train or fine-tune a reduced version of the global model, significantly lowering computational costs and allowing resource-constrained devices to participate [11, 43]. FedSpaLLM [11] is the first work to apply pruning techniques to LLMs within an FL framework, aiming to improve communication efficiency and mitigate system heterogeneity. It introduces a novel aggregation function that preserves important parameters by averaging only the non-zero elements across client models, thereby maintaining both sparsity and relevance. To further address limited client resources and the need to protect domain-specific knowledge in FL, PPC-GPT [43] compresses LLMs into task-specific small language models and employs a chain-of-thought distillation strategy to enable lightweight and effective local adaptation. By enabling lightweight model versions that retain essential performance, model pruning plays a central role in making FL practical for large-scale foundation models, particularly in cross-device environments.

Model Compression. Complementing pruning techniques, model compression strategies focus on minimizing communication cost in FFMs by reducing the size of model updates and memory usage without altering the model structure [193]. One widely adopted compression strategy is quantization, which reduces the precision of model parameters and gradients during transmission. By representing weights in lower bit-width formats (e.g., 8-bit or even binary), quantization dramatically reduces the size of model updates while still preserving most of the learning capacity. Formally, quantization maps the model parameter vector ω into a discrete set of values $\tilde{\omega}$, which can be expressed as

$$\min_{\tilde{\omega}} F(Q(\omega)) + \lambda \|\omega - Q(\omega)\|_2^2, \quad \tilde{\omega} = Q(\omega), \tag{7}$$

where $Q(\cdot)$ is the quantization function that projects the continuous values of ω into a discrete space with lower bit precision. For instance, in the case of 8-bit quantization, the function $Q(\omega)$ maps each parameter to one of 256 possible values. The objective is to minimize the loss F while controlling the quantization error, where λ balances the trade-off between model performance and compression error. This approach facilitates efficient communication, especially in FL scenarios with limited bandwidth or resource-constrained devices. Figure 3 (b) illustrates the process of model compression through quantization, highlighting its impact on reducing model update size and transmission cost.

In FFMs, this is particularly valuable because even partial updates of large-scale models can be prohibitively significant in their raw form. CE-FFT [185] is one of the first works to investigate transmission efficiency in FFMs by proposing a per-channel quantization method that leverages the structural patterns of activations and parameters within LoRA modules. Note that LoRA is a parameter-efficient fine-tuning technique that injects low-rank adaptation modules into pre-trained models, as introduced in Section 3.2. Similarly, the authors in [76] integrate quantization strategies with LoRA to enable efficient parameter updates while significantly reducing communication

overhead. To address quantization-induced bias, FedQLoRA [62] introduces a quantization-aware adapter mechanism that compensates for quantization errors and further reduces memory usage during inference. These advances highlight quantization as a promising direction for enabling scalable, efficient communication in FFM systems without compromising model performance. Another effective approach is knowledge distillation, which serves as a powerful framework for model compression in FFMs [29, 40, 106, 161]. This allows clients to benefit from the knowledge of the global model without storing or training the complete foundation model locally. FedID [106] first investigates the application of knowledge distillation to FL in the context of LLMs to mitigate the problem of misleading privileged knowledge caused by confirmation bias in previous works. By transferring the knowledge on an unlabeled public dataset, the proposed method reduces the communication overhead while achieving effectiveness in NLP tasks. Regarding knowledge transfer method, the authors in [80] propose the knowledge transfer-based FFMs framework that formulates problems of grounding FMs in FL, including corresponding objectives, representative knowledge transfer approaches, and privacy measurements. Recently, FedCKMS [161] introduces the cluster-based FL framework designed for efficient adaptation and fine-tuning of FMs across highly heterogeneous and resource-constrained edge devices. It combines partial training with cluster-aware knowledge distillation, enabling effective cross-cluster knowledge transfer from diverse sub-models to the central foundation model while significantly reducing communication and computation overhead. Overall, knowledge distillation in FFMs effectively transfers knowledge from large FMs to lightweight local models, enabling efficient personalization and training under strict resource and privacy constraints. The key enabling communication- and computation-efficient technologies are summarized in Table 4. Note that, in addition to model pruning and compression, meta-learning techniques such as MAML [45] can also be integrated with these strategies. This integration allows each client to not only adapt rapidly but also communicate and compute more efficiently with a compact model, thereby accelerating convergence and improving the scalability of personalized federated adaptation [38].

3.2 Efficient Adaptation Strategies

The scale of FMs presents a major barrier to their deployment in FL, where the communication bandwidth and device resources are severely constrained, especially on edge devices [138]. Full fine-tuning of FMs on each client is impractical due to the model size and communication constraints. To address this, a class of techniques known as PEFT has emerged that enables client-side model adaptation by updating only a small fraction of the parameters [55]. PEFT approaches are highly attractive in FL as they significantly reduce resource demands while allowing client-specific specialization.

LoRA. One of the most widely used PEFT strategies is LoRA [59], which introduces trainable low-rank matrices into the attention layer of the transformer model. Instead of updating the full weight matrix, LoRA enables personalization in federated settings by allowing each client to learn its own low-dimensional adaptation while keeping the shared backbone frozen. Formally, for a pre-trained weight matrix $\omega \in \mathbb{R}^{d\times k}$, each client learns a personalized increment $\Delta\omega$ defined as

$$\omega' = \omega + \Delta\omega = \omega + BA, \tag{8}$$

where $A \in \mathbb{R}^{d\times l}$ and $B \in \mathbb{R}^{l\times k}$ are trainable client-specific low-rank matrices with $l \ll \min(d, k)$, significantly reducing the number of trainable parameters. This low-rank decomposition explicitly characterizes the personalization mechanism: each client obtains a distinct adaptation BA that reflects its local data distribution, while all clients share the same frozen global weight ω. Typically, A is initialized from a normal distribution (e.g., $A \sim \mathcal{N}(0, \sigma^2)$) and B is initialized as zero, ensuring

that the initial output remains unchanged. The working mechanism of LoRA is illustrated in Appendix C.

LoRA is particularly well-suited for FL due to its modular structure and minimal communication overhead. Clients can locally fine-tune only the low-rank matrices while keeping the backbone of the foundation model frozen. During communication rounds, only these compact matrices are transmitted to the server, substantially reducing bandwidth usage. Several studies [145, 162, 183] have explored LoRA for federated fine-tuning approaches. FedIT [183] was the first work to apply LoRA in federated fine-tuning, demonstrating that transmitting low-rank updates can significantly reduce the communication cost. However, FedIT assumes that all clients use LoRA modules with same rank, which limits its practical applicability. In the practical systems, clients have heterogeneous computational budgets and therefore require different LoRA ranks to balance the personalization capacity. Due to this characteristic, mismatched shapes across LoRA matrices make the aggregation infeasible, preventing flexible deployment across devices. To address this limitation, subsequent works have proposed several strategies to overcome heterogeneous client ranks and aggregation inconsistency. For example, FLora [162] introduces an aggregation-noise-free federated fine-tuning approach that supports heterogeneous LoRAs by employing a stacking mechanism for aggregation. FFA-LoRA [145] tackles data heterogeneity in federated fine-tuning by keeping the pre-trained LLMs frozen and initializing the non-zero components of the LoRA modules randomly, thus enhancing the privacy under differential privacy and improving computational efficiency. Beyond the early LoRA-based federated works, several studies refine how LoRA updates are combined and initialized across heterogeneous clients. The authors in [52] propose selective aggregation of LoRA parameters, updating only layers or ranks whose client updates exhibit high cross-client consistency-thereby reducing drift from non-IID data and improving stability under partial participation. LoRA-FAIR [15] complements this idea with an aggregation-and-initialization refinement scheme that tackles the challenge of server-side aggregation bias in LoRA matrices and client-side initialization lag. These methods make LoRA aggregation more robust to statistical heterogeneity while preserving PEFT's communication advantages.

Prompt Tuning. Another effective PEFT method is *prompt tuning*, which adapts a foundation model by appending a small set of learned tokens, known as soft prompts, to the input. This method avoids changing internal model weights and steers the model towards downstream tasks using trainable embeddings [113]. Formally, given an input sequence x, prompt tuning prepends a learnable prompt vector $P \in \mathbb{R}^{l \times d}$, where l is the number of soft tokens and d is the model's hidden size. The new input to the model becomes:

$$F(\hat{x}); \quad \hat{x} = [P; x], \tag{9}$$

where only the parameters in P are updated during training, while the backbone model $F(\cdot)$ remains frozen. This design enables lightweight and task-specific adaptation with minimal overhead. The core working mechanism of prompt tuning is illustrated in Appendix D.

Therefore, the lightweight nature of prompt vectors makes prompt tuning highly memory-and communication-efficient, well-suited to the constraints of FL. In FL scenarios, prompt tuning enables each client to maintain its personalized prompt vectors while sharing a globally frozen foundation model. This setup not only reduces the communication burden, as only the tiny prompt vectors are exchanged, but also offers better privacy protection since the majority of the model remains untouched and is less likely to reveal sensitive local data. Several recent studies have explored the adaptation of prompt tuning to FL, addressing key challenges such as domain generalization and data heterogeneity, as well as application areas including vision-language tasks and weather forecasting [53, 91, 189]. FedPrompt [189] was among the first to introduce prompt learning into FL, aiming to accelerate global aggregation and address scenarios with limited user data. To address

Table 5. Taxonomy of Key Adaptation Strategies for Personalized Federated Intelligence.

Methods	Categories	Related Works
LoRA	Standard LoRA	Hu *et al.* (0.025% trainable parameters) [59]
	Federated LoRA	FedIT (0.26% params) [183], FLora (21.5% params) [162], FFA-LoRA [145], Guo *et al.* [52], LoRA-FAIR [15]
Prompt Tuning	Soft Prompts	Min *et al.* [113], FedPrompt (0.007-0.014% Comm.) [189], pFedPrompt [53], Li *et al.* [91]
	Textual (Hard) prompts	AutoPrompt [141], PromptSource [10]
Adapter Tuning	Single Adapter	Houlsby *et al.* (1.14% params)[58], FedAdapter (0.48-0.55% params) [22]
	Dual Adapter	FedDAT (0.75% Commu.) [29], FedDPA (0.06% Commu.) [174], FedPIA [133]

statistical heterogeneity in a personalized manner, pFedPrompt [53] introduces a client-specific attention module that generates locally tailored spatial visual features. To enable efficient model personalization, the authors in [91] deploy the unbalanced optimal transport to align local visual features with corresponding local prompts, effectively balancing global consensus and local personalization. Recently, the authors in [104] introduce a personalized MoE prompt mechanism in federated settings. Clients learn lightweight expert prompts and a routing policy that selects or blends experts per input/client distribution, improving transfer across skewed modalities while keeping communication small (only prompt and router parameters). In addition to soft prompts for prompt tuning, recent studies have also explored textual prompts in the context of foundation models. AutoPrompt [141] introduces a gradient-guided method for automatically generating discrete textual prompts, enabling language models to perform sentiment analysis. Notably, on the LAMA fact-retrieval benchmark, AutoPrompt achieves 43.3% top-1 precision, substantially outperforming the single-prompt baseline of 34.1%. Meanwhile, PromptSource [10] offers a collaborative environment for designing, sharing, and evaluating human-crafted textual prompts across diverse tasks, with more than 2,000 prompts covering around 170 datasets. Overall, prompt tuning offers a highly adaptable and communication-efficient mechanism for deploying FMs in FL. Its lightweight nature, modularity, and compatibility with privacy-preserving constraints make it a powerful tool for enabling PFI at scale.

Adapter Tuning. Beyond LoRA and prompt tuning, *adapter tuning* has also gained attention for enabling efficient and modular fine-tuning of FMs in FL [22, 29, 133, 174]. Adapter tuning represents another line of PEFT methods, where small bottleneck modules are inserted between the layers of a pre-trained transformer [58]. Specifically, an adapter module typically consists of a down-projection, a nonlinearity, and an up-projection, formulated as follows:

$$\text{Adapter}(h) = h + \omega_{\text{up}}\, f(\omega_{\text{down}}\, h), \tag{10}$$

where h is the hidden activation from the Transformer layer, $\omega_{\text{down}} \in \mathbb{R}^{m\times d}$ projects the features into a low-dimensional space ($m \ll d$), $f(\cdot)$ is a non-linear activation function (e.g., ReLU), and $\omega_{\text{up}} \in \mathbb{R}^{d\times m}$ projects back to the original dimension. The residual connection ensures that the pre-trained model's behavior is preserved when the adapter is initialized to near-zero impact. Only the adapter parameters are updated during fine-tuning, making the approach both parameter-efficient and modular. We illustrate the key design of the adapter architecture in Appendix E.

Therefore, this design is especially attractive in FL, where devices operate under limited resources and direct access to global model parameters is often restricted. In federated settings, adapter tuning enables clients to personalize the model by training only their local adapter modules while sharing or aggregating them selectively to improve generalization. To reduce NLP training costs, FedAdapter [22] employs a progressive training paradigm with trial-and-error profiling while training only adapter modules, enabling convergence within just a few hours-up to 155.5× faster than vanilla FedNLP and 48× faster than strong baselines. In the domain of vision-language tasks, FedDAT [29] is the first to propose a fine-tuning framework for heterogeneous federated

settings by leveraging a dual-adapter teacher model. Their approach regularizes client updates and employs mutual knowledge distillation to enable efficient knowledge transfer. Moreover, FedDPA [174] introduces a dual-personalization design that separates server-shared and client-specific adapter components, enabling global transfer while preserving local specialization. Recently, FedPIA [133] enhances the integration of FL and PEFT by enabling richer information exchange between server–client adapters and local–global adapters within each client, achieving F1-score improvements of 5.1% on Open-I and 2.73% on MIMIC over FedDAT across all clients and adapter configurations. Specifically, the method permutes client adapter neurons at each layer to align with globally initialized adapter neurons, guided by the theory of Wasserstein Barycenters, thereby improving cross-client consistency and personalization. Overall, adapter tuning offers a strong balance between model adaptability and communication efficiency, making it a practical choice for real-world FL systems involving foundation models across diverse user contexts and devices. We summarize the key adaptation technologies for enabling personalized federated intelligence in Table 5.

Trade-offs among PEFT techniques: Although PEFT methods such as LoRA, prompt tuning, and adapter tuning all aim to enable lightweight personalization of FMs, they exhibit distinct trade-offs in practical personalized federated intelligence deployments. Prompt tuning achieves the lowest communication and memory overhead by only optimizing a small set of input embeddings. However, its limited adaptation capacity may lead to suboptimal performance under severe statistical heterogeneity or complex domain shifts. Adapter-based methods introduce dedicated trainable modules within the model, providing stronger personalization robustness and stability across heterogeneous clients, at the cost of increased memory footprint and aggregation complexity. LoRA offers a flexible middle ground, where the choice of rank directly governs the balance between personalization expressiveness and efficiency: low-rank settings reduce communication overhead and overfitting risk, while higher-rank configurations improve adaptation accuracy but may amplify personalization drift in highly non-IID settings. Overall, choosing among PEFT strategies requires jointly considering deployment constraints, personalization granularity, and the degree of cross-client heterogeneity to achieve an optimal balance between efficiency and performance.

3.3 Summary and Lessons Learned

This section reviewed two complementary technique families, model compression and PEFT, that are essential for enabling efficient deployment of FMs. Both approaches mitigate the communication and computation constraints of edge devices while preserving personalization capability.

Takeaways: Model compression methods (e.g., pruning, quantization, and knowledge distillation) reduce model size and inference cost, facilitating efficient transmission and local execution. PEFT techniques, such as LoRA, prompt tuning, and adapter tuning, enable lightweight personalization by updating only a small subset of parameters. Together, these methods provide flexible and scalable adaptation of large models to heterogeneous clients.

Connection to PFI: Compression and PEFT are key enablers of PFI, as they significantly reduce the resource footprint of FMs and allow diverse clients to participate in federated training and inference while supporting privacy-preserving personalization. Open challenges include performance degradation under aggressive compression, limited expressiveness of PEFT under extreme client heterogeneity, dependency on proxy data for distillation, and difficulties in aggregating adapter- or LoRA-based updates. Dynamically selecting client-specific strategies in decentralized settings remains an important direction for future research.

4 Towards Trustworthy Personalized Federated Intelligence

Building upon efficient personalization, the next stage of PFI focuses on ensuring that adapted models remain trustworthy in real-world deployment. This section examines the evolving landscape of trustworthy PFI, beginning with fairness considerations and progressing through interpretability, hallucination mitigation, robustness against various attacks, and privacy protection, thereby establishing a reliable foundation for subsequent adaptive-driven refinement.

4.1 Mitigating Bias and Ensuring Fairness

Fairness in FMs refers to the model's ability to avoid discriminating against individuals or groups within society [82]. Since FMs are often trained on diverse datasets collected from numerous clients, they are susceptible to inheriting and amplifying biases present in the data, which can result in unfair or discriminatory outcomes. To mitigate such issues, we explore three complementary strategies to enhance fairness in PFI: model-level processing, client-side processing, and server-side processing.

Model-level Processing. Given the privacy constraints inherent to FMs, a complementary strategy to mitigate bias is to modify the FM architecture or adapt the training objective to include fairness-aware components. For example, a practical and widely adopted approach involves decoupling the model into shared base layers and a personalized head, such as FedPer [7], enabling collaborative learning of generalizable base layers while allowing client-specific adaptations.

Client-side Processing. In addition to model-level processing, client-side data preprocessing has been explored as a means for bias mitigation in centralized settings [132]. However, in FL, individual user data remains inaccessible to other parties, including servers. An alternative approach to tackle this challenge is integrating democratized learning [116], which can leverage client-level model characteristics to enhance fairness while preserving personalization. By exploiting similarities across clients' model updates or representations, it becomes possible to promote fairness among participating clients without compromising local preferences. A complementary line of work adopts proactive bias-aware aggregation, such as FedFair [27], which evaluates each client's contribution based on both bias metrics and personalized objectives before global aggregation. Clients with excessive bias can then be downweighted or excluded to prevent unfair influence on the shared model. However, such approaches may remain sensitive to non-IID data, where heterogeneity may cause certain clients to be misclassified as biased despite the absence of systematic unfairness [131]. More recent studies, such as FairDPFL-SCS [131], explicitly highlight this limitation and introduce dynamic client-selection strategies that balance fairness, accuracy, and personalized utility under non-IID data, yielding dramatic accuracy gains from 58.21% to 99.04% and thereby demonstrating the clear benefits of aggregating local updates.

Server-side Processing. Another important approach to mitigating bias involves applying client-wise evaluation and penalization, post-aggregation correction, and fair routing in MoE architectures on the server side to adjust model behaviors or representations and promote fairness. By replacing unfair word embeddings within the models, the model bias can be effectively reduced, thereby enhancing fairness [16]. One notable approach is adjusting the output of FMs without altering their parameters, which involves leveraging the features learned by FMs to modify the model output according to desired specifications. This innovative technique, as demonstrated in [90], involves interpreting the meaning of feature vectors and adapting model outputs accordingly to address bias and ensure safety content in the system. Overall, the coordinated examination of key enabling technologies across the client, model, and server sides is expected to collectively benefit PFI, resulting in reduced bias and enhanced fairness.

4.2 Interpretable Capabilities

Interpretable capabilities are key to building trust in PFI, where models operate on sensitive, user-specific data. To this end, we focus on two complementary approaches: traditional XAI techniques and emerging reasoning-based methods. XAI enhances transparency by revealing how models arrive at decisions, either at a global model level or for individual predictions. Meanwhile, reasoning models improve interpretability by generating structured intermediate steps, aligning model behavior with human-like logic. Together, these two selected techniques provide clearer insights into personalized model behavior, making the decision in PFI more transparent and trustworthy.

XAI. Various strategies can be proposed to enhance model performance in the context of FL within FMs [146], but these often result in opaque models with decision-making processes that are difficult to interpret. This lack of transparency can undermine user trust and accountability [102]. XAI techniques offer a solution by providing insights into how models arrive at their predictions, helping to bridge this transparency gap. Based on the scope of explanation, XAI methods can be broadly categorized into two types: local explanations, which focus on individual predictions, and global explanations, which aim to interpret the model's overall behavior. The core principles and workflows of both local and global explainability approaches are illustrated in Appendix F.

Popular methods such as GradCAM [137], AttentionViz [178], and SHAP [103] provide local explanations by illustrating how specific features influence individual predictions. For instance, GradCAM highlights the regions of the input that had the most influence on the model's decisions, while AttentionViz leverages attention mechanisms to visualize how input features are weighted during processing. Additionally, SHAP enhances interpretability by assigning Shapley values to input features, quantifying their contributions to the predictions based on game-theoretic principles. Building on this, FedSA [117] adapts SHAP to federated settings by estimating each client's contribution, thereby enabling the detection of and defense against malicious participants in the federation. In contrast, global explanations, such as model distillation [147], GAM [23], and LIME [129], offer a broader understanding of the model's decision-making processes across all data, helping users grasp how the model operates as a whole. For instance, LIME [129] treats the model as a black box and approximates its local decision boundary by fitting an interpretable surrogate model on perturbed samples around a specific input. Formally, LIME solves the following optimization problem:

$$\arg\min_{g \in G} \; F(f, g, \pi_x) + \Omega(g), \tag{11}$$

where f is the original black-box model, $g \in G$ is an interpretable surrogate model (e.g., linear regression), $F(f, g, \pi_x)$ measures the fidelity of g in approximating f near the input x using a locality-aware weighting function π_x, and $\Omega(g)$ is a regularization term that penalizes model complexity to ensure interoperability. Additionally, the distill-and-compare approach proposed by [147] allows for auditing black-box risk scoring models by distilling them into transparent student models and comparing these to ground-truth models, thereby offering valuable insights into the black-box model's decision-making process. Therefore, integrating these XAI methods into the development of the PFI process shows great promise in ensuring that the model's decision-making process remains transparent and trustworthy to users.

Reasoning Methods. FMs can enhance their reasoning capabilities through the integration of various techniques, including chain-of-thought (CoT) [43], self-consistency with CoT (CoT-SC) [159], and tree-of-thought (ToT) [175]. The difference between them is as shown in Appendix G. CoT facilitates a systematic approach to problem-solving, guiding the model through logical steps rather than directly jumping to conclusions. This method enhances the model's proficiency in handling complex tasks requiring multiple steps, such as mathematical problems, logic puzzles, and decision-making processes. By incorporating CoT, FMs are encouraged to provide intermediate

steps to justify their answers, thereby offering more insight into the reasoning behind the model's outputs. Building upon CoT, CoT-SC was introduced to enhance the reasoning process by exploring multiple paths simultaneously. In CoT-SC, several CoTs are generated during the inference phase of FMs, allowing them to navigate through various logical steps towards the correct solution. These CoTs are generated independently, and the final output is determined based on the most frequently occurring solution among the CoTs. On the other hand, ToT takes this process a step further by fostering connections among these independent CoT-SC pathways. In ToT, solutions are not just evaluated independently; they also share explanations, collaboratively refining reasoning to formulate the best possible answer. By aligning different perspectives, ToT enables FMs to generate more comprehensive explanations and identify the optimal decision for a given problem.

In the context of applying reasoning abilities within PFI, we utilize RL with feedback to enhance the reasoning capabilities of FMs. However, privacy concerns present challenges in implementing techniques such as RLHF [118] and RLAIF [12] within PFI. A more detailed discussion of these issues will follow in the next section. Conversely, the pure RL approach [51] has demonstrated promising reasoning potential for FMs. Unlike methods that require training a separate reward model for integration into PFI, pure RL directly generates rewards using predefined rules. Furthermore, integrating pure RL into PFI enables a more adaptive learning process. By leveraging predefined reward criteria, FMs can refine their reasoning without relying on external validation or manually labeled data. This independence makes pure RL particularly suitable for PFI, where privacy and decentralization are paramount considerations.

4.3 Mitigating the Challenges of Hallucinations

Although FMs can already provide strong generalization capabilities, they still tend to produce false or fabricated details, also known as hallucinations. Such hallucination reduces the reliability of FMs by generating content that is nonsensical or unfaithful to the source material [110]. Based on the alignment failures between model outputs and either real-world facts or the provided inputs, we discuss two types of hallucination, including factuality hallucination and faithfulness hallucination [63].

Factuality hallucinations arise when FMs produce outputs that contradict real-world facts or provide unverifiable claims. These can be further divided into two subtypes: factual contradictions, which involve inaccuracies about entities or relationships (e.g., misattributing inventions), and factual fabrications, which consist of unverifiable statements or exaggerated claims. In contrast, faithfulness hallucinations occur when FMs fail to align with user instructions, provided context, or logical reasoning. These deviations are categorized into three subtypes: instruction inconsistency, where outputs unintentionally misalign with directives; context inconsistency, where responses contradict user-provided contextual details; and logical inconsistency, which involves contradictions in reasoning or discrepancies between intermediate steps and final conclusions. To mitigate the challenges of hallucinations in the development of PFI, various strategies are typically employed across data preparation, training methodologies, and inference processes [63]. For instance, in data preparation, techniques such as data augmentation and the careful curation of high-quality training data can help reduce the occurrence of hallucinations [63]. During training, methods like RLHF [118] and RLAIF [12] can help align model outputs more closely with human expectations and real-world facts. In the inference phase, approaches such as RAG [68] and contrastive decoding [160] can ensure that the generated content remains both factual and contextually relevant. However, within the constraints of the PFI scenario, the lack of access to client data limits the feasibility of approaches reliant on data. For instance, applying data filtering [63] to ensure high-quality and accurate information requires clients to possess strong domain expertise. This is impractical, as data is collected from a diverse range of devices, making it impossible to guarantee its consistency

or reliability. As a result, minimizing hallucinations during the training phase becomes a critical focus for improving the performance and reliability of FMs in PFI.

Effective techniques for improving reasoning in PFI include RLHF [118] and RLAIF [12]. These methods help refine the reasoning capabilities of FMs by leveraging feedback-driven optimization. For an in-depth discussion and comparative analysis of RLHF and RLAIF, readers are referred to [122]. However, implementing RLHF in PFI presents significant privacy challenges, as it typically requires sampling and training a reward model using client datasets [122]. A possible solution for this challenge is collaborative training that can effectively optimize both the FM and the reward model. In this approach, each client independently generates reasoning data from their local dataset and trains a reward model alongside the FM. Given the large number of clients, the aggregated reward model reflects the collective intent of the broader community, ensuring a more balanced and generalized training process. However, one limitation of this approach is the requirement for additional labeled data, which can be expensive and resource-intensive. In real-world scenarios, obtaining consistent labeled data across multiple clients poses significant challenges due to data heterogeneity and privacy constraints. A scalable alternative, such as RLAIF [12], can help reduce reliance on manual annotation during rewording model training. However, this technique often leverages off-the-shelf LLMs, which are prone to generating inaccurate or fabricated outputs, especially when fine-tuned on low-quality datasets. Therefore, further empirical studies are necessary to address and minimize hallucination risks in FMs within PFI.

4.4 Enhancing Robustness Against Attacks and Privacy Protection

Beyond mitigating bias and improving interpretability, enhancing robustness against adversarial attacks is a critical component of FL. Such attacks can manipulate model updates, degrade performance, or alter behavior across FL frameworks, creating serious security concerns in real-world deployments. While safety is a fundamental baseline requirement for all FL systems, it is especially crucial in the context of PFI, where FMs demand higher accuracy and reliability given their broad and sensitive applications. Addressing robustness and safety challenges is therefore essential, as adversarial threats pose unique risks to the trustworthy personalization of FMs. These attacks encompass adversarial examples [121], where subtle input perturbations are used to mislead model predictions; backdoor attacks [49], in which malicious clients inject triggers to manipulate an FM's behavior on specific inputs; model poisoning attacks [44], where adversarial participants deliberately corrupt FM updates during federated training; and inference attacks [60], in which adversaries attempt to extract sensitive information from shared model updates or predictions.

4.4.1 Adversarial Examples. The increasing adoption of FMs in FL introduces new vulnerabilities, as their high-dimensional input space, large parameter size, and modular adaptation layers create a substantially larger attack surface compared to traditional FL models [61]. This makes it critical to explore defense mechanisms tailored to the unique challenges of FL-FMs. To counter adversarial examples, FatCC [121] combines global feature contrastive learning with adversarial training, providing an effective defense mechanism in traditional FL settings. However, its effectiveness in the context of FMs remains uncertain, as the large scale and complexity of FMs can amplify vulnerabilities. Moreover, full-scale adversarial training becomes computationally prohibitive for FMs due to their size. To address these challenges, recent approaches have explored alternative strategies. For instance, FedEAT [119] applies embedding-space adversarial training combined with robust aggregation (geometric median) to improve the adversarial robustness of federated LLMs with minimal performance loss. Other lightweight methods, such as pre-trained model-guided adversarial fine-tuning [158] and adversarial adaptation techniques like AdvLoRA [72], aim to

enhance FM robustness efficiently without incurring the high cost of full-scale adversarial training.

4.4.2 Backdoor and Poisoning Attacks. In addition to adversarial examples, FL-FMs are also susceptible to other malicious interventions, such as backdoor and poisoning attacks, which require dedicated defense mechanisms. To mitigate backdoor attacks, SDFC [64] employs Fisher calibration to identify and regulate distribution-sensitive parameters based on parameter consistency, thereby prioritizing clients whose updates align with the global distribution and effectively suppressing the influence of malicious participants in heterogeneous federated settings. Beyond parameter-level manipulations, recent studies such as [26] also highlight the threat of backdoor injection in RAG, where attackers poison the retrieval corpus, for example, by inserting trigger-containing documents that cause the FM to output targeted responses whenever the trigger appears, thus enabling covert and persistent manipulation of downstream FM behaviors. Another line of defense leverages the modular structure of FL-FMs, recognizing that securing these models is substantially more challenging, particularly because attackers can exploit FM vulnerabilities to inject backdoors through synthetically generated data. To address this, the authors in [13] propose a data-free defense strategy that constrains internal activations to remain within safe, expected ranges, thereby suppressing backdoor behaviors while preserving the functionality of the FM. These activation constraints are optimized jointly with FL training using synthetic data, resulting in an effective and scalable defense mechanism tailored to the unique challenges of FL-FM settings. To defend against model poisoning attacks, [176] introduces an internal auditor that analyzes encrypted gradients using Gaussian mixture models and Mahalanobis distance, achieving robust aggregation with minimal computational and communication overhead.

4.4.3 Inference Attacks. Beyond active attacks on model behavior, inference attacks [60] pose another serious threat to user privacy, as adversaries attempt to extract sensitive information from shared model updates or predictions. In traditional FL, studies such as deep leakage from gradients (DLG) [192] have shown that an honest-but-curious server can reconstruct users' private training data by optimizing dummy inputs to match the shared gradients, revealing that gradient exchange alone can expose sensitive information. In the context of FL-FMs, recent work [134] has further demonstrated that attackers can perform gradient-inversion attacks on PEFT workflows, exploiting even lightweight adapter gradients to reconstruct high-fidelity local training samples. To counter such attacks, methods like FLSG [39] have been proposed, which defend against passive label inference attacks by generating and using Gaussian-distributed gradients while maintaining lower computational costs. Additionally, FedShield-LLM [112] proposes a secure and efficient federated fine-tuning framework that combines fully homomorphic encryption (FHE)-protected and pruned LoRA updates to defend against inference attacks while significantly reducing computational and communication overhead for resource-constrained clients. Although these defenses do not fully address the underlying attack issues, their proposed mechanisms show potential for balancing robustness and efficiency, making them relatively suitable for deployment in personalized federated settings involving large-scale models.

4.4.4 Security and Privacy. At a broader level, the continued adoption of FL-FMs highlights the need for a unified security and privacy framework, such as confidential federated learning, that can jointly strengthen robustness and privacy across diverse threat vectors. In this direction, secure aggregation [143] ensures that individual client updates remain hidden from the server even when training high-dimensional FM adapters, such as LoRA-based updates in federated tuning of LLMs. Differential privacy [163] further enhances protection by injecting carefully calibrated noise into updates, offering formal privacy guarantees against inference attacks in federated large

models. Moreover, trusted execution environments [114] provide hardware-backed isolation that safeguards both computations and sensitive parameters in FL-FM pipelines, protecting them from untrusted servers or malicious insiders and thereby enabling secure and privacy-preserving PFI. We summarize the key trustworthiness technologies for enabling personalized federated intelligence in Appendix H.

4.5 Summary and Lessons Learned

This section has examined five core dimensions for enabling trustworthy PFI: bias mitigation, fairness, interpretability, hallucination mitigation, and robustness. Bias and fairness address disparities arising from pre-training and deployment through multi-level mitigation strategies. Interpretability enhances transparency via XAI and reasoning-based generation, while hallucination mitigation improves factuality using data filtering and feedback-driven optimization (e.g., RLHF and RLAIF). Robustness is reinforced against adversarial, backdoor, poisoning, and inference attacks through lightweight and privacy-preserving defenses such as AdvLoRA and Fisher calibration. AdvLoRA enhances robustness to adversarial examples in FL-FMs, while Fisher calibration suppresses backdoor behaviors by identifying distribution-sensitive parameters. Although they defend against different classes of threats, both approaches align with and contribute to the overarching goal of trustworthy PFI.

Takeaways: Bias, fairness, interpretability, hallucination mitigation, and robustness collectively enable equitable, transparent, factual, and secure personalization, forming the foundation of trustworthy PFI.

Connection to PFI: These dimensions are essential for trustworthy PFI, ensuring fair and accountable personalization, faithful generation, and resilience against adversarial, backdoor, and privacy-invasive attacks. However, open challenges include limited access to sensitive attributes for fairness correction, privacy–explainability trade-offs, supervision scarcity for hallucination mitigation in decentralized settings, and efficiency–robustness trade-offs.

5 Towards Adaptive Personalized Federated Intelligence

Beyond efficient and trustworthy deployment, the final stage of PFI lies in continuous adaptation to evolving user contexts. While efficient and trustworthy mechanisms lay the foundation for deployment, static models may still struggle with outdated knowledge. To address this, integrating RAG offers a promising solution by enabling adaptive and context-aware generation. This section explores the role of RAG in PFI by examining its foundations and design space.

5.1 Fundamentals of Retrieval-Augmented Generation

RAG [88] is a hybrid framework that integrates neural retrieval with conditional generation or prediction by explicitly conditioning the model on retrieved external evidence items. Here, an *evidence item* d denotes a generic retrieved unit of external information (e.g., passage, structured record, table row, feature vector, or multimodal snippet) used to condition prediction or generation. In the context of PFI, RAG can be understood as decomposing personalized intelligence into two cooperative functions: a retriever that identifies information relevant to the client's current input, and a predictor/generator that produces the desired output using both the input and the retrieved evidence items. Accordingly,

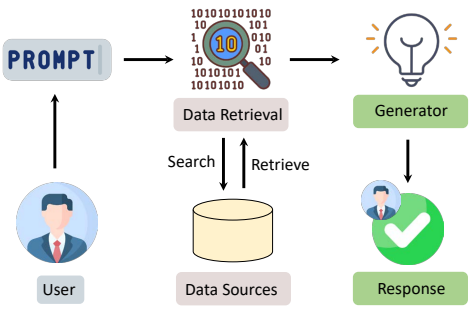


Fig. 4. The working mechanism of RAG [37].

we can represent the personalized model for client k as $\omega_k = \{\omega_k^R, \omega_k^G\}$, where ω_k^R parameterizes the retriever $\mathcal{R}_{\omega_k^R}$ and ω_k^G parameterizes the predictor/generator $\mathcal{G}_{\omega_k^G}$.

Different from purely parametric models that rely solely on internal weights, RAG augments inference by retrieving relevant evidence items at inference time. The retriever is typically implemented as a dense bi-encoder [81], where both the input and the evidence items are encoded into a shared embedding space. The retriever score between an input x and an evidence item d parameterized by ω_k^R is computed as follows:

$$\mathcal{R}_{\omega_k^R}(x, d) = \mathrm{sim}(E_x(x; \omega_k^R), E_d(d; \omega_k^R)) = E_x(x; \omega_k^R)^\top E_d(d; \omega_k^R), \tag{12}$$

where $E_x(\cdot)$ and $E_d(\cdot)$ denote the input and evidence item encoders, respectively, and $\mathrm{sim}(\cdot)$ represents dot product similarity. The retriever selects the top-N evidence items $\{d_1, d_2, \ldots, d_N\}$ with the highest scores for a given input. To obtain a retrieval probability distribution, we normalize the scores over the retrieved set using a softmax function with temperature τ:

$$P(d_i|x; \omega_k^R) = \frac{\exp(\mathcal{R}_{\omega_k^R}(x, d_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{R}_{\omega_k^R}(x, d_j)/\tau)}. \tag{13}$$

In practice, the top-N items are retrieved, and the scores are renormalized over this set. Once the relevant evidence items are retrieved, the predictor/generator $\mathcal{G}_{\omega_k^G}$ produces the final output y. Depending on the task, $\mathcal{G}_{\omega_k^G}$ may correspond to a conditional language model, a multimodal network, or a task-specific predictor that consumes (x, d_i). The output is conditioned on the original input x and the retrieved evidence items. To aggregate evidence from multiple retrieved evidence items, the likelihood of producing the output given the personalized model ω_k is computed by marginalizing over the retrieved evidence items as follows:

$$P(y|x; \omega_k) = \sum_{i=1}^{N} P(y|x, d_i; \omega_k^G)\, P(d_i|x; \omega_k^R), \tag{14}$$

where $P(y|x, d_i; \omega_k^G)$ denotes the probability of producing y given the input x and retrieved evidence item d_i. The training objective of RAG is to maximize the likelihood of the correct task output. Within the PFI framework, this objective defines the local loss function $F_k(\omega_k)$ for client k, computed as the empirical risk over the client's local dataset $\mathcal{D}_k$:

$$F_k(\omega_k) = \frac{1}{|\mathcal{D}_k|} \sum_{(x,y)\in\mathcal{D}_k} -\log \sum_{i=1}^{N} P(y|x, d_i; \omega_k^G)\, P(d_i|x; \omega_k^R). \tag{15}$$

The retrieval probability $P(d_i|x; \omega_k^R)$ serves as a mixture weight over retrieved evidence items, allowing the predictor/generator to synthesize evidence items into reliable and context-aware outputs. Note that in PFI, adaptation may occur either by optimizing (parts of) ω_k^R and/or ω_k^G via FL, or by keeping parameters fixed and updating the external knowledge/index used for retrieval. The working mechanism of RAG at inference time is shown in Fig. 4.

5.2 Design Space for Adaptive Personalized Federated Intelligence

Building on the foundational principles of RAG outlined above, this subsection investigates how retrieval-based mechanisms can be effectively adapted to support user-centric personalization in PFI. Below, we structure the design space into two complementary dimensions: client-side personalization and server-side optimization.

5.2.1 Client-side Personalization. To enable personalization on the client side, various lightweight and privacy-preserving techniques such as prompt conditioning, knowledge indexing, and retriever adaptation can be developed.

Prompt conditioning. Prompt conditioning is an inference-time mechanism for steering a *frozen* FM by prepending or otherwise augmenting the input with a task- or user-specific prompt, rather than updating model parameters [21]. When the prompt includes instructions and/or exemplars, the FM may exhibit *in-context learning*, i.e., apparent task adaptation induced purely by the provided context without gradient-based optimization [21].

Formally, given a query x, a client-specific prompt p_k, and a frozen FM with parameters ω, generation is conditioned on the concatenated input:

$$y \sim p_\omega(y \mid [p_k; x]), \qquad \omega \text{ is fixed.} \tag{16}$$

Here, $[p_k; x]$ denotes token-level concatenation of the prompt and query. The prompt p_k may be a natural-language instruction, a set of exemplars (few-shot context), or a retrieved text prefix (e.g., policy, persona, or domain hints) prepended to the input [21, 88]. Such conditioning can control task specification, output format, and domain emphasis without modifying the backbone parameters.

Prompt conditioning is particularly attractive in federated or multi-client settings, since client-specific behavior can be adjusted via client-side context selection rather than exchanging full-model updates. Recent *federated in-context* frameworks explicitly exploit this property to reduce communication by sharing natural-language context instead of model parameters[157, 166]. Prompt conditioning also integrates naturally with RAG: retrieved evidence is injected into the input and thus serves as additional contextual conditioning for generation [88]. Moreover, prompts can shape retrieval intent (e.g., query rewriting/augmentation) and guide how retrieved evidence is interpreted and formatted [107]. An example system architecture for federated prompt conditioning is presented in Appendix I.

Local Knowledge Indexing. A central challenge in deploying FFMs lies in achieving user-level personalization while safeguarding privacy. Although FMs offer strong generalization, they often falter in highly contextual settings such as personalized document understanding, niche domain tasks, or individual knowledge retrieval, where access to user-specific information is essential [50]. Traditional cloud-based solutions that require uploading user data for fine-tuning or indexing are often constrained by privacy regulation, policy, or bandwidth. To overcome this, *local knowledge indexing* provides a promising alternative that adheres to the *model to data* principle [154], where computation occurs locally at the data source.

In RAG frameworks [88], local indexing enables each client to build and maintain a private, queryable repository of documents, including emails, notes, and clinical records. These documents are embedded into vector representations using a local encoder and stored locally on the client; when stronger protection is required, retrieval and/or generation can be executed inside a Trusted Execution Environment (TEE) (confidential computing), so the plaintext context is only exposed within the protected boundary [2]. During inference, the retriever performs a similarity search over this private index to retrieve the top-k relevant documents. The retrieved evidence can then be consumed by a generator that is executed locally on the client using globally shared weights, so the private context does not leave the device and the model-to-data principle is preserved. Alternatively, when generation must be offloaded, the generator can be executed within a trusted execution environment at an edge tier, which keeps the retrieved context within a protected boundary [2].

Recent systems such as ERAGent [140] demonstrate that retrieval-augmented assistants can improve task relevance and personalization by leveraging historical interactions (e.g., a memory knowledge base and learned user profiles) alongside external retrieval. Additionally, modular toolkits such as LlamaIndex have gained popularity for rapidly assembling and experimenting

with end-to-end RAG pipelines. In decentralized settings, C-FedRAG [2] and FRAG [188] explore secure retrieval across distributed data silos: C-FedRAG leverages confidential computing/TEEs to protect contextual data during orchestration and LLM inference, while FRAG supports encrypted approximate nearest-neighbor search over distributed vector databases using (multi-party) homomorphic encryption. Altogether, local knowledge indexing can enhance the personalization capabilities of FFMs by anchoring responses in user-owned data, enabling privacy-friendly, adaptive, and domain-specific intelligence. When integrated with federated RAG workflows, it becomes a cornerstone technology for building adaptive PFI.

Retriever Adaptation. In federated settings, users often operate across heterogeneous domains such as healthcare, law, or education, and client data are typically non-IID across participants, which can lead to retrieval mismatches caused by variations in vocabulary, semantics, and document structures. While prompt tuning provides a lightweight mechanism to steer model behavior, it is often insufficient to address domain-specific retrieval errors, particularly when the underlying embedding space of the retriever is misaligned with the user's corpus in a RAG pipeline [88]. To address this, clients can locally fine-tune only the *retriever* component of the RAG pipeline using parameter-efficient methods such as LoRA [59] and adapter modules [58]. These techniques introduce small, trainable layers into pre-trained retrievers without modifying the backbone model. Because the update size remains minimal, retriever adaptations can be securely aggregated using FL protocols like FedAvg [111] or secure aggregation [18], enabling scalable personalization without compromising privacy or bandwidth efficiency.

In addition, this localized retriever tuning enhances the relevance of retrieved documents while keeping the global generator unchanged, thus preserving the generator's inference cost and memory footprint (with only a small overhead from the added PEFT modules), which are critical requirements for edge deployments [58, 59]. Recent in-domain RAG adaptation studies show that retriever-side adaptation can improve retrieval quality and downstream grounded generation, especially under domain shift [108]. Moreover, this strategy aligns with privacy-friendly design principles. Because only small adapter/LoRA parameters are communicated (instead of raw data), the update size is bandwidth-efficient. In practice, privacy can be strengthened by protecting these updates with secure aggregation protocols [18]. Appendix J illustrates an example where only the LoRA modules are updated, while all other components remain fixed. The modularity of this setup also allows asynchronous or partial retriever updates across clients [115], making it suitable for deployment in low-resource or heterogeneous federated environments. In summary, retriever adaptation via LoRA or adapter modules empowers clients to improve retrieval accuracy in a lightweight, private, and communication-efficient manner, enabling scalable deployment of personalized FFMs across diverse domains.

5.2.2 Server-side Optimization. While client-side personalization enables tailored responses and privacy-preserving adaptation, global coordination remains essential to ensure fairness, generalization, and robustness across diverse users. In FFMs, the server or regional edge tier is responsible not only for aggregating updates but also for enriching model capabilities through shared knowledge. Importantly, the role of retrieval here is *RAG-inspired* rather than standard inference-time RAG: the server uses retrieval mechanisms to *select*, *route*, or *weight* auxiliary data and training signals that improve coordination across heterogeneous clients, without accessing raw client content. By leveraging RAG at the server level, we can design intelligent orchestration strategies that optimize training dynamics and enhance sample efficiency without compromising user data. In other words, retrieval is used as a coordination primitive for *optimization* (e.g., task matching, clustering, distillation, curriculum reweighting), rather than solely as an evidence injector for generation.

Task Matching. In real-world deployments, clients often vary significantly in domain, task complexity, and data availability, ranging from richly annotated clinical corpora to sparse educational notes or niche financial records. To personalize global learning signals without accessing raw client data, *retrieval-inspired* task matching is a promising server-side coordination strategy that enhances contextual alignment during model aggregation. In this paradigm, the server maintains a broad, heterogeneous corpus of publicly available or pseudo-public data, such as open domain texts, instructional datasets, or synthetic knowledge bases. At each communication round, the server utilizes coarse routing signals (e.g., cluster IDs, group-level statistics, or aggregated metadata) to retrieve semantically related examples from this pool that reflect the dominant domain/task (or estimated difficulty) of a *group* of clients [48]. Unlike standard RAG, which retrieves evidence to condition *inference* [88], these retrieved examples can be used to augment the next round of training through server-side data-assisted objectives, such as distillation on public/unlabeled data [89, 96], curriculum-style reweighting based on estimated difficulty [151], or representation-shaping regularizers (e.g., model-level contrastive alignment).

This design introduces a form of implicit personalization by ensuring that each client receives globally aggregated updates that are contextually relevant to its local domain. For example, if a client cluster exhibits difficulty in a biomedical classification task, the server retrieves and integrates related medical samples from the public/pseudo-public pool to reinforce learning during the global update. This targeted augmentation can reduce global–local mismatch, improve optimization stability under non-IID data, and accelerate convergence when the server-side pool provides adequate coverage of newly emerging domains. Rather than claiming "fairness by default," this mechanism is best viewed as improving *relevance* and *sample efficiency*, and it can be paired with explicit fairness objectives when needed [94]. Importantly, since task matching operates over abstract representations (e.g., gradients or embeddings) rather than raw data, it reduces direct exposure but does not by itself guarantee privacy, as gradients can still leak information via inversion attacks [192]. Consequently, practical deployments often necessitate, stronger privacy requires combining routing logic with secure aggregation and/or client-level differential privacy [18]. Moreover, the mechanism is especially valuable in open-ended deployments where new clients with previously unseen domains continuously join the federation. In such cases, retrieval-based task matching serves as a lightweight yet effective personalization mechanism, offering adaptive, domain-aware support without requiring direct fine-tuning or explicit supervision.

Knowledge Sharing. In FL, personalization frequently conflicts with global generalization due to statistical heterogeneity across clients. Users may belong to vastly different application domains, such as clinical diagnostics [125], legal document processing, wireless environments [126] or educational tutoring, where a single global model often fails to generalize effectively. Cluster-wise knowledge sharing offers a middle ground by grouping clients into clusters with similar training objectives or data distributions, e.g., update-direction similarity as in FedGroup [36] or principal angles between client data subspaces as in PACFL [150]. In text-centric settings, corpus-level similarity (e.g., cosine distance over TF-IDF vectors or latent embeddings) is also a common heuristic when only privacy-safe descriptors are shared. In a clustered RAG-based design, each cluster maintains a domain-specific RAG sub-model that comprises a retriever and an adapter module. The retriever is fine-tuned to retrieve documents relevant to the cluster's specific domain, while the adapter provides lightweight personalization over a shared generator. This design allows intra-cluster model updates to remain focused and semantically aligned. Standard aggregation algorithms such as FedAvg or FedProx [93] are used within each cluster to synchronize the sub-models. When exchanging knowledge across clusters, one can avoid sharing raw documents by transferring higher-level signals (e.g., distilled outputs or adapter-level representations), and further strengthen privacy via secure aggregation or differential privacy. Depending on the system goal,

the generator may remain globally shared (optionally updated under the same protections), while retrievers/adapters remain cluster-specific.

In addition, the RAG framework is particularly valuable in this clustered setting because it decouples knowledge retrieval from language generation. While the generator remains globally shared across all clusters, each cluster-specific retriever ensures that input queries are grounded in domain-relevant knowledge. For instance, the retriever in a biomedical cluster retrieves clinical notes or research abstracts, while the one in a legal cluster fetches case law or statutes. This modularity enables domain-aligned personalization without incurring the computational cost or privacy risk of full-model fine-tuning. Additionally, RAG's retrieval-based conditioning improves factual grounding, and can mitigate hallucinations when retrieval quality is high, and enables continual adaptation as the client corpus evolves. Empirical studies support this hybrid strategy. For example, FedGroup [36] clusters clients based on update similarity and achieves notable gains in both accuracy and convergence, improving absolute test accuracy by +14.1% on FEMNIST over FedAvg [111], +3.4% on Sentiment140 over FedProx [93], and +6.9% on MNIST over FeSEM [101]. Meanwhile, PACFL [150] forms clusters based on principal angles between data subspaces to ensure distributional similarity. In a clustered RAG setting, these ideas can be interpreted as aligning both the client distribution (via clustering) and the contextual evidence used for generation (via cluster-specific retrieval). In summary, cluster-wise knowledge sharing with RAG enables federated systems to provide domain-specific personalization at scale. By tuning the retriever within each cluster while retaining a globally shared generator, it balances specialization and generalization. This architecture aligns well with the goals of PFI, where personalization, privacy, and communication efficiency must coexist in decentralized and heterogeneous environments.

Continual Pre-Training. Another core limitation of static FMs in FL is their gradual degradation in relevance, particularly in dynamic domains such as healthcare, finance, and education. A typical remedy is to continually pre-train FMs on domain-specific datasets to keep them up-to-date and aligned with evolving data distributions, as illustrated in Appendix K. However, traditional FL methods often lack mechanisms to integrate evolving knowledge across clients without incurring high communication overhead or violating data privacy. To address this, RAG offers a lightweight and privacy-preserving means of keeping the global generator up-to-date while enhancing personalization across domain-diverse clients. In RAG-assisted fine-tuning, the server periodically updates the shared generator on either: (a) pseudo-labeled data produced from its own retrieval-augmented pipeline via self-distillation [68], or (b) highly relevant documents dynamically retrieved from curated external sources (e.g., domain-specific databases, medical literature, or trusted web content) [54]. These documents are not randomly selected, but semantically aligned with the distribution of prior client tasks or updates, enabling the server to capture emergent knowledge relevant to clusters of users. This improves the generator's downstream alignment with personalized user contexts, even without direct exposure to raw user data.

The integration of RAG into continual pretraining pipelines can support three key objectives for federated personalization. First, it allows adaptive knowledge updates without relying on centralized annotations or fine-tuning per user, which is crucial in privacy-sensitive environments. Second, it enhances factual grounding by repeatedly training on retrieved hard negatives and diverse contexts, thereby reducing hallucination and domain mismatch. Third, by limiting updates to selectively retrieved, task-relevant content, the approach improves sample efficiency and ensures that model refinement reflects actual user needs rather than generic global trends. This server-side process indirectly fuels personalization at the client level. When a client in a niche domain (e.g., legal research or rare disease diagnosis) queries the generator, the model, fine-tuned on semantically similar retrievals, delivers more accurate, relevant, and context-aware outputs, despite never being fine-tuned on that specific user. Recent studies, such as Few-Shot RAG [68], have shown that this

retrieval-centric adaptation not only enhances factual precision but also boosts personalization fidelity in low-resource settings.

Although retrieval-augmented continual pretraining can improve freshness and grounding, it can also introduce *retrieval-induced drift* when the retrieval store evolves over time. This drift arises when imperfect, biased, stale, or adversarially influenced evidence is repeatedly injected into the context during inference and/or reused as training signal during continual updates, gradually amplifying spurious correlations and increasing hallucination and overconfident generations if unreliable or conflicting evidence is not detected and handled. Moreover, poisoning the retrieval corpus can induce *retrieval backdoors* that selectively steer downstream outputs while leaving the base model unchanged, highlighting the need for security-aware corpus maintenance [171]. To directly address these risks, robust *content curation* is essential for long-running RAG systems: practical mechanisms include provenance-aware indexing with source vetting/whitelisting, versioning and temporal validity checks for evolving documents, de-duplication and conflict detection to avoid repeatedly reinforcing erroneous claims, and continuous auditing/traceback to identify and quarantine suspicious or poisoned passages [182].

A widely adopted mitigation direction is to make *retrieval-quality awareness* explicit *before* generation and *before* any model update. Corrective RAG (CRAG) methods estimate retrieval quality and trigger corrective actions, such as query rewriting, additional retrieval, or selective context recomposition, when evidence is unreliable, as exemplified by CRAG and RQ-RAG [25, 172]. Complementarily, self-reflective RAG approaches encourage models to decide when retrieval is needed and to critique evidence and generations, improving faithfulness under noisy retrieval (e.g., SELF-RAG, SimRAG, and Self-BioRAG) [8, 71, 170]. In our framework, these safeguards naturally extend to continual pretraining by *gating* updates on high-confidence evidence and by excluding low-trust or conflicting contexts identified by the curation and quality-checking stages. We summarize the key RAG-enabled technologies for personalized federated intelligence in Appendix L.

Table 6 summarizes how major personalization mechanisms trade off communication overhead, personalization fidelity, and compatibility with heterogeneous clients. In practice, *prompt tuning* is most effective in large-scale cross-device deployments where uplink bandwidth and on-device compute are tightly constrained, and personalization mainly targets format/style/intent control. *LoRA* is preferable when clients exhibit stronger non-IID domain/task shifts and can afford lightweight local backpropagation, as it offers higher behavioral fidelity by updating internal representations while keeping payloads small. *RAG* is most effective when personalization is driven by user-owned or rapidly changing knowledge; local indexing and retrieval improve grounding and freshness, and can be combined with PEFT (e.g., LoRA or prompt tuning) when both knowledge and behavioral/style adaptation are required.

Note that, although significant efforts have been made to explore the core stages of the PFI pipeline, including efficient personalization, trustworthy adaptation, and adaptive refinement via RAG, we acknowledge that [186] also offers a complementary perspective on personalizing FMs through various approaches, such as prompting methods, representation learning, and RLHF. These techniques, developed from different perspectives, can also improve the design of personalized FL frameworks capable of accommodating individual user preferences.

5.3 Summary and Lessons Learned

This section has examined how RAG frameworks can support adaptive PFI by enabling lightweight, privacy-preserving, and domain-aligned personalization across client-side and server-side designs, without requiring full-model retraining or raw data exchange.

Table 6. Comparative analysis of key personalization mechanism in PFI.

Tech.	What is updated	Comm. overhead	Fidelity	Hetero. compat.	Most effective scenario
LoRA	Trainable low-rank adapter matrices inserted into selected projections (e.g., attention/MLP); base weights frozen	**Low** (communicate only adapter deltas; size scales with rank and chosen layers)	**High** (modifies internal representations; strong for domain/task shift vs pure prompting)	**Medium** (naive averaging assumes identical adapter shapes; hetero ranks/layers need padding, clustering, or separate heads)	Non-IID clients where you need *behavioral/domain* adaptation with limited uplink; devices can afford light backprop on a few layers
Prompt	Trainable soft prompts / prefix vectors (optionally layer-wise prefix tuning); base model frozen	**Very low** (only a small set of prompt vectors; minimal round payload)	**Medium** (excellent for format/style/task cues; limited capacity for large distribution shifts)	**High** (tiny memory/compute footprint; robust to device diversity and partial participation)	Massive cross-device settings with tight uplink/compute where quick personalization is needed (e.g., style, tone, and intent) without deep model changes
RAG	User/private corpus → embeddings/index; retrieval pipeline; generator often shared/frozen (retriever may be updated periodically)	**Low–Medium** (low if indexing stays local; higher if sharing/training retriever or syncing embeddings)	**High for knowledge** (grounding, freshness, and source control), **medium for style** unless paired with PEFT	**Medium–High** (depends on storage/retrieval support; can be local, edge-assisted, or clustered)	When personalization is mainly *user-owned or rapidly changing knowledge* and reducing hallucinations/keeping answers up-to-date matters; add PEFT if tone/behavior must also adapt

Takeaways: RAG enables effective personalization through both client-side adaptation (e.g., prompt conditioning and local retrieval) and server-side coordination (e.g., task matching and cluster-wise knowledge sharing). Its modular design, which decouples retrieval from generation, allows efficient and flexible adaptation in federated settings.

Connection to PFI: RAG naturally aligns with the core principles of PFI, including personalization, privacy preservation, and communication efficiency. By supporting per-user or per-cluster adaptation under the *model-to-data* paradigm, RAG provides a practical foundation for scalable and adaptive federated intelligence. Open challenges include balancing personalization and generalization across heterogeneous users, ensuring efficiency and fairness at scale, and designing privacy-preserving retrieval mechanisms suitable for resource-constrained devices. In addition, continual server-side adaptation must carefully manage retrieved content to avoid model drift and hallucination.

6 Future Directions

In the following section, we discuss promising future directions, including Meta-PFI, quantum-enabled PFI, and sustainable PFI.

6.1 Meta-PFI

Integrating autonomous capabilities into PFI opens new avenues for enhancing adaptability and intelligence. We term this emerging paradigm as meta-personalized federated intelligence (Meta-PFI). Meta-PFI operates within a dynamic user-agent-world interaction loop: users (human or avatar) provide real-time behavioral data and feedback; agents (embodied federated agents) autonomously adapt strategies based on user input or environmental perception; and the world (physical or virtual) shapes both data collection and learning dynamics.

To enable context-aware federated adaptation, virtual and mixed-reality environments (VR, AR, MR, XR), as well as digital twins, can serve as structured and immersive ecosystems for interaction and data generation [5]. These environments facilitate the collection of fine-grained user feedback and the development of rich agent perception, allowing personalized insights from diverse users and contexts to be integrated into the global FM without sharing raw data. Specifically, in this paradigm, embodied federated agents form perception–action loops within virtual or physical

environments: they locally observe user interactions (e.g., gaze, gestures, and speech) and environmental dynamics through these immersive technologies, fine-tune their local FMs for personalized reasoning, and periodically transmit model updates instead of raw data to a central aggregator. This mechanism enables coordinated and privacy-preserving learning across heterogeneous users while maintaining personalization and adaptability. Recent advances in embodied AI research provide concrete empirical grounding for this direction. For example, Habitat 3.0 [120] provides a large-scale embodied simulation environment that enables realistic perception–action loops for training and evaluating embodied agents. RoboCat [20] further demonstrates how foundation models can be adapted into robotic and embodied systems through multi-task and multi-embodiment learning, achieving rapid adaptation to novel tasks using only a small number of target examples. Overall, these systems exemplify how embodied agents can continuously learn from multimodal sensory data, which, to some extent, aligns with Meta-PFI's vision of using federated adaptation pipelines to connect user-specific interactions with FM refinement.

6.2 Quantum-Enabled PFI

Quantum-enabled PFI offers a promising alternative by leveraging quantum parallelism, cryptography, and machine learning to enhance optimization, security, and efficiency. Quantum-secure hardware, such as quantum mask circuits [4], can serve as a powerful hardware component in the context of PFI. By integrating these circuits into edge devices within PFI, sensitive elements-like user data, model parameters, and personalization layers-are protected against side-channel attacks, including power analysis, electromagnetic emission, and timing attacks, which are especially threatening in distributed, heterogeneous environments. Overall, unlike traditional algorithmic defenses (e.g., differential privacy or secure aggregation), quantum-enabled circuits [78] offer hardware-level protection that complements software-based methods. This hybrid approach ensures end-to-end robustness, enhancing both the privacy and trustworthiness of PFI during training and inference.

To address vulnerabilities associated with centralized aggregation in PFI, it is crucial to rethink server and client selection strategies in a quantum-aware context [109]. A promising direction involves the development of a quantum-probabilistic framework [31] that leverages quantum randomness and quantum interference to enable secure, unpredictable, and highly dynamic aggregation processes. In this approach, quantum circuits can generate secure random selection vectors to enable decentralized, entropy-driven aggregation. This may mitigate server collusion and biased sampling while enhancing resilience and eliminating single points of failure in traditional centralized designs. However, the limitation of quantum hardware in terms of qubit count or low gate fidelity prevents the direction from large-scale model aggregation. To overcome this challenge, hybrid quantum-classical machine learning approaches have been proposed in [30], where a classical pre-trained model is used to map the input image to a much lower dimension before it is fed into the quantum circuit for classification. Similarly, authors in [32] also proposed the hybrid model to accommodate a larger number of clients within the network. On the other hand, the authors in [67] designed a 4-qubit quantum neural network and trained it using the FL framework, which still achieves a relatively high performance. Furthermore, the work conducted privacy-preserving aggregation algorithms and quantum gates for encryption to further protect client data. Another approach to resolve the hardware constraint is variational quantum algorithms, which have emerged as the leading strategy to obtain quantum advantage on Noisy Intermediate-Scale Quantum (NISQ) computers [24]. Another promising direction is the integration of quantum wireless networks [153], enabling ultra-secure, low-latency communication across distributed clients and servers. Leveraging technologies such as entanglement distribution and quantum repeaters can extend the distance of quantum communication. Additionally, integrating quantum wireless communication with edge

intelligence and context-aware networking would enhance PFI's robustness and efficiency, facilitating secure AI development and beyond ecosystems. Regardless of these promising advancements, it faces a couple of limitations before deploying in real-world applications, such as the fragility of communication qubits, the stochastic success probability of entanglement swapping in quantum networks, and also the lack of a fully functional physical quantum repeater [86]. Additionally, the security enhancement from cryptographic and quantum techniques can lead to an increase in computational resources and make it infeasible for resource-constrained edge devices, which motivates the shift in focus toward the hybrid post-quantum cryptography research [149].

6.3 Sustainable PFI

Sustainable PFI represents a promising paradigm at the intersection of personalized foundation models and environmental sustainability, emphasizing energy-aware model optimization, carbon-aware scheduling, and the utilization of renewable energy.

To ensure scalable and environmentally responsible deployment of PFI at the edge, energy-aware model optimization is essential, especially as PFI scales across heterogeneous devices with diverse computational and battery capacities [95]. For instance, [65] measures the energy consumption of quantized LLMs running on a low-resource edge device and demonstrates that even the inference of LLMs on edge hardware can incur substantial energy and latency costs, highlighting the real energy burden of on-device LLM deployment. Similarly, [168] shows that balancing GPU frequency and batch size can reduce the energy-delay product by 12.4%–29.9% on an embedded platform, underlining the need for energy-efficient configurations when running LLMs at the edge. Moreover, combining model optimization with carbon-aware scheduling can further reduce the carbon footprint. For instance, GreenScale [84] introduces a carbon-aware scheduling framework that accounts for the time and location-based carbon intensity of edge-cloud infrastructures, enabling optimized task placement for applications such as AR and VR, and reducing carbon emissions by up to 29.1%. In addition, renewable energy integration in sustainable PFI can involve leveraging solar or wind-powered edge devices to enhance green energy availability, allowing for the planning of training rounds in advance. For example, [6] examines renewable energy-powered edge clouds and demonstrates how workload allocation across distributed edge sites can reduce operating costs and carbon emissions while maintaining quality of service. These studies collectively highlight the importance of designing personalized models that are environmentally sustainable, motivating future research on adaptive lightweight architectures, carbon-aware scheduling, and renewable energy utilization.

7 Conclusion

This survey envisions artificial personalized intelligence (API) as a natural complement to artificial general intelligence (AGI), emphasizing intelligence that is inherently user-centric, context-aware, and continuously adaptive. Within this landscape, personalized federated intelligence (PFI) emerges as a foundational paradigm, providing a principled framework for delivering large model capabilities to edge environments while preserving personalization and privacy. By revisiting the foundations of FL and FMs, we have proposed unified taxonomies that illuminate their structural components and the opportunities created by their convergence. Through the three stages of the PFI pipeline, we have synthesized: (1) efficiency-driven methods that reduce computation and communication cost through model compression and parameter-efficient fine-tuning, (2) trust-oriented mechanisms covering bias mitigation, fairness, interpretability, hallucination reduction, and robustness, and (3) adaptive personalization techniques that leverage user heterogeneity and contextual signals through lightweight, retrieval-driven, and modular adaptation mechanisms. Together, these insights clarify how PFI can serve as a compelling pathway towards the emerging era of API. Looking

ahead, we outline several promising directions, including Meta-PFI, quantum-enabled learning, and sustainable intelligence. Beyond the scope of PFI, we believe that other research communities will likewise play a pivotal role in shaping the evolution of API, which opens up intriguing opportunities for future research.

References

[1] Josh Achiam, Steven Adler, et al. 2023. Gpt-4 technical report. *ArXiv:2303.08774* (2023).

[2] Parker Addison et al. 2024. C-fedrag: A confidential federated retrieval-augmented generation system. *ArXiv:2412.13163* (2024).

[3] Shivam Aggarwal et al. 2023. Chameleon: Dual memory replay for online continual learning on edge devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 43, 6 (2023), 1663–1676.

[4] Noor Ul Ain et al. 2025. Secure Quantum-based Adder Design for Protecting Machine Learning Systems Against Side-Channel Attacks. *Applied Soft Computing* 169 (2025), 112554.

[5] Moayad Aloqaily et al. 2022. Integrating digital twin and advanced intelligent technologies to realize the metaverse. *IEEE Consumer Electronics Magazine* 12, 6 (2022), 47–55.

[6] Duong Thuy Anh Nguyen et al. 2024. Optimal Workload Allocation for Distributed Edge Clouds with Renewable Energy and Battery Storage. In *Proc. International Conference on Computing, Networking and Communications*. 700–705.

[7] Manoj Ghuhan Arivazhagan et al. 2019. Federated learning with personalization layers. *ArXiv:1912.00818* (2019).

[8] Akari Asai et al. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

[9] Muhammad Awais et al. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 4 (2025), 2245–2264.

[10] Stephen Bach et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *Association for Computational Linguistics: System Demonstrations*.

[11] Guangji Bai et al. 2025. Fedspallm: Federated pruning of large language models. In *Proc. Conf. Nations Americas Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*

[12] Yuntao Bai, Saurav Kadavath, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *ArXiv:2212.08073* (2022).

[13] Xiaohuan Bi and Xi Li. 2025. Securing Federated Learning Against Backdoor Threats with Foundation Model Integration. In *IEEE International Conference on Information Reuse and Integration and Data Science*.

[14] Jieming Bian, Yuanzhe Peng, Lei Wang, Yin Huang, and Jie Xu. 2025. A Survey on Parameter-Efficient Fine-Tuning for Foundation Models in Federated Learning. *ArXiv:2504.21099* (2025).

[15] Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. 2025. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *IEEE/CVF International Conference on Computer Vision*.

[16] Tolga Bolukbasi, Kai-Wei Chang, et al. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

[17] Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *ArXiv:2108.07258* (2021).

[18] Keith Bonawitz et al. 2017. Practical secure aggregation for privacy-preserving machine learning. In *ACM SIGSAC Conference on Computer and Communications Security*.

[19] Sebastian Borgeaud et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proc. ICML*.

[20] Konstantinos Bousmalis, Giulia Vezzani, et al. 2023. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation. *Transactions on Machine Learning Research* (2023).

[21] Tom Brown et al. 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* (2020).

[22] Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. 2023. Efficient federated learning for modern nlp. In *Annual International Conference on Mobile Computing and Networking*.

[23] Rich Caruana et al. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[24] Marco Cerezo et al. 2021. Variational quantum algorithms. *Nature Reviews Physics* 3, 9 (2021), 625–644.

[25] Chi-Min Chan et al. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *First Conference on Language Modeling (COLM)* (2024).

[26] Harsh Chaudhari et al. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *ArXiv:2405.20485* (2024).

[27] Xin Che, Jingdi Hu, Zirui Zhou, Yong Zhang, and Lingyang Chu. 2024. Training Fair Models in Federated Learning without Data Privacy Infringement. In *IEEE International Conference on Big Data*.

[28] Chen Chen, Jie Fu, and Lingjuan Lyu. 2023. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325* (2023).

[29] Haokun Chen et al. 2024. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Association for the Advancement of Artificial Intelligence*.

[30] Samuel Yen-Chi Chen and Shinjae Yoo. 2021. Federated quantum machine learning. *Entropy* 23, 4 (2021), 460.

[31] S Chowdhury et al. 2019. A probabilistic approach to quantum inspired algorithms. In *IEEE Int. Electron Devices Meet.*

[32] Hevish Cowlessur, Chandra Thapa, Tansu Alpcan, and Seyit Camtepe. 2025. A hybrid quantum neural network for split learning. *Quantum Machine Intelligence* 7, 2 (2025), 76.

[33] Damai Dai, Chengqi Deng, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *Association for Computational Linguistics* (2024).

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[35] Nan Du et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*.

[36] Moming Duan, Duo Liu, et al. 2021. FedGroup: Efficient clustered federated learning via decomposed data-driven measure. In *Proc. IEEE Intl Conf on Parallel & Distributed Processing with Applications,*.

[37] EDUCBA. 2025. Retrieval Augmented Generation (RAG). https://www.educba.com/retrieval-augmented-generation/.

[38] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*.

[39] Kai Fan, Jingtao Hong, Wenjie Li, Xingwen Zhao, Hui Li, and Yintang Yang. 2023. FLSG: A novel defense strategy against inference attacks in vertical federated learning. *IEEE Internet Things Journal* 11, 2 (2023), 1816–1826.

[40] Tao Fan et al. 2025. Fedmkt: Federated mutual knowledge transfer for large and small language models. In *COLING*.

[41] Tao Fan, Hanlin Gu, Xuemei Cao, et al. 2025. Ten challenging problems in federated foundation models. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[42] Tao Fan, Yan Kang, et al. 2023. Fate-llm: A industrial grade federated learning framework for large language models. *ArXiv:2310.10049* (2023).

[43] Tao Fan, Guoqiang Ma, Yuanfeng Song, Lixin Fan, et al. 2025. PPC-GPT: federated task-specific compression of large language models via pruning and chain-of-thought distillation. In *Empirical Methods in Natural Language Processing*.

[44] Minghong Fang et al. 2024. Byzantine-robust decentralized federated learning. In *ACM SIGSAC Conference on Computer and Communications Security*.

[45] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. International Conference on Machine Learning*.

[46] Jianhai Fu et al. 2024. Lite-sam is actually what you need for segment everything. In *Proc. Eur. Conf. Comput. Vis.*

[47] Zhenxiang Gao et al. 2024. Examining the potential of ChatGPT on biomedical information retrieval: fact-checking drug-disease associations. *Annals of biomedical engineering* 52, 8 (2024), 1919–1927.

[48] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. 2020. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*.

[49] Xueluan Gong, Yanjiao Chen, Qian Wang, and Weihan Kong. 2023. Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions. *IEEE Wireless Communications* 30, 2 (2023), 114–121.

[50] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. 2024. Federated learning for medical image analysis: A survey. *Pattern recognition* 151 (2024), 110424.

[51] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* 645, 8081 (2025), 633–638.

[52] Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. 2025. Selective Aggregation for Low-Rank Adaptation in Federated Learning. In *International Conference on Learning Representations*.

[53] Tao Guo, Song Guo, and Junxiao Wang. 2023. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *ACM Web Conference*.

[54] Kelvin Guu et al. 2020. Retrieval augmented language model pre-training. In *Proc. Int. Conf. Mach. Learn.*

[55] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Transactions on Machine Learning Research* (2024).

[56] Kaiming He et al. 2022. Masked autoencoders are scalable vision learners. In *IEEE/CVF Comput. Vis. Pattern Recognit.*

[57] Xingxin He, Yifan Hu, Zhaoye Zhou, Mohamed Jarraya, and Fang Liu. 2025. Few-Shot Adaptation of Training-Free Foundation Model for 3D Medical Image Segmentation. *ArXiv:2501.09138* (2025).

[58] Neil Houlsby et al. 2019. Parameter-efficient transfer learning for NLP. In *ICML*. 2790–2799.

[59] Edward J Hu et al. 2022. Lora: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.* (2022).

[60] Hongsheng Hu et al. 2022. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.* 54, 11s (2022), 1–37.

[61] Jiahui Hu, Dan Wang, Zhibo Wang, Xiaoyi Pang, Huiyu Xu, Ju Ren, and Kui Ren. 2024. Federated large language model: Solutions, challenges and future directions. *IEEE Wireless Communications* (2024).

[62] Zhiwei Hu, Liang Zhang, Shan Dai, Shihua Gong, and Qingjiang Shi. 2025. FedQLoRA: Federated Quantization-Aware LoRA for Large Language Models. *Open Review* (2025).

[63] Lei Huang et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2025).
[64] Wenke Huang, Mang Ye, Zekun Shi, Bo Du, and Dacheng Tao. 2024. Fisher calibration for backdoor-robust heterogeneous federated learning. In *European Conference on Computer Vision*.
[65] Erik Johannes Husom et al. 2025. Sustainable llm inference for edge ai: Evaluating quantized llms for energy efficiency, output accuracy, and inference latency. *ACM Transactions on Internet of Things* (2025).
[66] Alex Iacob et al. 2024. Worldwide Federated Training of Language Models. In *International Workshop on Federated Foundation Models in Conjunction with NeurIPS 2024*.
[67] Nouhaila Innan, Muhammad Al-Zafar Khan, Alberto Marchisio, Muhammad Shafique, and Mohamed Bennai. 2024. FedQNN: Federated Learning using Quantum Neural Networks. In *International Joint Conference on Neural Networks*.
[68] Gautier Izacard et al. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.
[69] Sami Jaghouar, Jack Min Ong, et al. 2024. INTELLECT-1 Technical Report. *ArXiv:2412.01152* (2024).
[70] Nidhal Jegham, Marwan Abdelatti, and Abdeltawab Hendawi. 2025. Visual Reasoning Evaluation of Grok, Deepseek Janus, Gemini, Qwen, Mistral, and ChatGPT. *ArXiv:2502.16428* (2025).
[71] Minbyul Jeong, Jiwoong Sohn, et al. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* 40, 1 (2024), 119–129.
[72] Yuheng Ji, Yue Liu, et al. 2025. Enhancing Adversarial Robustness of Vision-Language Models through Low-Rank Adaptation *(ICMR '25)*. Association for Computing Machinery, New York, NY, USA, 550–559.
[73] Feibo Jiang, Cunhua Pan, Li Dong, Kezhi Wang, Merouane Debbah, et al. 2025. A Comprehensive Survey of Large AI Models for Future Communications: Foundations, Applications and Challenges. *ArXiv:2505.03556* (2025).
[74] Xiaopeng Jiang and Cristian Borcea. 2023. Complement sparsification: Low-overhead model pruning for federated learning. In *AAAI Conference on Artificial Intelligence*.
[75] Zhida Jiang et al. 2022. Fedmp: Federated learning through adaptive model pruning in heterogeneous edge computing. In *International Conference on Data Engineering*.
[76] Zhu JianHao, Changze Lv, Xiaohua Wang, Muling Wu, Wenhao Liu, Tianlong Li, et al. 2024. Promoting Data and Model Privacy in Federated Learning through Quantized LoRA. In *Association for Computational Linguistics*.
[77] Ce Ju et al. 2020. Privacy-preserving technology to help millions of people: Federated prediction model for stroke prevention. *International Joint Conferences on Artificial Intelligence Workshop* (2020).
[78] Keyi Ju, Hui Zhong, Xinyue Zhang, Xiaoqi Qin, and Miao Pan. 2024. Quantum Computing Catalyzes Future Quantum Networks: Efficiency and Inherent Privacy. *IEEE Network* (2024).
[79] Peter Kairouz et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.
[80] Yan Kang et al. 2025. Grounding foundation models through federated transfer learning: A general framework. *ACM Transactions on Intelligent Systems and Technology* 16, 4 (2025), 1–54.
[81] Vladimir Karpukhin et al. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Empirical Methods in Natural Language Processing*.
[82] Davinder Kaur et al. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Computing Survey* 55, 2 (2022).
[83] Salman Khan et al. 2022. Transformers in vision: A survey. *ACM Computing Survey* 54, 10s (2022).
[84] Young Geun Kim et al. 2023. Greenscale: Carbon-aware systems for edge computing. *ArXiv:2304.00404* (2023).
[85] Alexander Kirillov et al. 2023. Segment anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[86] Vinay Kumar, Claudio Cicconetti, Marco Conti, and Andrea Passarella. 2025. Quantum internet: Technologies, protocols, and research challenges. *International Journal of Networked and Distributed Computing* 13, 2 (2025), 22.
[87] Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, and Choong Seon Hong. 2025. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion* (2025).
[88] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Adv. Neural Inf. Process. Syst.*
[89] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. In *Workshop on Federated Learning for Data Privacy and Confidentiality in NeurIPS*.
[90] Hang Li et al. 2024. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[91] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. 2024. Global and local prompts cooperation via optimal transport for federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[92] Shenghui Li et al. 2024. Synergizing foundation models and federated learning: A survey. *ArXiv:2406.12844* (2024).
[93] Tian Li et al. 2020. Federated optimization in heterogeneous networks. In *Machine Learning and Systems*.
[94] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair resource allocation in federated learning. In *International Conference on Learning Representations*.

[95] Yixuan Li et al. 2022. Energy-aware edge association for cluster-based personalized federated learning. *IEEE Transactions on Vehicular Technology* 71, 6 (2022), 6756–6761.

[96] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*.

[97] Zheng Lin, Guanqiao Qu, Qiyuan Chen, Xianhao Chen, Zhe Chen, and Kaibin Huang. 2025. Pushing Large Language Models to the 6G Edge: Vision, Challenges, and Opportunities. *IEEE Communications Magazine* 63, 9 (2025), 52–59.

[98] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv:2405.04434* (2024).

[99] Yang Liu et al. 2024. Vertical federated learning: Concepts, advances, and challenges. *IEEE Trans. Knowl. Data Eng.* (2024).

[100] Yang Liu, Tao Fan, Tianjian Chen, Qian Xu, and Qiang Yang. 2021. Fate: An industrial grade platform for collaborative learning with data protection. *Journal of Machine Learning Research* 22, 226 (2021), 1–6.

[101] Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. 2023. Multi-center federated learning: clients clustering for better personalization. *World Wide Web* 26, 1 (2023), 481–500.

[102] Luis M Lopez-Ramos, Florian Leiser, Aditya Rastogi, Steven Hicks, Inga Strümke, Vince I Madai, et al. 2024. Interplay between Federated Learning and Explainable Artificial Intelligence: a Scoping Review. *ArXiv:2411.05874* (2024).

[103] Scott M Lundberg et al. 2017. A unified approach to interpreting model predictions. In *Adv. Neural Inf. Process. Syst.*

[104] Jun Luo, Chen Chen, and Shandong Wu. 2025. Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models. In *International Conference on Learning Representations*.

[105] Jun Ma, Yuting He, et al. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.

[106] Xinge Ma et al. 2023. Fedid: Federated interactive distillation for large-scale pretraining language models. In *Empirical Methods in Natural Language Processing*.

[107] Xinbei Ma et al. 2023. Query rewriting in retrieval-augmented large language models. In *Proc. Empir. Methods Nat. Lang. Process.*

[108] Kelong Mao et al. 2024. RAG-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Empirical Methods in Natural Language Processing, Miami, Florida, USA.*

[109] Aakar Mathur et al. 2025. When Federated Learning Meets Quantum Computing: Survey and Research Opportunities. *IEEE Commun. Surv. Tutor.* (2025).

[110] Joshua Maynez et al. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Association for Computational Linguistics*.

[111] Brendan McMahan et al. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*.

[112] Md Jueal Mia and M Hadi Amini. 2025. FedShield-LLM: A Secure and Scalable Federated Fine-Tuned Large Language Model. *ArXiv:2506.05640* (2025).

[113] Bonan Min et al. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Survey* (2023).

[114] Fan Mo et al. 2021. PPFL: Privacy-preserving federated learning with trusted execution environments. In *Annual International Conference on Mobile Systems, Applications and Services*.

[115] John Nguyen et al. 2022. Federated learning with buffered asynchronous aggregation. In *International Conference on Artificial Intelligence and Statistics*.

[116] Minh N.H. Nguyen et al. 2021. Distributed and Democratized Learning: Philosophy and Research Challenges. *IEEE Computational Intelligence Magazine* 16, 1 (2021), 49–62.

[117] Khaoula Otmani, Rachid El-Azouzi, and Vincent Labatut. 2024. FedSV: Byzantine-Robust Federated Learning via Shapley Value. In *IEEE International Conference on Communications*.

[118] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* (2022).

[119] Yahao Pang et al. 2025. FedEAT: A Robustness Optimization Framework for Federated LLMs. *ArXiv:2502.11863* (2025).

[120] Xavier Puig et al. 2024. Habitat 3.0: A Co-Habitat for Humans, Avatars, and Robots. In *International Conference on Learning Representations*.

[121] Yu Qiao et al. 2025. Logit Calibration and Feature Contrast for Robust Federated Learning on Non-IID Data. *IEEE Transactions on Network Science and Engineering* 12, 2 (2025), 636–652.

[122] Yu Qiao, Phuong-Nam Tran, Ji Su Yoon, Loc X Nguyen, Eui-Nam Huh, Dusit Niyato, et al. 2025. Deepseek-inspired exploration of RL-based LLMs and synergy with wireless networks: A survey. *ACM Computing Survey* (2025).

[123] Alec Radford et al. 2018. Improving language understanding by generative pre-training. *Preprint* (2018).

[124] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

[125] Avi Deb Raha et al. 2024. Attention to Monkeypox: An Interpretable Monkeypox Detection Technique Using Attention Mechanism. *IEEE Access* (2024).
[126] Avi Deb Raha et al. 2025. Security Risks in Vision-Based Beam Prediction: From Spatial Proxy Attacks to Feature Refinement. *IEEE Transactions on Wireless Communications* (2025).
[127] Nikhila Ravi et al. 2024. Sam 2: Segment anything in images and videos. *in Proc. IEEE CVPR* (2024).
[128] Chao Ren et al. 2025. Advances and open challenges in federated foundation models. *IEEE COMST* (2025).
[129] Marco Tulio Ribeiro et al. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *ACM International Conference on Knowledge Discovery and Data Mining*.
[130] Carlos Riquelme et al. 2021. Scaling vision with sparse mixture of experts. *Adv. Neural Inf. Process. Syst.* (2021).
[131] Fahad Sabah et al. 2025. FairDPFL-SCS: Fair Dynamic Personalized Federated Learning with strategic client selection for improved accuracy and fairness. *Information Fusion* 115 (2025).
[132] Alexandre Sablayrolles et al. 2020. Radioactive data: tracing through training. In *Proc. Int. Conf. Mach. Learn.*
[133] Pramit Saha et al. 2025. Fedpia–permuting and integrating adapters leveraging wasserstein barycenters for finetuning foundation models in multi-modal federated learning. *AAAI Conference on Artificial Intelligence* (2025).
[134] Hasin Us Sami et al. 2025. Gradient Inversion Attacks on Parameter-Efficient Fine-Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[135] Lorenzo Sani et al. 2025. Photon: Federated LLM Pre-Training. In *Machine Learning and Systems*.
[136] Steffen Schneider et al. 2019. wav2vec: Unsupervised pre-training for speech recognition. *ArXiv:1904.05862* (2019).
[137] Ramprasaath R Selvaraju et al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *IEEE International Conference on Computer Vision*.
[138] Ahmed Sharshar et al. 2025. Vision-Language Models for Edge Networks: A Comprehensive Survey. *IEEE Internet of Things Journal* 12, 16 (2025), 32701–32724.
[139] Xuan Shen et al. 2025. Numerical pruning for efficient autoregressive models. In *Proc. AAAI Conf. Artif. Intell.*
[140] Yunxiao Shi et al. 2024. Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems. *Frontiers in Artificial Intelligence and Applications* (2024).
[141] Taylor Shin et al. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *in Proc. of Conference on Empirical Methods in Natural Language Processing* (2020).
[142] Viswanath Sivakumar et al. 2024. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. *Advances in Neural Information Processing Systems* (2024).
[143] Jinhyun So et al. 2022. Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning. *Machine Learning Research* (2022).
[144] Chen Sun et al. 2019. Videobert: A joint model for video and language representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[145] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. 2024. Improving LoRA in Privacy-preserving Federated Learning. In *International Conference on Learning Representations*.
[146] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems* 34, 12 (2022), 9587–9603.
[147] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *AAAI/ACM Conference on AI, Ethics, and Society*.
[148] Gemini Team et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv:2312.11805* (2023).
[149] Togu Novriansyah Turnip et al. 2025. Towards 6G Authentication and Key Agreement Protocol: A Survey on Hybrid Post Quantum Cryptography. *IEEE Commun. Surv. Tutor.* (2025).
[150] Saeed Vahidian et al. 2023. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *AAAI Conference on Artificial Intelligence*.
[151] Saeed Vahidian et al. 2023. When do curricula work in federated learning?. In *IEEE/CVF Int. Conf. Comput. Vis.*
[152] Francois-Lavet Vincent et al. 2018. An introduction to deep reinforcement learning. *Found. Trends Mach. Learn.* 11, 3-4 (2018).
[153] Chonggang Wang and Akbar Rahman. 2022. Quantum-enabled 6G wireless networks: Opportunities and challenges. *IEEE Wireless Communications* 29, 1 (2022), 58–69.
[154] Jianyu Wang et al. 2021. A field guide to federated optimization. *ArXiv:2107.06917* (2021).
[155] Jiaqi Wang and Xi Li. 2025. Federated Foundation Models Raise New Concerns of Robustness, Privacy, and Fairness. *Trustworthy FMs Workshop In conjunction with ICCV* (2025).
[156] Kevin Wang, Junbo Li, Neel P Bhatt, Yihan Xi, Zhangyang Wang, et al. 2024. OnThePlanning Abilities of OpenAI's o1 Models: Feasibility, Optimality, and Generalizability. In *Language Gamification-NeurIPS Workshop*.
[157] Ruhan Wang et al. 2025. Federated In-Context Learning: Iterative Refinement for Improved Answer Quality. *International Conference on Machine Learning* (2025).
[158] Sibo Wang et al. 2024. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*.

[159] Xuezhi Wang et al. 2022. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.
[160] Xintong Wang et al. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Proc. Association for Computational Linguistics*.
[161] Xianda Wang, Yaqi Qiao, Duo Wu, Chenrui Wu, and Fangxin Wang. 2025. Cluster Based Heterogeneous Federated Foundation Model Adaptation and Fine-Tuning. In *AAAI Conference on Artificial Intelligence*.
[162] Ziyao Wang et al. 2024. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*.
[163] Kang Wei et al. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security* 15 (2020), 3454–3469.
[164] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics* 14, 2 (2023), 513–535.
[165] Herbert Woisetschläger, Alexander Erben, et al. 2024. A Survey on Efficient Federated Learning Methods for Foundation Model Training. In *International Joint Conferences on Artificial Intelligence*.
[166] Panlong Wu, Kangshuo Li, et al. 2024. Federated in-context llm agent learning. *ArXiv:2412.08054* (2024).
[167] Xiaodong Wu et al. 2024. Unveiling security, privacy, and ethical concerns of ChatGPT. *J. Inf. Intell.* 2, 2 (2024).
[168] Hao Xu et al. 2025. Camel: Energy-aware llm inference on resource-constrained devices. *ArXiv:2508.09173* (2025).
[169] Mengwei Xu et al. 2025. Resource-efficient algorithms and systems of foundation models: A survey. *Comput. Surveys* 57, 5 (2025), 1–39.
[170] Ran Xu et al. 2025. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. In *Association for Computational Linguistics*.
[171] Jiaqi Xue et al. 2024. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *ArXiv:2406.00083* (2024).
[172] Shi-Qi Yan et al. 2024. Corrective retrieval augmented generation. *ArXiv:2401.15884* (2024).
[173] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 1–19.
[174] Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. 2024. Dual-personalizing adapter for federated foundation models. *Advances in Neural Information Processing Systems* 37 (2024), 39409–39433.
[175] Shunyu Yao et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* (2023).
[176] Yazdinejad et al. 2024. A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Transactions on Information Forensics and Security* 19 (2024), 6693–6708.
[177] Mang Ye et al. 2025. Vertical federated learning for effectiveness, security, applicability: A survey. *Comput. Surveys* 57, 9 (2025), 1–32.
[178] Catherine Yeh et al. 2024. AttentionViz: A Global View of Transformer Attention. *IEEE Trans. Vis. Comput. Graph.* 30, 1 (2024), 262–272.
[179] Liping Yi et al. 2024. Fedpe: Adaptive model pruning-expanding for federated learning on mobile devices. *IEEE Transactions on Mobile Computing* (2024).
[180] Dongxiao Yu et al. 2025. Pruning-based Adaptive Federated Learning at the Edge. *IEEE Trans. Comput.* (2025).
[181] Sixing Yu et al. 2023. Federated foundation models: Privacy-preserving and collaborative learning for large models. In *Proc. Computational Linguistics, Language Resources and Evaluation*.
[182] Baolei Zhang et al. 2025. Traceback of poisoning attacks to retrieval-augmented generation. In *ACM on Web Conference*.
[183] Jianyi Zhang et al. 2024. Towards building the federatedgpt: Federated instruction tuning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
[184] Peiyuan Zhang et al. 2024. Tinyllama: An open-source small language model. *ArXiv:2401.02385* (2024).
[185] Pengyu Zhang et al. 2025. CE-FFT: Communication-Efficient Federated Fine-Tuning for Large Language Models via Quantization and In-Context Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
[186] Zhehao Zhang et al. 2024. Personalization of large language models: A survey. *Trans. Mach. Learn. Res.* (2024).
[187] Zhuoyang Zhang, Han Cai, and Song Han. 2024. Efficientvit-sam: Accelerated segment anything model without performance loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[188] Dongfang Zhao. 2024. Frag: Toward federated vector database management for collaborative and secure retrieval-augmented generation. *arXiv preprint arXiv:2410.13272* (2024).
[189] Haodong Zhao et al. 2023. FedPrompt: Communication-Efficient and Privacy-Preserving Prompt Tuning in Federated Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
[190] Xunyi Zhao. 2024. *Optimizing Memory Usage when Training Deep Neural Networks.* Ph. D. Dissertation.
[191] Ce Zhou et al. 2024. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics* (2024).
[192] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Adv. Neural Inf. Process. Syst.*

[193] Xunyu Zhu et al. 2024. A survey on model compression for large language models. *TACL* 12 (2024), 1556–1577.
[194] Weiming Zhuang, Chen Chen, and Lingjuan Lyu. 2023. When foundation model meets federated learning: Motivations, challenges, and future directions. *Workshop on FL in the Age of Foundation Models in Conjunction with NeurIPS* (2023).