TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Soyarar, E., Aydoğan, R., Buzcu, B., & Calvaresi, D. (2025). Explaining Federated Learning-Based Movie Recommendations. In *Proceedings of the 2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)* (pp. 729-734). IEEE. https://doi.org/10.1109/MetroXRAINE66377.2025.11340199

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3315-0279-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/METROXRAINE66377.2025.11340199

Explaining Federated Learning-based Movie Recommendations

1st Ege Soyarar
Dept. of Computer Science
Özyeğin University
Istanbul, Turkiye
ege.soyarar@ozu.edu.tr

2nd Reyhan Aydoğan
Dept. of AI and Data Engineering
Özyeğin University
Istanbul, Turkiye
Interactive Intelligence Group
Delft University of Technology
The Netherlands
reyhan.aydogan@ozyegin.edu.tr

3rd Berk Buzcu
Institute of Informatics
HES-SO Valais-Wallis, Switzerland
Sierre, Switzerland
berk.buzcu@hevs.ch

4th Davide Calvaresi
Institute of Informatics
HES-SO Valais-Wallis, Switzerland
Sierre, Switzerland
davide.calvaresi@hevs.ch

***Abstract*—**
The widespread adoption of recommender systems across industries such as entertainment, healthcare, and e-commerce has heightened concerns related to privacy, transparency, and overall trustability of these systems. The traditional centralized recommender systems risk mismanagement of user privacy and they lack interpretability, conflicting with emerging regulatory standards outlined by the EU AI Act and EU Data Act. To address these arising challenges head on, we propose an innovative recommender system framework that blends concepts from Federated Learning (FL) for the aspects of privacy and Explainable AI (XAI) to increase system trustability through transparency. FL facilitates decentralized model training, preserving user privacy by ensuring that personal data remains on local user devices while aggregating the global model updates with data scrubbed of information centrally. To enhance transparency and user trust, we borrow the post-hoc explanation strategies from the XAI literature and we leverage Large Language Models (LLMs) to harmonize the explanations with clear, understandable sentences for recommendations toward end-users. This combined solution balances privacy preservation, regulatory compliance, personalized recommendations, and interpretability, significantly enhancing recommender system design and adoption.

***Index Terms*—Explainable AI, Federated Learning, Personalized Recommender Systems, Large Language Models**

I. Introduction

The increasing reliance on recommender systems in various industries, including entertainment, healthcare, and e-commerce has raised significant concerns regarding privacy, transparency, and user trust. Traditional recommender systems often rely on large centralized datasets and models, which pose privacy risks and do not provide a clear understanding of how recommendations are generated. With the introduction of the EU AI Act and EU Data Act, regulatory focus is increasing on ensuring data processing and AI systems are transparent, accountable, and respect user privacy [4], [17]. These regulations emphasize the need for explainability in AI-driven decisions, the protection of personal data, and clear processes of user data on model training. Additionally, the majority of recommender systems work as "black boxes" providing users with recommendations without offering clear explanations of the underlying rationale, leading to reduced trust and satisfaction [22]. As these systems become more complex and integrated into everyday life, there is a growing need to design solutions that not only protect user privacy but also offer transparency into the decision-making process. Therefore, the development of recommender systems that balance privacy, scalability, and interpretability while complying with these regulatory standards is a critical challenge in the field of recommender systems.

Primarily, we address privacy concerns in recommender systems. Federated Learning (FL) has emerged as a promising solution that enables decentralized model training via unidentifiable aggregation without sharing user data [10]. In FL, user data remains on local devices while only model updates are exchanged with a central server [26]. The server then aggregates these updates, often using secure aggregation techniques, to produce a global model that benefits from distributed knowledge while maintaining data locality. This decentralized architecture significantly reduces the risk of data leakage and aligns with regulations about data and AI [25], [27]. Beyond preserving privacy, FL also enables effective personalization in recommender systems by allowing models to adapt to individual user behavior locally [8]. Since user preferences are unique, local training helps the model capture fine-tuned behavioral patterns that would otherwise be diluted in a centralized approach [24].

The second major challenge in modern AI systems, particularly in recommender systems, is the lack of transparency,

979-8-3315-0279-9/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.

often referred to as the "black-box" problem [16]. This issue arises when models generate predictions without offering insight into how decisions are made, which poses significant challenges for various stakeholders. Users may struggle to trust recommendations they do not understand, developers face difficulties in debugging or improving opaque models, and regulators are concerned about accountability and compliance with emerging legal standards [1]. Explainable AI (XAI) has emerged as a promising solution to this problem, aiming to make AI decisions more interpretable and understandable. In the context of recommender systems, XAI provides human-readable justifications for why specific items are recommended or not. These explanations not only increase user trust and satisfaction but also improve engagement and system usability, while helping mitigate risks associated with black-box decision-making [7], [28].

To address these challenges holistically, we propose a Federated Learning empowered Recommender System integrated with Explainable AI, designed to meet privacy, transparency, and personalization requirements simultaneously. Our research offers a Recommender System via Federated Learning approach to learn user preferences across distributed clients with a model. In this work, we used Synthetic Behavior Data related to movie watching. The dataset contains synthetic users and behavior data which is generated by using a probabilistic modeling approach that assigns diverse demographic, geographic, linguistic, and behavioral attributes, enabling realistic simulation of movie-watching behaviors and preferences. While all users' data contributes to global model, personalization is provided by user embeddings which are updated locally based on individual behavior, capturing nuanced preferences in a decentralized manner. By the end of model update iterations, all local models generate recommendations to own user. For explainability, we extract the trained neural network's rules by using Deep EXplanations with Rule Extraction (DEXiRE) [6] framework. Then, we incorporate a explanation generation to serve linguistic explanations of rules using Large Language Models (LLMs) [15]. Historical behavior of user, generated recommendation and important features are passed as inputs to local LLM in order to retrieve an explanation. Generated recommendations were evaluated using Precision, Recall, and F1-Score, while explanations were assessed based on metrics like accuracy, understandability and clarity. All the things considered, we show the applicability of our approach through experimental evaluations.

II. Related Work

Conventional recommender systems often rely on centralized data collection and deep learning models to personalize content for users. This centralized paradigm can leverage rich semantic information for recommendations [21], but it raises serious privacy concerns as user data is aggregated on company servers. Federated learning (FL) has emerged to mitigate such issues by training models in a decentralized manner across user devices, keeping personal data local [27]. Applied to recommendation tasks, FL allows collaborative model training without directly sharing raw user interactions, thus preserving privacy while still benefiting from collective learning [24]. Numerous FL frameworks, such as TensorFlow Federated and Flower [3], have been developed to facilitate scalable and customizable FL deployments across diverse domains, including recommender systems [12]. Early frameworks on federated recommendation systems demonstrated that techniques like federated collaborative filtering can achieve comparable personalization to centralized methods with significantly enhanced privacy safeguards [24]. Comprehensive surveys of FL highlight challenges such as statistical heterogeneity (not independent and identically distributed (non-IID) data across clients) and system scalability [11], [27]. Recent works address these challenges; for example, FedGP employs genetic programming to evolve aggregation strategies in non-IID federated settings, improving recommendation performance on heterogeneous data distributions [18]. Moreover, FedBN mitigates feature-shift non-IID data by leaving each client's batch-normalization parameters entirely local and averaging only the remaining weights, which markedly accelerates convergence and boosts accuracy under heterogeneous distributions [11]. FedProx augments FedAvg with a proximal regularization term and tolerates variable, inexact local updates, yielding provably more stable and accurate convergence under simultaneous statistical and systems heterogeneity in federated networks [13]. Likewise, advances in federated recommender systems continue to refine personalization, fairness, and efficiency under data decentralization constraints [8]. These efforts show that decentralized training can retain utility while respecting user privacy, a balance increasingly demanded by modern regulations. In fact, the regulatory landscape (e.g., the EU's upcoming AI Act) is scrutinizing data-driven personalization, assigning clearer responsibility and compliance requirements for FL deployments [25]. Such perspectives on legal accountability in FL emphasize that technological solutions for privacy (like federated recommenders) must align with governance and transparency expectations [25]. This convergence of technical and regulatory considerations underpins the importance of privacy-preserving, yet accountable, recommender system designs in the federated era.

Beyond privacy, explainability has become a crucial aspect of modern recommender systems, aiming to make model decisions transparent and user-interpretable. A rich body of literature on explainable recommendations surveys various techniques for illuminating why a particular item was recommended [28]. Many approaches focus on post-hoc explanations, where one first trains a high-accuracy model and then derives an explanation from it without altering its internal mechanisms [28]. Among the most popular post-hoc explanation techniques are LIME and SHAP, which provide instance-level interpretability by estimating feature importance for individual predictions. LIME approximates a complex model locally with simpler interpretable models, such as linear regressions, to explain why a particular decision was made [20], while SHAP leverages Shapley values from cooperative game theory to fairly attribute contributions of each

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.

feature to a prediction across all possible feature combinations [14]. Additionally, rule-extraction method, DEXiRE, can distill a trained neural network's behavior into propositional if-then rules, offering human-understandable insights into the model's decision logic [6]. Such post-hoc strategies allow complex models (e.g., deep learning recommenders) to be accompanied by explanations, thereby improving user trust. Indeed, trustworthiness frameworks for recommender systems explicitly identify transparency and explainability as key pillars alongside accuracy and fairness [7]. To further enhance accessibility of explanations, researchers have turned to large language models (LLMs) as explanation generators. Recent work proposes using LLMs to translate a model's reasoning or output into natural-language justifications that are easy for non-experts to understand [15]. These LLM-based explanation techniques can dynamically adjust to the user's preferences, effectively personalizing the explanation itself and making it more convincing and relatable [5]. By leveraging the vast contextual knowledge of LLMs, even intricate recommendation rationales can be communicated in a user-friendly way, bridging the gap between complex model logic and human comprehension. Furthermore, initial efforts have integrated explainability into FL environments. For example, Fed-XAI frameworks train models in a distributed fashion while simultaneously ensuring each federated model component remains interpretable [2]. Such approaches have been explored in domains like networked vehicles, where a federated model provides recommendations or decisions with local explanations, aligning with both privacy requirements and the need for transparency in safety-critical applications [19]. By combining FL with explainable AI techniques, these systems aim to maintain user trust and regulatory compliance, ensuring that recommendations are not only privacy-preserving and personalized but also transparent and accountable.

III. Methodology

The proposed methodology unites privacy-first federated learning with post-hoc explanation generation to deliver trustworthy movie recommendations. It is driven by two feedback loops: (i) a communication-efficient federated training loop that learns a global model from distributed user data without ever centralizing personal records, and (ii) an explanation loop that converts the model's internal reasoning into natural-language justifications by combining symbolic rule extraction with a Large Language Model (LLM). Fig 1 depicts the overall pipeline; the numbered circles on the diagram correspond to the sequence of operations described in further sections.

Algorithm 1 presents the full pipeline for federated recommendation with explanation generation. The process starts by initializing the global model G and defining the number of rounds R. In each round, the global model is distributed to all users (Line 2). Each user u_i trains a local model L_i on their data (Line 4) and sends the update to the server (Line 5). The server aggregates these updates to refine the global model (Line 7). After training, each user retrains their local model (Line 9) and extracts a decision rule R_i using symbolic

Algorithm 1 Federated Recommendation with Explanation

Require: U: Set of users with local data
G: Initial global recommendation model
R: Number of rounds

1: **for** round = 1 to R **do**
2: DistributeModelToClients(G)
3: **for** each user u_i in U **do**
4: L_i = TrainLocalModel(u_i)
5: SendModelUpdate(L_i, server)
6: **end for**
7: G = AggregateModelUpdates()
8: **end for**
9: **for** each user u_i in U **do**
10: L_i = TrainLocalModel(u_i)
11: R_i = ExtractRule(L_i, movie)
12: P_i = ConstructPrompt(R_i, u_i, label)
13: GetExplainedRecommendation(P_i)
14: **end for**

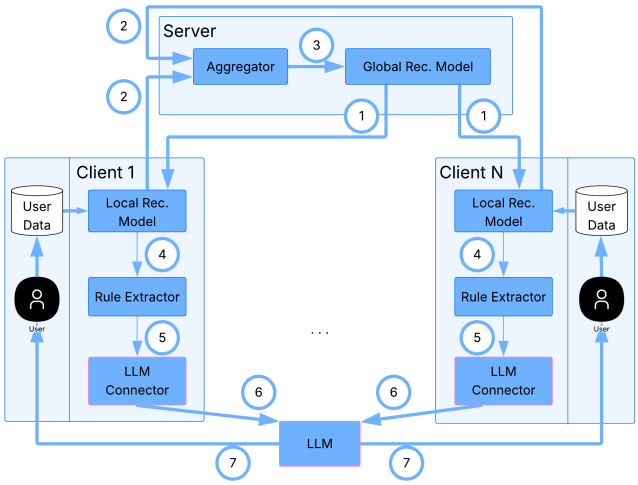

Fig. 1. Proposed System Architecture

rule extraction (Line 10). A prompt P_i is then constructed with the rule, user data, and label (Line 11) and sent to a language model to generate a natural language explanation (Line 12).

A. Dataset

Since real-world datasets like MovieLens [9] or Netflix Prize lack detailed user behaviors, profiles, and interaction patterns, and collecting such data is often costly and challenging, we use the Synthetic Behavior Generation (SBG) dataset introduced in prior work [23]. SBG simulates behavior by assigning each user a probabilistic profile covering demographic, geographic, and preference-based attributes. The generation pipeline involves: (1) stochastic profiling with dependency constraints (e.g., users under 18 are always "Single"), (2) day-by-day simulation based on seasonality, time, and watch tendency, (3) satisfaction scoring from weighted features like genre similarity and mood, and (4) label assignment based on a user-specific satisfaction threshold.

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.

B. Federated Recommender System

In the first phase of the system, the global recommender model is deployed from the server to multiple client devices (step 1 in the Fig 1). This initial model serves as a foundation for personalized learning on each client. The global model architecture consists of neural network layers designed to capture complex patterns in user behavior while remaining lightweight enough for efficient deployment across diverse client devices. Each client maintains a local copy of this model, which is then fine-tuned using user data. The training process occurs entirely on user devices, with each client utilizing only their local data to update model parameters. This approach ensures that sensitive user information never leaves the device, addressing privacy concerns inherent in traditional centralized recommender systems. During local training, the model parameters are adjusted to minimize prediction error on the local dataset, capturing user behaviors. The local training process employs a combination of collaborative and content-based filtering techniques adapted for the federated setting, allowing the model to learn both from user behaviors and item features. Following local training, clients transmit model parameters to the server (step 2). The server employs an aggregation mechanism to combine these updates from multiple clients into a global model improvement. The aggregation process (step 3) utilizes techniques such as weighted averaging of model parameters. We implement and evaluate three aggregation strategies: FedAvg, FedProx, and FedBN. Federated Averaging (FedAvg) serves as our baseline approach, performing weighted averaging of model parameters based on the quantity of local data at each client. To address challenges with non-IID data distributions, we incorporate Federated Proximal (FedProx), which adds a proximal term to the local optimization objective, preventing client models from deviating too far from the global model. Additionally, we implement Federated Batch Normalization (FedBN), which maintains client-specific batch-normalization parameters while sharing other model parameters, effectively handling domain shifts between clients. The updated global model is then redistributed to all clients, initiating another round of local training and global aggregation. This iterative process continues for a predetermined number of rounds or until convergence criteria are met. The global model gradually improves by incorporating insights from diverse user behaviors. Upon completion of the federated learning process, each client device maintains a personalized version of the recommender model that combines global knowledge with local adaptations. Hence the model generates recommendations based on both the global model's shared knowledge and the user's individual behavior patterns.

C. Explanation Generation

The explanation pipeline is a two–step procedure that marries *rule extraction* with *LLM verbalisation*. The goal is to transform the opaque neural recommender into concise and friendly justifications. The first step in explanation generation involves extracting interpretable rules from the trained neural network using the Deep EXplanations with Rule Extraction (DEXiRE) [6] approach (step 4 in the Fig 1). This technique transforms the high-dimensional representations learned by the neural network into a set of human-interpretable rules that approximate the model's decision-making process. The rule extraction process identifies the most influential features and their relationships that led to specific recommendations, creating a symbolic representation of the model's reasoning. The decision rules are form of IF-THEN (e.g. IF ((imdbRating $>$ 5.85) AND (season_Spring $>$ 0.50)) THEN recommended).

These extracted rules provide a foundation for generating natural language explanations but are typically expressed in a formal, logical structure that may not be immediately accessible to end-users. To bridge this gap, our system incorporates Large Language Models (LLMs) to transform these technical rules into clear, contextual explanations (steps 5-7). The LLM connector module on each client device prepares the necessary inputs for explanation generation, including the extracted matching rule, content name, and target label (recommended or not recommended). We feed the inputs to an instruction-tuned LLM together with the prompt template in Fig 2.

```
You are an assistant that explains movie
recommendations to users in a friendly, simple
way.
The system uses IF--THEN rules, where:
- IF part has conditions about the user or
the movie (like mood, rewatch count, duration,
ratings, etc.)
- THEN part says if the movie is recommended or
not.
Movie ID: {movie_id}
Rule: {rule}
Write a short, user-friendly explanation for
why this movie was recommended or not.
Only return the explanation for the user.
```

Fig. 2. Prompt template used for generating user-friendly recommendation explanations.

Once the LLM has generated the explanation, it is sent directly to the user. The user sees a sentence based only on the rule that triggered the recommendation. For example:

> "You are likely to enjoy *Film X* because it matches two of your favourite genres (HORROR, ACTION) and is spoken in languages you understand. Enjoy your movie."

IV. RESULTS

We evaluate our approach using the Synthetic Behavior Generation (SBG) dataset. Dataset includes 15 synthetic users and 2,940 user–movie interactions. To simulate federated learning, we used a two-phase setup. A global model was first pretrained on data from 3 clients to avoid cold start, then distributed to all users. Federated training ran for 10 rounds across 15 clients using FedAvg, FedBN, and FedProx. The entire pipeline was implemented with the Flower, enabling seamless client–server coordination and custom training logic.

A. Experimental Results

To assess the effectiveness of different federated learning strategies, we conducted a comparative analysis of FedAvg,

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.

FedBN, and FedProx over 10 rounds using 15 clients. Fig 3 illustrates the average performance across four evaluation metrics: accuracy, precision, recall, and F1-score. All strategies show consistent performance improvement across rounds, indicating stable convergence. Among the strategies, FedProx and FedBN exhibit slightly superior results in the later rounds, particularly in recall and F1-score, where both strategies surpass FedAvg by the final rounds. FedProx achieves the highest F1-score and recall in the end, suggesting its effectiveness in capturing positive instances. FedBN, on the other hand, demonstrates balanced performance across all metrics and appears to generalize well. While FedAvg starts with relatively strong precision, its improvements plateau in later rounds. Overall, these results indicate that FedProx and FedBN provide better generalization and robustness.

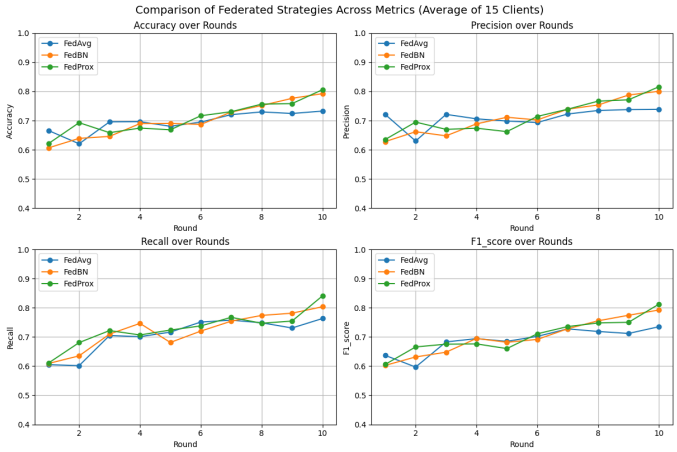

Fig. 3. Comparison of FL strategies across four evaluation metrics

To assess pretraining's impact with FedProx, users were grouped as pretrained or non-pretrained. As Table I shows, pretrained users began with higher accuracy and improved slightly, while non-pretrained users started lower but gained more. This suggests pretraining gives a better starting point, but non-pretrained users still improve considerably.

To judge how convincingly the system justifies its suggestions we use the LLM-based evaluation framework [23]. An LLM evaluator (GPT-4o mini) is prompted with (i) the user's behavior history, (ii) movie name, (iii) the binary relevance label (recommended or not), and (iv) the textual explanation. LLM evaluates explanations according to the three metrics which are accuracy, understandability, convincingness.

The Table II illustrates the average explanation quality scores—accuracy, convincingness, and understandability—across user groups clustered by the number of watched movies. Users were grouped into three categories based on their test set size: Low (14–24 samples), Medium (25–67), and High (68–130). This evaluation experiment is conducted using FedProx, as it demonstrated the best performance among the federated learning strategies tested. Notably, accuracy improves with data scale, increasing from 6.4 in the Low group to 7.4 in the High group, indicating that LLMs can provide more accurate evaluations when user profiles are rich. Convincingness remains consistently high across all groups (7.7–7.8), suggesting that LLMs find most explanations generally persuasive, regardless of user data size. Understandability remains relatively stable across user groups, without showing a clear pattern in relation to the amount of user data. Overall, these results highlight that explanation accuracy benefits significantly from larger user histories, while clarity and persuasiveness remain relatively stable.

TABLE I
ACCURACY GAIN: PRETRAINED VS. NON-PRETRAINED USERS

Group	*First Accuracy*	*Last Accuracy*	*Increase Rate (%)*
Pretrained	0.75	0.83	10.67
Not Pretrained	0.59	0.80	35.59

Accuracy values are averaged across users in each group.

TABLE II
AVERAGE EXPLANATION METRICS BY USER GROUP

User Group	*Accuracy*	*Convincingness*	*Understandability*
Low	6.4	7.8	7.3
Medium	7.1	7.7	6.9
High	7.4	7.7	6.9

Metrics are averaged across users clustered by number of watched movies.

To go beyond mere numerical evaluation and gain deeper insight into the reasoning behind the LLM's judgments, we retrieved the textual explanations generated by the LLM during the scoring process. Below are two representative examples—one high-scored and one low-scored—to illustrate the strengths and weaknesses observed during evaluation. High-scored: The recommendation for 'The Prestige' received an 8 for convincingness as it provided good reasoning based on IMDb score, language match. The explanation was easy to understand (9 for understandability). It accurately predicted the user's preference, hence an 9 for accuracy. Low-scored: For 'Stories of Our Lives', the convincingness was rated 4 due to a lack of strong reasoning for its unlikelihood. The explanation was understandable (8), but it did not align well with the user's history, leading to a low accuracy score of 3.

Consequently, high-scoring explanations align with user preferences and show meaningful feature-level reasoning, while low-scoring ones lack justification or contextual relevance despite being syntactically clear.

V. CONCLUSION

The widespread adoption of recommender systems in critical domains such as e-commerce, healthcare, and entertainment has highlighted significant concerns regarding user privacy, transparency, and trustworthiness of AI platforms. Traditional centralized recommender systems, while effective in delivering personalized content, inherently risk user privacy by requiring the aggregation of sensitive behavioral data on central servers. Moreover, their opaque "black-box" nature leaves users without meaningful insights into how recommendations are generated, creating a gap between system outputs and user understanding. This lack of transparency not only diminishes user trust but also raises compliance challenges

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.

under emerging regulatory frameworks such as the EU AI Act and EU Data Act, which mandate explainability and data protection in AI systems. To address these multifaceted challenges, this paper introduced an innovative framework that synergizes Federated Learning (FL) and Explainable AI (XAI) to create a privacy-preserving, interpretable, and user-centric recommender system. By adopting FL, our approach eliminates the need for centralized data collection, instead enabling collaborative model training across distributed devices while keeping user data localized. This decentralized paradigm not only aligns with modern data protection regulations but also enhances personalization by allowing models to adapt to individual user behaviors without compromising privacy. Complementing this, our integration of XAI techniques—specifically, rule extraction and natural language explanation generation using LLMs—provides users with clear, contextual justifications for recommendations, bridging the gap between algorithmic decisions and human interpretability. The combined strengths of FL and XAI in our framework thus offer a robust solution that advances the state-of-the-art in recommender systems by simultaneously addressing privacy preservation, regulatory compliance, and user trust through transparency.

Future work includes exploring adaptive explanation generation with real-time human feedback to improve quality and relevance. This could involve interactive interfaces where users rate or adjust explanations, creating a feedback loop to refine LLM outputs. The framework may also extend to high-stakes domains like healthcare and finance, where privacy and explainability are essential. More efficient federated learning techniques could improve handling of heterogeneous data while preserving privacy. Lastly, aligning the system with evolving AI regulations would support compliance and trust. These enhancements aim to build robust, user-centered explainable recommenders that learn from interaction.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

[2] José Luis Corcuera Bárcena, Mattia Daole, Pietro Ducange, Francesco Marcelloni, Alessandro Renda, Fabrizio Ruffini, and Alessio Schiavo. Fed-xai: Federated learning of explainable artificial intelligence models. In *XAI. it@ AI* IA*, pages 104–117. Udine, 2022.

[3] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. Flower: A friendly federated learning research framework, 2022.

[4] Federica Casolari, Chiara Buttaboni, and Luciano Floridi. The eu data act in context: A legal assessment. *International Journal of Law and Information Technology*, 31(4):399–412, February 2024.

[5] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Kai Zheng, Defu Lian, and Enhong Chen. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, June 2024.

[6] Victor Contreras, Niccolo Marini, Lora Fanda, Gaetano Manzo, Yazan Mualla, Jean-Paul Calbimonte, Michael Schumacher, and Davide Calvaresi. A dexire for extracting propositional rules from neural networks via binarization. *Electronics*, 11(24), 2022.

[7] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. A survey on trustworthy recommender systems. *ACM Trans. Recomm. Syst.*, 3(2), November 2024.

[8] Marko Harasic, Felix-Sebastian Keese, Denny Mattern, and Adrian Paschke. Recent advances and future challenges in federated recommender systems. *International Journal of Data Science and Analytics*, 17(4):337–357, 2024.

[9] 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.

[10] Danish Javeed, Muhammad Shahid Saeed, Prabhat Kumar, Alireza Jolfaei, Shareeful Islam, and A. K. M. Najmul Islam. Federated learning-based personalized recommendation systems: An overview on security and privacy challenges. *IEEE Transactions on Consumer Electronics*, 70(1):2618–2627, 2024.

[11] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study, 2021.

[12] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020.

[14] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[15] Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. Xai for all: Can large language models simplify explainable ai?, 2024.

[16] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. Explainable recommendation: when design meets trust calibration. *World Wide Web*, 24(5):1857–1884, 2021.

[17] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. Taking ai risks seriously: A new assessment model for the ai act. *AI & Society*, 39(5):2493–2497, October 2024.

[18] Elia Pacioni, Francisco Fernández De Vega, and Davide Calvaresi. Fedgp: Genetic programming for evolutionary aggregation in federated learning with non-iid data. In Pablo García-Sánchez, Emma Hart, and Sarah L. Thomson, editors, *Applications of Evolutionary Computation*, pages 419–434, Cham, 2025. Springer Nature Switzerland.

[19] Alessandro Renda, Pietro Ducange, Francesco Marcelloni, Dario Sabella, Miltiadis C. Filippou, Giovanni Nardini, Giovanni Stea, Antonio Virdis, Davide Micheli, Damiano Rapone, and Leonardo Gomes Baltar. Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking. *Information*, 13(8), 2022.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[21] Sunny Sharma, Vijay Rana, and Vivek Kumar. Deep learning based semantic personalized recommendation system. *International Journal of Information Management Data Insights*, 1(2):100028, 2021.

[22] Clara Siepmann and Mohamed Amine Chatti. Trust and transparency in recommender systems, 2023.

[23] Ege Soyarar, Berk Buzcu, Davide Calvaresi, and Reyhan Aydoğan. Llm-based evaluation methodology of explanation strategies. In *Explainable and Transparent AI and Multi-Agent Systems*, Cham, 2025. Springer Nature Switzerland.

[24] Muhammad Ammad ud din, Elena Ivannikova, Suleiman A. Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. Federated collaborative filtering for privacy-preserving personalized recommendation system, 2019.

[25] Herbert Woisetschläger, Simon Mertel, Christoph Krönke, Ruben Mayer, and Hans-Arno Jacobsen. Federated learning and ai regulation in the european union: Who is responsible? – an interdisciplinary analysis, 2024.

[26] Liu Yang, Ben Tan, Vincent W. Zheng, Kai Chen, and Qiang Yang. *Federated Recommendation Systems*, pages 225–239. Springer International Publishing, Cham, 2020.

[27] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

[28] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101, March 2020.

Authorized licensed use limited to: TU Delft Library. Downloaded on February 05,2026 at 07:40:12 UTC from IEEE Xplore. Restrictions apply.