



Leveraging Generative AI to Improve Comprehensibility in Social Recommender Systems

Md Ashaduzzaman

University of Nebraska Omaha
College of Information Science & Technology
Omaha, NE, USA
aashaduzzaman@unomaha.edu

Chun-Hua Tsai

University of Nebraska Omaha
College of Information Science & Technology
Omaha, NE, USA
chunhuatsai@unomaha.edu

Abstract

Generative AI, particularly Large Language Models (LLMs), has revolutionized human-computer interaction by enabling the generation of nuanced, human-like text. This presents new opportunities, especially in enhancing explainability for AI systems like recommender systems, a crucial factor for fostering user trust and engagement. LLM-powered AI-Chatbots can be leveraged to provide personalized explanations for recommendations. Although users often find these chatbot explanations helpful, they may not fully comprehend the content. Our research focuses on assessing how well users comprehend these explanations and identifying gaps in understanding. We also explore the key behavioral differences between users who effectively understand AI-generated explanations and those who do not. We designed a three-phase user study with 17 participants to explore these dynamics. The findings indicate that the clarity and usefulness of the explanations are contingent on the user asking relevant follow-up questions and having a motivation to learn. Comprehension also varies significantly based on users' educational backgrounds.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Human-centered computing** → *User studies*.

Keywords

Large Language Models, User comprehension, Personalized recommendation explanations, AI-Chatbots, Follow-up questions.

ACM Reference Format:

Md Ashaduzzaman and Chun-Hua Tsai. 2025. Leveraging Generative AI to Improve Comprehensibility in Social Recommender Systems. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '25)*, June 16–19, 2025, New York City, NY, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3708319.3733696>

1 Introduction

Recommender systems reduce information overload by offering personalized recommendations in entertainment and online shopping [57]. These systems assess user-item interactions and supplementary details like item descriptions, user profiles, and reviews to estimate the chances of a user showing interest in particular

items using machine learning or data-driven approaches [17]. The recommendations generated could significantly improve the user experience by providing pertinent recommendations, empowering people to make informed decisions on various aspects of their daily lives, including high-stake healthcare decisions [28, 45]. However, the internal logic of how the system makes or generates these recommendations is usually inscrutable or not transparent to its users, causing criticism of the black box or lack of explainability issue, which can undermine user trust and experience [1, 42].

Methods to increase the explainability of AI-driven recommender systems are being developed in the research area of Explainable AI (XAI) [5, 44]. These methods range from global or local explanations (or interpretability) of AI models' predictions [43, 55] to enhancing human-centered XAI [44]. Many different types of explanations have been proposed and studied for their effects on users, such as feature-based [4, 29], example-based [11], counterfactual [48], conceptual [21], and many others [40, 49, 51]. Numerous studies have demonstrated the efficacy of these works across various aspects of user perception and experience, encompassing trust [10, 52], understanding [24, 46], usability [27, 41], and performance [16, 18].

An essential goal of explanation techniques is to foster users' understanding of complex AI systems [14, 44]. User understanding could be measured in two ways. The subjective understanding represents a user's own perception, usually measured by post-task questionnaires [7, 23]. Previous research has revealed a paradoxical phenomenon wherein users can perceive the effectiveness of AI-generated explanations yet struggle to comprehend their underlying meaning [14]. This lack of understanding is particularly pronounced among non-experts or users with limited domain-specific knowledge [22]. The objective understanding, a test to verify users' understanding of a model's inner workings, could be done through proxy tasks inviting users to simulate the system behavior and reproduce the output [7, 15]. Users' performance on this task can be used as a proxy for their understanding, but this method might be limited to only simpler tasks and computable by humans [44].

Furnkranz et al. argues that comprehensibility in XAI systems needs to be evaluated. Comprehension refers to how well users grasp explanations and can apply them to new tasks [19]. Comprehensibility is a more complete cognitive process that can be measured through three aspects: retention (understanding and recalling the AI model's predictions), transfer (applying the gained knowledge to further tasks), and recall (reproducing the knowledge) [25]. To achieve this, users may need to adapt and build their mental models interactively. According to Hoffman et al., [26], understanding users' mental models is crucial for evaluating their



This work is licensed under a Creative Commons Attribution 4.0 International License. *UMAP Adjunct '25*, New York City, NY, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1399-6/25/06
<https://doi.org/10.1145/3708319.3733696>

comprehension of XAI, with several methods proposed for measurement. Providing structure or scaffolding is crucial to support users in explaining their thoughts and reasoning. This scaffolding can take various forms, such as guided questions or tasks encouraging users to reflect on their understanding and articulate their mental models. The primary goal of scaffolding is to help users express their knowledge and reasoning processes more effectively.

LLMs with billions of parameters, have revolutionized fields such as NLP [9]. These transformer-based models, trained on diverse textual data, excel in language comprehension and generation, producing more human-like responses [13]. LLMs also demonstrate strong generalization and reasoning abilities, allowing them to adapt to new tasks. Prompt engineering offers reusable solutions to improve output generation and interaction with LLM, enabling more accurate and context-specific responses, especially in conversational AI [56]. To enhance the explainability of recommendations, ashaduzzaman et al. [6] integrates LLMs' conversational abilities, enabling users to receive personalized suggestions alongside clear explanations of the system's reasoning. Users found these explanations engaging and intuitive, valuing the transparency that fosters trust and confidence in the system. However, concerns emerged about whether users genuinely comprehend the provided explanations, as their understanding appeared subjective, highlighting a potential gap between perceived interest and actual comprehension.

Our study focuses on addressing two key research questions. First, we aim to determine whether users comprehend the explanations provided by the AI-Chatbot integrated into the recommender system and to what extent this comprehension occurs (**RQ1**). Understanding the depth and accuracy of user comprehension is critical for evaluating the effectiveness of AI-generated explanations. Second, we seek to identify the key behavioral differences between users who successfully comprehend these explanations and those who do not (**RQ2**). Exploring these differences will provide valuable insights into user interactions with AI systems, helping to design more effective and user-centered recommender systems. To objectively evaluate users' understanding of recommender system explanations, we introduce a novel three-phase user study framework that systematically evaluates user comprehension of AI-generated explanations by employing scaffolding techniques, such as think-aloud problem-solving and structured interviews, to capture users' mental models, an approach that is rigorous and difficult to replicate. In the first phase, users interact with the recommender system without LLM explanations to establish baseline knowledge. In the second and third phases, LLM explanations are introduced to evaluate their impact on user understanding. In each phase, users first interact with the recommender system, either with or without the chatbot, followed by a quiz assessing their knowledge of the system. After the quiz, a post-quiz interview is conducted to capture users' mental models. Our key findings indicate that users comprehend explanations more effectively when they can ask the right follow-up questions. Additionally, explanations should be tailored to the user's background knowledge, as novice users tend to spend more time understanding the system's core concepts, while more knowledgeable users are better equipped to grasp the concepts.

2 Related Work

Providing clear and understandable justifications for recommendations, even if they are not perfectly accurate, can increase user trust and overall system acceptance [20]. Kouki et al. [30] offers design guidelines for hybrid interfaces that integrate multiple explanation styles (e.g., textual, visual) that enhance user understanding and satisfaction with recommender systems. Explanation strategies should align with user preferences, knowledge, and expectations. Tailoring explanations to individual users increases engagement and satisfaction. Shokeen et al. [47] highlights the integration of social data, such as user relationships and interactions, to enhance personalized recommendations. The research focuses on the importance of explainability, ensuring that recommendations are interpretable and socially relevant.

Explanation techniques aim to enhance users' understanding of recommender or ML systems, distinguishing between subjective understanding (perception-based, measured via questionnaires) and objective understanding (actual comprehension, assessed through proxy tasks) [44]. Subjective understanding often relies on post-task questionnaires, using standardized or adapted methods to gauge users' perceived clarity of explanations [15]. Objective understanding employs tasks like forward simulation (predicting model output) [3], relative simulation (choosing the highest prediction score) [8], and manipulation tasks (adjusting inputs to achieve specific outcomes) [48]. These tasks test users' ability to interpret a model's inner workings. Additionally, user studies can assess users' understanding of explanations, serving as a sanity check to verify the clarity and intelligibility of conceptual explanations [21].

In the XAI context, a mental model reflects a user's understanding of an AI system, inferred through empirical evidence and guided methods [26]. These include Think-Aloud problem-solving tasks, where participants verbalize their thought process during a task [2]; concurrent question-answering [36]; retrospection tasks, where users reflect on their reasoning post-task [39]; structured interviews that expand retrospection with guided questions [32]; and prediction tasks, where users forecast outcomes and justify their reasoning [3]. These approaches rely on the researcher's skill in question design and interviewing to probe specific aspects of users' mental models effectively.

Miller et al. [38] advocates for incorporating social science research into explainable AI, emphasizing that solving computational challenges alone won't address trust issues in AI. It suggests that while early work in AI explanation is valuable, newer theories relevant to social psychology can enhance the design and implementation of intelligent agents capable of explanation. The authors [25] conclude that while explainability is important, it does not guarantee comprehension. XAI techniques like SHAP can bridge the gap, but more research is needed to explore other XAI methods and their impact on user understanding.

Mavrepis et al. [37] concludes that Large Language Models, such as x-[p]lAI[n], can significantly enhance the accessibility of XAI methods by providing audience-specific, simplified explanations. Wu et al. [54] proposes Polyjuice, a tool for generating counterfactual explanations using GPT-2. Li et al. [31] shows how explanations generated by LLMs can improve smaller models' reasoning abilities. Yang et al. [33] leverages GPT-3 to generate factual sentences

used for bottleneck models in image classification. Cambria et al. [12] focuses on discussing XAI issues related to LLMs. Turpin et al. [53] discusses the unpredictability of LLMs and the need for better interpretability as these models scale. Liu et al. [34] surveys how LLMs align with human values and intentions, covering areas like reliability, safety, fairness, and robustness. The study by Ashaduzzaman et al. [6] leverages LLMs to provide engaging, transparent explanations for personalized recommendations, fostering user trust, but notes a potential gap between perceived interest and actual comprehension. This work addresses the limitation of subjective user comprehension by introducing a rigorous three-phase user study framework that evaluates genuine understanding of LLM-generated explanations, identifying behavioral differences and providing actionable design implications for transparent, user-centered recommender systems.

3 Methodology

3.1 Advisor Recommender System and Dataset

We developed an advisor recommender system integrated with an AI-Chatbot to explain recommendations. Designed for prospective graduate students, the system matches students with potential supervisors based on shared research interests and publications. By streamlining the search process and aligning students with compatible mentors, it fosters more effective and fulfilling research collaborations. We adopted a similar system architecture but addressed the limitations of the previous work [6]. Our enhancements include displaying advisors' affiliations and publication lists alongside recommendations, offering clear user instructions, and refining prompt design with data retrieval generation (RAG) to deliver more coherent recommendation explanations.

Our proposed system generates recommendations using data from the Citation Network Dataset [50]. This dataset includes details on conference papers, authors, and attendees, such as names, publications, affiliations, and positions. For our research, we randomly selected data from 2,500 researchers, focusing on names, IDs, affiliations, research keywords, and paper titles. Limiting the dataset ensures quick system responses and avoids complexity, aligning with our goal of delivering clear and explainable recommendations.

Figure 1 showcases the user interface of the Advisor Recommender System, where users enter their name and research interest keywords, with suggested options provided. Upon submission, text and topic similarity models generate the top three advisor recommendations for each model. Results are organized into two tabs: the "Text Similarity" tab lists advisors with affiliations and publication details based on the text similarity model, while the "Topic Similarity" tab displays recommendations based on the topic similarity model. Each table cell is clickable for extended details. An AI-Chatbot integrated at the bottom in phases 2 and 3 provides explanations for recommendations and addresses user queries via an LLM-powered API (gpt-4o-mini). The app and associated resources are hosted on GitHub and deployed through Streamlit Cloud. The GitHub repository is available at ¹.

¹https://github.com/asadayon/Advisor_Recommendation

Advisor Recommender System

This study aims to assess users' understanding of social recommendation explanations provided by a large language model (LLM) integrated into this System.

Enter your Name

Ashad

Enter keywords of your research interest (i.e. Machine Learning, Natural Language Processing, Cybersecurity, Cryptography, Data Science, Human Computer Interaction, Robotics, Software Engineering, Cloud Computing, etc)

Machine Learning

Submit

Text Similarity Topic Similarity

Top 3 recommended advisor based on Text Similarity of keywords:

Rankin	Name	Publication	Affiliation
1	Blaine Nelson	Can machine learning be secure?, Revisiting proba	University of California, Berkeley;Unive
2	S. J. Russell	Analogical and inductive reasoning	Univ. of California, Berkeley;The Turing
3	Mohamed Amine	LM-DTM: An Environment for XML-Based, LIP/PAPI-	RWTH Aachen University;PROLEARN N

Figure 1: User interface of the Advisor Recommender System.

3.2 Recommendation Models and Explanations

3.2.1 Preference elicitation. The text similarity model uses count vectors from publication keywords to represent textual data, applying cosine similarity to compare user-input keywords with those from authors' publications. The top three authors with the highest similarity scores are recommended as advisors, aligning research interests for meaningful collaborations.

For the topic similarity model, Latent Dirichlet Allocation (LDA) categorizes publication keywords into 30 topics, creating thematic profiles for authors based on their involvement probabilities in each topic. User input keywords are matched to these profiles using the gensim library to recommend the top three authors with closely aligned research themes. This approach ensures recommendations based on thematic relevance.

3.2.2 Generating Explanations. Prompt engineering involves crafting and refining input prompts to guide model responses, ensuring accuracy, relevance, and coherence [35]. In the proposed prompt engineering architecture, three roles are defined: system, assistant, and user. Mavrepis et al. [37] developed prompt designs for enhancing XAI using Chain of Thought techniques. Building on their guidelines, we developed prompts for the system role, as shown in Table 1. These prompts define the expected input, the system's approach to explaining recommendations, and how to address follow-up questions. To enhance user interaction, contextual details such as the user's name, research interests, recommendations from both models, advisors' research interests, similarity scores, publication titles, and affiliations are integrated into the prompts. This retrieval augmentation ensures that the generative model has relevant and factual input to guide its outputs. The user role captures user input or queries, while the assistant role serves as an intermediary between the user and the system (see figure 2) for a conversation between user and the AI-Chatbot.

Table 1: Prompts for the ‘system’ role for advisor recommender system explanation

Objective: Provide explanations of the reasoning behind advisor recommendations based on text similarity, and LDA topic similarity. The goal is to enhance users’ technical comprehension of how their research interests align with those of the recommended advisors.

Input: You are given a username and three recommended advisor names to the user based on each model. You are also provided with user research interests, and research interests of three recommended advisors.

Retrieved Info: {User name + User research interest keywords + Top 3 recommended advisor list, ranking, affiliation, publication titles, similarity scores based on Text and Topic similarity.}

Expected Outcome: Users should gain a clear understanding of why specific advisors were recommended, with explanations tailored to their research interests. Users should be able to grasp the concepts of cosine similarity, topic similarity via LDA, and accurately interpret recommendation scores.

Guidelines: Do not provide overly technical jargon unless asked by the user. Do not give lengthy explanations; keep responses short, concise, and user-friendly. Do not assume the user understands complex mathematical concepts; provide examples when necessary.

Hello Ashad! I see that you are interested in Machine Learning. Based on your research interests, here are your top 3 recommended advisors from each model:

Cosine Similarity Model:

1. Blaine Nelson - Cosine similarity score: 0.9113
2. S. J. Russell - Cosine similarity score: 0.7253
3. Mohamed Amine Chatti - Cosine similarity score: 0.6804

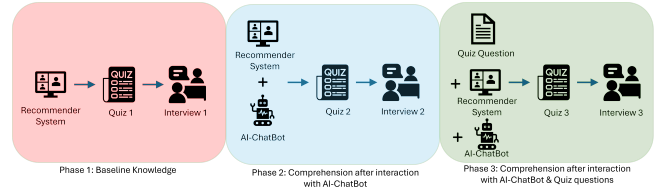
LDA Topic Modeling:

1. Diego Scardaci - Similarity score: 0.9961
2. M. Ortega - Similarity score: 0.9961
3. Josephine Reid - Similarity score: 0.9961

What does a similarity score of 0.9113 mean for Nelson? Also Can you tell me about Nelson's research?

A similarity score of 0.9113 for Nelson indicates that there is a very high level of similarity between your research interests in Machine Learning and Nelson's expertise. This score suggests a strong alignment between your interests and Nelson's research focus, making them a highly recommended advisor for you in the field of Machine Learning.

As for Nelson's research, they specialize in Machine Learning within adversarial environments and secure learning techniques. Nelson's work includes investigating proper learning techniques, the security of machine learning, and adversarial machine learning challenges. They have contributed to various publications covering topics such as evading anomaly detection through attacks on Principal Component Analysis (PCA), understanding risks in learning under adversarial conditions, and exploring strategies for classifier evasion. Nelson holds affiliations with the University of California, Berkeley; University of Tubingen, Germany; and the Wilhelm Schickard Institute for Computer Science in Tuebingen, Germany.

Figure 2: Example of LLM enhanced explanation of Machine Learning related advisor by AI-Chatbot.**Figure 3: Study Design for Evaluating User Comprehension of AI-Driven Advisor Recommendations.**

3.3 User Study Design

This study employed a *Within-Subjects* design to evaluate user comprehension of AI-driven advisor recommendations provided by an Advisor Recommendation System integrated with an LLM-powered chatbot. The experiment was divided into three phases (see figure 3), each designed to assess participants’ understanding of the system progressively. Below are the details of the study design:

3.3.1 Phase 1: Baseline Knowledge. Participants began by interacting with the Advisor Recommendation System through a web page where they input a username and research interest keywords related to computer science. The system used a text and topic similarity-based model to generate a list of academic advisors. Participants reviewed these recommendations, noting any questions or observations to establish their *baseline understanding* of the system’s recommendations.

After this interaction, participants completed a multiple-choice quiz to evaluate their comprehension of the recommendations and the system’s logic. The quiz was hosted on a separate platform, prohibiting the use of external resources to ensure accuracy. Feedback was provided for each question after completion of the quiz, indicating whether the selected answers were correct or incorrect, but the correct answers were not revealed for incorrect responses.

Finally, participants were interviewed to gather qualitative insights into their understanding of the recommendations. The interview also assessed their educational background in areas such as machine learning, statistics, or mathematics and explored any gaps between their expressed understanding and the actual workings of the recommendation system.

3.3.2 Phase 2: Comprehension After Interaction with the AI-ChatBot. In the second phase, participants interacted with an enhanced version of the Advisor Recommendation System integrated with an AI-ChatBot. After entering their username and research keywords, the system displayed the same recommendations as in Phase 1, supplemented by natural language explanations from the AI-ChatBot. These explanations clarified the reasoning behind the recommendations and provided summaries of each. Participants were encouraged to ask follow-up questions about the recommendations and the underlying model’s logic.

After this interaction, participants completed a randomized version of the same multiple-choice quiz from Phase 1 to mitigate order effects and reduce the potential for memorization. This allowed for the assessment of changes in comprehension resulting from the chatbot interaction. A post-quiz interview followed, focusing on

how the chatbot interaction influenced participants' understanding, any challenges encountered, and their evolving perceptions of the system. Feedback was collected on how the chatbot improved comprehension and how it could be further enhanced.

3.3.3 Phase 3: Comprehension After Interaction with AI-ChatBot and Quiz Questions. In the final phase, participants interacted with the recommendation system and chatbot again, but this time, they were informed of specific quiz questions they would need to answer afterward. This adjustment aimed to engage participants more actively by encouraging them to seek out relevant information during their interaction. Following this interaction, participants completed a final multiple-choice quiz designed as a cumulative assessment of their understanding after all system interactions.

A final interview was conducted to reflect on participants' learning journey, including how their comprehension evolved, their confidence in interpreting AI-driven recommendations, and the learning strategies they employed. Participants also provided feedback on how the system and chatbot integration could further support user learning and enhance educational systems. The post-experiment interview was not a task assigned to the participants, but rather aimed at collecting and understanding user feedback and experience. This *Within-Subjects Design* allowed for repeated assessments of participants' comprehension across different interaction levels, providing valuable insights into the effectiveness of the Advisor Recommendation System and the integrated AI-Chatbot in improving user understanding.

3.4 Data Collection and Analysis

3.4.1 Post-study Quiz. The quiz consists of 22 multiple-choice questions designed to evaluate users' understanding of text similarity and topic similarity in the context of an advisor recommender system. The quiz questions were categorized into three levels of difficulty to reflect the cognitive demands and complexity of the topics covered. The Foundational Knowledge Group, comprising 6 questions, tested basic concepts such as the purpose of text similarity-based models and the significance of cosine similarity scores. These questions primarily required direct recall and minimal cognitive processing, emphasizing foundational understanding. The Application and Integration Group, consisting of 11 questions, focused on intermediate complexity, challenging participants to apply theoretical knowledge to practical scenarios. These included tasks like calculating cosine similarity between vectors or determining preprocessing steps for text similarity analysis. The Advanced Analysis Group, with 5 questions, represented the highest difficulty level, requiring deep analytical reasoning and algorithmic understanding. These questions involved interpreting LDA outputs and addressing topic modeling challenges, reflecting advanced reasoning and problem-solving skills.

Each user is required to complete the quiz three times, with each attempt featuring a randomized order of questions and options. The quiz is administered through a web application built using Streamlit, providing a seamless and interactive experience for the users. The scoring system provides immediate feedback upon completion of the quiz. Users are prompted to enter their username and attempt number before submitting their answers. Once submitted, the application displays all the questions along with the

selected answers. Correct answers are highlighted in green with a tick mark, while incorrect answers are highlighted in red with a cross sign. Additionally, a percentage-based horizontal bar graph visually represents the user's performance, with green indicating the percentage of correct answers and red indicating the percentage of incorrect answers. The total number of correct answers out of the total number of questions is also displayed, offering users a clear overview of their performance. This quiz design aims to reinforce users' understanding of key concepts by providing multiple attempts and detailed feedback, thereby enhancing their learning experience.

3.4.2 Interviews. The interviews were conducted via Zoom meetings, with participants consenting to keep their cameras on, and were integral to gathering qualitative insights throughout the study. Conducted in a semi-structured format, the interviews took place after each of the three quiz phases, with additional questions asked during the interaction stages. In the first interview, participants were asked about their familiarity with recommender systems and relevant background knowledge. The second interview focused on their experience with the integrated AI-Chatbot, assessing the helpfulness of explanations, satisfaction levels, any encountered difficulties, and their learning process. The final interview explored how participants' understanding evolved, their overall learning experience, the usefulness of quiz questions, challenges in obtaining desired answers, and their perceptions of the AI-Chatbot as a co-learning agent.

3.4.3 Data Analysis. The interviews were recorded through Zoom's cloud system, where they were automatically transcribed by the platform. The first author reviewed the transcriptions and made slight adjustments for clarity and readability, ensuring the accuracy of participants' responses. Both authors then conducted a thematic analysis of the data, systematically reviewing the transcripts to identify emerging themes and patterns relevant to the study's focus on user comprehension of AI-driven recommendations. The analysis process involved multiple rounds of coding and discussion to ensure a thorough interpretation of the data within the context of the research. A MySQL database server logged quiz interactions for each participant, capturing user IDs, phase count, timestamps, selected answers for each question, and overall scores. Repeated Measures ANOVA was employed to evaluate differences in quiz scores across the three phases. This method was chosen as it accounts for repeated measurements on the same participants, controlling for individual variability and reducing error variance. Assumption checks for normality (Shapiro-Wilk test) and sphericity (Levene's test) were conducted to ensure the validity of the analysis.

3.4.4 Participants. 17 graduate students from four different universities in the United States were recruited for this study through email advertisement. Participants voluntarily agreed to take part in the study and signed informed consent forms prior to participation. The group consisted of individuals from diverse educational backgrounds, including Computer Science, Data Science, Management Information Systems, and Pharmaceutical Sciences. The participant pool was composed of 13 males and 4 females, representing a racially diverse group, with an age range of 25 to 38 years. Sixteen participants completed the entire study, while one participant did

not finish. The user studies ranged from 60 to 111 minutes per participant, with an average duration of 89 minutes. As compensation for their time, each participant received a \$20 Amazon gift card. This study received ethical approval from the Institutional Review Board (IRB) at our university, ensuring that it adheres to the highest standards of academic integrity and respect for human subjects.

4 Findings

4.1 Quantitative Findings

Firstly, we assessed the progression of participants' knowledge with the interaction of chatbot across three phases during the user study. Repeated Measures ANOVA revealed significant differences in quiz scores across phases ($F = 22.52$, $p < 0.001$), with paired t-tests confirming significant improvements between each phase. The post-hoc Tukey's HSD test showed no significant difference between Phase 1 and Phase 2 ($p = 0.0503$) but did reveal a significant difference between Phase 1 and Phase 3 ($p = 0.0004$). However, paired t-tests indicated a statistically significant increase in quiz scores from Phase 1 to Phase 2 ($t = -3.52$, $p = 0.0031$), suggesting that while the knowledge gain from the first treatment was smaller, it was still meaningful. Assumptions of normality ($p = 0.26$) and sphericity ($p = 0.135$) were met, validating the results. The analysis shows that participants' knowledge significantly improved across the three phases of the study. Both the Repeated Measures ANOVA and the paired t-tests demonstrated that quiz scores increased over time, particularly between Phase 1 and Phase 3. While there was a noticeable improvement from the baseline (Phase 1) to the second treatment phase (Phase 3), the results suggest that the first treatment (Phase 2) had a smaller but still significant impact on knowledge. The assumptions of normality and sphericity were met, supporting the validity of the ANOVA results.

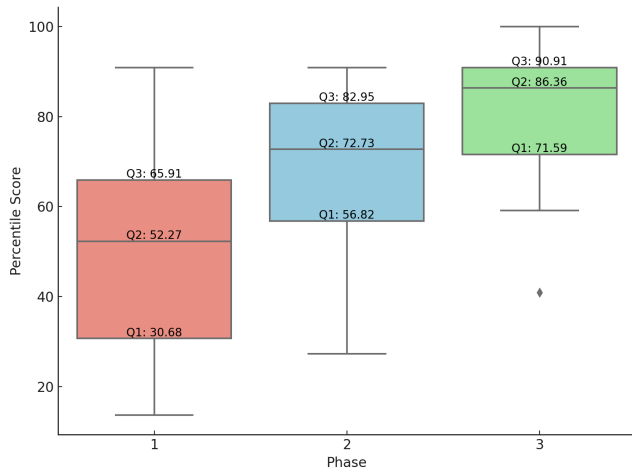


Figure 4: Users' score changes over the phases

Figure 4 illustrates the percentile score improvements of participants across the three phases of the study. The percentile scores, calculated as the ratio of a participant's quiz score to the total possible score (22), show a clear upward trend in knowledge acquisition. In Phase 1 (Baseline), the median score was 52.27%, with a

wide interquartile range (30.68% to 65.91%), reflecting lower overall performance and high variability among participants. In Phase 2 (Treatment 1), the median score increased significantly to 72.73%, with a narrower interquartile range (56.82% to 82.95%), indicating improved performance and more consistent knowledge levels. Phase 3 (Treatment 2) demonstrated the highest performance, with a median score of 86.36% and most participants scoring above 71.59%, with reduced variability. These findings demonstrate the effectiveness of AI-driven interventions in facilitating knowledge improvement, with Phase 3 yielding the most substantial gains and uniformity in participant performance.

Participant performance analysis revealed clear trends across the three difficulty levels of the quizzes. In the Foundational Knowledge Group participants demonstrated the strongest performance, with an average accuracy rate of 88.57%. This indicates a solid understanding of basic concepts and definitions. The Application and Integration Group saw a moderate drop in accuracy, with an average success rate of 68.17%. This reflects the increased cognitive demand of applying theoretical knowledge to practical scenarios, such as calculating cosine similarity or selecting preprocessing steps. The most challenging category, the Advanced Analysis Group had the lowest accuracy, with an average success rate of just 50.00%. These questions required advanced reasoning, highlighting the need for deeper algorithmic understanding among participants.

4.2 Qualitative Findings

4.2.1 Theme 1: The Approach. This theme reflects the strategies and best practices users employed to enhance their learning and comprehension when interacting with the generative AI-powered social recommender system. A key takeaway from participants' experiences is the importance of iterative learning and refining the way they ask questions to achieve more accurate and insightful responses.

Participants highlighted the iterative nature of asking questions as crucial for improving comprehension. As Participant 1 remarked, "If you don't know the right questions to ask, you're probably not going to get the right answers." This realization underscores the necessity of adjusting one's approach and gradually learning how to phrase inquiries more effectively to achieve clearer answers. Participant 6 echoed this sentiment, emphasizing that "once I actually understood what the real answers were and how to adjust my questions for answers, and how to refine what I was asking, it became clearer," reflecting a learning curve in interacting with the system.

Additionally, participants stressed the value of following along with the AI's reasoning process. Participant 3 described how "showing the steps in a calculation" allowed them to track where things might go wrong, which is essential for verifying the system's output. This aligns with the broader idea that the more transparent the AI's reasoning is, the easier it is for users to verify and trust the output. Consequently, many users adopted the practice of requesting explanations or step-by-step breakdowns to ensure correctness. As Participant 3 emphasized, "I should always ask it to show the work" to verify whether the system's output is accurate.

The ability to layer knowledge by asking follow-up questions and building on previously provided information also stood out as a powerful strategy. Participant 3 noted, "You learn a base concept,

and then you can immediately ask a follow-up or clarification, and it gives you more knowledge." This incremental learning process fostered a sense of progression and satisfaction, as described by Participant 11, who stated that *"being able to dive deeper with the chatbot can really enhance the learning experience."*

Finally, users appreciated how the chatbot could break down complex calculations or provide specific details on why recommendations were made, as Participant 13 noted: *"Explaining what each score was kind of helped me visualize that better because I prefer numbers."* This deeper dive into numerical reasoning offered users a richer context for understanding the system's decisions, allowing for more meaningful engagement with the recommendations.

4.2.2 Theme 2: Design Needs. The user studies revealed a clear need for design elements that promote user comprehension and engagement with AI-driven social recommender systems. Participants emphasized the value of reminders, confirmations, and real-time prompts to support their understanding, particularly when they encounter uncertainty or knowledge gaps.

Participant P1 described the desire for a conversational system that could guide users through their thought process when they arrive at incorrect answers. This *dialectical system* would provide explanations and corrections, enabling users to better grasp the logic and learn more effectively. P1 explained, *"You got this answer wrong. Okay, why did you get that answer wrong? Here's what my thought process was. Then it would be able to say, 'Okay, I understand why you're thinking that way. But here's where the logic goes wrong. Here's what you need to think instead because this logic is more fluid and accurate.' And this is why it's the right answer."* Such feedback would not only correct mistakes but also encourage deeper cognitive engagement, helping users develop more fluid, accurate reasoning. Participants like P2 highlighted the importance of immediate feedback, suggesting that AI should assist in pinpointing gaps in understanding without delay. P2 said, *"It will be helpful for identifying where we are missing something, so you don't have to wait for hours to find out what's missing. It will assist you right away."* This real-time support would streamline learning, reducing the time spent on confusion and enhancing the learning experience. Participant P3 envisioned a co-learning agent that could provide real-time definitions or explanations during lectures, allowing users to remain engaged and avoid missing key concepts. According to P3, *"If you're sitting in class and the teacher says a word you don't know, it would be helpful if a bot was listening with closed captioning on. You could click on the word, and the definition would pop up. That way, you wouldn't fall behind or miss anything because you didn't understand a term."* By offering quick, accessible information at the point of need, AI could prevent knowledge gaps from growing, helping learners stay on track.

The need for comprehensive responses was emphasized by P12, who expected AI-Chatbots to deliver complete, detailed answers to queries. P12 noted, *"If I want to know something and there's a lot of similarities, I expect the AI-Chatbot to give me all the information I need when I type a query. If it only provides a little bit of information, there might be some details I don't know."* Partial responses could leave users with lingering confusion, underscoring the necessity for thorough explanations that cover all relevant information. Finally, participants like P13 and P15 reflected on the importance of

active learning techniques, such as writing down information to reinforce memory, and the need for AI systems to help with recall and decision-making when faced with uncertainty. P13 mentioned, *"If I had written things down, it probably would have been a lot more helpful, instead of just reading and trying to memorize, because memorization doesn't always work."* Similarly, P15 expressed, *"I could remember tokenization and vector normalization, but I wasn't sure about stop word removal. That's why I was thinking it should be all of the above."* They indicated that AI should provide both reminders and contextual prompts to guide users through complex topics, ensuring they can make informed decisions without second-guessing their understanding.

4.2.3 Theme 3: The Evidence. This theme highlights how users' comprehension improved through interactions with AI-generated explanations, particularly by asking follow-up questions to clarify their understanding. Participants described how compression of information, combined with logical deductions and iterative questioning, helped them navigate the AI system's recommendations more effectively. Their responses indicate a gradual improvement in understanding, with the AI providing additional layers of context as needed.

For example, Participant 1 reflected on their learning process, noting significant growth in comprehension over time. Initially, their understanding was minimal, relying largely on guesswork: *"If I look back and compare my progress, the first time around it was all guesswork, with 0% understanding. I just made logical assumptions and deductions."* However, by asking targeted questions, their ability to comprehend the material improved: *"The second time, I was able to answer based on the limited number of questions I could ask, using quite a bit of logic—maybe around 50%."* They noted that their understanding eventually reached *"about 80% compared to when I first started."* This highlights the role of iterative inquiry in deepening understanding, especially when initial explanations were insufficient.

Participant 8 underscored the value of clarification in their understanding of technical concepts: *"I asked this question to confirm if it's the same question about how cosine similarity would be calculated. So, that's why I asked, and yes, it's the same thing. Based on this, I think I now understand how it will be calculated."* This demonstrates the importance of follow-up questions to validate assumptions and achieve clarity.

Participant 11 emphasized how detailed responses from the AI reinforced their existing knowledge. They felt that the system provided comprehensive descriptions of concepts they were already familiar with: *"It gave me a more detailed description of something I thought I knew the answer to, while also verifying if the AI system's response matched mine in my area of expertise."* The overlap between the AI's recommendations and the participant's domain knowledge further enhanced their understanding: *"Even without reading the full articles, I could see there was a lot of overlap with [...]."*

Similarly, Participant 13 reported a better grasp of the underlying mathematical principles after interacting with the system: *"I have a better understanding of the math and calculations going on behind the scenes regarding the similarities in topic and text that I didn't have complete context on before."* The system provided much-needed

context, which improved their understanding of why specific recommendations were made.

Lastly, P16 highlighted the importance of clarity in understanding the strength and ranking of recommendations. By asking what specific scores meant and where they originated, the user was able to gain a clearer picture of the logic behind the recommendation: *"When I asked what that particular score means and where it comes from, it made things a lot clearer to me."* This improved their trust in the system, underscoring the significance of transparency in AI-generated explanations and how it enhances user comprehension.

4.2.4 Theme 4: User Mental Models. The user study revealed how participants' mental models were influenced by their prior knowledge, particularly in mathematical concepts and technical terminology, and how they adapted these models when interacting with the AI-driven recommendation system.

Participants with a background in statistics or technical fields leveraged their existing knowledge to make sense of the recommendations. For example, P3 mentioned using their *"statistical background"* to logically eliminate options in a multiple-choice format, relying on intuition and prior experience to make the most reasonable choice. Similarly, P13 described using a rough estimation approach to interpret numerical outputs, stating that *"closer to 1 means a perfect match"* and using this understanding to navigate the recommendations.

However, participants without prior familiarity with certain technical terms, like P7, found themselves needing additional support. P7 noted that *"technical jargons"* made the process challenging, but acknowledged the AI bot's role in bridging the gap: *"the AI bot helped me to a greater extent within a shorter span of time."* This demonstrates how generative AI can assist users in rapidly developing an understanding, even in areas they are unfamiliar with.

Several participants, including P15 and P16, actively sought to expand their mental models by researching unfamiliar concepts. P15, for instance, encountered terms like cosine similarity, which they didn't understand initially. By searching for explanations and then returning to the AI bot for further clarification, P15 was able to grasp the underlying concept, stating, *"I searched for it, learned exactly what it is."* Similarly, P16 expressed a desire to comprehend the recommendation process in greater depth, initially assuming that it was based solely on keyword matching but later discovering the more complex data science principles involved.

5 Discussion

The ANOVA results and boxplot analysis of participants' percentile score improvements across the three phases, based on quantitative findings, suggest that users' comprehension improved over time with increased interaction with the AI chatbots and a better understanding of the explanations (RQ1). The qualitative findings further support this conclusion, revealing that participants who actively refined their questioning strategies and sought clarifications were better able to understand complex recommendations. As reflected in themes such as The Approach and The Evidence, iterative questioning, combined with layered learning, allowed users to progressively build their knowledge. Participants highlighted the importance of tailoring questions to extract clearer explanations and engaging with the reasoning process behind the AI's

recommendations. For instance, Participant 3's remark, *"I should always ask it to show the work,"* illustrates how prompting the AI for transparency fostered a clearer understanding of the system's logic. Thus, comprehension appears to be an outcome of both interacting with the AI-Chatbot and the ability to ask effective questions, confirming that users' active participation is essential to understanding AI-driven recommendations.

The study identified key behavioral differences between users who effectively comprehend AI-generated explanations and those who do not (RQ2). The most pronounced difference is related to the users' approach to interacting with the chatbot. Users who demonstrated higher comprehension levels were more likely to engage in iterative questioning and actively seek clarification when encountering ambiguities. Participants who lacked technical expertise initially struggled to understand explanations but gradually improved their comprehension by refining their questioning methods and exploring the reasoning behind the recommendations. This iterative process, highlighted by several participants, is critical in helping users develop an accurate mental model of the recommendation system's logic. Another behavioral difference is evident in users' approaches to understanding the technical keywords. Those with less domain knowledge, such as P7 and P15, augmented their understanding by researching unfamiliar terms and returning to the chatbot for clarification. In contrast, participants who did not engage in such practices or relied on surface-level answers exhibited less improvement in their quiz scores and often struggled to make sense of the recommendations. Additionally, participants with a background in technical fields demonstrated greater ease in navigating the recommendations by applying prior knowledge to interpret the AI's outputs. For example, P3 used their statistical background to make logical deductions when faced with multiple-choice questions, while P13 leveraged their knowledge of cosine similarity to grasp the system's ranking process. These users employed a mental model that allowed them to comprehend AI explanations more easily, underscoring the importance of prior knowledge in enhancing understanding. However, even participants without such expertise were able to improve their understanding by engaging in clarifying interactions with the chatbot, illustrating that comprehension is achievable with appropriate support mechanisms. This behavior highlights that users who are proactive in bridging knowledge gaps through further interactions with the AI-Chatbot tend to achieve a deeper comprehension of the explanations, emphasizing the importance of engagement and curiosity in enhancing understanding.

Design Implications The findings suggest several design improvements to enhance user comprehension in AI-driven social recommender systems. First, the inclusion of real-time feedback and prompts, as suggested by participants, can help users identify and correct misunderstandings more efficiently. For instance, a system that offers immediate feedback, as envisioned by P1, could guide users through their thought process and correct logical errors in real-time. Additionally, providing detailed step-by-step explanations and context-aware prompts, especially for users without technical backgrounds, can facilitate a more comprehensive understanding of recommendations. Participants also highlighted the need for conversational features that allow users to follow along with the AI's reasoning, ask for clarification, and progressively build their knowledge. Designing systems that encourage iterative

questioning and offer layered explanations will likely improve comprehension, as evidenced by participants' positive experiences when using such strategies. Furthermore, the integration of active learning techniques, such as asking follow-up questions or visualizing concepts through step-by-step breakdowns, could support users in deepening their understanding of complex topics and AI reasoning processes. In summary, we propose the following recommendations for designing systems that prioritize user comprehension and engagement:

- **Promote Active User Engagement:** Encourage users to ask relevant follow-up questions by designing chatbot interfaces that prompt clarifying queries or provide examples of effective questions. This can enhance comprehension by fostering an interactive dialogue tailored to the user's needs.
- **Adapt Explanations to User Backgrounds:** Leverage user profile data, such as educational background or familiarity with AI systems, to dynamically adjust the complexity and tone of explanations. For instance, users with limited technical knowledge may benefit from simpler, more intuitive explanations.
- **Provide Contextual Cues:** Include visual or textual cues within the chatbot interface to highlight key aspects of the recommendation rationale, such as the primary factors influencing a suggestion. This can improve clarity and help users focus on critical information.
- **Evaluate Comprehension Iteratively:** Integrate real-time comprehension checks, such as brief quizzes or feedback prompts, to assess how well users understand explanations and refine the system's responses accordingly.

6 Conclusion

This study demonstrates that user comprehension of AI-generated explanations is closely tied to the depth of interaction and engagement with the system. Behavioral differences between users who comprehend AI explanations and those who do not stem from the iterative nature of questioning, AI-Chatbot utilization, and reliance on prior knowledge. To enhance user comprehension, AI-driven recommender systems should incorporate design elements that encourage iterative learning, provide real-time feedback, and offer comprehensive, context-aware explanations. These improvements will not only foster better user understanding but also increase trust in AI-generated recommendations, making such systems more effective for a broader range of users.

Our study provides valuable insights into the role of generative AI in enhancing the comprehensibility of explanations in social recommender systems, but several limitations must be acknowledged. First, the small sample size of 17 participants limits the robustness and generalizability of our findings. A larger and more diverse sample is needed to strengthen the validity of the results. Second, the study design introduces potential learning and ordering biases, as the incremental mixed-method approach may have influenced participants' performance across phases. Although our preliminary study [6] justified a non-randomized order due to the high cognitive demand of the task, we recognize that this may introduce confounding factors. To mitigate memorization effects, we implemented mechanisms such as shuffling the order of the

22 quiz questions and their answer choices across phases, alongside incorporating chatbot interactions and follow-up interviews. These measures facilitated observable knowledge improvement, as supported by qualitative and quantitative data. However, learning and ordering biases may still persist, and future studies employing between-subject designs with larger sample sizes and randomized conditions are necessary to obtain more definitive results. Third, our findings are specific to social recommender systems, and their applicability to other recommendation contexts may be limited.

Future research should focus on implementing and testing design strategies to enhance user comprehension in AI-driven recommender systems by promoting active engagement through chatbot interfaces that prompt clarifying questions, adapting explanations to users' educational backgrounds and AI familiarity for tailored simplicity, incorporating visual or textual cues to highlight key recommendation rationales, and integrating iterative comprehension checks like real-time quizzes or feedback prompts to refine system responses.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2153509.

References

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [2] Sabrina G Anjara, Adrianna Janik, Amy Dunford-Stenger, Kenneth Mc Kenzie, Ana Collazo-Lorduy, Maria Torrente, Luca Costabello, and Mariano Provencio. 2023. Examining explainable clinical decision support systems with think aloud protocols. *Plos one* 18, 9 (2023), e0291443.
- [3] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2020. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848* (2020).
- [4] Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5277–5285.
- [5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Ben- netot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [6] Md Ashaduzzaman, Thi Nguyen, and Chun-Hua Tsai. 2024. Explaining Social Recommendations Using Large Language Models. In *International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence*. Springer, 73–84.
- [7] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 248–266.
- [8] Judy Borowski, Roland S Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA Wallis, Matthias Bethge, and Wieland Brendel. 2020. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606* (2020).
- [9] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [10] Zana Bućinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.
- [11] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.
- [12] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. Xai meets llms: A survey of the relation between explainable ai and large language models. *arXiv preprint arXiv:2407.15248* (2024).
- [13] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the

- potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter* 25, 2 (2024), 42–61.
- [14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
 - [15] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 307–317.
 - [16] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 510–518.
 - [17] Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, et al. 2022. A comprehensive survey on trustworthy recommender systems. *arXiv preprint arXiv:2209.10117* (2022).
 - [18] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.
 - [19] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. 2020. On cognitive preferences and the plausibility of rule-based models. *Machine Learning* 109, 4 (2020), 853–898.
 - [20] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.
 - [21] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in neural information processing systems* 32 (2019).
 - [22] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
 - [23] Lijie Guo, Elizabeth M Daly, Ozmur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building trust in interactive machine learning via user contributed interpretable rules. In *Proceedings of the 27th international conference on intelligent user interfaces*. 537–548.
 - [24] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).
 - [25] Lukas-Valentin Herm, Jonas Wanner, Franz Seubert, and Christian Janiesch. 2021. I Don't Get IT, but IT seems Valid! The Connection between Explainability and Comprehensibility in (X) AI Research.. In *ECIS*.
 - [26] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
 - [27] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
 - [28] Anthony Jameson, Martijn C Willemsen, Alexander Felfernig, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. Human decision making and recommender systems. *Recommender systems handbook* (2015), 611–648.
 - [29] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Walach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
 - [30] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 84–88.
 - [31] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726* (2022).
 - [32] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
 - [33] Yang Liu. 2023. The importance of human-labeled data in the era of LLMs. *arXiv preprint arXiv:2306.14910* (2023).
 - [34] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374* (2023).
 - [35] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786* (2021).
 - [36] Tauseef Ibne Mamun. 2021. *Investigating the Impact of Online Human Collaboration in Explanation of AI Systems*. Master's thesis. Michigan Technological University.
 - [37] Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. 2024. XAI for all: Can large language models simplify explainable AI? *arXiv preprint arXiv:2401.13110* (2024).
 - [38] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
 - [39] Sidra Naveed, Gunnar Stevens, and Dean Robin-Kern. 2024. An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI. *Applied Sciences* 14, 23 (2024), 11288.
 - [40] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.
 - [41] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
 - [42] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48 (2020), 137–141.
 - [43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
 - [44] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* (2023).
 - [45] Hanna Schäfer, Santiago Hors-Fraile, Raghav Pavan Karumur, André Calero Valdez, Alan Said, Helma Torkamaan, Tom Ulmer, and Christoph Trattner. 2017. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health*. 157–161.
 - [46] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.
 - [47] Jyoti Shokeen and Chhavi Rana. 2020. A study on features of social recommender systems. *Artificial Intelligence Review* 53, 2 (2020), 965–988.
 - [48] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do users benefit from interpretable vision? a user study, baseline, and dataset. *arXiv preprint arXiv:2204.11642* (2022).
 - [49] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. 2022. Intuitively assessing ml model reliability through example-based explanations and editing model inputs. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 767–781.
 - [50] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner: Extraction and Mining of Academic Social Networks. In *KDD'08*. 990–998.
 - [51] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31 (2021), 591–627.
 - [52] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
 - [53] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).
 - [54] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288* (2021).
 - [55] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. 2018. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 1563–1570.
 - [56] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
 - [57] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).