PFR-BENCH: BENCHMARKING PERSONALIZATION IN FEDERATED RECOMMENDER SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies on personalized federated recommendation have explored diverse personalization designs, but inconsistent implementations and evaluation protocols make existing results difficult to compare. More importantly, accuracy-centric evaluation provides limited insight into how these designs behave across users and under federated conditions. We therefore introduce PFR-Bench, a unified benchmark that standardizes six representative methods across ten datasets under a common experimental protocol. Beyond aggregate recommendation accuracy, PFR-Bench examines how personalization varies across heterogeneous users, depends on local interaction evidence, and behaves when client information becomes unreliable. Our evaluation reveals substantial differences that are obscured by overall performance alone. In particular, methods that use limited local evidence effectively are not necessarily those that remain stable when that evidence becomes unreliable, suggesting that local evidence should be modeled jointly through its quantity and reliability. We further find that stronger use of cross-client knowledge can compensate for insufficient local evidence while also increasing exposure to manipulated information from other clients, highlighting the need to balance collaborative benefit against propagation risk. PFR-Bench provides a reproducible basis for further evaluation and development of personalization mechanisms in federated recommendation. Code and benchmark resources are available at `https://anonymous.4open.science/r/PFR-BENCH-AD63/`.

1 INTRODUCTION

Federated recommender systems Chai et al. (2020); Yang et al. (2020); Sun et al. (2024); Wang et al. (2024) are built on the premise that users can benefit from collaborative learning without exposing their raw interaction data. In recommendation, however, such collaboration is complicated by substantial heterogeneity in user preferences: knowledge learned across clients can improve generalization, but may not align equally well with each user's local behavior. Personalization Zhang et al. (2025); Tan et al. (2022) therefore becomes central to federated recommendation, where the key problem is to balance cross-client knowledge sharing with user-specific adaptation rather than emphasizing either side in isolation. Recent PFR methods therefore instantiate this balance through markedly different personalization mechanisms, which determine how local and cross-client information are used to construct user-specific recommendations.

Existing PFR studies typically evaluate individual methods within method-specific experimental settings Zhang et al. (2023); Li et al. (2024); Zhang et al. (2024); Li et al. (2025); Chen et al. (2026); Zhang et al. (2026). Because these studies rely on different implementations, baselines, and evaluation protocols, their reported gains are not measured against a shared empirical reference Rendle et al. (2019); Sun et al. (2022). Without such a reference, it is difficult to determine whether an observed advantage arises from the personalization design itself or from the conditions under which it is evaluated, and thus whether that advantage would persist under other settings. Establishing a common experimental basis is therefore necessary to control for these study-specific factors and attribute observed differences more reliably to the personalization designs themselves.

Even with a common experimental basis, however, overall recommendation performance remains insufficient for evaluating personalization in PFR. Personalization is intended to accommodate heterogeneous user preferences, while aggregate performance can conceal substantial variation in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the gains achieved for different users Abdollahpouri et al. (2020); Rahmani et al. (2022). Moreover, similar overall performance may arise from personalization designs that rely differently on local user evidence and behave differently when client information becomes unreliable Chen et al. (2022); Wu et al. (2022). Evaluating personalization therefore requires looking beyond average accuracy to understand for whom a design is effective, what local evidence it depends on, and how reliably it operates through federated collaboration.

Taken together, these issues point to the need for a more systematic evaluation of personalization in federated recommendation. We present PFR-Bench, a unified benchmark for comparing and evaluating personalization in federated recommendation. To establish a common empirical basis, PFR-Bench standardizes the implementation and experimental protocol of six representative methods across ten datasets. With this basis in place, the evaluation moves beyond aggregate recommendation performance to examine how personalization varies across heterogeneous users, depends on local interaction evidence, and behaves under the efficiency and reliability constraints of federated collaboration. This provides a unified basis for systematically characterizing the capabilities and limitations of different personalization designs.

Results from PFR-Bench show that aggregate recommendation accuracy obscures substantial differences in where personalization gains arise and how different designs depend on local and collaborative information. Fine-grained evaluation reveals marked variation across users, local-data conditions, recommendation behaviors, and federated reliability. More importantly, two cross-experiment patterns emerge. First, the value of local evidence depends jointly on its quantity and reliability: methods that work effectively with limited interactions are not necessarily those that remain stable when those interactions become unreliable. Second, stronger use of cross-client knowledge can compensate for insufficient local evidence while also increasing exposure to manipulated information from other clients. These findings suggest that effective PFR design requires a more complete characterization of local evidence and a deliberate balance between the benefit and propagation risk of collaborative knowledge. The main contributions of this work are summarized as follows:

- **A unified benchmark for personalized federated recommendation.** We introduce PFR-Bench, which standardizes the implementation and experimental protocol of six representative PFR methods across ten datasets, providing a reproducible basis for systematic comparison.
- **A broader evaluation of personalization beyond aggregate recommendation performance.** PFR-Bench evaluates not only overall recommendation quality, but also how personalization varies across users, depends on local interaction evidence, and behaves under the efficiency and reliability constraints of federated collaboration. This provides a more complete empirical view of personalization than accuracy-centric evaluation alone.
- **Cross-experiment insights into personalization design.** By jointly analyzing different benchmark dimensions, PFR-Bench shows that local evidence should be characterized by both its quantity and reliability, while collaborative knowledge provides personalization benefits together with potential propagation risks. These findings motivate more informed PFR designs that jointly model the value of local evidence and explicitly regulate the degree of cross-client information incorporation.

2 PFR-Bench: Scope and Design

2.1 Problem Setting and Benchmark Scope

Personalized federated recommendation considers users whose interaction histories remain distributed across local clients. Rather than training users independently, PFR enables clients to exploit cross-user information while retaining user-specific recommendation models or representations, making the key problem how cross-client knowledge is incorporated into user-specific recommendation. Existing PFR methods address this problem through different personalization designs, differing in what information is shared and how collaborative knowledge is adapted to individual users.

Formally, let $\mathcal{U} = \{u_1, \ldots, u_N\}$ denote the set of users participating as federated clients, and let $\mathcal{I}$ denote the item set. Each user u locally retains recommendation data $\mathcal{D}_u$, with $\mathcal{D} = \{\mathcal{D}_u\}_{u \in \mathcal{U}}$ denoting the collection of user data. We represent a PFR method by a user-specific local component θ_u^{loc}, learned or maintained from local information, and a collaborative component θ_u^{col}, obtained through cross-client learning. The personalized recommendation model is then expressed as

$$\mathcal{M}_u = \Phi_u\big(\theta_u^{\mathrm{loc}}, \theta_u^{\mathrm{col}}\big), \tag{1}$$

Table 1: Comparison of personalization mechanisms in representative PFR methods.

Method	Personalized Components	Collaborative Knowledge Construction	Local–Collaborative Integration
PFedRec	Prediction function; item representations	Aggregation of item representations	Post-tuning aggregated item representations on local data
GPFedRec	Prediction function; item representations	Graph-guided aggregation of item representations	Global initialization and regularization toward graph-aggregated item representations
FedCA	Item representations	Aggregation of item representations from similar and complementary users	Interpolation between aggregated and local item representations
FedRAP	Item representations	Aggregation of item representations	Additive combination of global and local item representations
FedDAE	Local encoder; gating network	Aggregation of global encoder and decoder parameters	Gated fusion of global and local encoder outputs
FedEM	Local item representations; merging adapter	Similarity-weighted aggregation of item representations	Adapter-controlled merging of aggregated and local item representations

where the two components and the composition function Φ_u are method-specific, allowing different forms of parameter sharing and personalization without assuming a particular model architecture.

Within this formulation, PFR-Bench focuses on federated recommendation methods that explicitly introduce user-specific personalization into collaborative learning. We preserve method-specific differences in the local and collaborative components and their integration, as these constitute the personalization designs being compared rather than factors to be standardized. All methods are evaluated in the ID-based setting to provide a common basis for systematic comparison.

2.2 Representative Methods and Personalization Mechanisms

Following the formulation in Section 2.1, we compare PFR methods by how they instantiate the local component θ_u^{loc}, the collaborative component θ_u^{col}, and their integration through Φ_u, corresponding to what remains user-specific, what is learned across clients, and how the two are combined. We select six representative methods spanning substantially different realizations of these components, enabling comparison under a common conceptual basis rather than as isolated architectures.

For the six selected methods, we characterize personalization along three complementary aspects: **(1) personalized components, (2) collaborative knowledge construction, and (3) local–collaborative integration.** These aspects capture the main mechanism-level differences across methods and are not mutually exclusive, as personalization may occur at multiple stages. We highlight representative methods for each aspect below, with the complete cross-method mapping provided in Table 1.

(1) Personalized Components. PFR methods differ in which model components are maintained or adapted in a user-specific manner. PFedRec Zhang et al. (2023) personalizes the prediction function and item representations; FedRAP Li et al. (2024) separates shared and user-specific item representations; FedDAE Li et al. (2025) uses distinct shared and personalized encoders; and FedEM Chen et al. (2026) maintains local and aggregated item representations with client-specific merging. These designs show that personalization can be introduced at different levels, from representations to model components.

(2) Collaborative Knowledge Construction. PFR methods differ in how collaborative knowledge is constructed for personalization. GPFedRec Zhang et al. (2024) uses a user-relation graph for graph-guided aggregation from related users, whereas FedCA Zhang et al. (2026) aggregates item representations from both similar and complementary clients. These designs make aggregation itself a source of personalization by tailoring the collaborative knowledge available to each client.

(3) Local–Collaborative Integration. PFR methods further differ in how local and collaborative information are integrated. FedRAP additively combines local and shared item representations,

FedDAE uses a personalized gate to fuse two representation branches, and FedEM adaptively merges local and aggregated item representations through a client-specific adapter. These mechanisms range from direct addition to learned adaptive integration.

PFR-Bench preserves method-specific personalization mechanisms while standardizing only method-independent experimental factors, enabling fair comparison without imposing a common model architecture across different PFR methods.

2.3 Benchmark Construction and Evaluation Protocol

PFR-Bench includes ten recommendation datasets spanning diverse domains, scales, and interaction densities, with substantial variation in users, items, interactions, and sparsity. Detailed dataset statistics are summarized in Table 2. To ensure consistent comparison, we standardize all method-independent factors, including preprocessing, leave-one-out splitting, negative sampling, embedding dimension, and early stopping, while tuning method-specific hyperparameters on validation data. For the overall benchmark, each test instance is evaluated with 99 randomly sampled negatives using HR@10 and NDCG@10 He et al. (2017); Zhang et al. (2023). Fine-grained analyses adopt protocols tailored to their respective objectives, as specified in corresponding sections. Experimental details and hyperparameter ranges are given in Appendix A.

Table 2: Statistics of the datasets used in PFR-Bench.

Dataset	Interactions	Users	Items	Sparsity
MovieLens-100K Harper & Konstan (2015)	100,000	943	1,682	93.70%
MovieLens-1M Harper & Konstan (2015)	1,000,209	6,040	3,706	95.53%
LastFM-2K Cantador et al. (2011)	185,650	1,600	12,454	99.07%
FilmTrust Guo et al. (2013)	35,497	1,508	2,071	98.86%
Amazon-Book Hou et al. (2026)	388,483	9,043	7,877	99.45%
Amazon-Video Ni et al. (2019)	63,836	8,072	11,830	99.93%
Musical Instruments Hou et al. (2026)	16,562	1,093	664	97.72%
Movies & TV Hou et al. (2026)	165,236	3,622	3,939	98.84%
Baby Products Hou et al. (2026)	92,419	5,414	2,835	99.40%
Grocery & Gourmet Food Hou et al. (2026)	206,418	7,570	5,484	99.50%

3 Comprehensive Benchmark Evaluation

We conduct a comprehensive evaluation of PFR-Bench to characterize not only the overall recommendation performance of different personalization designs, but also how their advantages are realized under different user and federated conditions. We first establish an aggregate performance comparison, and then examine personalization across heterogeneous users and varying amounts of local interaction data. We further evaluate the efficiency and reliability of personalization under federated collaboration, revealing differences that are obscured by overall accuracy alone.

3.1 Overall Recommendation Performance

We compare the six representative PFR methods under the unified benchmark protocol across ten datasets, using the same data splits and evaluation procedure and reporting HR@10 and NDCG@10. Table 3 yields four main observations.

(1) Rich user histories favor methods whose personalization is learned directly from local behavior. PFedRec, GPFedRec, and FedDAE optimize personalized prediction functions or encoders from each user's interactions; on LastFM-2K, where users have over 100 interactions on average, these methods achieve comparatively strong performance, indicating that richer histories provide more direct support for such user-specific components. **(2) Methods that explicitly model and combine local and collaborative item representations show stronger performance when items are observed by more users.** FedCA, FedRAP, and FedEM perform well on MovieLens-100K, MovieLens-1M, and most Amazon subsets, where greater cross-user item coverage provides more reliable signals for learning both collaborative and personalized item representations. **(3) When both user histories and item observations are limited, the advantage of directly combining local and collaborative item representations becomes less evident.** On Amazon-Video, with only about eight interactions per user and five per item, FedCA, FedRAP, and FedEM show substantially weaker results, whereas PFedRec, GPFedRec, and FedDAE, which do not rely on such direct item-level fusion, remain comparatively stable. **(4) When items are observed by few users, the design of collaboration becomes more consequential.** On FilmTrust, FedCA and FedEM remain competitive while FedRAP performs notably worse; the former use selective or adaptive mechanisms, whereas FedRAP combines shared and local item views additively. This suggests that, with limited cross-user observations, selective or adaptive collaboration may exploit shared information more effectively than direct additive integration.

Table 3: Overall recommendation performance of representative PFR methods on ten benchmark datasets. The best result is shown in bold and the second-best is underlined.

Method	MovieLens-100K		LastFM-2K		FilmTrust		MovieLens-1M		Amazon-Book	
	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG
PFedRec	0.7253	0.4340	<u>0.8175</u>	**0.7191**	0.9162	0.8157	0.7295	0.4344	0.5097	0.2804
GPFedRec	0.7169	0.4332	0.7800	<u>0.6800</u>	0.9132	0.8093	0.7238	0.4329	0.5799	0.3327
FedDAE	0.5440	0.3053	**0.8325**	0.6790	<u>0.9169</u>	<u>0.8180</u>	0.5128	0.2808	0.5826	0.3252
FedCA	<u>0.8834</u>	<u>0.7659</u>	0.6812	0.6240	**0.9266**	**0.8383**	<u>0.8409</u>	0.6924	**0.7767**	**0.5303**
FedRAP	0.8727	0.7413	0.6498	0.5912	0.7665	0.5136	0.8243	<u>0.6954</u>	0.4752	0.2898
FedEM	**0.9608**	**0.8865**	0.7375	0.6596	0.8915	0.7955	**0.9379**	**0.8142**	<u>0.6702</u>	<u>0.4225</u>

Method	Amazon-Video		Musical Instruments		Movies & TV		Baby Products		Grocery & Gourmet Food	
	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG
PFedRec	0.6077	**0.3927**	0.2909	0.1622	0.3912	0.2116	0.3075	0.1633	0.2975	0.1622
GPFedRec	<u>0.6086</u>	<u>0.3918</u>	0.2681	0.1490	0.3600	0.1981	0.2913	0.1525	0.3569	0.1984
FedDAE	**0.6172**	0.3852	0.2653	0.1453	0.3736	0.1988	0.3251	0.1718	0.4285	0.2385
FedCA	0.1098	0.0509	0.8060	0.6388	0.5627	0.3666	<u>0.6760</u>	<u>0.4287</u>	**0.6267**	**0.3916**
FedRAP	0.1064	0.0465	<u>0.8253</u>	<u>0.6479</u>	<u>0.6657</u>	<u>0.4655</u>	0.5392	0.3218	0.4383	0.2515
FedEM	0.1054	0.0478	**0.8920**	**0.6901**	**0.7789**	**0.5279**	**0.8393**	**0.5712**	<u>0.6174</u>	<u>0.3790</u>

3.2 Personalization Effectiveness and Local Data Dependence

After examining aggregate performance, we characterize personalization across heterogeneous users and varying amounts of local interaction data. We first compare methods across users with different activity levels and preference characteristics, then examine whether their recommendations preserve users' historical interest patterns. We further control local training interactions to assess how different personalization designs depend on user-specific data. Together, these analyses characterize both the user-level effectiveness of personalization and its dependence on local interaction availability.

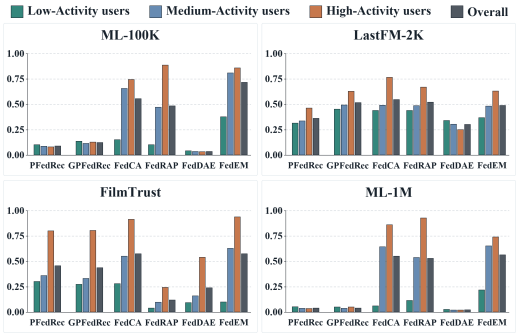
3.2.1 Personalization across Heterogeneous Users

User heterogeneity provides a natural lens for examining how personalization behaves across different users and whether recommendations preserve the preference patterns in their histories. We conduct these analyses on four representative datasets with different scales and interaction densities using full-ranking evaluation to avoid variation from sampled candidates. For the activity- and mainstreamness-based analyses, we report NDCG@10 in the main text, with the corresponding HR@10 results provided in Appendix B.

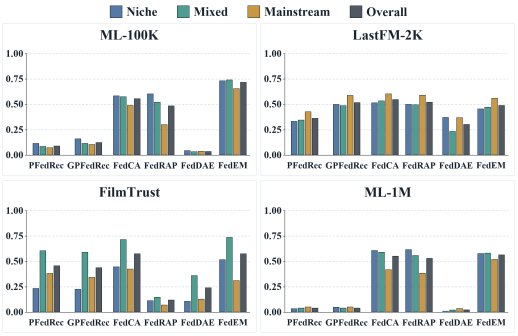
Performance across User Activity Levels. User activity reflects the amount of local evidence available for personalization Rahmani et al. (2022): richer histories support local preference learning, whereas sparse histories place greater demand on collaborative information. Within each dataset, we rank users by the number of training interactions and divide them into Low-Activity (bottom 25%), Medium-Activity (middle 50%), and High-Activity (top 25%) groups. We compare performance across these groups, with each method's overall performance included as a reference.

In Figure 1a, two observations emerge. **(1) The benefits of some personalization designs are concentrated among users with richer interaction histories.** FedCA, FedRAP, and FedEM perform substantially better for High-Activity than Low-Activity users, showing that their personalization advantages are more pronounced for users with richer observed behavior. **(2) Different personalization designs become competitive at different activity levels.** On MovieLens-1M, FedEM performs best for Low-Activity users, whereas FedRAP becomes strongest for High-Activity users; on FilmTrust, the advantage similarly shifts from PFedRec and GPFedRec to FedCA and FedEM. The reversals show that personalization designs have different strengths across activity levels.

Performance across User Interest Mainstreamness. Users also differ in the popularity of their interests Abdollahpouri et al. (2020); Rahmani et al. (2022). Popular interests receive more shared behavioral evidence, whereas niche interests place greater demands on personalization. We measure



(a) Performance across User Activity Levels.



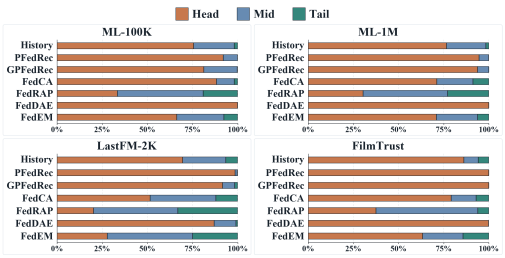
(b) Performance across User Interest Mainstreamness.

item popularity by training-set interaction frequency and characterize each user by the average popularity of items in their history. Users are then divided within each dataset into Niche (bottom 25%), Mixed (middle 50%), and Mainstream (top 25%) groups to examine whether personalization gains extend beyond mainstream preferences.

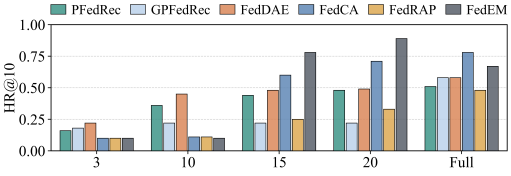
As shown in Figure 1b, two observations emerge. **(1) Mainstream preferences are not systematically easier to personalize.** Despite having more shared behavioral evidence, Mainstream users do not consistently achieve better performance: Niche or Mixed users are stronger for many methods on MovieLens, Mainstream users on LastFM-2K, and Mixed users on FilmTrust. Thus, more shared evidence alone does not guarantee better personalization; it must be translated into accurate user-specific ranking. **(2) The effect of interest mainstreamness is largely consistent across methods within the same dataset.** Most methods favor Mixed users on FilmTrust, Mainstream users on LastFM-2K, and Niche or Mixed users on MovieLens. This consistency suggests that dataset interaction and popularity patterns largely determine the direction of the mainstreamness effect, while personalization designs mainly determine its magnitude.

Preference Preservation in Recommendation Distribution. Beyond ranking performance, we examine whether recommendations preserve the preference composition in users' histories. This is important in PFR because popular items receive more cross-user exposure and may therefore be more strongly represented in collaborative knowledge. We group items into Head, Mid, and Tail by training-set popularity and compare their distributions in user histories and Top-10 recommendations.

As shown in Figure 2a, two observations emerge. **(1) PFR methods differ substantially in how much exposure they retain for less-popular items.** Relative to user histories, PFedRec, GPFedRec, and FedDAE shift recommendations toward Head items, FedRAP and FedEM retain substantially more Mid and Tail items, and FedCA generally produces smaller shifts. **(2) The direction of this deviation is largely consistent for the same method across datasets.** PFedRec, GPFedRec, and FedDAE repeatedly favor Head items, whereas FedRAP and FedEM consistently shift exposure toward less-popular items. This cross-dataset consistency suggests that how recommendation composition is reshaped is closely tied to the personalization design.



(a) Item Popularity Distribution in User Histories and Top-10 Recommendations.



(b) Recommendation Performance under Varying Local Interaction Counts.

3.2.2 Dependence on Local Interaction Availability

In PFR, user-specific preferences are learned from local interaction histories, while federated collaboration supplements this evidence with cross-user information. Different personalization mechanisms use local interactions differently: some rely on them to learn user-specific components, whereas

others use them mainly to adapt collaborative knowledge. Their sensitivity to local-data scarcity may therefore differ. Since the activity groups above contain different users, we conduct a controlled analysis that retains the same user population while restricting each user's training history to 3, 10, 15, or 20 interactions. Evaluation follows the same sampled-ranking protocol as the overall benchmark. The results are shown in Figure 2b, from which four observations emerge, with the corresponding NDCG@10 results provided in in Appendix C.

(1) Personalization mechanisms differ substantially in their robustness to limited local data. With only 3 or 10 interactions per user, PFedRec, GPFedRec, and FedDAE remain relatively stable, whereas FedCA, FedRAP, and FedEM degrade more sharply, showing that local-data scarcity affects different personalization designs unevenly. **(2) Additional local interactions yield different gains across personalization designs.** FedCA and FedEM improve sharply from 10 to 15 interactions, while GPFedRec changes little from 10 to 20 but improves substantially with full histories, possibly reflecting more reliable user-relation modeling from more complete local data. **(3) These different response patterns substantially change the relative advantages of PFR methods.** PFedRec and FedDAE remain competitive with sparse histories but are not necessarily strongest with richer histories, whereas FedEM becomes advantageous only with more local evidence. Evaluating methods only under full data therefore gives an incomplete view of personalization capability by obscuring their dependence on local-data availability. **(4) More local data does not always produce monotonic gains.** FedEM peaks at 20 interactions per user and then declines under the full setting, from 0.8884 to 0.6702 in HR and from 0.5703 to 0.4225 in NDCG. This may reflect a changing balance between local adaptation and collaborative information as local optimization strengthens, suggesting that additional local data can alter rather than simply strengthen personalization.

3.3 Personalization under Federated Collaboration

Personalization in PFR does not depend on local user data alone, but is also shaped by the information exchanged through federated collaboration. While Section 3.2 examines how personalization depends on local evidence, we next focus on the collaborative process itself. In particular, we examine the efficiency costs introduced by different personalization designs and their reliability when client information becomes noisy or deliberately manipulated.

3.3.1 Communication and Training Efficiency

Different personalization designs incur different communication and computation costs during federated training, while efficiency remains an important practical concern in federated recommendation Muhammad et al. (2020); Liu et al. (2024). We measure communication by the number of uploaded and downloaded parameters per round, and training efficiency by runtime per round and convergence rounds, capturing both per-round overhead and total training effort. Since our focus is on relative efficiency rather than cross-dataset variation, we conduct this analysis on FilmTrust.

Table 4: Communication and training efficiency.

Method	Comm. Params (Up/Down)	Time/Round (s)	Conv. Rounds
PFedRec	33,088/33,088	21.35	8
GPFedRec	33,088/66,176	29.34	24
FedDAE	205,396/205,396	60.90	25
FedCA	33,152/33,088	65.62	186
FedRAP	33,088/33,088	48.65	57
FedEM	33,088/33,088	44.72	232

As shown in Table 4, the results reveal two patterns. **(1) Similar per-round communication can lead to substantially different total communication costs.** PFedRec, FedRAP, and FedEM transmit nearly the same number of parameters per round but converge in 8, 57, and 232 rounds, respectively, producing an almost 29$\times$ difference in communication volume to convergence. Thus, when per-round communication is comparable, convergence becomes the dominant source of cost variation. **(2) PFR methods exhibit distinct efficiency bottlenecks.** GPFedRec incurs additional downlink communication through personalized embeddings, FedDAE has substantially higher per-round communication and computation costs, FedCA combines high runtime with slow convergence, and FedEM has moderate per-round cost but the longest convergence. These differences reveal method-specific efficiency profiles dominated by communication, computation, or convergence.

3.3.2 Reliability under Unreliable Client Information

Robustness to Noisy Local Interactions. User histories may contain accidental or preference-irrelevant interactions, making the local evidence used for personalization unreliable Wang et al.

Table 5: Performance under different interaction corruption ratios across four datasets. Each cell reports HR@10 / NDCG@10.

MovieLens-100K

Method	None	10%	30%	50%
PFedRec	.1707/.0900	.1485/.0743	.1018/.0418	.0933/.0521
GPFedRec	.2015/.1139	.1262/.0674	.1039/.0575	.0944/.0523
FedDAE	.0753/.0357	.0732/.0329	.0626/.0308	.0562/.0261
FedCA	.6288/.5431	.6310/.5479	.6617/.5747	.6723/.5838
FedRAP	.5705/.4887	.5493/.4765	.5376/.4693	.5249/.4600
FedEM	.8176/.7119	.8176/.7149	.7975/.6976	.8208/.7125

LastFM-2K

Method	None	10%	30%	50%
PFedRec	.5587/.3621	.5419/.3454	.5075/.4793	.4650/.4455
GPFedRec	.5544/.5142	.5425/.5185	.5012/.4750	.4562/.4317
FedDAE	.4044/.3138	.3575/.2712	.3087/.2318	.1087/.0547
FedCA	.5506/.5434	.5462/.5341	.5069/.4997	.4719/.4633
FedRAP	.5244/.5214	.5150/.4897	.4812/.4726	.4400/.4349
FedEM	.5144/.4880	.5044/.4740	.4625/.4416	.4169/.3957

FilmTrust

Method	None	10%	30%	50%
PFedRec	.6602/.4431	.6594/.3595	.5636/.2786	.4663/.2309
GPFedRec	.6340/.4439	.6033/.3337	.4656/.2386	.3593/.1792
FedDAE	.4162/.2425	.3832/.1904	.2987/.1407	.2620/.1178
FedCA	.7350/.5727	.7515/.6572	.7373/.5732	.6639/.5121
FedRAP	.2126/.1211	.1894/.1044	.1340/.0757	.1025/.0573
FedEM	.7171/.5748	.7238/.5875	.7275/.6021	.6864/.5658

MovieLens-1M

Method	None	10%	30%	50%
PFedRec	.0859/.0422	.0725/.0367	.0575/.0271	.0325/.0157
GPFedRec	.0886/.0433	.0702/.0329	.0439/.0211	.0363/.0177
FedDAE	.0409/.0213	.0377/.0191	.0366/.0181	.0310/.0151
FedCA	.6167/.5492	.6126/.5449	.6114/.5469	.6079/.5467
FedRAP	.5859/.5251	.5834/.5239	.5748/.5122	.5603/.4974
FedEM	.6675/.5574	.6715/.5620	.6508/.5439	.6339/.5358

Table 6: Attack performance under different poisoning ratios across four datasets. Each cell reports ASR@10/Nor-ASR@10.

MovieLens-100K

Method	None	10%	30%	50%
PFedRec	.0011/–	.1050/.0059	.6246/.4644	.9989/.9978
GPFedRec	.0011/–	.1124/.0141	.3330/.0484	.5334/.0678
FedDAE	.0000/–	.2884/.2096	.9968/.9954	.9989/.9978
FedCA	.0011/–	.1082/.0095	.3934/.1346	.7656/.5317
FedRAP	.0011/–	.1007/.0012	.3001/.0015	.5016/.0043
FedEM	.0032/–	.1060/.0070	.3022/.0045	.5281/.0572

LastFM-2K

Method	None	10%	30%	50%
PFedRec	.0006/–	1.0000/1.0000	1.0000/1.0000	1.0000/1.0000
GPFedRec	.0006/–	.2700/.1889	.5006/.2875	.6594/.3200
FedDAE	.0000/–	.1806/.1507	.4781/.3955	.6169/.5025
FedCA	.0006/–	.1006/.0007	.3081/.0116	.6162/.2325
FedRAP	.0006/–	.1006/.0007	.3006/.0009	.5006/.0012
FedEM	.0006/–	.1006/.0006	.3012/.0017	.5012/.0024

FilmTrust

Method	None	10%	30%	50%
PFedRec	.0007/–	.3376/.2643	.9671/.9530	.9753/.9506
GPFedRec	.0000/–	.4296/.3665	.9304/.9007	.9581/.9162
FedDAE	.0000/–	.2672/.1862	.9873/.9819	.9910/.9820
FedCA	.0007/–	.1003/.0008	.3772/.1110	.6519/.3038
FedRAP	.0195/–	.1826/.0922	.5150/.3077	.7650/.5300
FedEM	.0007/–	.0936/.0017	.3159/.0395	.5487/.1377

MovieLens-1M

Method	None	10%	30%	50%
PFedRec	.0002/–	.1094/.0105	.3424/.0605	.7531/.5076
GPFedRec	.0000/–	.0523/.0013	.2351/.0504	.6275/.3338
FedDAE	.0000/–	.0000/.0000	.1667/.1653	.3970/.3848
FedCA	.0005/–	.1008/.0009	.3222/.0317	.6570/.3139
FedRAP	.0007/–	.1002/.0002	.3002/.0002	.5002/.0003
FedEM	.0007/–	.1007/.0007	.3010/.0014	.5066/.0132

(2021); Gao et al. (2022). In PFR, such noise can affect both local adaptation and collaborative knowledge. We corrupt 10%, 30%, or 50% of each user's training interactions by replacing them with randomly sampled unobserved items, while keeping validation and test data unchanged, to examine robustness as signal quality deteriorates.

As shown in Table 5, two observations emerge. **(1) At low corruption ratios, most PFR methods remain relatively stable, but their robustness diverges as noise increases.** FedCA and FedEM remain comparatively stable under severe corruption, whereas PFedRec, GPFedRec, and FedDAE degrade progressively, making robustness differences increasingly pronounced. **(2) Robustness to noisy interactions is related to how local and collaborative item information is integrated.** FedCA and FedEM use selective or adaptive integration and remain comparatively stable, whereas FedRAP's direct additive combination shows less consistent robustness. This suggests that the integration strategy influences how strongly corrupted local signals affect personalized recommendation.

Robustness to Targeted Item-Promotion Attacks. Beyond random interaction noise, federated collaboration may expose personalization to deliberately manipulated client behavior, allowing injected preferences to affect both malicious and normal users. We randomly select 10%, 30%, or 50% of clients as malicious and inject a target-item interaction into their local training histories. The least-interacted item in each dataset is selected as the target to minimize the influence of existing popularity. We report ASR@10, the proportion of evaluated users whose Top-10 recommendations contain the target item, and Nor-ASR@10, the same rate among non-malicious users, to measure cross-client propagation, following prior targeted item-promotion attacks in federated recommendation Zhang et al. (2022); Wu et al. (2022); Liu et al. (2025).

As shown in Table 6, the near-zero ASR without poisoning and its general increase with the malicious-client ratio confirm effective target promotion. Three further observations emerge. **(1) Attack success does not necessarily imply cross-client propagation.** High ASR may mainly reflect compromised clients, whereas increased Nor-ASR indicates spillover to normal users. This distinction is clear

for FedRAP and FedEM on LastFM-2K and MovieLens-1M, where ASR rises substantially while Nor-ASR remains near zero. **(2) Personalization designs exhibit different propagation behaviors under the same poisoning level.** Under stronger attacks, PFedRec and FedDAE often show Nor-ASR approaching overall ASR, whereas FedRAP and FedEM largely confine the effect to malicious users and FedCA shows an intermediate pattern. This links how collaborative information is incorporated to how injected preferences spread across users. **(3) Propagation behavior also varies across datasets for the same method.** GPFedRec shows strong propagation on FilmTrust but much larger ASR–Nor-ASR gaps on MovieLens-100K and MovieLens-1M, with FedRAP exhibiting a similar cross-dataset change. Thus, resistance to malicious client information depends on both the personalization design and the recommendation setting. More details can be found in Appendix D.

4 Cross-Experiment Insights and Design Implications

Taken individually, the preceding experiments characterize different aspects of personalization in PFR, from user heterogeneity and local-data dependence to reliability under federated collaboration. Viewed together, they reveal distinctions that are not evident from any single experiment, allowing us to connect cross-experiment evidence and derive implications for future PFR design.

Local Evidence: Jointly Modeling Quantity and Reliability. The benchmark reveals two distinct profiles in how personalization designs use local evidence. When local interactions are progressively reduced, PFedRec, GPFedRec, and FedDAE remain comparatively effective with few observations, whereas FedCA and FedEM benefit more clearly from richer histories (Section 3.2.2, local-interaction availability analysis). Under interaction corruption, however, the pattern changes: FedCA and FedEM remain comparatively stable, while PFedRec, GPFedRec, and FedDAE become more sensitive to noise (Section 3.3.2, interaction-corruption analysis). Together, these results show that the value of local evidence cannot be inferred from either its amount or reliability alone. PFR designs should therefore **jointly model how much preference evidence is available and how trustworthy it is**, adapting personalization to their combined effect.

Collaborative Knowledge: Balancing Benefit and Propagation Risk. The benchmark also shows that personalization designs differ in how strongly they incorporate cross-client information. With limited local interactions, PFedRec and FedDAE remain comparatively effective, consistent with a greater ability to benefit from collaborative knowledge when local evidence is insufficient (Section 3.2.2, local-interaction availability analysis). Under targeted poisoning, however, these methods also exhibit substantial Nor-ASR, showing that manipulated information from other clients can more readily affect normal users (Section 3.3.2, targeted item-promotion analysis). Together, these patterns show that **greater openness to cross-client information shapes both collaborative benefit and exposure to harmful external influence**. Future PFR methods should therefore explicitly tune the degree of cross-client information incorporation to balance these two effects.

Taken together, these findings suggest that PFR should move beyond a fixed local–collaborative mechanism. Effective personalization requires jointly characterizing local evidence by its quantity and reliability, while regulating collaborative knowledge according to both its benefit and propagation risk. Together, these factors determine how PFR should balance local and cross-client information.

5 Conclusion

We presented PFR-Bench, a unified benchmark for systematically comparing personalization designs in federated recommendation through a common implementation and evaluation protocol for six representative methods across ten datasets. Our study shows that aggregate recommendation accuracy provides only a partial view of personalization: methods with similar overall performance can behave differently across users, depend differently on local information, and exhibit markedly different reliability under federated collaboration. Cross-experiment analysis further shows that local evidence should be assessed jointly by its quantity and reliability, while collaborative knowledge should be evaluated by both its benefit and the risk of propagating harmful information across users. PFR-Bench therefore provides a reproducible basis for identifying where different personalization designs are effective, what information they depend on, and where their limitations arise, while informing more principled use of local and collaborative knowledge.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising manuscript text, improving language clarity and concision, and refining the presentation of figures and tables. All research ideas, experimental design, implementation, analysis, and conclusions were developed and verified by the authors.

REFERENCES

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 726–731, 2020.

Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, pp. 387–388, New York, NY, USA, 2011. ACM.

Di Chai, Leye Wang, Kai Chen, and Qiang Yang. Secure federated matrix factorization. *IEEE Intelligent Systems*, 36(5):11–20, 2020.

Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. pfl-bench: A comprehensive benchmark for personalized federated learning. *Advances in Neural Information Processing Systems*, 35:9344–9360, 2022.

Jundong Chen, Honglei Zhang, Chunxu Zhang, Fangyuan Luo, and Yidong Li. Breaking the aggregation bottleneck in federated recommendation: A personalized model merging approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14547–14555, 2026.

Yunjun Gao, Yuntao Du, Yujia Hu, Lu Chen, Xinjun Zhu, Ziquan Fang, and Baihua Zheng. Self-guided learning to denoise for robust recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1412–1422, 2022.

Guibing Guo, Jie Zhang, and Neil Yorke-Smith. A novel bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2619–2625. AAAI Press, 2013.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Zhiwei Li, Guodong Long, and Tianyi Zhou. Federated recommendation with additive personalization. In *International Conference on Learning Representations*, volume 2024, pp. 11770–11787, 2024.

Zhiwei Li, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. Personalized federated collaborative filtering: A variational autoencoder approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18602–18610, 2025.

Langming Liu, Wanyu Wang, Xiangyu Zhao, Zijian Zhang, Chunxu Zhang, Shanru Lin, Yiqi Wang, Lixin Zou, Zitao Liu, Xuetao Wei, et al. Efficient and robust regularized federated recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1452–1461, 2024.

Minyi Liu, Chaofan Han, Liujun Luo, and Qiaohui Wu. Clean-label backdoor attack in deep learning based e-commerce recommendation system. In *2025 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence (DEBAI 2025)*. ACM, 2025. doi: 10.1145/3762249.3762312.

Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. Fedfast: Going beyond average for faster training of federated recommender systems. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1234–1242, 2020.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Hossein A Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. Experiments on generalizability of user-oriented fairness in recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*, pp. 2755–2764, 2022.

Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*, 2019.

Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, and Lizhen Cui. A survey on federated recommendation systems. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):6–20, 2024.

Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8206–8226, 2022.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Lingyun Wang, Hanlin Zhou, Yinwei Bao, Xiaoran Yan, Guojiang Shen, and Xiangjie Kong. Horizontal federated recommender system: A survey. *ACM Computing Surveys*, 56(9):1–42, 2024.

Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 373–381, 2021.

Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. Fedattack: Effective and covert poisoning attack on federated recommendation via hard sampling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4164–4172, 2022.

Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. In *Federated learning: Privacy and incentive*, pp. 225–239. Springer, 2020.

Chunxu Zhang, Guodong Long, Tianyi Zhou, Peng Yan, Zijian Zhang, Chengqi Zhang, and Bo Yang. Dual personalization on federated recommendation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/507. URL https://doi.org/10.24963/ijcai.2023/507.

Chunxu Zhang, Guodong Long, Tianyi Zhou, Zijian Zhang, Peng Yan, and Bo Yang. Gpfedrec: Graph-guided personalization for federated recommendation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4131–4142, 2024.

Chunxu Zhang, Guodong Long, Zijian Zhang, Zhiwei Li, Honglei Zhang, Qiang Yang, and Bo Yang. Personalized recommendation models in federated settings: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Honglei Zhang, Haoxuan Li, Jundong Chen, Sen Cui, Kunda Yan, Abudukelimu Wuerkaixi, Xin Zhou, Zhiqi Shen, and Yidong Li. Beyond similarity: Personalized federated recommendation with composite aggregation. *ACM Transactions on Information Systems*, 44(2):1–28, 2026.

Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Quoc Viet Hung Nguyen, and Lizhen Cui. Pipattack: Poisoning federated recommender systems for manipulating item promotion. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1415–1423, 2022.

A Implementation Details and Hyperparameter Settings

For a fair comparison, we set the negative sampling number $N = 4$, interaction filter threshold $t = 3$, and embedding dimension $d = 16$ for all methods. Besides, we utilize two common evaluation metrics for item ranking tasks: HR@K and NDCG@K where $K = 10$. We adopt the optimal settings of other hyperparameters as reported in their original papers. All models are trained to convergence with the early-stopping strategy to guarantee model convergence.

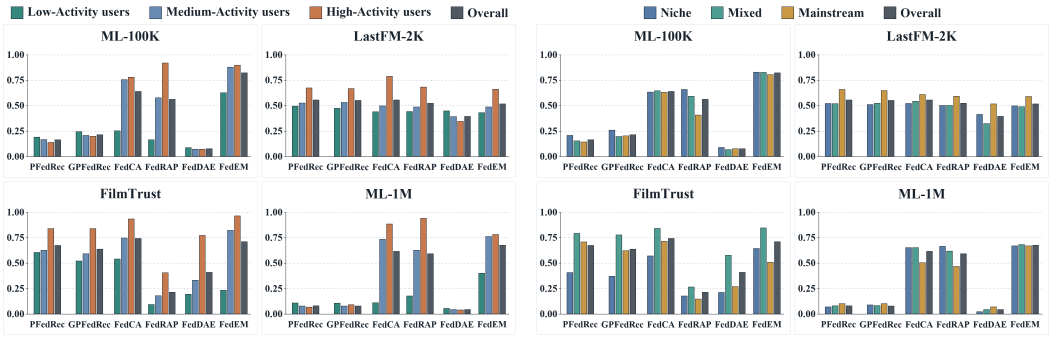
Hyperparameter tuning. For a fair comparison, we carefully tune the hyperparameters of each method via grid search. The search ranges are as follows:

- **PFedRec:** learning rate in $[0.01, 0.2]$, lr_eta in $[40, 200]$.
- **GPFedRec:** learning rate in $[0.01, 0.2]$, neighborhood threshold in $[0, 1]$, regularization coefficient in $[0, 1]$.
- **FedCA:** learning rate in $[0.01, 0.1]$, lr_structure in $[0.01, 0.1]$, lr_embedding in $[0.01, 0.1]$, α in $[0.2, 0.5]$, β in $[0.2, 0.5]$.
- **FedRAP:** learning rate in $[0.01, 0.1]$, α in $[0.01, 0.1]$, β in $[0.01, 0.1]$.
- **FedEM:** lr_embedding in $[0.01, 0.1]$, lr_adapter in $[0.01, 0.1]$, α in $[0.8, 1.0]$.
- **FedDAE:** learning rate in $[0.0005, 0.002]$, β in $[0.5, 1]$.

All hyperparameters are selected based on the validation performance, and the same experimental settings are applied across all datasets for consistency.

B Supplementary Results across User Groups

These results supplement the NDCG@10 analyses of user activity and interest mainstreamness in the main text. Figure 3a reports HR@10 across Low-, Medium-, and High-Activity users, while Figure 3b reports HR@10 across Niche, Mixed, and Mainstream users. Both analyses use the same four datasets, user partitions, and full-ranking evaluation protocol as their main-text counterparts, with overall performance included as a reference.



(a) Performance across User Activity Levels. (b) Performance across User Interest Mainstreamness.

Figure 3: HR@10 Performance across User Groups.

C Dependence on Local Interaction Availability

The main text reports the HR results of the controlled local-interaction analysis in Figure 2b. Here we provide the corresponding NDCG@10 results under the same experimental setting, i.e., the same user population, the same restricted training histories of 3, 10, 15, or 20 interactions, and the same sampled-ranking evaluation protocol. As shown in Figure 4, the NDCG trends are consistent with the HR trends reported in the main text: personalization mechanisms differ substantially in their robustness to limited local data, additional interactions yield uneven gains across designs, and more local data does not always produce monotonic improvements.

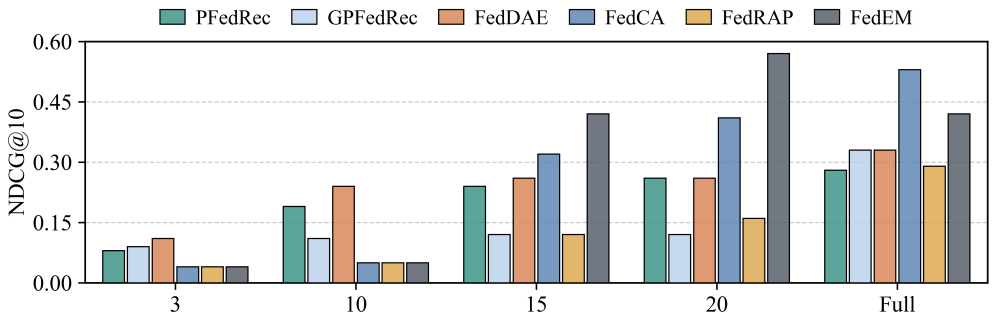


Figure 4: NDCG@10 performance under different local interaction budgets.

Table 7: Attack performance under different poisoning ratios across four datasets. Each cell reports AvgRank/Nor_AvgRank.

MovieLens-100K

Method	None	10%	30%	50%
PFedRec	1005.00/–	301.75/335.05	238.84/340.31	2.82/4.64
GPFedRec	1006.52/–	538.89/598.44	383.96/547.34	267.71/533.85
FedDAE	1467.41/–	20.98/23.19	2.18/2.68	1.52/2.04
FedCA	519.57/–	284.07/315.41	76.53/108.75	15.05/29.07
FedRAP	720.62/–	771.92/857.28	712.93/1016.66	555.61/1109.04
FedEM	840.21/–	699.79/777.16	452.53/645.16	249.93/498.33

LastFM-2K

Method	None	10%	30%	50%
PFedRec	7150.25/–	1.89/1.89	1.89/1.89	1.89/1.89
GPFedRec	7282.90/–	2008.97/2231.99	1407.59/2010.27	1428.31/2855.24
FedDAE	5890.53/–	1076.44/1161.79	746.71/992.29	519.14/904.17
FedCA	8187.59/–	4955.93/5506.48	1397.43/1995.90	235.03/469.05
FedRAP	9811.33/–	9640.21/10711.23	7453.76/10647.80	4976.26/9951.50
FedEM	5510.40/–	5154.20/5726.78	3886.71/5552.01	2780.60/5560.20

FilmTrust

Method	None	10%	30%	50%
PFedRec	410.79/–	107.01/103.32	69.26/98.18	46.76/91.31
GPFedRec	921.68/–	62.22/68.99	31.08/43.93	24.79/48.58
FedDAE	1341.10/–	41.56/46.04	24.22/34.14	17.63/34.26
FedCA	1157.98/–	594.52/660.14	96.72/137.63	18.70/36.40
FedRAP	364.29/–	259.59/288.18	154.58/220.21	67.37/133.74
FedEM	1462.63/–	979.95/1087.94	494.71/705.09	303.73/604.75

MovieLens-1M

Method	None	10%	30%	50%
PFedRec	2689.93/–	562.28/624.64	422.83/603.60	269.25/537.30
GPFedRec	3196.31/–	590.64/646.33	356.09/502.29	187.47/370.02
FedDAE	3506.81/–	289.56/289.92	20.58/20.82	11.90/11.60
FedCA	1385.95/–	847.74/941.82	288.62/411.88	88.54/176.08
FedRAP	1569.58/–	2400.78/2667.42	2324.29/3319.98	1698.55/3396.10
FedEM	2181.42/–	1782.76/1980.70	1112.20/1588.25	620.33/1239.27

D RELIABILITY UNDER UNRELIABLE CLIENT INFORMATION

To further characterize target promotion beyond Top-10 hit rates, we report AvgRank and Nor_AvgRank under the same poisoning settings. AvgRank denotes the average rank of the target item in the recommendation lists of all evaluated users, while Nor_AvgRank denotes its average rank among non-malicious users only.As shown in Table 7, AvgRank generally decreases as the malicious-client ratio increases, confirming that injected preferences progressively elevate the target item. The comparison between AvgRank and Nor_AvgRank further distinguishes attack effectiveness on compromised clients from spillover to normal users. For PFedRec and FedDAE, Nor_AvgRank often approaches AvgRank under stronger attacks, indicating substantial propagation to non-malicious users. In contrast, FedRAP and FedEM tend to preserve a larger gap, suggesting that their effects are largely confined to malicious clients, while FedCA exhibits an intermediate pattern. These trends are consistent with the ASR and Nor-ASR results in the main text and further show that propagation behavior depends on both the personalization design and the dataset.