Assessing the Quality of Data-Based Explanations in Recommender Systems: A Systematic Literature Review

Augustina Petraitytė, Asta Slotkienė

Vilnius University, Faculty of Mathematics and Informatics,
Universiteto g. 3, Vilnius, Lithuania
augustina.petraityte@mif.stud.vu.lt

Abstract. As recommender systems transition from "black boxes" to explainable models, assessing the quality of their explanations has become a critical research challenge. A systematic literature review has been performed to analyse how data-based explanation quality is evaluated in recent research (2021-2026). Findings reveal a significant reliance on system-oriented methods and metrics, while direct human-centred evaluation remains underrepresented.

Keywords: explainable AI (XAI), recommender systems, explanation quality, user trust, transparency, human-centred evaluation.

1 Introduction

Recommender systems are an essential component of many modern online platforms. They are widely used to help users discover relevant items such as products, movies, music, or news by analysing user preferences and behaviour. These systems aim to reduce information overload and support decision-making by suggesting items that are likely to be of interest to a particular user [1]. Recommender systems are used in e-commerce, media streaming services, social media and other domains.

Due to growing interest in explainable artificial intelligence (XAI) and explainable recommender systems, several studies have explored different methods for evaluating explanations. However, these studies use a variety of evaluation metrics and experimental approaches, making it difficult to compare results across different works [3].

It is important to note, that data-based explanations refer to explanations generated directly from underlying data patterns, feature importance and their relations. The generation process usually relies on post-hoc interpretability methods such as SHAP or LIME to justify recommendations, but not on rule-based or knowledge-based reasoning.

DOI: https://doi.org/10.15388/LMITT.2026.23

This systematic literature review analyses how the quality of data-based explanations in recommender systems is evaluated. The review focuses on identifying the evolution in emphasis on quality characteristics, the human-perceived quality of explanation analysis, and the evaluation methods and metrics used in the literature.

The rest of this paper is structured as follows. Section 2 provides background on recommender systems and XAI. Section 3 states the systematic literature review methodology. Section 4 highlights research results. Section 5 concludes this paper and outlines future work.

2 Background

Many modern recommendation algorithms rely on complex machine learning models that often operate as "black boxes", leading users to wonder why specific recommendations were made [2]. In the context of explainable recommendations, developing effective methods for delivering recommendations or explanations to users is equally important as transparent machine learning development, information retrieval, and data mining models [3]. The lack of transparency may reduce user trust and limit the acceptance of recommender systems. Therefore, improving transparency and interpretability has become an important research question in recent years.

To address these challenges, the concept of explainable recommender systems has gained significant attention. Explainable recommender systems aim to provide understandable justifications for recommendations, helping users understand how and why particular items were suggested [3]. Explanations can improve transparency, increase user trust, and support better decision-making by allowing users to understand the relationship between their preferences and the recommended items [4].

Although numerous explanation techniques have been developed, evaluating the quality of explanations remains a challenging task. Traditional recommender system evaluation focuses mainly on accuracy-related metrics such as precision or recall, which do not capture the effectiveness of explanations [3]. Explanation quality often involves human-centred aspects such as transparency, trust, clarity, and user satisfaction [4]. While it is easy to evaluate system by consistent metrics, it is difficult to base explanation quality on always changing human values.

3 Methodology

The systematic review process consists of several stages: preparation for review, identification, screening, eligibility, developing mapping and analysis (Figure 1). Such a structured approach was proposed by Kitchenham in 2007 [5], it is widely used in software engineering and computer science research to synthesise existing knowledge in a particular domain.

Specifying the research questions is one of the most important parts of any systematic review, it drives the entire methodology. Therefore, it is the first step of the review, we have raised three questions:

RQ1: How has the emphasis on quality characteristics evolved over the analysed period (2021–2026)?

RQ2: How is the human-perceived quality of explanations evaluated in recommender systems?

RQ3: What methods and metrics are used to evaluate explanation quality?

To ensure the relevance and quality of the selected studies, inclusion and exclusion criteria were defined. These criteria were applied during the screening process to determine which publications should be included in the final dataset of analysed papers. The criteria used in this systematic literature review are presented in Table 1.

Table 1. Inclusion and exclusion criteria.

Inclusion Criteria (IC)	Exclusion Criteria (EC)
IC1. Research had to be published after 2020 **IC2.** Publication must be written in English **IC3.** Research in the field of software engineering or computer science **IC4.** Research focus on recommender systems **IC5.** In the research explanation methods or explanations quality are evaluated	**EC1.** Duplicates of other publications of the same authors **EC2.** Publication was written in other language than English **EC3.** Publications inaccessible in full text

To identify relevant publications, a search query was constructed based on main concepts of this research: recommender systems, explainability, quality characteristics, and quality evaluation. Initially, the query included additional terms related to explanation properties such as faithfulness, fidelity, stability, robustness and consistency. However, including all these concept groups in the search query significantly reduces the number of

retrieved results. Therefore, the query was simplified to ensure that relevant studies would not be excluded.

Explanation context terms did not enforce XAI as a strict criterion during initial search to avoid relevant studies exclusion a head of time. XAI relevance was checked during full-text eligibility stage, to ensure that only research papers which address explainability mechanisms were included in the final analysis. This step was necessary to maximise coverage and maintain relevance in the selection process. The final search query is presented in Table 2.

Table 2. Search query (SQ).

Context	Parts of the Search Query (connected with AND)
Recommender system	(recommender* OR „recommendation system*“ OR XRS)
Explanation	(explainab* OR interpretab* OR „explainable AI“ OR XAI OR explanation*)
Quality characteristics	(understand* OR comprehensib* OR clarity OR transparen* OR trust* OR satisfaction*)
Quality evaluation	(evaluat* OR assess* OR metric* OR benchmark*)

The literature search was performed in two widely used academic databases: Google Scholar and Scopus. Scopus was selected because it contains many peer-reviewed publications related to recommender systems, XAI and machine learning. While Google Scholar increases coverage, as it captures the most recent advancements, which is important for a rapidly evolving XAI field. Databases like IEEE Xplore, ACM Digital Library and Clarivate Web of Science were excluded, because of free subscription functionalities limitations like advanced search and search wildcard numbers, making it hard to retrieve relevant research papers.

The identification was performed using limitations: research published after 2020 (IC1), publication written in English (IC2) and research is in the field of software engineering or computer science (IC3). 147 articles from Google Scholar and 507 articles from Scopus were selected. As the same search query was used in both databases, during the *screening* duplicated papers were removed (45 studies removed); records were screened by title, abstract and keywords, leaving 195 articles for further analysis; as a last screening step, articles were verified by inclusion and exclusion criteria (150 studies removed). The full text of each study was assessed for *eligibility*

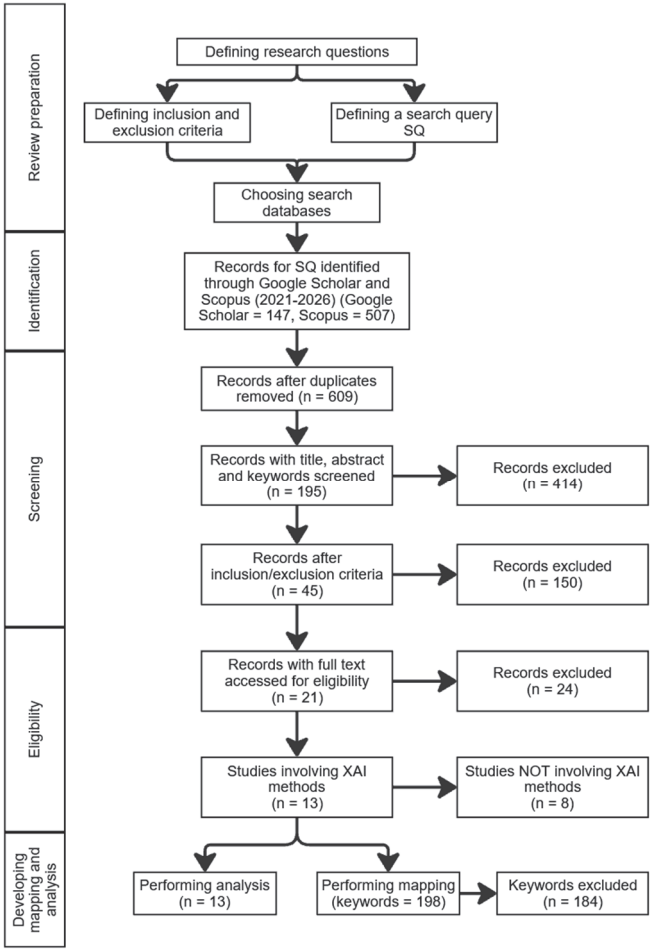

Figure 1. The flow diagram of the systematic literature review.

(excluding 24 articles). Each study was validated based on whether it includes XAI methodologies, leaving final 13 articles for deeper analysis (8 articles removed).

Domain-specific decision support systems which may look indirectly related to recommendation systems (like skin cancer diagnosis [14] and crop recommendation model [16]) were included in the analysis because

these systems function as expert recommender systems, suggesting specific actions or items based on complex user and environment data.

4 Results

This section presents the findings of the systematic literature review, structured according to the research questions.

Answering **RQ1** question, the analysis indicates a growing research interest in explanation quality in recommender systems over the period 2021-2026 (Figure 2). The number of publications has increased each year, with partial data from 2026 already exceeding those of earlier years such as 2021 and 2022. This suggests that the research community is increasingly prioritising user trust and transparency. As recommender systems become more integrated into daily decision-making, ensuring high-quality explanations has evolved from a secondary feature into a fundamental requirement. In terms of quality characteristics, the emphasis has shifted towards two main dimensions: system-oriented and human-centred. While awareness of human-related quality aspects has increased, the literature continues to prioritise technical performance over user experience, indicating an incomplete shift toward truly explainable systems.

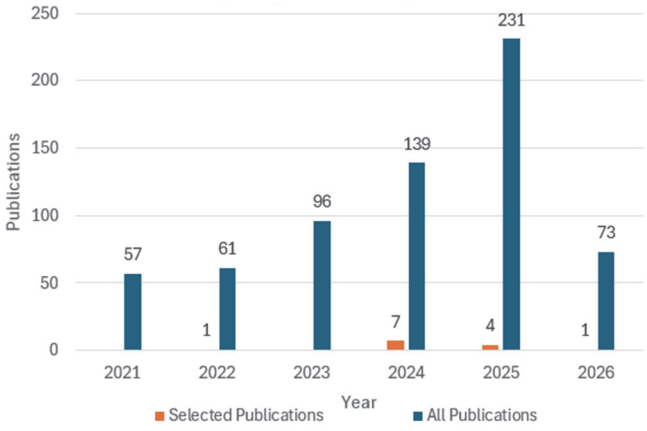

Figure 2. The trend of researching explanation quality in recommender systems (2021-2025; 2026 data is partial).

A keyword co-occurrence map (Figure 3) was generated to answer **RQ2**. This method was chosen for its ability to visualise associations between terms and organise them into thematic clusters, thereby enhancing the depth of analysis. Initially identified papers contained 198 unique keywords, less relevant terms were removed to focus on explanation quality in recommender systems. Firstly, the keyword list was filtered by frequency keyword a keyword, selecting a threshold of 2 mentions and retaining 25 keywords. Personalised terms for specific research fields (agriculture, diagnosis) and keywords indicating AI and machine learning frameworks (machine learning, deep learning, deep neural networks, learning algorithms) were removed from the analysis. After the removal, 14 relevant keywords remained. The map was created using the VOSviewer tool.

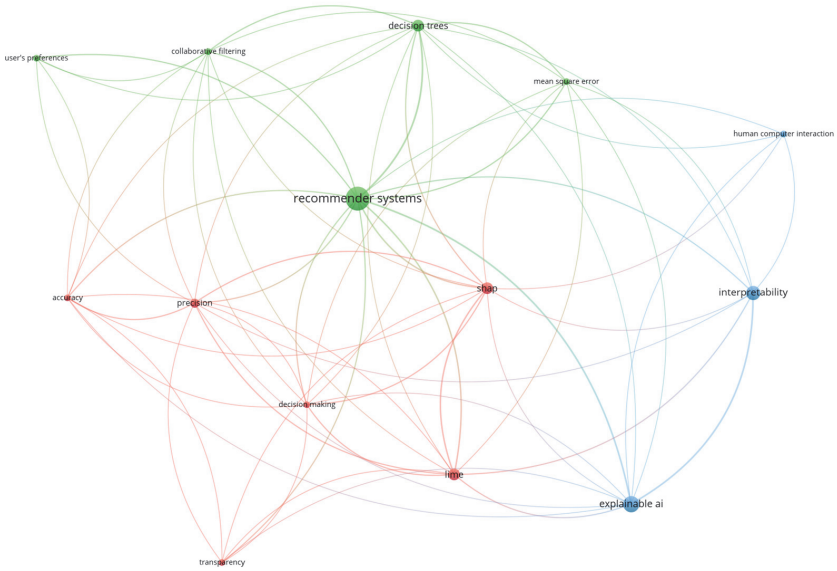

Figure 3. Keywords map.

Clusters in Figure 3 highlight that research on explanation quality in recommender systems is structured around three main thematic areas. The first cluster (green) focuses on recommender system models and technical performance, including collaborative filtering, decision trees, and error metrics such as mean squared error. The second cluster (red) bridges the

Table 3. Explanations quality evaluation methods and metrics.

Explainable Artificial Intelligence	Count. of Papers	References
Methods		
SHAP	10	[6][7][8][10][11][12][13][15][16][18]
LIME	8	[7][10][12][14][15][16][17][18]
LIRE	1	[9]
Grad-CAM++	1	[14]
System-oriented metrics		
Precision	9	[6][7][10][11][13][14][15][16][18]
Recall	9	[6][7][10][11][13][14][15][16][18]
F1-Score	8	[6][10][11][13][14][15][16][18]
Accuracy	5	[6][10][14][15][16]
ROC-AUC	3	[8][15][16]
MSE (Mean Squared Error)	3	[11][17][18]
AUC	2	[6][14]
MCC (Matthews Correlation Coefficient)	2	[13][15]
RMSE (Root Mean Square Error)	1	[7]
Model Fidelity	1	[9]
Top-N accuracy	1	[10]
Log-loss	1	[15]
MAE (Mean Absolute Error)	1	[17]
R2 (R-squared)	1	[17]
User-oriented metrics		
Trust	2	[8][12]
Effectiveness	2	[8][12]
Relevance	2	[10][12]
Weight of Advice (WOA)	1	[8]
Understandability	1	[10]

gap between performance and transparency. It includes traditional metrics such as accuracy and precision, alongside post-hoc explanation tools such as LIME and SHAP. The third cluster (blue) captures the human-centred dimension, focusing on interpretability, human-computer interaction, and XAI.

The distribution of these clusters suggests that, although human-centred concepts are present in the literature, they are less densely connected compared to technical and methodological themes. This indicates that evaluation of explanation quality is still driven by system-oriented perspectives, while user-focused evaluation remains comparatively underexplored.

Regarding **RQ3**, the analysis of evaluation methods and metrics (Table 3) reveals a heavy reliance on a few dominant post-hoc tools and traditional system-oriented metrics. Among XAI methods, SHAP and LIME are the most frequently mentioned, appearing in 10 and 8 papers respectively, while other tools remain marginal. The metrics used to assess explanation quality are overwhelmingly technical: precision and recall lead the list (9 papers each), followed closely by F1-score (8 papers) and accuracy (5 papers). In stark contrast, user-oriented metrics are significantly underused. Trust, effectiveness and relevance were each addressed in only 2 papers. This data confirms that while the industry is eager to "explain" recommendations, it primarily validates those explanations through the same algorithmic lenses used to measure the recommendations themselves.

5 Conclusion and Future Work

The results of this systematic literature review indicate that while the research community is increasingly vocal about the need for user trust and transparency, the practical evaluation of explanation quality remains in a transitional state. The keyword co-occurrence and metric distribution show a clear technical-heavy bias, with recommender system models and system-oriented metrics forming the core of current research. Although human-centred concepts such as interpretability and human-computer interaction have an established presence, they are not widely used in evaluating recommendation explanations. Ultimately, the field is successfully moving away from "black box" models, but it has yet to fully bridge the gap between technical fidelity and the actual human experience of understanding a recommendation.

This systematic literature review has limitations. The search was conducted only in two databases, possibly excluding relevant studies. Given the variability in terminology across explainable recommender systems, a more restrictive query could be used to make analysis more focused on the researched area.

To move beyond current evaluation strategies, future research should focus on a unified framework that integrates both system-oriented and human-perceived explanations in evaluating explanation quality. Moreover, to have a deeper understanding of existing research, the search of papers should be extended to cover more databases, such as IEEE Xplore, ACM Digital Library and Clarivate Web of Science.

References

[1] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems: Techniques, Applications, and Challenges". In: Recommender Systems Handbook, 3rd ed., 2021, pp. 1–35.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning" *arXiv preprint arXiv:1702.08608*, 2017.

[3] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives" *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[4] N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems" in *Recommender Systems Handbook*, Springer, 2010, pp. 479–510.

[5] B. Kitchenham, "Guidelines for performing systematic literature reviews in software engineering" 2007.

[6] S. Murindanyi et al., "Responsible artificial intelligence for music recommendation" in *Proc. Int. Conf. Data Science and Applications*, 2023, pp. 291–306.

[7] M. Huang et al., "An interpretable approach using hybrid graph networks and explainable AI for intelligent diagnosis recommendations in chronic disease care" *Biomedical Signal Processing and Control*, vol. 91, 2024, Art. no. 105913.

[8] V. Lyberatos et al., "Challenges and perspectives in interpretable music auto-tagging using perceptual features" *IEEE Access*, 2025.

[9] R. Yera, A. A. Alzahrani, and L. Martínez, "Exploring post-hoc agnostic models for explainable cooking recipe recommendations" *Knowledge-Based Systems*, vol. 251, 2022, Art. no. 109216.

[10] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of AI: Evaluation of explainability-enhanced recommendation systems" *Frontiers in Artificial Intelligence*, vol. 7, 2024, Art. no. 1410790.

[11] B. Venkateswarlu et al., "Cinematic curator: A machine learning approach to personalized movie recommendations" *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, 2024.

[12] M. Cesarini et al., "Explainable AI for text classification: Lessons from a comprehensive evaluation of post hoc methods" *Cognitive Computation*, vol. 16, no. 6, pp. 3077–3095, 2024.

[13] A. Chatterjee, M. A. Riegler, and P. Halvorsen, "Designing an ethical and explainable automatic coaching system for persuasive recommendations" *Multimedia Tools and Applications*, vol. 84, no. 41, pp. 50001–50035, 2025.

[14] A. S. N. Raju et al., "XAI-SkinCADx: A six-stage explainable deep ensemble framework for skin cancer diagnosis and clinical recommendations" *IEEE Access*, 2025.

[15] S. Alawadi et al., “A personalized and explainable federated learning approach for recommendation systems” in *Proc. IEEE Int. Conf. Edge Computing and Communications (EDGE)*, 2025, pp. 167–176.
[16] M. Bouni et al., “Interpretable machine learning techniques for an advanced crop recommendation model” *Journal of Electrical and Computer Engineering*, 2024, Art. no. 7405217.
[17] M. Y. Shams et al., “Enhancing crop recommendation systems with explainable artificial intelligence” *Neural Computing and Applications*, vol. 36, no. 11, pp. 5695–5714, 2024.
[18] P. Bairagi and S. Arora, “A transparent recommendation system for retail banking enabled by explainable AI” *Discover Artificial Intelligence*, 2026.