

Contextualizing Recommendation Explanations with LLMs: A User Study

Yuanjun Feng¹, Stefan Feuerriegel², Yash Raj Shrestha¹

¹University of Lausanne, Switzerland

²LMU Munich, Munich Center for Machine Learning (MCML), Germany
{yuanjun.feng, yashraj.shrestha}@unil.ch, feuerriegel@lmu.de

Abstract

Large language models (LLMs) are increasingly used in recommender systems to generate natural-language explanations for contextualized recommendations. This study examines how different types of LLM-generated explanations shape user perceptions and behavioral intentions in the context of movie recommendations. In a between-subjects online experiment ($N = 759$) and follow-up interviews ($N = 30$), we compare LLM-generated (a) generic explanations and (b) contextualized explanations. We find that contextualized explanations better address users' cognitive and affective needs and strengthen intention to watch the recommended movies. Interview findings further show that the effectiveness of contextualization depends on how well explanations align with users' genre and narrative preferences. References to prior movies can improve user perceptions and intent, but excessive detail may make explanations feel redundant. The benefits of contextualization are pronounced among users with higher engagement and greater trust in the system. Together, these findings highlight the potential of LLMs to support more user-centered recommendation experiences and offer practical guidance for designing explanations that improve trust, relevance, and engagement in entertainment platforms.

Code — <https://github.com/diana3135/Contextualizing-Recommendation-With-LLMs>

1 Introduction

Recommender systems influence human decision-making in daily life, such as purchasing products on e-commerce platforms and choosing movies on streaming services (Angelov et al. 2021; Goyani and Chaurasiya 2020). However, users often fail to understand why they receive specific recommendations. On streaming platforms like Netflix and Disney+, users are presented with recommended movies without any explanation of why those movies might be of interest. Without such explanations, users may find it difficult to judge whether the recommendation is helpful and decide on their consumption. Explanations help users make choices that reflect their preferences and improve decision quality (Vultureanu-Albisi and Badica 2021). As a result, there is a growing research interest in offering explanations behind recommendations (Ko et al. 2022; Zhang and Chen 2020).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The emergence of large language models (LLMs) presents potential for improving recommendation explanations (Wu et al. 2024; Hou, Leach, and Huang 2024). Some studies have adopted LLMs to produce natural and human-like explanations. However, current efforts often rely on generic explanation templates (Li, Zhang, and Chen 2020). These generic explanations overlook users' contexts or preferences, thereby limiting the benefits of the explanations and underutilizing the LLMs' reasoning capabilities. Existing research seldom provides a comprehensive understanding of how users perceive the received recommendations. Current work also rarely examines whether positive user perceptions translate into users' consumption intentions.

Motivated by this gap, our study adopts established explanation measurement standards (Balog and Radlinski 2020) and interprets user responses through the lens of the **Uses and Gratifications (U&G) framework** (Katz, Blumler, and Gurevitch 1973; Lin et al. 2013). The U&G framework categorizes users' diverse needs, including the desire for understanding and knowledge acquisition (**Cognitive needs**), the need for emotional connection and engagement (**Affective needs**), and the focus on practicality and task-oriented goals (**Utilitarian needs**). This framework provides a valuable perspective to understand how users actively select, consume, and engage with the received information. Our goal is to examine how different LLM-generated recommendation explanation types resonate with specific user needs and, in turn, influence their consumption intentions. Therefore, we designed the following research question:

How do LLM-generated explanations respond to users' diverse motivational needs (cognitive, affective, utilitarian) and affect their consumption intentions?

In this study, we conducted a pre-registered, between-subjects online experiment involving 759 subjects from the U.S. in the context of movie recommendations.¹ By drawing upon the U&G framework, we analyze users' perceptions and intentions towards three types of recommendations generated by GPT-4 Turbo: (1) recommendations with no explanations (**NE**), (2) recommendations with LLM-generated generic explanations (**GE**), and (3) recommendations with LLM-generated contextualized explanations (**CE**).² We also

¹The experimental protocol was pre-registered at OSF: <https://osf.io/j46vx/overview>

²We focus on real-world recommendation tasks based on view-

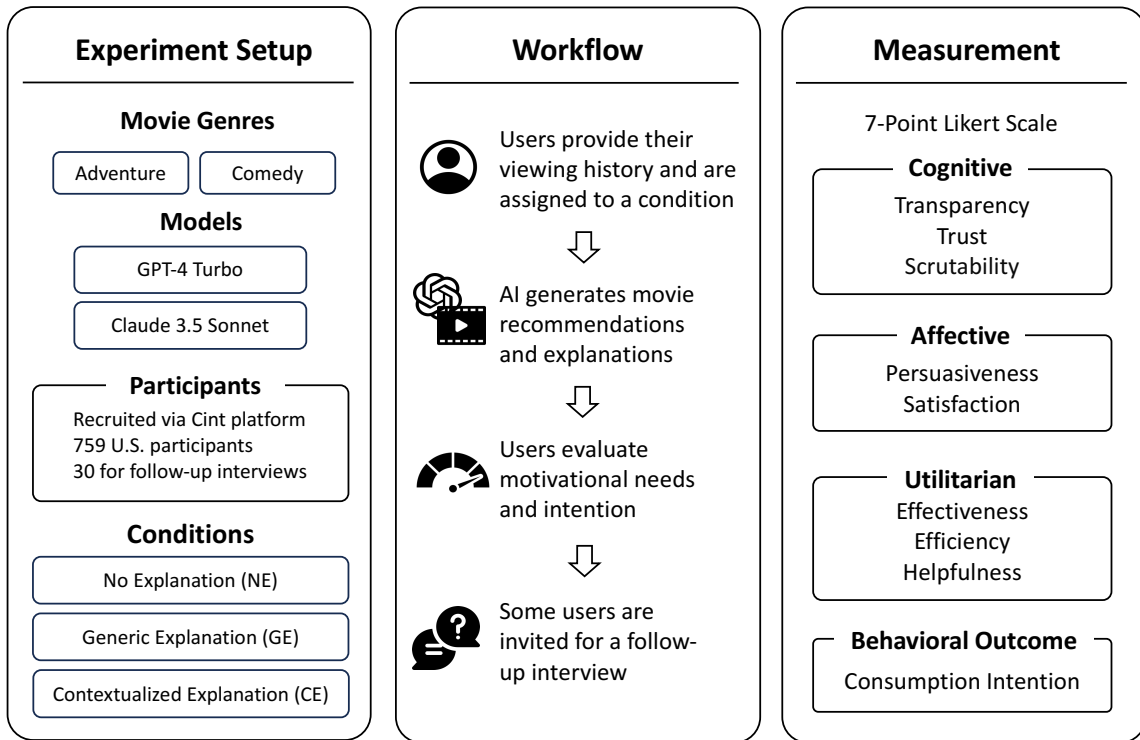


Figure 1: Overview of the study’s methodology, including the experimental setup, workflow, and the evaluation metrics.

drew upon user profile information and conducted follow-up interviews ($N = 30$) to interpret our results more deeply. To ensure the robustness of our study design and the generalizability of the results, we conducted supplementary experiments by using a different LLM (Claude 3.5 Sonnet) with an additional 140 users from the U.S. Figure 1 provides an overview of our study flow.

Contributions: This paper demonstrates how LLM-generated, contextualized explanations can more effectively address users’ diverse motivations and increase their consumption intentions. By integrating the U&G framework, our study shows how different explanations affect users’ cognitive, affective, and utilitarian needs. Moreover, we highlight the amplifying role of user engagement with the recommender system, revealing that users who already find the system beneficial gain the most from explanations. These findings bridge advancements in LLMs with user behavior and provide actionable insights to enhance user-centric recommender systems in social media contexts.

ing histories without relying on explicit or private user preferences (e.g., as in cold-start problems of recommendation websites). As such, we use the term *contextualized* rather than *personalized* throughout the paper. This terminology aligns with related work (Bar, Maarouf, and Feuerriegel 2024) and reflects practical usage on platforms such as IMDb, Yelp, and Google Maps, where recommendations are often informed by user activity (e.g., viewing or browsing history) but not by direct feedback like ratings.

2 Related Work

2.1 LLM-Generated Explanations in RecSys

In social media and online communities, recommendations help users find items aligned with their preferences. Initial research centered on improving predictive accuracy (Ma et al. 2023; Wu et al. 2023). However, growing attention to user trust and transparency has driven a shift toward explainable recommendations (Burkart and Huber 2021). Recent work emphasizes explainability to foster trust, transparency, and user satisfaction (Burkart and Huber 2021). Traditional recommendation techniques (e.g., matrix factorization with interpretable layers (Abdollahi and Nasraoui 2017) or attention-based neural models (Seo et al. 2017; Chen et al. 2018)) also integrated explainability.

The advent of LLMs has made it easy to generate explanations in natural language more flexibly. Building upon template-based methods (Sarwar et al. 2001; Pazzani and Billsus 2007; Yang 2016), studies have used LLMs to generate detailed item descriptions (Acharya, Singh, and Onoe 2023), visual explanations (Tsai and Brusilovsky 2019), sentiment-oriented narratives (Chen, Yan, and Wang 2019), and interactive user-friendly justifications (Deldjoo 2025). Previous works have also shown the unique potential of natural language explanations in mitigating information overload and facilitating community integration (Chang, Harper, and Terveen 2016). This growing trend of work underscores the potential for LLM-driven recommendation strategies in strengthening the recommendation processes.

2.2 User Perceptions of Explanations

Users increasingly rely on recommender systems to navigate overwhelming volumes of information. Grounded in the Uses & Gratifications framework, individuals seek to satisfy their cognitive, affective, and utilitarian needs when engaging with social media content (Katz, Blumler, and Gurevitch 1973; Korhan and Ersoy 2016; Bashardoust, Feuerriegel, and Shrestha 2024). Accordingly, explanations that clarify *why* a recommendation is made can reduce cognitive effort, foster trust, and enhance decision confidence (Mandl et al. 2011; Jameson et al. 2015).

Extant research also identifies several factors shaping user perceptions of explanations. For instance, users may prefer human-generated recommendations, particularly for hedonic products (Wien and Peluso 2021). Explanations that reflect prior user feedback or behavior tend to resonate with individuals seeking tailored, value-reinforcing content (Lu et al. 2023). Moreover, individual differences in user profiles, such as cognitive engagement level, highlight the importance of detailed and transparent justifications for users who prioritize control and understanding (Chatti et al. 2022). Therefore, aligning explanation strategies with varied user motivations and perceptions is essential for designing effective and user-centric recommender systems.

2.3 Empirical User Studies on Explanations

Empirical research has analyzed how different explanations in recommender systems impact users (see Table 1). These studies typically use controlled experiments and surveys to evaluate diverse explanation strategies. Silva et al. (2024) conducted a user study ($N = 94$) comparing AI-generated generic and personalized movie explanations. While both were well-received, personalization improved perceived effectiveness when recommendations were less relevant and enhanced persuasiveness in low-confidence scenarios. Similarly, Lubos et al. (2024) ran an online user study ($N = 97$), finding that LLM-generated explanations outperformed template-based baselines in trust and satisfaction due to their contextual richness and fluency. In another user study of a fitness plan recommender system ($N = 341$), Sun et al. (2023) showed that users trusted recommendations less when explanations revealed reliance on their social-media friends' activities, citing privacy and identity concerns. Instead, users preferred personalization based on their behavior. Together, these empirical findings provide insights into user perceptions and inform system design, but a comprehensive evaluation analyzing the cognitive, affective, and utilitarian needs is missing.

2.4 Research Gap

While the importance of recommendation explanations is widely acknowledged, research seldom examines how users interpret and respond to them. Furthermore, extant studies often rely on generic templates and have not fully explored the potential of LLM-generated contextualized explanations to address users' cognitive, affective, and utilitarian needs (see Table 1). To address this gap, our study provides the first comprehensive user evaluation of LLM-generated contextualized explanations through the lens of the U&G framework.

By examining how these explanations resonate with users' multifaceted needs and influence their intentions, we aim to offer deeper insights into designing recommendation explanations that are more aligned with user motivations.

3 Methodology

3.1 Overview

We conducted an online between-subjects experiment, where 759 subjects were randomly assigned to one of three experimental conditions: (1) *recommendations with no explanations (NE)*, (2) *recommendations with LLM-generated generic explanations (GE)*, and (3) *recommendations with LLM-generated contextualized explanations (CE)*.

In the experiment, users first selected movies they had previously watched from a given set to establish the viewing history, based on which LLMs generated movie recommendations and explanations depending on the experimental condition.³ After receiving the corresponding explanation, users provided a series of ratings reflecting how well the explanation met underlying motivational needs, as well as their resulting behavioral intentions:

- **Cognitive needs:** reflecting users' desire for understanding and credibility in the recommendation process, assessed through *Scrutability*, *Transparency*, and *Trust*.
- **Affective needs:** addressing the emotional and motivational dimensions of the user experience, evaluated through *Persuasiveness* and *Satisfaction*.
- **Utilitarian needs:** focusing on practical, goal-oriented aspects of the recommendation, captured by *Effectiveness*, *Efficiency*, and *Helpfulness*.
- **Consumption intentions:** reflecting users' behavioral outcome, captured by *UserIntent*.

Dimensions composing the needs and intentions are established for assessing explanations in prior user studies (Balog and Radlinski 2020). We organize the dimensions under the U&G framework to depict users' active pursuit of various needs from explanations. Given the statement for each dimension, users rated their agreement on a 7-point Likert scale from 1 ("Extremely Disagree") to 7 ("Extremely Agree"). The higher-level intentions and needs are averaged on the specific dimensions (e.g., *Cognitive* is the average of *Scrutability*, *Transparency*, and *Trust*). A summary of the main variables and survey questions is in Table 2. After the survey, we invited 30 users to conduct follow-up interviews to see their latent motivations across different explanations.

3.2 Subjects

Subjects were recruited from the U.S. via Cint.⁴ The online platform offers a built-in functionality for stratified sampling of subjects so that the sample is representative of the U.S.

³We used GPT-4 Turbo (gpt-4-1106-preview), which has knowledge of world events up to April 2023. For the experiment in the supplement, we used Claude 3.5 Sonnet (claude-3-5-sonnet-20240620).

⁴<https://www.cint.com/>

	Kunkel et al.	Silva et al.	Lubos et al.	Our study
Contextualized explanations		✓	✓	✓
Generic explanations	✓	✓	✓	✓
Cognitive needs	✓		✓	✓
Affective needs	✓	✓		✓
Utilitarian needs		✓	✓	✓
Consumption intentions	✓			✓

Table 1: Comparison between previous studies (Kunkel et al. 2019; Silva et al. 2024; Lubos et al. 2024) and our work.

Variable	Questions
COGNITIVE NEEDS	
<i>Transparency</i>	The text helps me to understand what the recommendation is based on.
<i>Trust</i>	The text helps me to trust the recommendation.
<i>Scrutability</i>	The text would allow me to give feedback on how well my preferences have been understood.
AFFECTIVE NEEDS	
<i>Persuasiveness</i>	The text makes me want to watch this movie.
<i>Satisfaction</i>	The text would improve how easy it is to pick a recommendation.
UTILITARIAN NEEDS	
<i>Effectiveness</i>	The text helps me to determine how well I will like this movie.
<i>Efficiency</i>	The text helps me to decide faster if I will like this movie.
<i>Helpfulness</i>	How helpful do you find the recommendation?
BEHAVIORAL OUTCOME	
<i>UserIntent</i>	How willing are you to watch this recommended movie?

Table 2: Summary of main variables and survey questions.

population. Subjects are remunerated with a fixed amount of around \$15 per hour.

As preregistered, we conducted a pilot study involving 50 subjects to assess the survey’s functionality and design. Informed by other user studies (Kang 2021; Lakens 2022), we performed a power analysis using G*Power 3.1 software (Faul et al. 2007, 2009). For multiple comparisons, we selected F -tests and chose a one-way ANOVA design with fixed effects (Lee and Lee 2018). We set the significance level (α) to 0.05 and the power ($1-\beta$) to 0.95, with the adjusted effect size set at 0.1. The power analysis based on the pilot study results suggested that we need at least 750 subjects for the experiment.

Initially, 1401 subjects completed the survey. We discarded responses from subjects failing an initial attention check. Subsequently, following prior research (Pennycook,

Cannon, and Rand 2018) and our pre-registration, a subset of subjects were excluded for three reasons: (i) subjects who self-reported to have answered randomly, (ii) failed the second attention check, (iii) responded excessively fast (i.e., the duration was less than around 6 seconds for responding to the user perceptions). These users still received remuneration after the exclusion. Eventually, due to our strict exclusion criteria, 759 of the 1,401 subjects qualified, exceeding our estimated sample size of 750 to ensure the results achieved the desired statistical power.

Among the 759 subjects, 387 are women, 370 are men, and two are non-binary. The sample has a mean age of 52.85 ($SD = 16.01$) and a range of 70 years (18 to 88). Most subjects have at least a high school diploma or equivalent.

3.3 Materials

We selected GPT-4 Turbo to generate movie recommendations for the main experiment. We conducted an additional experiment with Claude 3.5 Sonnet to validate the robustness of our study design and the generalizability of our findings (see Appendix C). We compiled all recommendations and explanations into a corpus to ensure consistent outcomes across different LLM calls. The corpus is generated based on the pre-defined strategy and prompts below. Therefore, users with the same viewing history and assigned conditions received the identical recommendation explanation.

Recommendation Strategy We implemented a sequential recommendation strategy following the algorithm in Table 5 in Appendix A. To reduce bias, users only rated movies they had not watched. For each user, we began the recommendation process with 10 distinct popular movies from the IMDb Top 250 list (IMDb.com, Inc. 2024b). If the user had already watched a movie, we proceeded to new recommended movies until they reached an unwatched one. The recommended movies watched by users were also added to their viewing history and used to generate new recommendations. For example, if a user has watched *Movie A* before, LLMs will recommend *Movie B*. If the user has watched it as well, then LLMs will combine both movies in the explanation for *Movie C*. This process continued until either an unwatched recommendation was found or a maximum of 4 recommendations were generated.

Rather than relying on explicit subjective feedback (e.g., likes or dislikes), we collected information about whether users had previously watched each movie. This design simulates a cold-start scenario (Lam et al. 2008; Chaimalas et al.

2023), where prior viewing history is unavailable and where the system starts with general popular choices before adapting to observed behavior. This approach mirrors real-world conditions, where explicit user feedback is often sparse or unavailable.

Recommender websites (such as IMDb, Rotten Tomatoes, Google Maps, Yelp, and TripAdvisor) are often user-independent and free of personalization, either because they lack access to user-specific information or because personalization is unwanted. They only have access to consumption data from their registered users, that is, actual review contributors. Moreover, the main selling point of recommender websites like IMDb and Yelp is the comparison of products or services. As a result, users are not required to sign in to use the website. Forcing users to register on the website would probably be counterproductive and decrease activity, as users might not want to sign up due to privacy reasons. Consistent with these, we collect users’ viewing histories, rather than explicit preferences, to generate recommendations and explanations.

Recommendation Generation We followed previous practices (Giray 2023) when designing the prompts. Table 3 shows the prompts used to generate different explanations. Figure 2 shows examples of contextualized and generic explanations. All explanations were limited to 100 words, aligning with the length of typical movie descriptions. To manage reading speed and cognitive load, each explanation was designed to include around three key features, such as the director, cast, plot, or scenes. Further, the length of the explanations was similar across the different conditions to ensure comparability.

The main difference between CE (contextualized explanations) and GE (generic explanations) is in the explicit connection: while both explanations incorporate various elements (e.g., feature, popularity, similarity) to support the recommendations, the contextualized explanations naturally integrate the recommended movies with the user’s viewing history, explicitly highlighting how the user’s inputs shape the recommendation. Meanwhile, the generic explanations focus exclusively on the recommended movies. All generated movies were cross-checked with the IMDb dataset (IMDb.com, Inc. 2024a) to ensure no hallucinations from the LLMs. Additionally, the corpus was validated by two experts with strong expertise in the field to ensure that all explanations used in the experiment were consistent and of fair quality.

3.4 Procedure

During the experiment, we collected users’ viewing history and assigned them to one of our three conditions. Based on the condition assignment and viewing history, we generated the LLM-generated recommendations for them and collected their evaluations. The entire procedure was conducted online, consisting of the following steps:

Step 1. Users visit a welcome page where they provide informed consent to proceed with the experiment.

Step 2. Users read a short paragraph about the movie recommendation and respond to an attention check. Users who

System Prompt

You are a movie recommendation assistant. Aim for clear, straightforward language that a broad audience can easily understand.

Output each recommendation into three segments: The movie’s details, formatted as “Title - Release Year - Director”; A concise summary within 40 words that outlines the storyline; An explanation within 100 words.

The explanation should focus on concrete aspects such as specific character arcs, notable plot developments, or memorable scenes.

User Prompt - CE

For {Movie}, provide one recommendation explanation for a user with a viewing history: {User viewing history}.

Highlight how the recommendation directly appeals to the provided viewing history.

User Prompt - GE

For {Movie}, write a recommendation explanation.

Table 3: System and user prompts used for generating recommendation explanations.

fail to pass are redirected to the end of the survey. The attention check also serves to assess English proficiency.

Step 3. Users read about the purpose and functionality of a recommender system.

Step 4. Users select movies they have seen from the movie shortlist \mathcal{M} .

Step 5. Users provide the frequency of their movie consumption, their usage frequency, and the perceived helpfulness of recommender systems.

Step 6. Users are randomly assigned to one of our three conditions and shown corresponding recommendations.

Step 7. Following each recommendation, users express their perceptions across the dimensions of cognitive, affective, and utilitarian needs on a 7-point Likert scale. Users also state the recommendation’s helpfulness and express their intentions to consume the movie. Steps 6 and 7 are repeated for all movies selected by users from \mathcal{M} .

Step 8. Users go through a second set of attention checks.

Step 9. Users provide demographic information, i.e., age, gender, and education.

Step 10. Users participate in an honesty test where they are asked whether they answered all questions honestly, while being guaranteed that compensation would not be contingent on their answers.

Step 11. Users receive a debrief of the experiment and are informed that LLMs power all movie recommendations.

Afterward, we reached out to 30 users for the follow-up interviews.⁵ We integrated their responses to help interpret the quantitative analysis results and investigate how explanations affect user motivational needs.

⁵Detailed interview questions in Appendix B.

	Example 1 – Ready Player One	Example 2 – Interstellar
CE	<p>"Ready Player One" is a visual spectacle that marries the thrill of competition found in "The Hunger Games" with the immersive and alternative reality akin to "Jumanji: Welcome to the Jungle." ... similar to Tris from "Divergent" or Thomas from "Maze Runner." ... Strap in for a movie that's part puzzle-solving, ..., and all Spielbergian magic!</p>	<p>"Interstellar" will propel you through space and time ... Imagine "Inception" meets the vastness of the cosmos. Matthew McConaughey and Anne Hathaway lead a desperate quest that combines the raw survival instinct seen in "Gladiator" with the innovative action sequences akin to "Mad Max: Fury Road." Nolan's knack for intricate storytelling...</p>
GE	<p>"Ready Player One" is a pop-culture-infused rollercoaster that feels like someone put Willy Wonka, Tron, and a Rubik's Cube in a blender. You'll embark on a high-stakes treasure hunt through the OASIS, a virtual reality universe with more layers than your grandma's famous lasagna. The film is a buffet of visual spectacles..</p>	<p>"Interstellar" is like a space burrito packed with sci-fi goodness, a touch of existential dread, and a side of "do I need a physics degree to get this?" Matthew McConaughey pilots through the galaxy... It's got more emotional father-daughter moments than a daddy-daughter dance... If you're into space odysseys, with a side of tear-jerking humanity, this cosmic ride is your ticket to the stars.</p>

Recommendation
 History Contexts
 Elements (e.g. Similarity, Popularity)

Figure 2: Examples of contextualized and generic explanations generated by GPT-4 Turbo.

3.5 Analyses

We conducted analyses based on user perceptions in the survey. Each participant could provide up to 10 ratings, resulting in a total of 2740 ratings from 759 participants in the main experiment (918 in NE, 911 in GE, and 911 in CE). We used hypothesis testing to identify differences in perceptions and consumption intentions across recommendation types, followed by linear mixed-effects regression to examine how user characteristics explain variation in both variables. Additionally, we conducted follow-up interviews to further interpret the quantitative results.

Hypothesis testing: We hypothesized that there are significant differences in the perceptions and consumption intentions between users receiving different recommendations. We conducted hypothesis testing on the four dependent variables of interest: *Cognitive*, *Affective*, *Utilitarian*, and *UserIntent* across the different conditions. Since the data did not meet the normality assumption (Lilliefors 1967), we used a *Mann-Whitney test* with *Holm correction* to identify pairwise differences between conditions (McHugh 2011). Hereafter, we use μ to report the mean value.

Linear mixed-effects regression. We used linear mixed-effects regression models to check further if any factors explained the variations in the variables of interest.⁶ We set NE as the reference condition so that our *Treatment* variables represent CE or GE. Given that each subject provides multiple ratings, we further included subject-specific random effects to control for between-subjects heterogeneity (Linden

⁶We also calculated the variance inflation factor (VIF) for all independent variables. All VIF values were below 5, indicating that multicollinearity should not bias the estimates (Akinwande, Dikko, and Samson 2015).

and Hönekopp 2021). Additionally, to control for the habituation effect caused by repetitive measurements (Gaito 1961; Jankowski 2021), we included a control variable for the recommendation display order (i.e., *DisplayOrder*). Our regression model is established as Equation 1:

$$\begin{aligned}
 Y_{ij} = & \beta_0 + \beta_1 Treatment_j + \beta_2 Age_j + \beta_3 Gender_j \\
 & + \beta_4 Education_j + \beta_5 MovieConsumption_j \\
 & + \beta_6 RecSysFrequency_j + \beta_7 RecSysUtility_j \\
 & + \beta_8 HistoryNum_j + \beta_9 DisplayOrder_j \\
 & + \beta_{10} HistoryNum_j \times MovieConsumption_j \\
 & + \beta_{11} HistoryNum_j \times RecSysUtility_j \\
 & + u_{0j} + \epsilon_{ij}
 \end{aligned} \quad (1)$$

where Y_{ij} denotes the dependent variables (i.e., *Cognitive*, *Affective*, *Utilitarian*, and *UserIntent*) for the i -th observation of the j -th subject, with intercepts β_0 , subject-level random effects u_{0j} , coefficients β_1 through β_{11} , and an error term ϵ_{ij} .

3.6 Ethical Considerations

This research adheres to rigorous ethical standards regarding the privacy and agency of all individuals who may be impacted by the study. The experimental design was formally reviewed and approved by the Ethics Commission at the University of Lausanne. All participant data were collected anonymously, with any potentially identifiable features removed during the preprocessing stage to ensure total de-identification. Furthermore, in compliance with transparency requirements for generative AI, all users were explicitly informed that both the movie recommendations and explanations were generated by LLMs.

4 Results

Below, we state the results of the hypothesis tests. The regression results are also shown in Table 4.

4.1 Cognitive Needs

We observed a significant difference (Holm-adjusted $p < 0.05$) in the distribution of the aggregated *Cognitive* measure between NE ($\mu = 4.94$), GE ($\mu = 4.97$), and CE ($\mu = 5.38$). Pairwise tests showed that CE significantly outperformed both NE and GE (Holm-adjusted $p < 0.05$), indicating that contextual explanations better support users' cognitive understanding. However, there was no statistical difference between NE and GE (Holm-adjusted $p > 0.05$), suggesting that generic explanations offer no additional cognitive benefit over providing no explanation.

In the linear mixed-effects regression, both *RecSysUtility* ($\beta = 0.304$, $p < 0.05$) and *MovieConsumption* ($\beta = 0.150$, $p < 0.05$) demonstrated positive, significant effects on *Cognitive*, indicating that participants who perceive the recommender system as more useful and who watch movies more frequently tend to rate higher in *Cognitive*. Besides, more movie contexts ($\beta = 0.048$, $p < 0.05$) in the explanation also better satisfy users' cognitive needs.

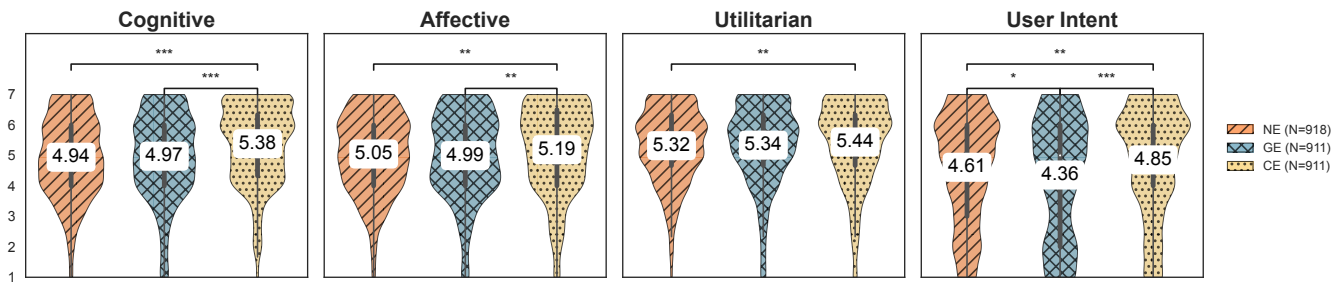


Figure 3: Ratings by explanation type (shown variables: *Cognitive*, *Affective*, *Utilitarian*, and *UserIntent*). Significance levels between conditions are: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In the interviews, several participants clearly expressed their trust and commended the clarity when receiving contextualized explanations.

“The thing that makes me trust the explanation was reading how the description had brought up other titles of movies that I saw in the recommendations, that had true things about the characters, and how the movies could be similar or how I could watch them already.” (Female, 29)

“They made the details of the movies sound more exciting and interesting and gave you more of a full rundown of theories so you don’t think it’s going to be one thing and turn out another.” (Female, 31)

On the other hand, sometimes users disliked the recommendations, but they still helped to filter out unwanted content, improving their sense of control and overall transparency.

“The movies it recommended didn’t interest me, so I don’t think I’ll be watching them. Felt like I understood the movies weren’t for me.” (Male, 32)

“I trusted the recommended movie, but I made my own mind up after reading the explanation.” (Male, 44)

4.2 Affective Needs

For the *Affective* dimension, CE ($\mu = 5.19$) was rated significantly higher than both NE ($\mu = 5.05$) and GE ($\mu = 4.99$) (Holm-adjusted $p < 0.05$). Still, no significant difference emerged between NE and GE (Holm-adjusted $p > 0.05$).

The linear mixed-effects regression highlights two user-engagement variables, *RecSysUtility* ($\beta = 0.295$, $p < 0.05$) and *MovieConsumption* ($\beta = 0.175$, $p < 0.05$), as positive, significant predictors of affective responses. This suggests that, regardless of the explanation style, users who view the recommender system as helpful and frequently watch movies are more likely to have positive feedback in affective needs. Besides, more movie context ($\beta = 0.069$, $p < 0.05$) in the explanation has a positive effect as well.

From the interview, we observed that the movie and explanation quality are important to users’ affective responses, no matter whether it contains the context. This also explains why users’ satisfaction is fairly similar when receiving CE or GE. Regarding the textual quality, the output of LLMs is consistent regardless of the explanation types.

“Because I am a movie buff, and frankly, the descriptions were very well laid out so that it gave me enough informa-

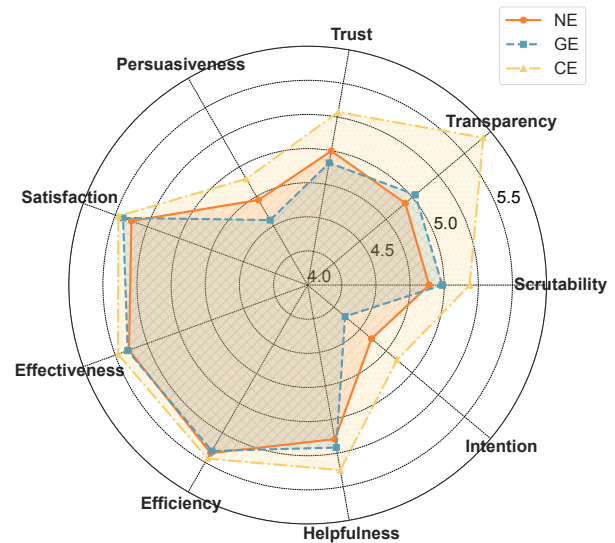


Figure 4: Radar chart of average user perception ratings in different dimensions by explanation types.

tion about the plot but also left enough unsaid that now I have to find out.” (Male, 39)

“Some of the word descriptions were a little over the top. The description was a little graphic.” (Female, 68)

“The explanation was well written, so it gave away a little more than I would normally like, but was just as good as a trailer in my mind.” (Male, 39)

4.3 Utilitarian Needs

In the *Utilitarian* dimension, CE ($\mu = 5.44$) was rated significantly higher than NE ($\mu = 5.32$) (Holm-adjusted $p < 0.05$). There was no significant difference between NE and GE or CE and GE ($\mu = 5.34$).

The linear mixed-effects regression showed *RecSysUtility* ($\beta = 0.271$, $p < 0.05$) and *MovieConsumption* ($\beta = 0.172$, $p < 0.05$) both positively influenced *Utilitarian*, implying that users who closely engage with recommender systems and movies tend to see greater practical value in the recommendations regardless of explanation style. Female participants also reported higher utilitarian scores (versus male,

$\beta = 0.292, p < 0.05$). More movie context increases utilitarian value ($\beta = 0.044, p < 0.05$).

In general, users felt that these explanations save time in deciding whether a movie fits their tastes. However, we observed that some users appreciated more details to support their decisions, while some preferred the explanations to be simplified.

“If I couldn’t clearly understand the explanation of the film, I thought that I was most likely not going to enjoy the movie. I trust my first instincts on things because it has saved me a lot of time in the past, by allowing me to avoid seeing a movie that would only waste time in my too short life.” (Female, 37)

“Sometimes too much is overwhelming.” (Male, 37)

4.4 Consumption Intentions

We found significant differences in the consumption intention across all conditions. CE ($\mu = 4.85$) outperformed GE ($\mu = 4.36$) and NE ($\mu = 4.61$) (Holm-adjusted $p < 0.05$). GE was also rated significantly lower than NE (Holm-adjusted $p < 0.05$).

Linear mixed-effects regression indicates that *RecSysUtility* ($\beta = 0.292, p < 0.05$) and *MovieConsumption* ($\beta = 0.125, p < 0.05$) both increase the likelihood of watching the recommended title. Older participants ($\beta = -0.147, p < 0.05$) expressed lower consumption intentions. Besides, more movie context ($\beta = 0.101, p < 0.05$) in the explanation has a positive effect as well.

Though users have overall positive feelings, their consumption intent is more attached to their taste. Many interviewees explicitly expressed that, if the movie is not their type, they will not watch it. This is not a consequence of our findings but of the accuracy of the algorithm behind the recommender systems. Hence, this still confirms the value of contextualization in motivating users’ decisions.

“The factor that motivated me the most in deciding whether or not to watch the movie was the correlation factor between the new movie and movies I have watched in the past.” (Female, 20)

“Sometimes the explanation was so intriguing I felt as though I wanted to know more about the movie. If the explanation is so boring, I don’t finish it, I’m most likely not going to watch the movie.” (Female, 37)

“If it seemed like the movie was for children, it was a no for me.” (Male, 32)

“This is becoming a little redundant, but again it explained it to the point I could see in my head, the picture and the action happening. Plus, anything with Will Smith has got to be awesome.” (Female, 61)

4.5 Interpretations

Based on the results above, we observed an overall advantage of LLM-generated contextualized explanations in addressing users’ motivational needs and increasing their consumption intentions. Figures 3 and 4 show that contextualized explanations successfully address users’ diverse needs. However, a key limitation remains: the explanations do not always align with users’ literary and genre preferences. Although we asked participants to evaluate the explanations

	<i>Cognitive</i>	<i>Affective</i>	<i>Utilitarian</i>	<i>UserIntent</i>
<i>Intercept</i>	4.959*** (0.088)	5.030*** (0.087)	5.172*** (0.084)	4.849*** (0.112)
<i>Treatment (GE)</i>	0.010 (0.102)	-0.032 (0.100)	0.064 (0.097)	-0.232 (0.127)
<i>Treatment (CE)</i>	0.418*** (0.102)	0.166 (0.100)	0.165 (0.097)	0.277* (0.127)
<i>Gender (Female)</i>	0.120 (0.088)	0.160 (0.086)	0.292*** (0.084)	-0.181 (0.110)
<i>Age</i>	-0.075 (0.046)	-0.089* (0.045)	0.013 (0.044)	-0.147* (0.057)
<i>Education</i>	0.038 (0.042)	0.028 (0.041)	0.003 (0.040)	0.040 (0.052)
<i>RecSysUtility</i>	0.304*** (0.063)	0.295*** (0.062)	0.271*** (0.060)	0.292*** (0.079)
<i>RecSysFrequency</i>	0.019 (0.072)	-0.004 (0.070)	-0.023 (0.068)	0.132 (0.089)
<i>MovieConsumption</i>	0.150** (0.047)	0.175*** (0.046)	0.172*** (0.045)	0.125* (0.059)
<i>DisplayOrder</i>	0.040 (0.024)	0.035 (0.028)	0.017 (0.025)	0.115** (0.042)
<i>HistoryNum</i>	0.048** (0.015)	0.069*** (0.018)	0.044** (0.016)	0.101*** (0.028)
<i>HistoryNum × MovieConsumption</i>	-0.028 (0.016)	0.007 (0.018)	-0.032* (0.016)	-0.047 (0.029)
<i>HistoryNum × RecSysUtility</i>	-0.019 (0.015)	-0.036* (0.018)	-0.013 (0.016)	-0.054 (0.028)

Table 4: Linear mixed-effects regression results on user response outcomes. Coefficients are reported with standard errors in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

rather than the recommended movies, interviews revealed that many users still evaluated the two together, resulting in lower scores for persuasiveness and consumption intention. When users turn to recommender systems, they often have specific goals, such as finding a movie to watch or exploring new options. Once the system delivers relevant recommendations, users may feel that their immediate needs are already satisfied. This likely explains why participants rated effectiveness, efficiency, and satisfaction high, even in the absence of explanations. Together, our findings highlight the value of LLM-generated explanations in building transparency and trust. However, to enhance user experience, explanations need to resonate with users’ tastes and expectations, especially in subjective domains like entertainment.

Demographics and prior engagement shape user perception: Based on the results in Table 4, users’ prior engagement significantly shapes their evaluation of explanations. Participants who consumed more movies rated explanations more positively across all dimensions, suggesting that domain familiarity enhances appreciation of contextualized content. Similarly, higher perceived usefulness of rec-

ommender systems strongly gave favorable evaluations.

Demographic effects were more selective. Female users gave higher utilitarian ratings, while older users rated affective and intention outcomes slightly lower. These findings suggest that both system engagement and user characteristics should be considered when designing the system.

More viewing history contexts, better perceptions: Table 4 reveals that the number of viewing-history items (*HistoryNum*) positively influenced users' evaluations in perception and intention. As detailed in Section 3.3, we limited the number of contexts to a maximum of four, which was guided by typical movie description lengths, native reading speeds, and cognitive load thresholds.

Importantly, the non-significant interaction effects between *HistoryNum* and both *MovieConsumption* and *RecSysUtility* suggest that the positive impact of including more movie contexts in explanations holds consistently, regardless of users' movie consumption frequency or perceptions of the recommender system's helpfulness. This implies that richer contextualization can enhance user evaluations without being contingent on prior engagement. Although richer context tends to improve perception and intent, the effects, based on the coefficients in the regression results, remain relatively small. Some users also flagged certain explanations as redundant or overly detailed. To prevent cognitive overload, we should carefully calibrate how much context to include, ensuring that explanations remain concise yet informative.

5 Heterogeneity Analysis

5.1 Claude 3.5 Sonnet

To supplement the main experiment, we expanded the experiment to a larger scope by leveraging another LLM (Claude 3.5 Sonnet). Given that GPT-4 Turbo is a high-performing baseline at the time of the study, we tested whether the superior performance of contextualization could be generalized to other LLMs. We generated another recommendation corpus using Claude 3.5 Sonnet and replicated the experiments on the U.S. samples ($N = 140$, 71 are women, 68 are men, and one is non-binary). The results are aligned with our main findings, validating the advantages of LLMs in generating contextualized explanations in terms of meeting users' motivational needs and increasing their consumption intent. The results and interpretations are provided in Appendix C.

5.2 Movie Genre

To further understand how user perceptions vary across recommendation contexts, we examined the role of movie genre, specifically *comedy* and *adventure*, with supplementary mixed-effects regression models (see Equation 2 in Appendix C). As shown in Table 7, comedy movies were consistently rated lower than adventure movies across affective, utilitarian, and consumption intention outcomes. We also observed that these genre effects were not moderated by explanation types. These supplementary results highlight genre as a key factor influencing user responses to LLM-generated recommendations. The influence of movie genre remains an important area for further investigation.

In summary, the supplementary experiment yielded results that supported the main experiment's findings and conclusions, thereby validating the robustness of our study design and the generalizability of the findings.

6 Discussion

6.1 Implications of Findings

Recommender systems play an important role in shaping user decisions, particularly influential in the context of web and social media. Given the far-reaching implications of these systems, there is growing demand from users and regulators for transparency, emotional connection, and efficiency in decision-making. This highlights the value of high-quality and context-aware recommendation explanations. However, many existing recommender systems do not or fail to provide explanations that address users' diverse needs (Li, Zhang, and Chen 2020; Lu et al. 2023).

Our study demonstrates that LLM-generated contextualized explanations can overcome these limitations by satisfying users' cognitive needs, ultimately increasing their consumption intentions. Also, our findings reveal the insufficiency of LLM-generated explanations in meeting users' utilitarian and affective needs, which raises concerns about the proper design and implications of LLM-generated contextualized content.

Build trust, but risk misrepresenting preferences: Our results show that contextualized explanations significantly address users' cognitive needs, particularly in terms of *Transparency*, *Trust*, and *Scrutability*. By referencing users' prior viewing history, LLM-generated explanations appear more tailored and meaningful, encouraging deeper engagement with recommendations.

Nevertheless, we also observed that LLMs tend to interpret viewing histories as an implicit preference signal. While this can increase perceived personalization, it also introduces a risk of over-interpretation, which is reflected in user interviews. In practice, real-world recommender systems often infer preferences from behavioral signals (e.g., clicks, watch time) rather than explicit feedback, which may not always align with users' true interests or sentiments. LLMs, when generating explanations based on such data, may reinforce these weak or ambiguous signals, thereby creating a disconnect between the explanation and the user's actual intent. This misalignment risks undermining the trust that contextualized explanations build. When users feel the system attributes preferences they do not hold, trust may erode over time. Therefore, in real-world deployments, systems that leverage LLMs for explanation generation should account for this interpretive bias and consider more nuanced strategies for referencing user history, particularly when preferences are inferred rather than explicitly stated.

Opportunities and challenges for the movie industry and users: Our findings show that LLM-generated movie recommendations not only increase users' intention to engage with suggested content but also make it easier for users to discover relevant viewing options tailored to their interests. By leveraging typical user viewing histories and the LLMs, without requiring extra effort or user input, these

recommendations simplify the process of finding enjoyable content. This user-friendly approach, combined with LLMs' ability to produce highly contextualized suggestions, delivers direct benefits to users and makes these tools valuable for streaming services and media platforms.

Integrating LLMs into recommendation systems can enhance operational efficiency, but it also introduces deployment costs that depend heavily on the chosen model and usage patterns. For example, OpenAI's GPT-4o mini is currently priced at \$0.15 per million input tokens and \$0.60 per million output tokens (OpenAI 2026) at the time of writing, so a typical interaction with 200 input tokens and 100 output tokens costs about \$0.00009. Using Netflix's 2025 year-end scale of 325 million paid memberships (Netflix, Inc. 2025), 50 such LLM interactions per day per membership would amount to roughly \$1.46 million in API costs per day. For ad-supported platforms such as YouTube, monetization is often discussed in terms of RPM, defined by YouTube as revenue per 1,000 video views (YouTube Help 2026), although earnings vary substantially by niche, geography, and monetization mix.

While LLMs present substantial opportunities to enhance recommendation quality and operational efficiency, successful deployment requires careful management of associated technical, economic, and ethical challenges. In particular, LLM-generated recommendations carry risks related to biased outputs (Feuerriegel et al. 2024). In movie recommendation contexts, biases may manifest as an underrepresentation of specific genres, cultures, or demographic perspectives. For example, the corpus generated in our study disproportionately favored Western-centric narratives. Additionally, perceived biases or cultural insensitivities arising from skewed training data or adversarial prompt manipulation can erode user trust and pose significant reputational risks to platforms.

6.2 Limitations

While our study offers valuable insights into user perceptions of LLM-generated explanations for movie recommendations, several limitations should be noted.

First, our findings may be limited in generalizability due to both the participant sample and the underlying models. We primarily recruited English-speaking users, which may not capture the full range of cultural and linguistic differences in user perceptions. Additionally, the effectiveness of the explanations likely depends on the capabilities of the underlying LLMs. Although we used state-of-the-art models in our study period, future advancements may further influence how users interpret and respond to recommendations and explanations.

Second, our study may oversimplify the explanation generation process. Future research should explore various explanation strategies to better understand their effectiveness across different contexts. While we employed a typical cold-start recommendation scenario, providing additional user information, such as demographics or expressed preferences, can more efficiently reveal LLMs' strong contextualization capabilities.

Lastly, our experimental design does not account for the longitudinal dynamics of real-world user interactions with recommender systems. While we capture users' immediate perceptions of explanations and recommendations, these impressions may evolve as users continue to engage with the recommended content over time. Additionally, due to our control over user inputs, participants did not have the opportunity to express specific or niche movie preferences to the LLMs, which may play a significant role in shaping their perceptions. Moreover, the broader long-term effects, such as the cumulative impact on content consumption patterns and industry outcomes, remain open for future investigation.

Despite these limitations, our study provides a strong foundation for further research into the role of LLM-generated explanations in recommender systems, emphasizing the importance of contextualization in improving user experience and engagement.

7 Conclusion

Our findings show how and why recommendation explanations meet cognitive, affective, and utilitarian motivations under the U&G perspective. Our study confirms that LLM-generated contextualized explanations effectively address users' cognitive and affective needs and increase their consumption intentions. Notably, recommender systems should also avoid overwhelming users with irrelevant information that can hinder decision-making or erode trust. Future recommender system designs that integrate contextualized LLM-generated explanations can more effectively serve a wide range of user motivations, ultimately improving user satisfaction and leading to actionable engagement with the recommended content.

Acknowledgments

This work was supported by the Swiss National Science Foundation (Grant: 215542) and the German Research Foundation (Grant: 543018872).

We used generative AI for language polishing and grammar correction. All claims, analyses, and citations were verified by the authors.

References

- Abdollahi, B.; and Nasraoui, O. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the ACM Conference on Recommender Systems*.
- Acharya, A.; Singh, B.; and Onoe, N. 2023. LLM based generation of item-description for recommendation system. In *Proceedings of the ACM Conference on Recommender Systems*.
- Akinwande, M. O.; Dikko, H. G.; and Samson, A. 2015. Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, 05(07): 754.
- Angelov, P. P.; Soares, E. A.; Jiang, R.; Arnold, N. I.; and Atkinson, P. M. 2021. Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5): e1424.

- Balog, K.; and Radlinski, F. 2020. Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Bar, D.; Maarouf, A.; and Feuerriegel, S. 2024. Generative AI may backfire for counterspeech. arXiv:2411.14986.
- Bashardoust, A.; Feuerriegel, S.; and Shrestha, Y. R. 2024. Comparing the willingness to share for human-generated vs. AI-generated fake news. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–21.
- Burkart, N.; and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70: 245–317.
- Chaimalas, I.; Walker, D. M.; Gruppi, E.; Clark, B. R.; and Toni, L. 2023. Bootstrapped personalized popularity for cold start recommender systems. In *ACM Conference on Recommender Systems*, 715–722.
- Chang, S.; Harper, F. M.; and Terveen, L. G. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the ACM Conference on Recommender Systems*.
- Chatti, M. A.; Guesmi, M.; Vorgerd, L.; Ngo, T.; Joarder, S.; Ain, Q. U.; and Muslim, A. 2022. Is more always better? The effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization*.
- Chen, C.; Zhang, M.; Liu, Y.; and Ma, S. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the World Wide Web Conference*.
- Chen, L.; Yan, D.; and Wang, F. 2019. User Evaluations on Sentiment-based Recommendation Explanations. *ACM Trans. Interact. Intell. Syst.*, 9(4): 20:1–20:38.
- Deldjoo, Y. 2025. Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems*.
- Faul, F.; Erdfelder, E.; Buchner, A.; and Lang, A.-G. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4): 1149–1160.
- Faul, F.; Erdfelder, E.; Lang, A.-G.; and Buchner, A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2): 175–191.
- Feuerriegel, S.; Hartmann, J.; Janiesch, C.; and Zschech, P. 2024. Generative AI. *Business & Information Systems Engineering*.
- Gaito, J. 1961. Repeated measurements designs and counterbalancing. *Psychological Bulletin*, 58(1): 46–54.
- Giray, L. 2023. Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, 51(12): 2629–2633.
- Goyani, M.; and Chaurasiya, N. 2020. A review of movie recommendation system: Limitations, survey and challenges. *Electronic Letters on Computer Vision and Image Analysis*, 19: 18–37.
- Hou, H.; Leach, K.; and Huang, Y. 2024. ChatGPT giving relationship advice – how reliable is it? *Proceedings of the International AAAI Conference on Web and Social Media*.
- IMDb.com, Inc. 2024a. IMDb Non-Commercial Datasets. <http://developer.imdb.com/non-commercial-datasets/>. Accessed: 2024-10-15.
- IMDb.com, Inc. 2024b. IMDb Top 250 movies. <http://www.imdb.com/chart/top/>. Accessed: 2024-10-15.
- Jameson, A.; Willemsen, M. C.; Felfernig, A.; de Gemmis, M.; Lops, P.; Semeraro, G.; and Chen, L. 2015. Human decision making and recommender systems. In *Recommender Systems Handbook*, 611–648.
- Jankowski, J. 2021. Habituation effect in social networks as a potential factor silently crushing influence maximisation efforts. *Scientific Reports*, 11(1): 19055.
- Kang, H. 2021. Sample size determination and power analysis using the G*Power software. *Journal of Educational Evaluation for Health Professions*, 18: 17.
- Katz, E.; Blumler, J. G.; and Gurevitch, M. 1973. Uses and gratifications research. *The Public Opinion Quarterly*, 37(4): 509–523.
- Ko, H.; Lee, S.; Park, Y.; and Choi, A. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1): 141.
- Korhan, O.; and Ersoy, M. 2016. Usability and functionality factors of the social network site application users from the perspective of uses and gratification theory. *Quality & Quantity*, 50(4): 1799–1816.
- Kunkel, J.; Donkers, T.; Michael, L.; Barbu, C.-M.; and Ziegler, J. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Lakens, D. 2022. Sample size justification. *Collabra: Psychology*, 8(1): 33267.
- Lam, X. N.; Vu, T.; Le, T. D.; and Duong, A. D. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the international conference on Ubiquitous information management and communication*, 208–211.
- Lee, S.; and Lee, D. K. 2018. What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, 71(5): 353–360.
- Li, L.; Zhang, Y.; and Chen, L. 2020. Generate neural template explanations for recommendation. In *Proceedings of the ACM International Conference on Information & Knowledge Management*.
- Lilliefors, H. W. 1967. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318): 399–402.
- Lin, Y.-R.; Margolin, D.; Keegan, B.; Baronchelli, A.; and Lazer, D. 2013. #Bigbirds never die: understanding social dynamics of emergent hashtags. *Proceedings of the International AAAI Conference on Web and Social Media*.

- Linden, A. H.; and Hönekopp, J. 2021. Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 16(2): 358–376.
- Lu, H.; Ma, W.; Wang, Y.; Zhang, M.; Wang, X.; Liu, Y.; Chua, T.-S.; and Ma, S. 2023. User perception of recommendation explanation: Are your explanations what users need? *ACM Transactions on Information Systems*, 41(2): 48:1–48:31.
- Lubos, S.; Tran, T. N. T.; Felfernig, A.; Polat Erdeniz, S.; and Le, V.-M. 2024. LLM-generated explanations for recommender systems. In *Adjunct Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization*.
- Ma, T.-m.; Wang, X.; Zhou, F.-c.; and Wang, S. 2023. Research on diversity and accuracy of the recommendation system based on multi-objective optimization. *Neural Computing and Applications*, 35(7): 5155–5163.
- Mandl, M.; Felfernig, A.; Teppan, E.; and Schubert, M. 2011. Consumer decision making in knowledge-based recommendation. *Journal of Intelligent Information Systems*, 37(1): 1–22.
- McHugh, M. L. 2011. Multiple comparison analysis testing in ANOVA. *Biochemia Medica*, 21(3): 203–209.
- Netflix, Inc. 2025. Netflix Investor Relations: Quarterly Earnings. <https://ir.netflix.net/>. Accessed: 2026-03-22.
- OpenAI. 2026. OpenAI API Pricing. <https://openai.com/api/pricing/>. Accessed: 2026-03-22.
- Pazzani, M. J.; and Billsus, D. 2007. Content-Based Recommendation Systems. In *The Adaptive Web*, volume 4321, 325–341.
- Pennycook, G.; Cannon, T. D.; and Rand, D. G. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology. General*, 147(12): 1865–1880.
- Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the International Conference on World Wide Web*.
- Seo, S.; Huang, J.; Yang, H.; and Liu, Y. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the ACM Conference on Recommender Systems*.
- Silva, Í.; Marinho, L.; Said, A.; and Willemsen, M. C. 2024. Leveraging ChatGPT for automated human-centered explanations in recommender systems. In *Proceedings of the International Conference on Intelligent User Interfaces*.
- Sun, Y.; Drivas, M.; Liao, M.; and Sundar, S. S. 2023. When recommender systems snoop into social media, users trust them less for health advice. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Tsai, C.-H.; and Brusilovsky, P. 2019. Evaluating visual explanations for similarity-based recommendations: User perception and performance. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization*.
- Vultureanu-Albisi, A.; and Badica, C. 2021. Recommender systems: An explainable AI perspective. In *Proceedings of the International Conference on INnovations in Intelligent SysTems and Applications*.
- Wien, A. H.; and Peluso, A. M. 2021. Influence of human versus AI recommenders: The roles of product type and cognitive processes. *Journal of Business Research*, 137: 13–27.
- Wu, L.; He, X.; Wang, X.; Zhang, K.; and Wang, M. 2023. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(05): 4425–4445.
- Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; Xiong, H.; and Chen, E. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5): 60.
- Yang, J. 2016. Effects of popularity-based news recommendations (“most-viewed”) on users’ exposure to online news. *Media Psychology*, 19(2): 243–271.
- YouTube Help. 2026. Understand Revenue per mille (RPM). <https://support.google.com/youtube/answer/9314357>. Accessed: 2026-03-22.
- Zhang, Y.; and Chen, X. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1): 1–101.

Reproducibility Checklist

1. General Paper Structure

- 1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) **yes**
- 1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) **yes**
- 1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) **yes**

2. Theoretical Contributions

- 2.1. Does this paper make theoretical contributions? (yes/no) **no**

If yes, please address the following points:

- 2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)
- 2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)
- 2.4. Proofs of all novel claims are included (yes/partial/no)
- 2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

- 2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)
- 2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)
- 2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

- 3.1. Does this paper rely on one or more datasets? (yes/no) **yes**

If yes, please address the following points:

- 3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) **yes**
- 3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) **yes**
- 3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) **yes**
- 3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) **yes**
- 3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) **yes**
- 3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying (yes/partial/no/NA) **NA**

4. Computational Experiments

- 4.1. Does this paper include computational experiments? (yes/no) **no**

If yes, please address the following points:

- 4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA)
- 4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no)
- 4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no)
- 4.5. All source code required for conducting and ana-

lyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no)

- 4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no)
- 4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA)
- 4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no)
- 4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no)
- 4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no)
- 4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no)
- 4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no)
- 4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA)

A Recommendation Corpus Generation

Table 5 details the procedure we use to construct the recommendation corpus \mathcal{C} from a given movie shortlist \mathcal{M} . For each movie $m_i \in \mathcal{M}$, the system generates a sequence of four recommendations, indexed by $k = 1$ to 4.

At each step:

- The variable m_i denotes a single movie from the input shortlist.
- The index k indicates the recommendation order (which, equivalently, is the number of movies in the users viewing history used to produce the recommendation).
- The system generates $r_{i,k}$, the k -th recommendation for movie m_i .

Recommendation Corpus Generation	
Input: Movie shortlist \mathcal{M}	
Output: Corpus \mathcal{C} containing recommendations and explanations	
1	for each $m_i \in \mathcal{M}$ do
	m_i : Represents a single movie in the shortlist \mathcal{M}
2	for $k = 1$ to 4 do
	k : The recommendation order (1st, 2nd, 3rd, or 4th)
3	if $k = 1$ then
4	Generate $r_{i,1}$ from m_i
	$r_{i,k}$: The k -th recommendation for the i -th movie
5	else
6	Generate $r_{i,k}$ from $\{m_i, r_{i,1}, \dots, r_{i,(k-1)}\}$
	Previous recommendations are incorporated to generate $r_{i,k}$
7	end if
8	Generate $e_{i,k,c}$ and $e_{i,k,g}$ for $r_{i,k}$
	$e_{i,k,c}$: Contextualized Explanation for $r_{i,k}$
	$e_{i,k,g}$: Generic Explanation for $r_{i,k}$
9	Save $r_{i,k}$, $e_{i,k,c}$, and $e_{i,k,g}$ in \mathcal{C}
	Corpus \mathcal{C} stores the recommendations and their explanations
10	end for
11	end for

Table 5: Recommendation corpus preparation steps. The corpus \mathcal{C} contains recommendations $r_{i,k}$ for each movie $m_i \in \mathcal{M}$, with contextualized explanations ($e_{i,k,c}$) and generic explanations ($e_{i,k,g}$) for each recommendation.

For the first recommendation ($k = 1$), the system conditions only on the original movie m_i . For subsequent recommendations ($k > 1$), it incorporates m_i and all previously generated recommendations ($r_{i,1}, \dots, r_{i,(k-1)}$) to produce $r_{i,k}$. This iterative approach models a realistic, context-aware recommendation process.

After generating each recommendation $r_{i,k}$, the system produces two types of explanations for users in different conditions:

- $e_{i,k,c}$: a *contextualized explanation* that references the earlier items in the recommendation sequence;
- $e_{i,k,g}$: a *generic explanation* that relates $r_{i,k}$ to the original movie m_i without incorporating context.

Each recommendation $r_{i,k}$ and its corresponding explanations ($e_{i,k,c}$ and $e_{i,k,g}$) are added to the final corpus \mathcal{C} . This procedure ensures that the corpus captures both sequential recommendation dynamics and consistency across explanation styles.

B Interview Questions

- “What made you trust or doubt the recommendation after seeing the explanation?”
- “How did the explanation affect your overall feeling about the movie? Why did it make you feel that way?”
- “How did the explanation assist (or fail to assist) you in understanding if the movie fit your interests or needs?”
- “After reading the explanation, what influenced you

most in deciding whether to watch the movie or not?”

- “How do you think your profile, such as your demographics or viewing habits, influenced how you felt about the explanation you received?”

C Supplementary Experiments

Claude 3.5 Sonnet

We replicated the experiment on a smaller U.S. sample ($N = 140$) using Claude 3.5 Sonnet. The results align with our main study, showing that contextualization improves overall explanation quality, particularly in terms of persuasiveness, transparency, and trust. Besides, contextualized explanations increase users’ consumption intentions. While some minor discrepancies emerged between this supplementary experiment and the main experiment, the core findings remain consistent. Due to the small sample size, some results are insignificant. We ran the same regression analysis as Equation 1.

Overall, the supplemental experiment supports the generalizability of our conclusions. Figure 6 shows user ratings of motivational needs and consumption intentions. Figure 5 provides a detailed dimension comparison. Table 6 presents the regression results.

Movie Genre

Observation of popular movie recommenders like Netflix and Amazon Video showed that most movies recommended tended to fall into widely favored genres such as *Comedy* and *Adventure*. Based on IMDb genre labels, we annotated

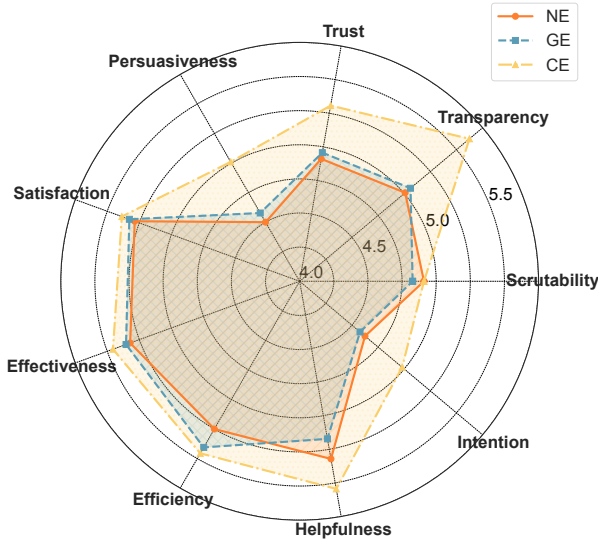


Figure 5: Radar chart of average user perception ratings in different dimensions by explanation types (Model: Claude 3.5 Sonnet).

each recommended movie in our experimental study as either *Comedy* or *Adventure*. To examine how genre influences user perceptions and consumption intentions, we performed the following regression analysis:

$$\begin{aligned}
 Y_{ij} = & \beta_0 + \beta_1 \text{Treatment}_j + \beta_2 \text{Age}_j + \beta_3 \text{Gender}_j \\
 & + \beta_4 \text{Education}_j + \beta_5 \text{MovieConsumption}_j \\
 & + \beta_6 \text{RecSysFrequency}_j + \beta_7 \text{RecSysUtility}_j \\
 & + \beta_8 \text{HistoryNum}_j + \beta_9 \text{DisplayOrder}_j \\
 & + \beta_{10} \text{MvGenre}_{ij} + \beta_{11} \text{MvGenre}_{ij} \times \text{Treatment}_j \\
 & + u_{0j} + \epsilon_{ij}
 \end{aligned} \tag{2}$$

where Y_{ij} denotes the dependent variables (i.e., *Cognitive*, *Affective*, *Utilitarian*, and *UserIntent*) for the i -th observation of the j -th subject, with intercepts β_0 , subject-level random effects u_{0j} , coefficients β_1 through β_{11} , and an error term ϵ_{ij} .

Regression results in Table 7 show that the genre of the recommended movie significantly influenced user perceptions. Specifically, movies categorized as *Comedy* were associated with lower ratings across affective ($\beta = -0.203$, $p < 0.05$), utilitarian ($\beta = -0.149$, $p < 0.05$), and behavioral intention outcomes ($\beta = -0.415$, $p < 0.05$), compared to *Adventure* movies. The effect of genre was not significantly moderated by explanation type (GE or CE), as the interaction terms were non-significant across all outcomes. These results suggest that genre perceptions persist regardless of contextualization strategies and that users may perceive adventure movies as more engaging or informative in recommendation explanations.

	<i>Cognitive</i>	<i>Affective</i>	<i>Utilitarian</i>	<i>UserIntent</i>
<i>Intercept</i>	4.955*** (0.198)	4.935*** (0.202)	5.280*** (0.204)	4.712*** (0.269)
<i>Treatment (GE)</i>	-0.038 (0.249)	0.017 (0.254)	-0.200 (0.257)	-0.041 (0.340)
<i>Treatment (CE)</i>	0.333 (0.255)	0.397 (0.260)	0.165 (0.263)	0.502 (0.348)
<i>Gender (Female)</i>	-0.022 (0.209)	-0.190 (0.213)	0.105 (0.215)	-0.377 (0.285)
<i>Age</i>	0.227* (0.110)	0.196 (0.112)	0.239* (0.113)	0.115 (0.150)
<i>Education</i>	-0.043 (0.104)	-0.079 (0.106)	-0.090 (0.107)	0.001 (0.142)
<i>RecSysUtility</i>	0.118 (0.156)	0.051 (0.159)	0.168 (0.160)	0.250 (0.214)
<i>RecSysFrequency</i>	0.227 (0.185)	0.230 (0.188)	0.166 (0.190)	0.188 (0.252)
<i>MovieConsumption</i>	0.099 (0.126)	0.044 (0.129)	-0.006 (0.130)	-0.092 (0.172)
<i>DisplayOrder</i>	-0.077 (0.048)	-0.013 (0.050)	-0.011 (0.042)	-0.119 (0.082)
<i>HistoryNum</i>	-0.001 (0.050)	-0.099 (0.052)	-0.003 (0.044)	-0.007 (0.084)
<i>HistoryNum × MovieConsumption</i>	0.027 (0.055)	0.058 (0.057)	0.007 (0.048)	0.039 (0.091)
<i>HistoryNum × RecSysUtility</i>	0.009 (0.057)	-0.040 (0.059)	-0.021 (0.050)	-0.006 (0.095)

Table 6: Linear mixed-effects regression results on user response outcomes (Model: Claude 3.5 Sonnet). Coefficients are reported with standard errors in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

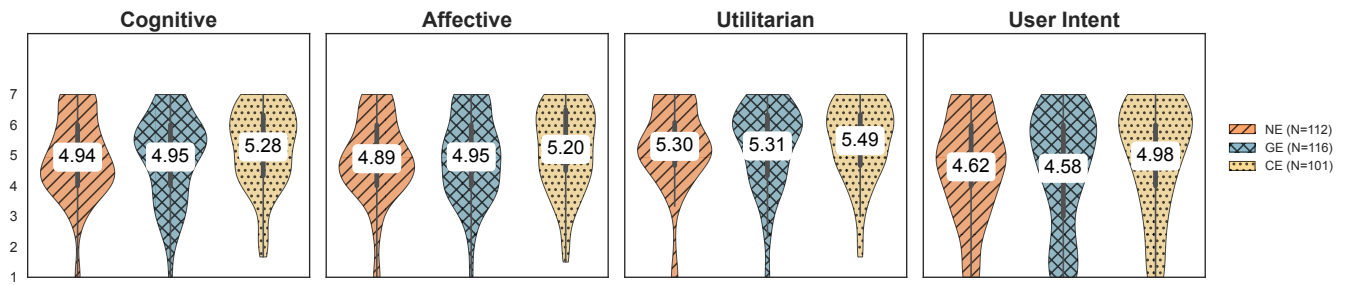


Figure 6: Ratings by explanation type (Variables: *Cognitive*, *Affective*, *Utilitarian*, and *UserIntent*; Model: *Claude 3.5 Sonnet*). Significance levels between conditions are indicated: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	<i>Cognitive</i>	<i>Affective</i>	<i>Utilitarian</i>	<i>UserIntent</i>
<i>Intercept</i>	4.946*** (0.090)	5.045*** (0.090)	5.189*** (0.086)	4.879*** (0.117)
<i>Treatment (GE)</i>	0.006 (0.106)	-0.058 (0.106)	0.034 (0.102)	-0.268 (0.138)
<i>Treatment (CE)</i>	0.438*** (0.106)	0.135 (0.106)	0.125 (0.102)	0.188 (0.139)
<i>Gender (Female)</i>	0.116 (0.088)	0.160 (0.086)	0.291*** (0.084)	-0.183 (0.110)
<i>Age</i>	-0.075 (0.046)	-0.089* (0.045)	0.013 (0.044)	-0.148* (0.057)
<i>Education</i>	0.037 (0.042)	0.028 (0.041)	0.002 (0.040)	0.038 (0.052)
<i>RecSysUtility</i>	0.306*** (0.063)	0.298*** (0.062)	0.273*** (0.060)	0.297*** (0.079)
<i>RecSysFrequency</i>	0.018 (0.072)	-0.004 (0.070)	-0.025 (0.069)	0.130 (0.089)
<i>MovieConsumption</i>	0.152** (0.047)	0.174*** (0.046)	0.174*** (0.045)	0.126* (0.059)
<i>DisplayOrder</i>	0.036 (0.024)	0.034 (0.028)	0.016 (0.025)	0.113** (0.042)
<i>HistoryNum</i>	0.045** (0.015)	0.068*** (0.018)	0.042** (0.015)	0.098*** (0.028)
<i>MvGenre (Com)</i>	-0.096 (0.062)	-0.203** (0.073)	-0.149* (0.064)	-0.415*** (0.111)
<i>MvGenre (Com) × Treatment (GE)</i>	0.014 (0.072)	0.064 (0.085)	0.077 (0.074)	0.096 (0.132)
<i>MvGenre (Com) × Treatment (CE)</i>	-0.051 (0.071)	0.068 (0.084)	0.095 (0.073)	0.206 (0.131)

Table 7: Linear mixed-effects regression results on user response outcomes with split movie genres. Coefficients are reported with standard errors in parentheses. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.