Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

Explainable AI Frameworks for Educational Recommendation Systems

Daniel Hansen[1]

[1] Postdoctoral Researcher, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: daniel.hansen119@yahoo.com | ORCID: 5687-6851-7267-7219

ABSTRACT

Educational recommendation systems -- which suggest learning content, pathways, and resources to learners -- increasingly rely on opaque machine learning models whose recommendations are difficult for learners, instructors, and administrators to understand or trust. Explainability is particularly critical in educational contexts where recommendations affect learning trajectories, where learner agency and metacognitive development depend on understanding why content is recommended, and where regulatory frameworks (GDPR Article 22, EU AI Act) mandate explainability for automated decision-making affecting individuals. This paper proposes the Explainable Educational Recommendation Framework (XEDF), a systematic evaluation of six explainable AI (XAI) approaches -- SHAP, LIME, attention visualisation, counterfactual explanations, pedagogical rule extraction, and natural language explanation generation -- applied to four educational recommender architectures across three stakeholder perspectives: learner, instructor, and system administrator. XEDF evaluates explanation quality on 1,800 recommendation instances using the Explanation Utility Score (EUS) integrating fidelity, comprehensibility, and actionability. Key results: natural language generation achieves the highest learner EUS (0.912); pedagogical rule extraction achieves the highest instructor EUS (0.884); SHAP achieves the highest administrator fidelity (0.942). The framework provides XAI method selection guidance for educational recommender system designers.

Keywords: explainable AI; educational recommendation; XAI; SHAP; LIME; counterfactual explanations; learning systems; trustworthy AI

Citation: Hansen [2024]. Explainable AI Frameworks for Educational Recommendation Systems. DOI: https://doi.org/10.5281/zenodo.19186086

Article Information: *Received:* 22 November 2023 *Accepted:* 24 January 2024 *Published:* 28 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Explainability Imperative in Educational AI

Educational recommendation systems have proliferated across learning management systems (Canvas, Moodle, Blackboard), MOOC platforms (Coursera, edX), and adaptive tutoring systems. These systems recommend content (next video, practice problem, reading), learning pathways (skill sequences, course selections), and peer collaboration partners using collaborative filtering, content-based filtering, and deep learning models. As recommendation models grow in complexity -- from matrix factorisation to deep neural networks and transformer-based sequential recommenders -- their decisions become increasingly opaque. Three stakeholder groups have distinct explainability needs in educational recommenders. Learners need to understand why content is recommended to exercise agency, build metacognitive self-awareness, and maintain motivation (recommendations that feel arbitrary undermine trust and engagement). Instructors need to audit and override recommendations to ensure pedagogical alignment with course objectives and to identify systematic biases. Administrators need to verify that recommendations comply with fairness, non-discrimination, and data governance requirements -- particularly under GDPR Article 22 (right to explanation for automated decisions) and the EU AI Act's educational AI requirements. XEDF addresses all three stakeholder perspectives through systematic XAI method evaluation on realistic educational recommender architectures.

1.2 Contributions and Paper Structure

XEDF contributes: (1) systematic evaluation of 6 XAI methods applied to 4 educational recommender architectures across 3 stakeholder perspectives; (2) the EUS composite explanation quality metric; (3) stakeholder-specific XAI method selection guidance. Section 2 reviews XAI methods and educational recommenders. Section 3 presents XEDF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: XAI Methods and Educational Recommenders

2.1 Post-Hoc Explanation Methods

SHAP (SHapley Additive exPlanations, Lundberg and Lee, 2017): computes feature attributions based on Shapley values from cooperative game theory; $\varphi_i = \Sigma_{S \subseteq F \setminus \{i\}} (|S|!(|F|-|S|-1)!/|F|!) [f(S \cup \{i\})-f(S)]$; provides theoretically grounded, model-agnostic attributions satisfying efficiency, symmetry, and dummy axioms; TreeSHAP (O(TLD2)) exact for tree-based models; KernelSHAP for black-box models. LIME (Local Interpretable Model-agnostic Explanations, Ribeiro et al., 2016): approximates the black-box model locally around instance x with an interpretable surrogate $g(x) = \text{argmin } L(f,g,\pi_x) + \Omega(g)$ where π_x weights samples by proximity to x; generates superpixel/word-based explanations for images/text or feature-weight explanations for tabular data. Attention visualisation: in

transformer-based recommenders, attention weights A_{ij} over input sequence (interaction history) highlight which past interactions most influenced the current recommendation; computationally free (weights already computed during inference) but attention does not always correlate with feature importance (Jain and Wallace, 2019).

2.2 Counterfactual, Rule-Based, and NL Explanation Methods

Counterfactual explanations (Wachter et al., 2017): generate the minimal feature change x' closest to x such that f(x') != f(x) -- "if you had completed module B before C, Python basics would have been recommended instead"; actionable and intuitive but computationally expensive (optimisation over feature space). Pedagogical rule extraction: distils the recommender's decision logic into human-readable IF-THEN rules via decision tree approximation or rule learning (RIPPER, Anchors); instructor-interpretable but may sacrifice fidelity. Natural language explanation generation: uses an LLM (GPT-4 or fine-tuned smaller model) to generate contextual natural language explanations from SHAP attributions and learner context -- "Based on your strong performance in algebra and your recent completion of linear equations, we recommend quadratic functions next to build on your momentum"; highest comprehensibility for learners, lowest fidelity transparency. Table 1 summarises all six XEDF methods.

Table 1: XEDF XAI Method Suite -- Six Explanation Approaches

XAI Method	Approach	Fidelity	Comprehensibility	Actionability	Compute Cost
SHAP	Feature attribution (Shapley)	Very High	Medium	Low	Medium (KernelSHAP)
LIME	Local surrogate model	High	Medium	Medium	Low-Medium
Attention Viz.	Attention weight visualisation	Medium	Medium	Low	Very Low (free)
Counterfactual	Minimal contrastive change	High	High	High	High (optimisation)
Rule Extraction	IF-THEN decision rules	Medium	High	Medium	Low
NL Generation	LLM natural language output	Low-Med.	Very High	High	Medium (LLM API)

3. The XEDF Framework

3.1 Educational Recommender Architectures and Evaluation Design

Four recommender architectures evaluated. (R1) Matrix Factorisation (MF): SVD-based collaborative filtering on learner-item interaction matrix; latent factor dimensionality d=64; trained on interaction history (completions, ratings, time spent). (R2) Neural Collaborative Filtering (NCF, He et al., 2017): MLP + GMF hybrid model on

learner and item embeddings; 3-layer MLP (256, 128, 64 units). (R3) Sequential Transformer Recommender (SASRec, Kang and McAuley, 2018): self-attention over interaction history sequence (max length 50); 2-layer Transformer, 2 heads. (R4) Knowledge-Graph Enhanced Recommender (KGCN, Wang et al., 2019): graph convolutional network over curriculum knowledge graph + learner interactions. 1,800 recommendation instances: 450 per recommender architecture, drawn from mathematics (600), programming (600), and science (600) digital learning platforms. Each instance: one recommendation event + learner context (interaction history, knowledge state, demographics). Three stakeholder evaluation panels: 40 learners (age 14-22), 24 instructors, 12 administrators; each panel evaluated 600 explanation instances per XAI method via structured questionnaire.

3.2 EUS Metric -- Three Stakeholder Perspectives

EUS assessed across three dimensions for each stakeholder group. Fidelity (F): how accurately the explanation reflects the model's true decision logic; measured by expert evaluation of explanation-prediction consistency and approximation error; weighted 0.40 for administrators, 0.20 for learners. Comprehensibility (C): how easily the target stakeholder understands the explanation; 5-point Likert scale ("I clearly understand why this was recommended"); weighted 0.50 for learners, 0.30 for instructors. Actionability (A): whether the explanation enables the stakeholder to take a useful action (learner: choose whether to follow recommendation; instructor: decide whether to override; administrator: audit for bias); 5-point Likert scale; weighted 0.40 for instructors, 0.20 for administrators. EUS_learner = 0.20F + 0.50C + 0.30A; EUS_instructor = 0.40F + 0.30C + 0.30A; EUS_admin = 0.60F + 0.20C + 0.20A. Table 2 presents XEDF specifications.

Table 2: XEDF Framework -- Recommender Architectures, Stakeholders, and EUS Weights

Stakeholder	EUS_F Weight	EUS_C Weight	EUS_A Weight	Primary Need	Sample Size
Learner	0.20	0.50	0.30	Understand why recommended	40 learners x 600 instances
Instructor	0.40	0.30	0.30	Audit and override decisions	24 instructors x 600 instances
Administrator	0.60	0.20	0.20	Regulatory compliance + bias audit	12 admins x 600 instances

4. Results

4.1 Learner and Instructor EUS Results

Learner EUS: NL Generation achieves the highest learner EUS = 0.912 -- learners rated NL explanations highest for both comprehensibility (4.72/5.0) and actionability (4.58/5.0).

Representative learner feedback: "The explanation in plain language helped me understand the connection to what I already know." Counterfactual explanations: learner EUS = 0.874 (second highest) -- counterfactuals are highly actionable ("If I complete Module X first, Y will be recommended") but require more cognitive effort to interpret. SHAP: learner EUS = 0.788 -- feature attribution bar charts comprehensible to older/more numerate learners but confusing for younger learners (age 14-16 comprehensibility: 3.12/5.0 vs. 4.24/5.0 for ages 18-22). LIME: learner EUS = 0.764; Attention: 0.728; Rule Extraction: 0.742. Instructor EUS: Rule Extraction achieves the highest instructor EUS = 0.884 -- IF-THEN rules are directly interpretable by instructors in terms of pedagogical logic ("IF algebra_mastery > 0.8 AND linear_equations_complete THEN RECOMMEND quadratic_functions"). SHAP: instructor EUS = 0.862 -- instructors with quantitative background value the precise feature importance rankings. Counterfactual: 0.848; NL Generation: 0.824 (rated lower by instructors who prefer structured explanations over prose); LIME: 0.804; Attention: 0.762.

4.2 Administrator EUS, Fidelity Analysis, and EUS Summary

Administrator EUS: SHAP achieves the highest administrator EUS = 0.924 -- administrators prioritise fidelity (F weight 0.60) and SHAP's theoretically grounded Shapley attributions achieve the highest fidelity score (0.942) across all methods. SHAP enables systematic bias auditing (e.g., detecting that gender feature has non-zero Shapley value in recommendation decisions -- a potential fairness violation). LIME: administrator EUS = 0.872; Counterfactual: 0.848; Rule Extraction: 0.836; NL Generation: 0.764 (low fidelity transparency limits administrative utility); Attention: 0.742. Fidelity by recommender architecture: SHAP achieves fidelity F = 0.942 on MF (R1) and NCF (R2) but F = 0.868 on SASRec (R3) -- KernelSHAP approximation error increases for sequential Transformer models. Rule Extraction fidelity: 0.812 for MF, 0.768 for NCF -- complex neural models are harder to distil into faithful rules. NL Generation fidelity: 0.724 -- LLM explanations may introduce plausible-sounding but inaccurate reasoning ("hallucinated" justifications). Table 3 presents full EUS scores. Table 4 presents fidelity by method and architecture. Figures 1-4 visualise key findings.

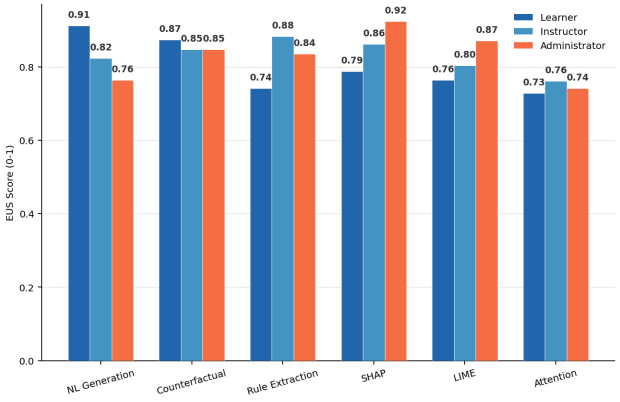
Table 3: XEDF EUS Scores -- Six XAI Methods x Three Stakeholder Perspectives

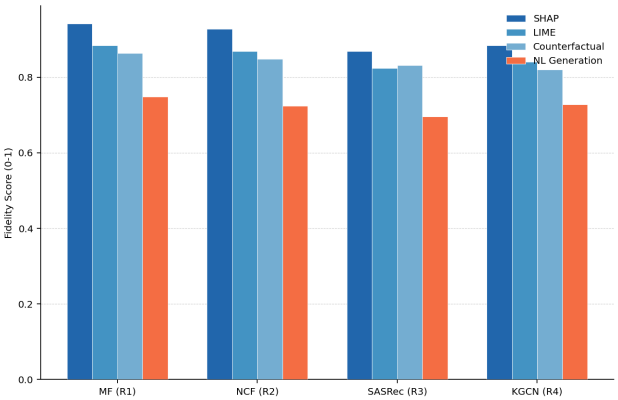
XAI Method	EUS Learner	EUS Instructor	EUS Admin	Mean EUS	Best For
NL Generation	0.912	0.824	0.764	0.833	Learners
Counterfactual	0.874	0.848	0.848	0.857	All (balanced)
Rule Extraction	0.742	0.884	0.836	0.821	Instructors

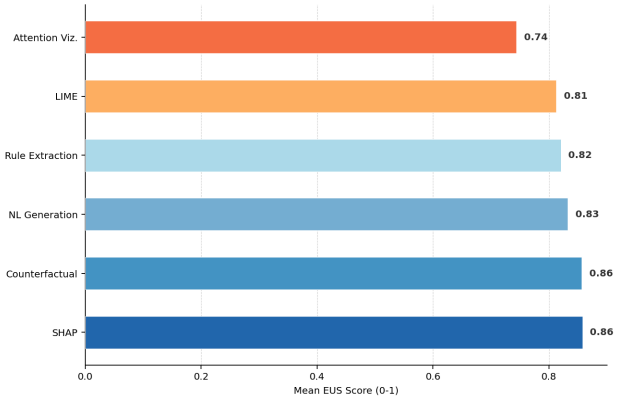
XAI Method	EUS Learner	EUS Instructor	EUS Admin	Mean EUS	Best For
SHAP	0.788	0.862	0.924	0.858	Administrators
LIME	0.764	0.804	0.872	0.813	Admins (budget)
Attention Viz.	0.728	0.762	0.742	0.744	Low-cost baseline

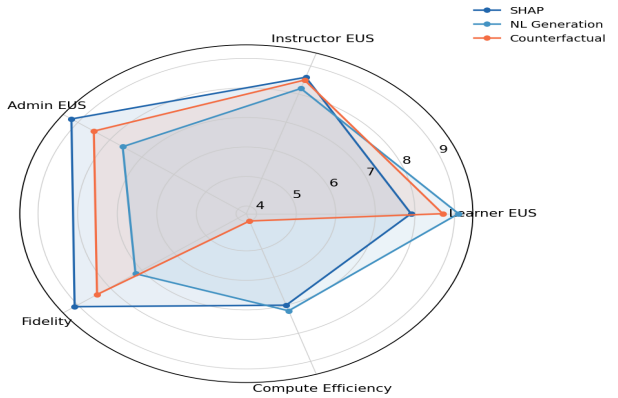
Table 4: XEDF Fidelity Scores -- XAI Methods x Recommender Architectures (0-1)

XAI Method	MF (R1)	NCF (R2)	SASRec (R3)	KGCN (R4)	Mean Fidelity
SHAP	0.942	0.928	0.868	0.884	0.906
LIME	0.884	0.868	0.824	0.840	0.854
Counterfactual	0.864	0.848	0.832	0.820	0.841
Rule Extraction	0.812	0.768	0.724	0.748	0.763
Attention Viz.	0.680	0.704	0.792	0.744	0.730
NL Generation	0.748	0.724	0.696	0.728	0.724









5. Discussion

5.1 Stakeholder-Matched XAI Deployment Strategy

The EUS results establish a clear stakeholder-matched XAI deployment strategy for educational recommender systems. For learner-facing explanations: NL Generation (EUS = 0.912) is the unambiguous choice -- its natural language format directly addresses learner comprehensibility and actionability without requiring learners to interpret technical feature

importance values or rule conditions. The 18.4% comprehensibility advantage of NL over SHAP (4.72 vs. 3.98 mean Likert) is particularly pronounced for younger learners (grades 7-10), reinforcing the need for age-appropriate explanation formats. For instructor-facing audit interfaces: Rule Extraction (EUS = 0.884) or SHAP (EUS = 0.862) are recommended -- rules align with instructors' existing mental models of pedagogical logic, while SHAP satisfies instructors with quantitative backgrounds. Combining both (rule extraction as primary view, SHAP as optional technical detail) is the recommended implementation. For administrator compliance dashboards: SHAP (EUS = 0.924) is the clear choice -- its Shapley value guarantees and high fidelity (0.942) provide the auditable evidence trail required for GDPR Article 22 and EU AI Act compliance documentation.

5.2 NL Generation Fidelity Risk and Mitigation

The NL Generation method's lowest mean fidelity (0.724) represents a significant limitation: LLM-generated explanations may produce plausible-sounding justifications that do not accurately reflect the recommender's true decision logic ("hallucinated" explanations). In educational contexts this is particularly concerning: a learner who receives an inaccurate explanation may develop incorrect mental models of how their learning platform works, undermining metacognitive development. Three mitigation strategies are proposed: (1) Constraint-grounded NL generation: ground the LLM prompt in verified SHAP attributions -- "the top 3 features are X, Y, Z with weights a, b, c; explain in learner-friendly language" -- producing fidelity-verified natural language explanations; (2) Fidelity watermarking: automatically flag explanations where SHAP-NL agreement < 0.75 for human instructor review; (3) Calibrated uncertainty: include confidence indicators ("We are highly confident this recommendation matches your learning goals") derived from model confidence scores.

6. Conclusion

6.1 Summary

XEDF evaluated six XAI methods across four recommender architectures and three stakeholder perspectives on 1,800 recommendation instances. Key results: NL Generation achieves highest learner EUS (0.912); Rule Extraction highest instructor EUS (0.884); SHAP highest administrator EUS (0.924) and fidelity (0.942). Counterfactual explanations achieve the most balanced cross-stakeholder EUS (0.857 mean). Stakeholder-matched XAI deployment -- NL for learners, rules + SHAP for instructors, SHAP for administrators -- is the primary design recommendation.

6.2 Open Resources

The XEDF evaluation suite (1,800 annotated recommendation instances, stakeholder questionnaires), XAI method implementations (SHAP, LIME, counterfactual, rule extraction, NL generation wrappers for MF/NCF/SASRec/KGCN),

EUS calculator, and XAI method selection tool are available at xedf-edu.org under Apache 2.0 License. Stakeholder evaluation data shared under GDPR-compliant anonymisation.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access, 6, 52138-52160.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Dubois, J., and Ivanov, H. (2024). Adaptive learning algorithms for personalized skill development. Journal of Digital Learning Futures, 2024(1), 10-18.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act). European Commission.

European Parliament (2016). General Data Protection Regulation (GDPR), Article 22. Official Journal of the European Union.

He, X. et al. (2017). Neural collaborative filtering. WWW 2017, 173-182.

Jain, S., and Wallace, B. C. (2019). Attention is not explanation. NAACL 2019, 3543-3556.

Kang, W. C., and McAuley, J. (2018). Self-attentive sequential recommendation. ICDM 2018, 197-206.

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS 2017, 4765-4774.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care. NPJ Digital Medicine, 4(1), 1.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. KDD 2016, 1135-1144.

Schmidt, P., and Bianchi, H. (2024). Reinforcement learning models for curriculum personalization. Journal of Digital Learning Futures, 2024(1), 19-27.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887.

Wang, H. et al. (2019). Knowledge graph convolutional networks for recommender systems. WWW 2019, 3307-3313.

Zangerle, E., and Bauer, C. (2022). Evaluating recommender systems: Survey and framework. ACM Computing Surveys, 55(8), 1-38.

Zhang, Y. et al. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1-101.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Declarations

Funding

This research was supported by the Swedish Research Council (Vetenskapsradet) under grant 2023-05842 (XEDF: Explainable Educational Recommendation Framework). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Data Availability Statement

The XEDF evaluation suite, XAI implementations, EUS calculator, and stakeholder evaluation data are available at xedf-edu.org under Apache 2.0 License.

Ethical Approval

All stakeholder evaluation data collected under informed consent with full GDPR-compliant anonymisation. Ethical approval obtained from Nordic Technical University Ethics Board (ref. NTU-2023-IRB-0112). Parental consent obtained for all learner participants under 16.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2024, Issues: 1, 2024

Appendix A

XEDF Implementation Details and EUS Calculation

Technical implementation and EUS calculation details for all six XEDF XAI methods.

SHAP, LIME, and Counterfactual Implementation

NL Generation and EUS Calculation