A Survey of Personalized Federated Foundation Models for Privacy-Preserving Recommendation

Zhiwei Li[1], **Guodong Long**[1], **Chunxu Zhang**[2], **Honglei Zhang**[3], **Chengqi Zhang**[4], **Jing Jiang**[1]
[1] Australian Artificial Intelligence Institute, University of Technology Sydney
[2] College of Computer Science and Technology, Jilin University
[3] School of Computer Science and Technology, Beijing Jiaotong University
[4] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University
zhw.li@outlook.com, guodong.long@uts.edu.au

Abstract

Integrating Foundation Models (FMs) into recommendation systems is an emerging and promising research direction. However, centralized paradigms face growing pressure from privacy concerns and strict regulatory requirements. Federated learning offers a viable solution that enables collaborative model refinement while keeping raw user data on local devices or organizational silos. Yet, applying FMs in this setting creates a fundamental tension, where the system must balance the leverage of global knowledge with the necessity of capturing user personality. This survey provides a comprehensive overview of Personalized Federated Foundation Models for privacy-preserving recommendation, and reviews recent progress in this emerging field. We first analyze personalization techniques that function effectively under federated settings. Furthermore, we discuss the adaptation of foundation models to such federated architectures to balance generalization with user-specific needs for achieving privacy-preserving recommendation. In contrast to existing reviews, our work specifically emphasizes the architectural intersection of federation, personalization, and foundation models.

1 Introduction

Recommendation services constitute a fundamental component of the modern digital ecosystem, and rely on aggregating fine-grained interaction data to infer user preferences [Ko *et al.*, 2022]. However, this centralized paradigm faces severe systemic constraints with intensifying imperatives for data sovereignty and privacy. Regulatory frameworks like GDPR [Voigt and Von dem Bussche, 2017] enforce strict constraints on data usage. Simultaneously, industrial developments highlight the friction between global service delivery and local data governance. Notable instances, such as the scrutiny surrounding TikTok [Aharazi, 2024], underscore the principle of data sovereignty, which requires user information to remain physically stored within national jurisdictions. In addition, Australia's Social Media Minimum Age framework [Commonwealth of Australia, 2024] mandates rigorous controls over processing data from minors.

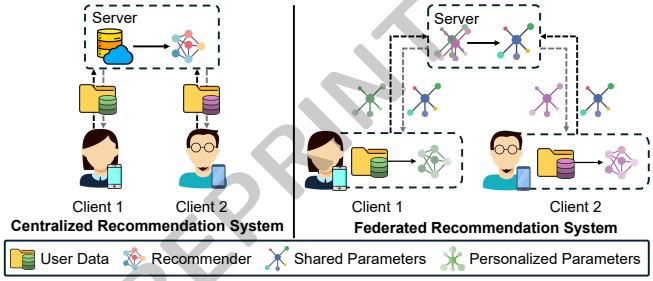


Figure 1: Comparison of Centralized RS (Left), where all user data is server collected, versus FRS (Right), where raw user data and local models remain on client devices, preserving data locality.

These pressures push the industry to seek alternatives that expand beyond their home markets into a global environment, decoupling service quality from centralized data location.

To mitigate these risks, Federated Learning (FL) [McMahan *et al.*, 2017] offers a distributed paradigm to establish a privacy-aware and regulation-compliant framework for recommendation at scale. Applying this paradigm to Recommendation Systems (RSs) enables data localization with a decentralized service architecture, leading to Federated Recommendation Systems (FRSs). The conceptual framework is illustrated in Fig. 1. Unlike centralized RSs which train the models on aggregated logs collected by the platform, FRSs force each client to train a local model using the user's private data, thereby maintaining data locality [Sun *et al.*, 2024b]. In FRSs, the central server functions solely as a coordinator to aggregate model updates, such as gradients, rather than collecting raw data. FRSs thus offer a viable pathway to navigate the tension between data utilization and regulatory compliance at scale.

While FRSs solve the issue of data locality for privacy, modern recommendations require deep semantic understanding. Foundation Models (FMs) demonstrate remarkable capabilities in discerning complex patterns and forging transferable representations from extensive datasets [Zhou *et al.*, 2025], which is particularly useful when client data are sparse or heterogeneous. FMs generalize well from sparse data and adapt efficiently to downstream tasks [Zhang *et al.*, 2024a], thereby acting as a robust semantic backbone of RSs. Integrating FMs into the federated infrastructure leads

to Federated Foundation Models (FFMs) [Yu *et al.*, 2023]. FFMs leverage pre-trained knowledge to enhance local learning within the decentralized framework, embodying a distinct approach to collaborative intelligence across heterogeneous clients [Ren *et al.*, 2025].

However, simply deploying FFMs is insufficient for recommendation tasks. Recommendations demand handling highly idiosyncratic user traits, not just generic semantic knowledge. A single global FFM cannot capture the diverse and distinct personalities of all users. This necessitates Personalized Federated Foundation Models (PFFMs) for privacy-preserving recommendations. PFFMs must architecturally decouple shared global knowledge from unique individual characteristics. Effective PFFMs must therefore actively *learn to preserve* user personality. This involves treating user traits as sensitive semantic structures to be isolated within the local adaptation process. This survey focuses on this critical intersection of personalization, federation, and foundation models for privacy-preserving recommendation.

To our knowledge, this is the first survey dedicated to Personalized Federated Foundation Models for recommendations, and makes the following contributions to the field:

- We formally define the scope of this emerging area. We establish a novel perspective centered on the active preservation of user personality semantics, rather than passive data protection for recommendation.
- We formulate a principled framework for achieving this preservation, and detail the necessary mechanisms to architecturally decouple generalized knowledge from individualized adaptations within federated systems.
- We conduct a principled analysis of fundamental challenges, and dissect theoretical conflicts between gradient-based utility optimization and geometric constraints necessary for semantic privacy boundaries.
- We outline critical future directions, and propose new paradigms to bridge current gaps in modeling stable user traits under strict privacy conditions.

2 Related Works

2.1 Differentiation from Federated Recommendation Surveys

Existing surveys on FRSs primarily synthesize methods based on collaborative filtering and neural networks [Yang *et al.*, 2020; Sun *et al.*, 2024b; Qi *et al.*, 2025; Zhang *et al.*, 2025b], and center on optimization strategies for communication efficiency and cryptographic protocols. However, these works predate the widespread adoption of FMs, and treat privacy primarily as a cryptographic problem of hiding gradients. Thus, they fail to address how pre-trained semantic priors fundamentally alter the personalization landscape of distributed recommendation.

Our survey examines the integration of extensive knowledge into federated architectures. We move beyond simple parameter aggregation and personalization. Our analysis focuses on how to leverage the representational capacity of FMs individually for each client within a federated framework. We specifically address the gap in using these models to preserve user personality for privacy-preserving recommendations, rather than just protecting raw data transmission.

2.2 Differentiation from Foundation Model Surveys

Conversely, recent reviews on FMs for recommendation focus predominantly on centralized paradigms [Liu *et al.*, 2023; Wu *et al.*, 2024]. These works categorize techniques such as prompt engineering and instruction tuning under the assumption of data aggregation on a central server. They analyze utility maximization but largely overlook the theoretical constraints of distributed deployment, and rarely discuss the architectural implications of keeping sensitive user traits strictly local while updating the massive global backbone of FMs.

Our work bridges this critical gap by providing the first structured taxonomy of Personalized Federated Foundation Models for privacy-preserving recommendations. We emphasize the necessary *architectural decoupling* required in this new era. Our survey uniquely highlights the tension between global semantic generalization and the rigorous isolation of idiosyncratic user personalities. We frame this not merely as a deployment issue, but as a fundamental shift towards active semantic preservation in recommendation systems.

3 Problem Settings

This section reviews recommendation systems and the preference–personality dichotomy, the role of foundation models, and the motivation for privacy-preserving federated foundation paradigms.

3.1 Preliminaries

Let $\mathcal{U}$ be the set of users and $\mathcal{I}$ be the set of items. A user $u \in \mathcal{U}$ has historical interaction data $\mathcal{D}_u$. The full dataset of observed interactions is $\mathcal{D} = \bigcup_{u\in\mathcal{U}} \mathcal{D}_u$, where each interaction is a tuple (u, i, r_{ui}). r_{ui} indicates feedback from user u for item i. For recommendation modeling, we use Θ to represent the learnable model parameters, comprising global parameters θ_g shared across users, and user-specific parameters θ_u. θ_u captures individual characteristics and is refined using $\mathcal{D}_u$. In the context of FFMs, ϕ_{FM} denotes the base FM parameters. To adapt such FMs for specific tasks and domains, we denote global adaptable components collaboratively fine-tuned with ϕ_g. Correspondingly, ϕ_u represents user-specific adaptable components refined locally for user u.

3.2 Personalization in Recommendation

Recommendation Systems

Recommendation Systems (RSs) aim to predict preferences $\hat{r}_{ui}$ and discover relevant items for users [Ko *et al.*, 2022]. A general RS objective learns model parameters Θ by minimizing a loss $\mathcal{L}$ over $\mathcal{D}$, often with regularization $\Omega(\Theta)$ [Zhang *et al.*, 2019]:

$$\min_{\Theta} \sum_{(u,i,r_{ui})\in\mathcal{D}} \mathcal{L}(r_{ui}, \hat{r}_{ui}(\theta_g, \theta_u)) + \Omega(\Theta). \tag{1}$$

Effective personalization hinges on accurately modeling user-specific characteristics captured by θ_u in Eq. (1) [He

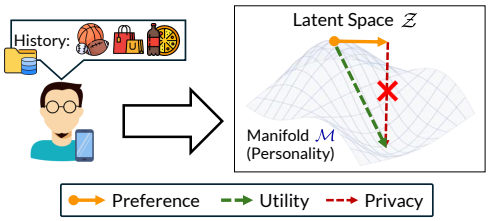


Figure 2: Interpretation of the dichotomy between preference and personality. User personality forms a stable **base manifold** $\mathcal{M}$ with the interaction history in the latent space $\mathcal{Z}$. Transient preferences appear as **vectors** (orange arrows) supported by this manifold.

et al., 2023]; however, centralized RSs aggregate all user data $\mathcal{D}_u$ for training, raising privacy concerns about the personality traits embedded in θ_u [Meng *et al.*, 2018].

Recommendation in Federated Settings

To mitigate such privacy risks, FL offers a distributed paradigm for collaborative refinement across users while keeping their raw data localized. **Federated Recommendation Systems** (FRSs) apply this to maintain data locality in recommendations [Sun *et al.*, 2024b], as shown in Fig. 1. In FRSs, each user $u \in \mathcal{U}$ contributes to training with their local data $\mathcal{D}_u$. The joint objective across clients in FRSs often learns with the shared θ_g and the user-specific θ_u:

$$\min_{\theta_g, \{\theta_u\}_{u\in\mathcal{U}}} \sum_{u\in\mathcal{U}} \frac{|\mathcal{D}_u|}{|\mathcal{D}|} \sum_{(u,i,r_{ui})\in\mathcal{D}_u} \mathcal{L}_u(r_{ui}, \hat{r}_{ui}(\theta_g, \theta_u)) + \Omega(\Theta), \tag{2}$$

where $\mathcal{L}_u$ is the local loss for user u. While FRSs satisfy baseline compliance requirements, the federated architecture introduces new theoretical challenges. First, client interaction data $\mathcal{D}_u$ is often sparse and statistically heterogeneous, hindering the training of robust models that can generalize across diverse user groups [Li *et al.*, 2024b]. Importantly, though the server does not have access to clients' raw data, model updates themselves may still leak sensitive information under certain attack settings [Chai *et al.*, 2020].

Dichotomy of Preference and Personality

Effective privacy preservation of FFMs necessitates distinguishing two semantic layers within the user representation θ_u for recommendations. We formalize this distinction through the geometry of the latent representation space $\mathcal{Z}$. **Preference** reflects transient interests conditioned on the context. Mathematically, it manifests as high-frequency variance in the user's interaction data $\mathcal{D}_u$. Conversely, **Personality** functions as the invariant prior that governs the stationary distribution of the user's long-term behaviors [Dhelim *et al.*, 2022]. We postulate that personality constitutes a low-dimensional **base manifold** $\mathcal{M}$ within the latent space $\mathcal{Z}$. As shown in Fig. 2, preference p_u functions as a directional vector learned as a perturbation originating from $\mathcal{M}$, capturing the direction of change relative to the user's long-term stable baseline.

This geometric hierarchy enables mathematical decoupling. The recommendation task in Eq. (2) primarily relies on the orientation of p_u relative to item embeddings, while

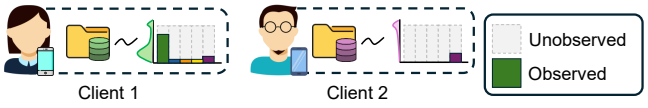


Figure 3: Core data challenges in FRSs. Each client observes only a small fraction of the items, leaving the majority unobserved locally. This inherent data sparsity, combined with non-IID distributions of user behavior across clients, creates significant heterogeneity that hampers both local personalization and global model aggregation.

the exact coordinates of the manifold $\mathcal{M}$ are not strictly necessary for item scoring. We thus designate the span of $\mathcal{M}$ as the *sensitive subspace* $\mathcal{S} \subseteq \mathcal{Z}$, and treat utility and personality reconstruction as orthogonal tasks: p_u is updated for accuracy while orthogonality to $\mathcal{S}$ blocks leakage of the underlying manifold structure to potential adversaries.

3.3 Foundation Models

Foundation Models (FMs) represent a paradigm shift in knowledge representation. FMs possess extensive pre-trained parameters ϕ_{FM} trained on diverse datasets, enabling them to discern complex patterns and generate potent representations [Bommasani *et al.*, 2021], while demonstrating robust generalization and adaptability [Zhang *et al.*, 2024a]. For recommendations, adaptation involves fine-tuning a parameter set ϕ of the model using downstream task-specific data $\mathcal{D}'$, which is defined as:

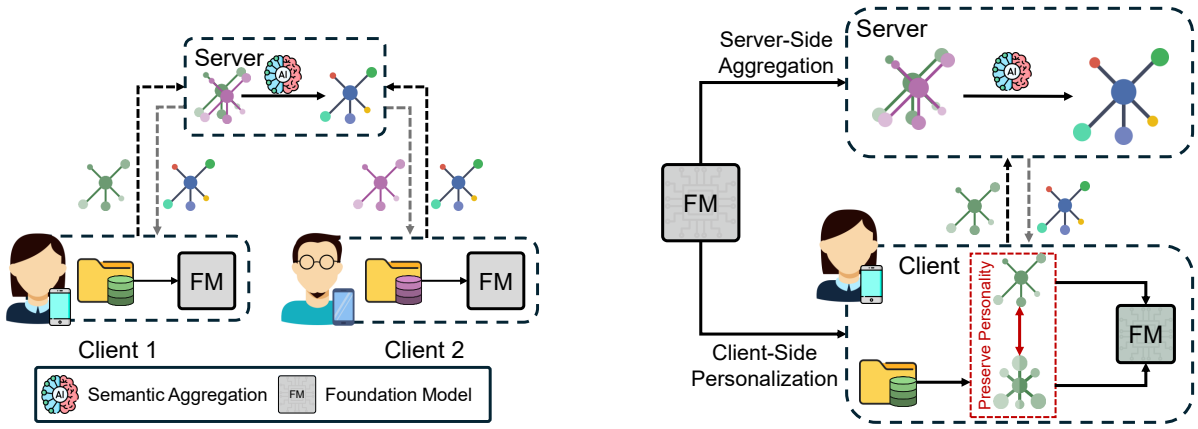
$$\min_{\phi} \mathcal{L}(f_{\text{FM}}(\phi_{\text{FM}}, \phi; \mathcal{D}')) + \Psi(\phi). \tag{3}$$

where f_{FM} is the FM, ϕ denotes the adaptable parameters, and $\Psi(\phi)$ is an optional regularizer. This adaptability transfers general knowledge from ϕ_{FM} to the recommendation context, enabling deep and semantically rich user representations even with extremely sparse user-item interactions [Bommasani *et al.*, 2021].

3.4 Federated Foundation Models

Despite the benefits of data locality and personalized parameters θ_u, FRSs still face hurdles in capturing and preserving personality. As shown in Fig. 3, local data $\mathcal{D}_u$ for each client in FRSs suffers from **Data Sparsity**, where the minimal density makes learning accurate θ_u difficult [Zhang *et al.*, 2024c]. Furthermore, **Statistical Heterogeneity** complicates deriving generalizable patterns [Qi *et al.*, 2023]. Critically, FRSs do not inherently *learn to preserve* personality characteristics from exposure during aggregation [Chai *et al.*, 2020]. These profound limitations highlight the urgent need for more advanced and privacy-preserving modeling paradigms tailored to recommendation.

FMs offer powerful representational capabilities to address this need [Han *et al.*, 2026], motivating our exploration of Federated Foundation Models (FFMs) for recommendations in Fig. 4a, in light of the FRS limitations in capturing user characteristics and protecting sensitive personality information. We aim to harness FMs for deep personality understanding within a federated structure that safeguards these inherent user traits, necessitating a paradigm shift towards systems that actively learn to preserve the sensitive aspects of each user's personality.



(a) Federated Foundation Models

(b) Personalized Federated Foundation Models

Figure 4: Evolution of the collaborative learning paradigm in recommendations. (a) The standard FFM approach aggregates local updates into a unified global model. It assimilates all learned patterns across clients without structural disentanglement, with no parameters dedicated to per-user personalization. (b) The proposed PFFM framework introduces active semantic defense. It utilizes the FMs as a shared semantic basis. Mechanisms such as constrained local optimization and semantic-aware aggregation architecturally decouple shared knowledge from individual personality. This ensures that the global model learns generalizable patterns while strictly isolating sensitive user personality.

4 Personalized Federated Foundation Models for Recommendation

To realize privacy-preserving recommendations, the field is witnessing a paradigm shift toward Personalized Federated Foundation Models (PFFMs), as shown in Fig. 4. This moves the focus from passive data protection to active semantic defense [Kang *et al.*, 2025]. While standard FL ensures data locality, it inherently fails to prevent the inference of high-level attributes from model updates [Ren *et al.*, 2025]. We argue that the core objective of PFFMs is to *learn to preserve personality* via active semantic disentanglement: leveraging FMs to infer personality from sparse interactions while orthogonalizing sensitive traits from shared information, yielding recommendations that are both personality-aware and privacy-respectful for every user.

4.1 Foundation Model as Semantic Basis

Unlike NLP inputs with explicit semantics, RSs process sparse behavioral signals consisting of non-semantic item identifiers, creating a fundamental semantic gap in the field.

As illustrated in Fig. 4b, the PFFM paradigm posits that the FM serves as a semantic basis distributed across the federated network, establishing a continuous latent space $\mathcal{Z}$ that maps discrete interaction histories for recommendations. Here $\mathcal{Z}$ does not merely encode item similarity but constitutes a complex behavioral manifold whose user coordinates encode intrinsic dispositions—stable personality traits that govern interaction probability distributions.

PFFMs leverage pre-trained base parameters ϕ_{FM} of FMs to serve as a high-dimensional semantic basis, providing extensive world knowledge [Ren *et al.*, 2025]. This allows the mapping of discrete interaction history $\mathcal{D}_u$ into a continuous latent space $\mathcal{Z}$. Within it, global adaptable components ϕ_g, collaboratively learned across users, align the generic semantic basis with the recommendation domain [Zhang *et al.*, 2025a]. In contrast, user-specific components ϕ_u function as low-rank perturbations applied to the global state, encoding the stable personality traits of user u which deviate from the global consensus [Zhang *et al.*, 2024d]. Consequently, this process transforms into a constrained disentanglement problem that seeks to minimize a joint objective:

$$\min_{\phi_g,\{\phi_u\}} \sum_{u\in\mathcal{U}} \frac{|\mathcal{D}_u|}{|\mathcal{D}|} \sum_{(u,i,r_{ui})\in\mathcal{D}_u} \mathcal{L}_u(r_{ui}, \hat{r}_{ui}(\phi_{\text{FM}}, \phi_g, \phi_u)). \tag{4}$$

The overall framework in Eq. (4) therefore orchestrates the federated optimization of global components and facilitates subsequent local refinement for personalization, with the challenge of maximizing utility while strictly bounding the semantic leakage inherent in ϕ_u.

4.2 Client-Side Personalization

Eq. (4) represents a dual-goal problem: representations must be maximally sufficient for utility yet minimally sufficient for personality inference, which we cast as a constrained optimization with orthogonality between shared parameters and the sensitive subspace:

$$\begin{aligned} \min_{\phi_g,\{\phi_u\}} \quad & \sum_{u\in\mathcal{U}} \frac{|\mathcal{D}_u|}{|\mathcal{D}|} \sum_{(u,i,r_{ui})\in\mathcal{D}_u} \mathcal{L}_u(r_{ui}, \hat{r}_{ui}(\phi_{\text{FM}}, \phi_g, \phi_u)), \\ \text{s.t.} \quad & \Psi_\perp(\phi_u, \phi_g) \le \epsilon, \forall u \in \mathcal{U}. \end{aligned} \tag{5}$$

Eq. (5) aggregates the loss over fine-grained interactions with a pivotal constraint: $\Psi_\perp(\phi_u, \phi_g)$ requires the joint representation formed by ϕ_g and ϕ_u to maintain a projection onto the sensitive personality subspace $\mathcal{S}$ below a threshold ϵ. By constraining local ϕ_u to be orthogonal to the sensitive directions of ϕ_g, PFFMs ensure that computed global updates remain disentangled from local user-specific profiles,

transforming personalization from a soft trade-off into a hard constraint-satisfaction problem.

4.3 Server-Side Aggregation

While local optimization in Eq. (5) mitigates leakage at the source, global robustness requires server-side intervention. Although the server cannot access ϕ_u, it receives global updates $\Delta\phi_g^{(u)}$ derived from local data [Li *et al.*, 2024b]. We thus redefine aggregation as semantic filtering rather than algebraic averaging, using the fixed ϕ_{FM} as a semantic discriminator to audit incoming gradients.

We propose a semantic-aware aggregation protocol: the server inspects the alignment of $\Delta\phi_g^{(u)}$, and an aggregation function g_{agg} projects these updates onto the known semantic basis of the FM, filtering out components highly correlated with sensitive attribute clusters. The update rule for ϕ_g is:

$$\phi_g^{t+1} \leftarrow \phi_g^t - \eta \cdot G_\perp \left(\{\Delta\phi_g^{(u),t+1}\}_{u\in\mathcal{U}}, \{\zeta_u^t\}_{u\in\mathcal{U}}; \phi_{\text{FM}} \right). \tag{6}$$

Here $\Delta\phi_g^{(u),t+1}$ is the global-parameter update from client u, and ζ_u^t denotes distributional statistics uploaded to assist semantic analysis. The operator $G_\perp$ uses ϕ_{FM} to analyze ζ_u^t, identifies sensitive semantic directions in the update batch, and projects out the corresponding components of $\Delta\phi_g^{(u)}$ before aggregation. This guarantees that ϕ_g assimilates generalizable patterns while rejecting personality fingerprints carried by gradients, jointly constituting—with Eq. (5)—our active-preservation paradigm.

4.4 Learning to Preserve Personality

The proposed framework establishes a fundamental distinction from standard FRSs, which lies in the controllability of the latent representation space. Traditional FRSs train user representations in the subspace $\mathcal{S}$ of personality traits from random initialization. Without a prior basis, the resulting latent space $\mathcal{Z}$ is isotropic and rotationally invariant [Maiorca *et al.*, 2023]. In such a manifold, the basis of $\mathcal{S}$ is unknown and time-varying, and the gradient update vector $\Delta\theta$ derived from local data represents a coupled signal. Consequently, the projection of the update onto $\mathcal{S}$ is undefined, and separating generic preferences from sensitive personality is mathematically intractable—structural isolation alone therefore fails to prevent this semantic leakage.

PFFMs advocate a paradigm shift from structural to semantic privacy boundaries, predicated on a fixed semantic reference frame uniquely provided by the pre-trained FFM. This prior breaks the rotational invariance and allows for the rigorous definition of $\mathcal{S}$ within $\mathcal{Z}$, so the privacy boundary in recommendations is reformulated as:

$$\langle \Delta\phi_g, \zeta_u \rangle \to 0, \quad \forall u \in \mathcal{U}. \tag{7}$$

Even if $\Delta\phi_g$ is intercepted, its projection onto the semantic directions of personality is theoretically nullified, shifting the privacy boundary from an architectural heuristic to a verifiable geometric constraint and preserving privacy both by where the data resides and by how the representation is constructed.

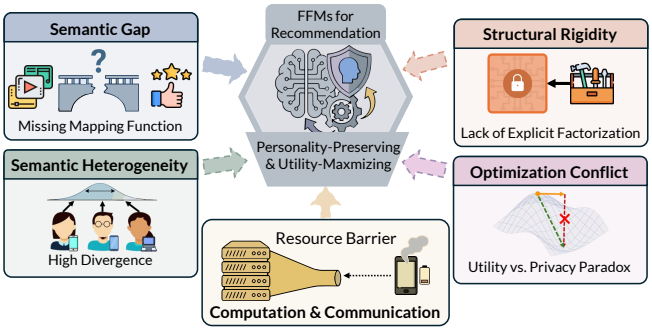


Figure 5: Impediments to realizing Personalized Federated Foundation Models for privacy-preserving Recommendations.

4.5 Taxonomy of Instantiating the Framework

To map existing research into the above framework, we categorize PFFMs based on the specific architectural realization of the decoupled parameters ϕ_u and ϕ_g [Ren *et al.*, 2025].

Adapter-based Disentanglement. In this category, ϕ_g constitutes the frozen backbone of the FM plus shared adapter modules [Zhang *et al.*, 2025a]. The user-specific ϕ_u is instantiated as local lightweight adapters injected into transformer layers [Sun *et al.*, 2024a], structurally enforcing $\Psi_\perp$ via separate matrices. Privacy is preserved by keeping local LoRA matrices on-device while only shared adapters or statistical prototypes are aggregated [Yang *et al.*, 2024], satisfying Eq. (7) by construction.

Prompt-based Projection. Here the FM ϕ_{FM} remains static. The personalization ϕ_u is modeled as a set of continuous learnable vectors prepended to the input embeddings [Liu *et al.*, 2023]. These prompts act as a geometric transformation function $f : \mathcal{X} \to \mathcal{Z}$. It projects user behavior into the semantic space of the FM [Li *et al.*, 2026; Su *et al.*, 2024]. Since prompts are detached from the model weights, they offer a natural boundary for the semantic isolation described in Eq. (7) [Luo *et al.*, 2025a].

Instruction-tuned Alignment. This stream focuses on semantic alignment of the global model ϕ_g through local instruction tuning on $\mathcal{D}_u$ [Zhuang *et al.*, 2023]. Here ϕ_u is the local-update offset diverging from the global gradient [Wang *et al.*, 2026]; since ϕ_u and ϕ_g share physical parameters, advanced aggregation in Eq. (6) is critical for filtering personality-sensitive gradients [Wang *et al.*, 2025a].

5 Open Problems

Realizing FFMs that learn to preserve personality involves overcoming fundamental contradictions in model optimization and data representation. We identify five core problems, illustrated in Fig. 5, that hinder current progress for recommendations.

5.1 Semantic Gap

A primary theoretical hurdle is the dissonance between observed data and the latent ground-truth personality p_u. In the FFM setting, ϕ_u is optimized to maximize the likelihood of

observed interactions $\mathcal{D}_u$ as a surrogate for recovering p_u:

$$\max_{\phi_u} P(\mathcal{D}_u \mid \phi_u; \phi_{\text{FM}}, \phi_g). \tag{8}$$

However, $\mathcal{D}_u$ is sparse and noisy, often reflecting transient interests rather than stable traits p_u. The core challenge is the absence of a verified mapping $f : \mathcal{Z} \to \mathcal{P}$ from the learned latent space $\mathcal{Z}$ to the psychological space $\mathcal{P}$. Without ground-truth labels for p_u, the model relies on spurious correlations [Dhelim *et al.*, 2022], yielding a ϕ_u where transient noise is indistinguishable from stable personality features, so maximizing Eq. (8) does not guarantee that ϕ_u accurately encodes p_u, creating a persistent semantic gap.

5.2 Structural Rigidity

Current architectures lack the structural inductive bias required for recommendations, especially in the federated context [Zhang *et al.*, 2024b]. Most FMs largely repurpose models pre-trained on centralized textual corpora to minimize sequence entropy, whereas recommendations rely on collaborative signals to minimize $\mathcal{L}_u$. Ideally, model parameters should be decomposable into a shared knowledge module and a private personality module:

$$\phi = \phi_g \oplus \phi_u, \quad \text{s.t. } \phi_g \perp \phi_u. \tag{9}$$

However, ϕ_{FM} is monolithic and does not admit the explicit factorization in Eq. (9) [Yang *et al.*, 2024]; forcefully adapting text-centric models to this decoupled structure incurs suboptimal performance and high computational costs, preventing physical isolation of sensitive personality from the shared global model [Melis *et al.*, 2019; Zhu *et al.*, 2019].

5.3 Semantic Heterogeneity

FFMs face a unique form of heterogeneity, which we term *Semantic Heterogeneity*, distinct from statistical non-IIDness in traditional FL [Lin *et al.*, 2022]. It arises from the varying degrees of alignment between the FM's pre-trained prior knowledge P_{FM} and the local behavioral distribution P_u of different users [Wang *et al.*, 2025a]. Some users align well with the FM's generalized knowledge, while others exhibit idiosyncratic patterns that contradict its priors. We formalize this as the variance of semantic divergence across $\mathcal{U}$:

$$\sigma^2 = \text{Var}_{u \in \mathcal{U}} \left(D_{KL}(P_u(\mathbf{z}) \| P_{\text{FM}}(\mathbf{z})) \right). \tag{10}$$

A high σ^2 implies that a single global ϕ_g cannot accommodate all users: forcing alignment with ϕ_g dilutes the personality of high-divergence users, while excessive local deviation ϕ_u causes catastrophic forgetting of the FM's general capabilities, creating a fundamental dilemma in selecting the appropriate aggregation weight for each user.

5.4 Optimization Conflict

FFMs face an inherent paradox between utility maximization and privacy constraints [Zhang *et al.*, 2023b]. As discussed in Sec. 4, preserving personality requires gradient updates orthogonal to the sensitive subspace $\mathcal{S}$, i.e., the local gradient $\nabla \mathcal{L}_u$ must satisfy:

$$\|\text{Proj}_{\mathcal{S}}(\nabla \mathcal{L}_u)\| = 0. \tag{11}$$

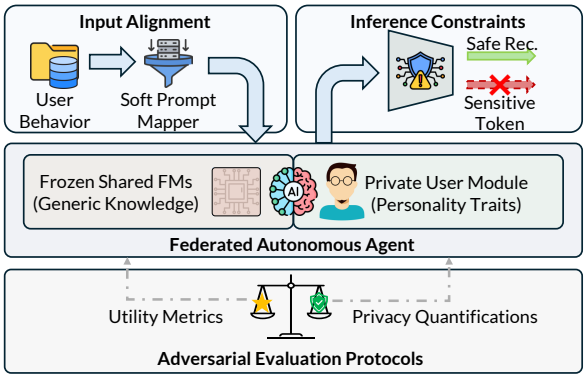


Figure 6: Framework for future directions in personality-preserving Federated Foundation Models for Recommendations.

Eq. (11) means the projection of $\nabla \mathcal{L}_u$ onto $\mathcal{S}$ is zero; however, utility maximization relies on leveraging personality traits p_u to predict behavior, so if $\nabla \mathcal{L}_u$ provides useful information for personalization, it implies a correlation:

$$|\text{Corr}(\nabla \mathcal{L}_u, p_u)| > 0. \tag{12}$$

Since p_u lies within $\mathcal{S}$, the two conditions are contradictory; standard gradient descent cannot satisfy both simultaneously without degrading utility, leading to an inevitable optimization conflict where the pursuit of accuracy inherently compromises the semantic orthogonality required for privacy.

5.5 Computational and Communication Overhead

Beyond theoretical paradoxes, deploying FFMs faces rigid physical constraints. Clients in federated settings typically possess limited computational capacity and battery life [Yang *et al.*, 2019]. However, FM inference and fine-tuning require substantial FLOPs and memory bandwidth [Sun *et al.*, 2024a]; even with Parameter-Efficient Fine-Tuning (PEFT), the forward pass still loads the massive ϕ_{FM} into memory [Yang *et al.*, 2024], creating a resource barrier. Communication bandwidth further restricts the frequency and dimension of $\Delta\phi_g$ transmitted between clients and server [Vu *et al.*, 2024], conflicting with the depth of semantic reasoning required for personality analysis; standard compression [Dantas *et al.*, 2024] often degrades the ability to capture subtle personality nuances during transmission.

6 Future Directions

The formal challenges in Sec. 5 suggest that incremental improvements are insufficient; we advocate new research paradigms in the federated context. As shown in Fig. 6, these directions explicitly address the mathematical contradictions above, aiming for personality-aware recommendations while actively *learning to preserve* the sensitive aspects of user personality.

6.1 Federated Autonomous Agents

The semantic gap between transient feedback and stable personality arises from the limitations of single-step prediction. **Federated Autonomous Agents** offer a viable solution by

reformulating the interaction paradigm [Deng *et al.*, 2025]. We advocate treating the FFM as a local agent with policy π, modeling user interactions as a Markov decision process [Lian *et al.*, 2024]. Rather than maximizing feedback likelihood, the agent optimizes a state-value function $V_\pi(s_t)$:

$$V_\pi(s_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k} \mid s_t, p_u)\right]. \quad (13)$$

Here, γ is a discount factor and s_t explicitly includes the latent personality variable p_u, while $R(\cdot)$ enforces consistency in the agent's behavior a_t over t, so the agent learns to encode stable personality traits as necessary components for coherent long-term planning, shifting from fitting noisy logs to learning a coherent policy $\pi(a|s_t, p_u)$. Recent agent-based recommenders, including multi-step planners and tool-using LLM agents [Lian *et al.*, 2024], demonstrate this trajectory; their federated extensions must additionally guarantee that p_u never crosses the local boundary, e.g., by replacing remote rollouts with locally simulated trajectories that disclose only utility statistics.

6.2 Parameter-Efficient Modular Design

To overcome the structural rigidity of monolithic models, research must adopt Parameter-Efficient Modular Design that physically decouples general knowledge from user-specific traits [Zhang *et al.*, 2024b]. We advocate treating personality parameters as independent lightweight modules such as adapters or hyper-networks [Yang *et al.*, 2024]: the massive ϕ_{FM} remains frozen and shared while only $\{\phi_u\}$ are updated locally. Promising instantiations include per-user low-rank LoRA factors, hyper-networks that generate ϕ_u conditioned on user embeddings, and modular routing schemes that isolate personality experts from the shared backbone, all of which keep the gradient dynamics of personality strictly separated from ϕ_{FM} and thus enable rigorous encryption or differential privacy on ϕ_u alone without degrading FM utility.

6.3 Personalized Prompt Learning

Semantic heterogeneity across users motivates **Personalized Prompt Learning**: instead of modifying weights, the system learns user-specific soft prompts that act as continuous mappings [Luo *et al.*, 2025a], projecting each user's behavioral manifold into the foundation model's semantic space [Li *et al.*, 2024a]. Designs such as mixture-of-prompts gating, hierarchical prompt prototypes shared across similar users, and meta-prompt initialization for fast local adaptation [Su *et al.*, 2024] each trade off expressivity, communication cost, and personality privacy differently. Minimizing the alignment error locally through prompts reduces the heterogeneity of uploaded representations and allows the global model to aggregate insights effectively while respecting the distinct semantic topology of each personality.

6.4 Inference-Time Privacy Constraints

The optimization conflict between utility and privacy is hard to resolve during training due to collinear gradients [Wang *et al.*, 2025b]. **Inference-Time Privacy Constraints** offer a pragmatic alternative by shifting the defense to the decoding stage: in generative recommendation, the system enforces privacy without altering parameters [Luo *et al.*, 2025b] by integrating constraints into beam search or sampling, suppressing tokens with high mutual information w.r.t. sensitive attributes. Concrete instantiations include logit reweighting against sensitive token clusters, constrained nucleus sampling that masks personality-revealing continuations, and gradient-free differential privacy injected at the decoder, so that outputs remain useful while statistically invariant to private personality traits. Compared with training-time defenses, this strategy decouples privacy budgets from optimization trajectories and is therefore robust to subsequent fine-tuning of the shared model.

6.5 Adversarial Evaluation Protocols

The lack of standardized metrics hinders the verification of personality-preserving FFMs, motivating rigorous **Adversarial Evaluation Protocols**. We advocate a dual-metric system: utility is measured by standard ranking metrics, while privacy is quantified via *Personality Inference Attacks* in which an adversarial model reconstructs the user's base-manifold coordinates from exposed model updates or outputs [Zhang *et al.*, 2023a; Chang *et al.*, 2024]. Privacy success is then scored by the inverse of attacker accuracy or the reduction in mutual information, which a robust system must minimize while maximizing recommendation utility. Building reproducible benchmarks further requires pairing standard recommendation corpora with personality labels obtained from Big-Five inventories or LLM-based persona elicitation, so that the trade-off frontier between utility and personality leakage can be tracked across methods. Robust evaluation must cover both gradient-side leakage during training and output-side inference at deployment, since these expose distinct facets of personality and warrant distinct defenses.

7 Conclusion

This survey provides a comprehensive analysis of Personalized Federated Foundation Models for privacy-preserving recommendation, examining how the architecture resolves the conflict between generalizable knowledge and personalized adaptation within federated settings. A key insight is the necessity of shifting from passive data protection to active semantic preservation: foundation models can infer nuanced personality traits, but decentralized protocols must structurally isolate these attributes from shared parameters to ensure privacy. Advancing this paradigm requires rigorous mechanisms that maintain semantic boundaries around user identity, ensuring trustworthy recommendation systems that protect both user privacy and personality. We hope the taxonomy of architectural realizations and the geometric privacy boundary articulated here serve as a foundation for next-generation federated recommendation systems that genuinely empower users with both relevance and self-determination over their digital identity, and that the open problems we identify catalyze cross-disciplinary work bridging federated optimization, foundation-model adaptation, and privacy-aware semantic geometry.

References

[Aharazi, 2024] Bshaer Kameil Aharazi. Contextualizing tiktok controversies: Critical discourse analysis of platform privacy debates. 2024.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Chai *et al.*, 2020] Di Chai, Leye Wang, Kai Chen, and Qiang Yang. Secure federated matrix factorization. *IS*, 36(5):11–20, 2020.

[Chang *et al.*, 2024] Hongyan Chang, Brandon Edwards, Anindya S Paul, and Reza Shokri. Efficient privacy auditing in federated learning. In *USENIX Security*, pages 307–323, 2024.

[Commonwealth of Australia, 2024] Commonwealth of Australia. Online safety amendment (social media minimum age) act 2024. Federal Register of Legislation, 2024. Act No. 127 of 2024; Title ID: C2024A00127; Effective: 10 Dec 2024; Registered: 12 Dec 2024.

[Dantas *et al.*, 2024] Pierre Vilar Dantas, Waldir Sabino Da Silva, Lucas Carvalho Cordeiro, and Celso Barbosa Carvalho. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 54(22):11804–11844, 2024.

[Deng *et al.*, 2025] Yongxin Deng, Xihe Qiu, Xiaoyu Tan, and Yaochu Jin. Fedslate: A federated deep reinforcement learning recommender system. *IEEE TETCI*, 2025.

[Dhelim *et al.*, 2022] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. A survey on personality-aware recommendation systems. *Artificial Intelligence Review*, pages 1–46, 2022.

[Han *et al.*, 2026] Renda Han, Xiaobao Wang, Luzhi Wang, Wenxin Zhang, Guangzhen Yao, and Hongxiang Liang. A graph foundation model for unified anomaly detection. In *WWW*, pages 523–534, 2026.

[He *et al.*, 2023] Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. A survey on user behavior modeling in recommender systems. *arXiv preprint arXiv:2302.11087*, 2023.

[Kang *et al.*, 2025] Yan Kang, Tao Fan, Hanlin Gu, Xiaojin Zhang, Lixin Fan, and Qiang Yang. Grounding foundation models through federated transfer learning: A general framework. *ACM TIST*, 16(4):1–54, 2025.

[Ko *et al.*, 2022] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[Li *et al.*, 2024a] Guanghao Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. Visual prompt based personalized federated learning. *TMLR*, 2024.

[Li *et al.*, 2024b] Zhiwei Li, Guodong Long, and Tianyi Zhou. Federated recommendation with additive personalization. In *ICLR*, 2024.

[Li *et al.*, 2026] Zhiwei Li, Guodong Long, Jing Jiang, Chengqi Zhang, and Qiang Yang. Federated vision-language-recommendation with personalized fusion. In *AAAI*, volume 40, pages 23337–23345, 2026.

[Lian *et al.*, 2024] Jianxun Lian, Yuxuan Lei, Xu Huang, Jing Yao, Wei Xu, and Xing Xie. Recai: Leveraging large language models for next-generation recommender systems. In *WWW*, pages 1031–1034, 2024.

[Lin *et al.*, 2022] Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. Fednlp: Benchmarking federated learning methods for natural language processing tasks. In *NAACL*, pages 157–175, 2022.

[Liu *et al.*, 2023] Peng Liu, Lemei Zhang, and Jon Atle Gulla. Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems. *TACL*, 11:1553–1571, 2023.

[Luo *et al.*, 2025a] Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *ICLR*, 2025.

[Luo *et al.*, 2025b] Xinjian Luo, Ting Yu, and Xiaokui Xiao. Prompt inference attack on distributed large language model inference frameworks. In *SIGSAC*, pages 1739–1753, 2025.

[Maiorca *et al.*, 2023] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *NeuRIPS*, 36:55394–55414, 2023.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282. PMLR, 2017.

[Melis *et al.*, 2019] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *IEEE SP*, pages 691–706. IEEE, 2019.

[Meng *et al.*, 2018] Xuying Meng, Suhang Wang, Kai Shu, Jundong Li, Bo Chen, Huan Liu, and Yujun Zhang. Personalized privacy-preserving social recommendation. In *AAAI*, volume 32, 2018.

[Qi *et al.*, 2023] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *ACM MM*, pages 3099–3107, 2023.

[Qi *et al.*, 2025] Xin Qi, Meixuan Li, Sijin Zhou, Wei Feng, and Zhuang Qi. Federated learning for science: A survey on the path to a trustworthy collaboration ecosystem. *Authorea Preprints*, 2025.

[Ren *et al.*, 2025] Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Bo Zhao, Liping Yi, Alysa Ziying Tan, Yulan Gao, Anran Li, Xiaoxiao Li, et al. Advances and open challenges in federated foundation models. *IEEE ComST*, 2025.

[Su *et al.*, 2024] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *AAAI*, volume 38, pages 15117–15125, 2024.

[Sun *et al.*, 2024a] Guangyu Sun, Umar Khalid, Matias Mendieta, Pu Wang, and Chen Chen. Exploring parameter-efficient fine-tuning to enable foundation models in federated learning. In *IEEE BigData*, pages 8015–8024. IEEE, 2024.

[Sun *et al.*, 2024b] Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, and Lizhen Cui. A survey on federated recommendation systems. *TNNLS*, 2024.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[Vu *et al.*, 2024] Minh Vu, Truc Nguyen, My T Thai, et al. Analysis of privacy leakage in federated large language models. In *AISTATS*, pages 1423–1431. PMLR, 2024.

[Wang *et al.*, 2025a] Huan Wang, Haoran Li, Huaming Chen, Jun Yan, Jiahua Shi, and Jun Shen. Fedsc: Federated learning with semantic-aware collaboration. In *ACM SIGKDD*, pages 2938–2949, 2025.

[Wang *et al.*, 2025b] Yubo Wang, Min Tang, Nuo Shen, Shujie Cui, and Weiqing Wang. Privacy risks of llm-empowered recommender systems: An inversion attack perspective. In *ACM RecSys*, pages 812–821, 2025.

[Wang *et al.*, 2026] Xiaobao Wang, Renda Han, Ronghao Fu, and Di Jin. Federated graph-level clustering network with dual knowledge separation. In *ICLR*, 2026.

[Wu *et al.*, 2024] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

[Yang *et al.*, 2019] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM TIST*, 10(2):1–19, 2019.

[Yang *et al.*, 2020] Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. *Federated Learning: Privacy and Incentive*, pages 225–239, 2020.

[Yang *et al.*, 2024] Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *NeuRIPS*, volume 37, pages 39409–39433. Curran Associates, Inc., 2024.

[Yu *et al.*, 2023] Sixing Yu, J Pablo Muñoz, and Ali Jannesari. Federated foundation models: Privacy-preserving and collaborative learning for large models. *arXiv preprint arXiv:2305.11414*, 2023.

[Zhang *et al.*, 2019] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM CSUR*, 52(1):1–38, 2019.

[Zhang *et al.*, 2023a] Shijie Zhang, Wei Yuan, and Hongzhi Yin. Comprehensive privacy analysis on federated recommender system against attribute inference attacks. *IEEE TKDE*, 36(3):987–999, 2023.

[Zhang *et al.*, 2023b] Xiaojin Zhang, Yan Kang, Kai Chen, Lixin Fan, and Qiang Yang. Trading off privacy, utility, and efficiency in federated learning. *ACM TIST*, 14(6):1–32, 2023.

[Zhang *et al.*, 2024a] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

[Zhang *et al.*, 2024b] Chunxu Zhang, Guodong Long, Hongkuan Guo, Xiao Fang, Yang Song, Zhaojie Liu, Guorui Zhou, Zijian Zhang, Yang Liu, and Bo Yang. Federated adaptation for foundation model-based recommendations. *arXiv preprint arXiv:2405.04840*, 2024.

[Zhang *et al.*, 2024c] Chunxu Zhang, Guodong Long, Tianyi Zhou, Zijian Zhang, Peng Yan, and Bo Yang. When federated recommendation meets cold-start problem: Separating item attributes and user interactions. In *WWW*, pages 3632–3642, 2024.

[Zhang *et al.*, 2024d] Honglei Zhang, He Liu, Haoxuan Li, and Yidong Li. Transfr: Transferable federated recommendation with pre-trained language models. *arXiv preprint arXiv:2402.01124*, 2024.

[Zhang *et al.*, 2025a] Chunxu Zhang, Guodong Long, Hongkuan Guo, Zhaojie Liu, Guorui Zhou, Zijian Zhang, Yang Liu, and Bo Yang. Multifaceted user modeling in recommendation: A federated foundation models approach. In *AAAI*, volume 39, pages 13197–13205, 2025.

[Zhang *et al.*, 2025b] Chunxu Zhang, Guodong Long, Zijian Zhang, Zhiwei Li, Honglei Zhang, Qiang Yang, and Bo Yang. Personalized recommendation models in federated settings: A survey. *IEEE TKDE*, 2025.

[Zhou *et al.*, 2025] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 16(12):9851–9915, 2025.

[Zhu *et al.*, 2019] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *NeuRIPS*, 32, 2019.

[Zhuang *et al.*, 2023] Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.