Large Language Models as Conversational Movie Recommenders: A User Study

Ruixuan Sun
Grouplens Research, University of Minnesota
Minneapolis, Minnesota, United States

Xinyi Li
Northwestern University
Evanston, Illinois, United States

Avinash Akella
Grouplens Research, University of Minnesota
Minneapolis, Minnesota, United States

Joseph A. Konstan
Grouplens Research, University of Minnesota
Minneapolis, Minnesota, United States

ABSTRACT

This paper explores the effectiveness of using large language models (LLMs) for personalized movie recommendations from users' perspectives in an online field experiment. Our study involves a combination of between-subject prompt and historic consumption assessments, along with within-subject recommendation scenario evaluations. By examining conversation and survey response data from 160 active users, we find that LLMs offer strong recommendation explainability but lack overall personalization, diversity, and user trust. Our results also indicate that different personalized prompting techniques do not significantly affect user-perceived recommendation quality, but the number of movies a user has watched plays a more significant role. Furthermore, LLMs show a greater ability to recommend lesser-known or niche movies. Through qualitative analysis, we identify key conversational patterns linked to positive and negative user interaction experiences and conclude that providing personal context and examples is crucial for obtaining high-quality recommendations from LLMs.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Users and interactive retrieval**; **Recommender systems**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Large Language Model, Generative AI, Recommender System, Human-AI Interaction

ACM Reference Format:
Ruixuan Sun, Xinyi Li, Avinash Akella, and Joseph A. Konstan. 2024. Large Language Models as Conversational Movie Recommenders: A User Study. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large language models (LLMs) have been developed extensively and applied to various domains to assist with a range of human tasks [6, 10, 30]. Previous studies have examined the integration of LLMs into recommender systems (RecSys) using two main approaches: 1) Using LLMs in the RecSys training process, such as testing fine-tuning techniques for candidate generation [62], generating latent space or item descriptions [2, 48, 57], and designing personalized prompts for downstream recommendation tasks [17]; 2) Leveraging LLMs to simulate user agents [22] or directly evaluate recommendation results [14, 56].

However, there is a shortage of studies focusing on real user evaluations of LLM-based recommenders, an essential but challenging step in gauging practical recommender system (RecSys) performance [26]. Given their interactive and question-answering nature [55], LLMs are ideal for tasks that aim to make recommendations more transparent to users [5]. Real user evaluations can offer deeper insights into recommendation quality by focusing on human-centric values like interpretability, trustworthiness, and why users like or dislike certain recommended items [25, 42]. These evaluations can help clarify the underlying characteristics that contribute to user satisfaction with recommendations.

In this study, we explore the gap by understanding the quality of recommendations generated by pre-trained open-source LLMs under differently-personalized system prompts and three distinct scenarios. With hundreds of active users recruited from an online movie recommender system, we develop a chatbot user interface and assess their natural interaction experience to get personalized recommendations from LLMs in an online field experiment. By analysing survey response and users' conversation patterns, we address the following three research questions:

***RQ1**: How do users perceive the LLM recommender compared to a classic recommender experience?*

***RQ2**: How do different prompts, scenarios, or users' native consumption factors influence perceived LLM recommendation qualities?*

***RQ3**: What are some effective interaction strategies users can employ with the LLM to attain more satisfactory recommendations?*

We outline the main contributions of this paper as follows:

- We confirm that out-of-the-box LLMs excel in providing explainable recommendations and creating an interactive user experience. However, they struggle with generating personalized, novel, diverse, serendipitous, and trustworthy movie recommendations compared to traditional recommenders.

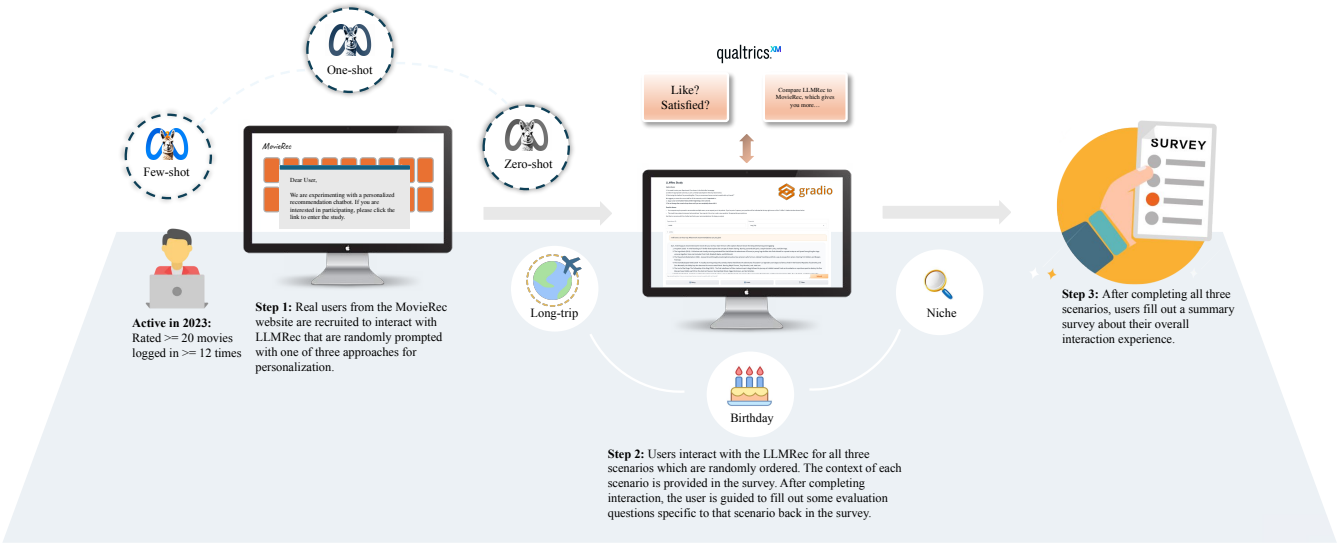

Figure 1: Overview of the LLM recommender user study design.

- We find that different prompting techniques do not significantly impact user-perceived recommendation quality and experience. We also reveal that user consumption history plays a significant role in how they appreciate recommendations generated by LLMs. Moreover, LLMs better accommodate niche or unpopular movie recommendations, compared to more personalized or ask-for-others requests.
- An analysis of user conversation patterns shows that providing context or examples is crucial for improving recommendation quality and user satisfaction with LLM recommenders.

In the rest of this paper, we first review relevant literature, then share details of the study procedure, followed by presenting the results of RQs and analysis with both statistical and qualitative data from the survey response. Finally, we synthesize the findings and design implications for future LLM-RecSys applications.

2 RELATED WORK

2.1 Large Language Models in Recommendation

Trained on vast textual corpora, LLMs have demonstrated remarkable capabilities on a variety of tasks relying on textual instructions [44]. Their applications extend beyond natural language processing (NLP) [43] into fields like education [4, 24] and medicine [50]. To harness the full potential of LLMs, several studies have examined prompting techniques to improve their performance. These techniques include one-shot or few-shot prompting, where the model learns from provided examples [34], and more advanced strategies such as 'Chain of Thought' (CoT) [54] and 'Tree of Thoughts' (ToT) [58], which introduce structured reasoning to generate responses.

As LLMs' capabilities grow, an increasing body of research focuses on their use in RecSys. This work often frames recommendation tasks as language-based problems and adapts LLM architectures [17, 28, 29, 63]. Another emerging trend is applying LLMs to RecSys tasks in a zero-shot setting. For example, Liu et al. and Dai et al. examined ChatGPT's potential for RecSys tasks [13, 32], while He et al. explored LLMs as zero-shot recommenders in conversational settings [20]. Sanner et al. found that LLMs are competitive with cold-start recommenders, especially with natural language preferences, and that few-shot prompting outperforms zero-shot [46]. However, these studies mainly focused on offline evaluations, using LLMs solely as recommendation engines while neglecting online user experiences.

To bridge the gap between theoretical assessments and real-world applications of LLM-based recommenders, we conduct live experiments with actual users of an operating movie recommender site. These experiments involve users interacting with LLMs via a real-time chatbot interface to get recommendations. Our study examines not only the impact of prompting techniques on offline RecSys tasks, as outlined in [31], but also how different prompting approaches—specifically zero-shot, one-shot, and few-shot—affect user experiences in practical recommendation scenarios. By gathering direct feedback from users, we gain valuable insights into their engagement and evaluate the LLM's recommendation performance in real-world conditions.

2.2 RecSys User Experience

User experience studies in recommendation systems (RecSys) are crucial for creating solutions that meet user needs and expectations. Researchers like Pu et al. and Knijnenburg et al. proposed human-centric frameworks to evaluate RecSys user experiences [25, 41]. Agner et al. examined the effectiveness of machine-learning-based streaming media recommendation systems, focusing on users' mental models [3]. Liu and Wang investigated how different types of explanations could improve user trust and satisfaction through online experiments [33]. Narducci et al. explored user interactions with chatbot recommenders powered by PageRank algorithms [38]. However, these studies did not specifically focus on user experiences when interacting with LLMs in RecSys.

Granada et al. conducted a study on an LLM-based conversational RecSys, assessing its performance with personalized and non-personalized recommendations. They evaluated accuracy, safety, and fairness [18]. Silva et al. leveraged ChatGPT to generate personalized movie recommendation explanation to users and evaluated user perception of personalization, effectiveness, and persuasiveness with a user study [48]. In contrast, our study focuses on user experiences with recommendations generated directly by the LLM, without relying on predefined candidate lists. Along with evaluating the impact of personalized and non-personalized recommendations, we also examine how different prompting techniques affect user experiences. Using metrics like diversity, novelty, serendipity, trustworthiness, and explainability, we ask users to compare recommendations from a baseline RecSys with those from an LLM-based system. This helps understand how to best integrate LLMs into existing RecSys to improve user experiences.

3 STUDY DESIGN

As illustrated in Fig. 1, our study design can be split into three phases: 1) participant recruitment and personalized system prompt generation; 2) interaction with LLM-based recommender under different scenarios; and 3) scenario-based and summary surveys including various evaluation metrics.

3.1 Participants and Personalized Prompts

We partnered with a movie recommendation platform (**MovieRec**[1]) to select qualified participants for our study. To ensure that users had substantial experience with recommender systems, providing a baseline for comparison in survey questions, we recruited only active users who logged into the system at least 12 times and rated more than 20 movies in 2023. Using these criteria, we identified 3,031 qualified users from the platform's database. Recruitment messages and access to the questionnaire were displayed prominently on users' homepages upon their initial login during the experiment period. Participation was entirely voluntary, with no incentives provided, promoting more thoughtful responses to survey questions. Our study design was determined as a non-human subject study by our institutional review board (IRB).

For the recommendation task, we used the pre-trained Llama2-7b-Chat [51], an open-source model released by Meta in 2023, designed for conversational dialogues [2]. Due to privacy concerns related to user data, we hosted the model on a local server instead of using a cloud-based API. We differentiated three prompting techniques based on the level of personalized user context provided, as outlined in Fig. 2. Beyond the shared general guidelines, the zero-shot prompt only contained the top three rated genres based on each user's historical ratings. The one-shot prompt included three additional lists that each contains one movie most recently liked (top 10% of user's all ratings), disliked (bottom 10% of user's all ratings), and recommended for the user based on the platform's existing recommendation model. The few-shot prompt extended this by listing four movies in each category. This design was inspired by previous research, suggesting that providing more than five example movies could lead to unstable outcomes [47]. Additionally, we categorized each movie's popularity into three tiers: high, medium, and low, based on their respective percentiles in the MovieRec database based on its total number of user ratings (high: top 20%, medium: 60%-80%, low: below 60%).

3.2 User Interface and Scenarios

For user interaction, we developed the chatbot interface using Gradio [3], as shown in Appendix Fig. A1. Throughout the rest of the paper, we'll refer to this interface as **LLMRec**. Users were encouraged to interact with the chatbot for as long as they needed in each scenario. To mitigate potential latency issues arising from multiple users making inference requests, we ran pilot tests to ensure the model could handle real-time inference for at least two users at once. Additionally, we included a disclaimer message and an estimated wait time on the chatbot interface to keep users informed about potential generation queues during their interactions.

Upon entering the survey, users were greeted with an informational page followed by three common movie recommendation scenarios: 1) *Ask-for-others – Friend's or Family's Birthday*; 2) *Personalization – Long Trip*; and 3) *Unpopular – Something Niche*. Detailed description of each scenario is included in Fig. 3. To mitigate potential position bias, the order of scenarios was randomized for each user within the questionnaire. This ensured that users' prior experiences did not influence their responses to subsequent scenarios.

3.3 Evaluation Metrics

When finishing their conversation with LLMRec in each scenario, users were directed back to the survey to answer a set of questions. These questions covered: 1) their general appreciation of the recommendation scenario, focusing on enjoyment and satisfaction; and 2) a comparison between their experiences with MovieRec and LLMRec, measured by factors such as *Personalization, Diversity, Novelty, Serendipity, Trustworthiness*, and *Explainability*. The detailed phrasing of the survey questions for each scenario is shown on the left side of Fig. 4, inspired by previous designs from Ekstrand et al. and Nguyen [15, 39].

After completing the scenario evaluations, users were directed to a summary page to provide feedback on their overall interaction experience with LLMRec. Drawing inspiration from chatbot survey designs in Chaves and Gerosa's work, we asked users to rate various aspects of chatbot interaction, including *Responsiveness, Understandability, Helpfulness*, and whether they would like to see LLMRec integrated into the MovieRec platform in the future [11]. The survey concluded with an optional free-response question, allowing users to share additional, unstructured insights.

4 RESULTS

Our experiment ran for one month from 02/13/2024 until 03/14/2024. In total, 449 users enrolled in the study (i.e. opening the questionnaire), 178 submitted the questionnaire, and we build our analysis with 160 unique users who have completed all quantitative questions except the optional free response. The distribution of prompting techniques is relatively equal in size, with 56 users in zero-shot, 48 users in one-shot, and 56 users in few-shot bucket. The median

[1]The platform's real name is anonymized and will be revealed upon publication of this paper.
[2]Download link: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[3]https://www.gradio.app/

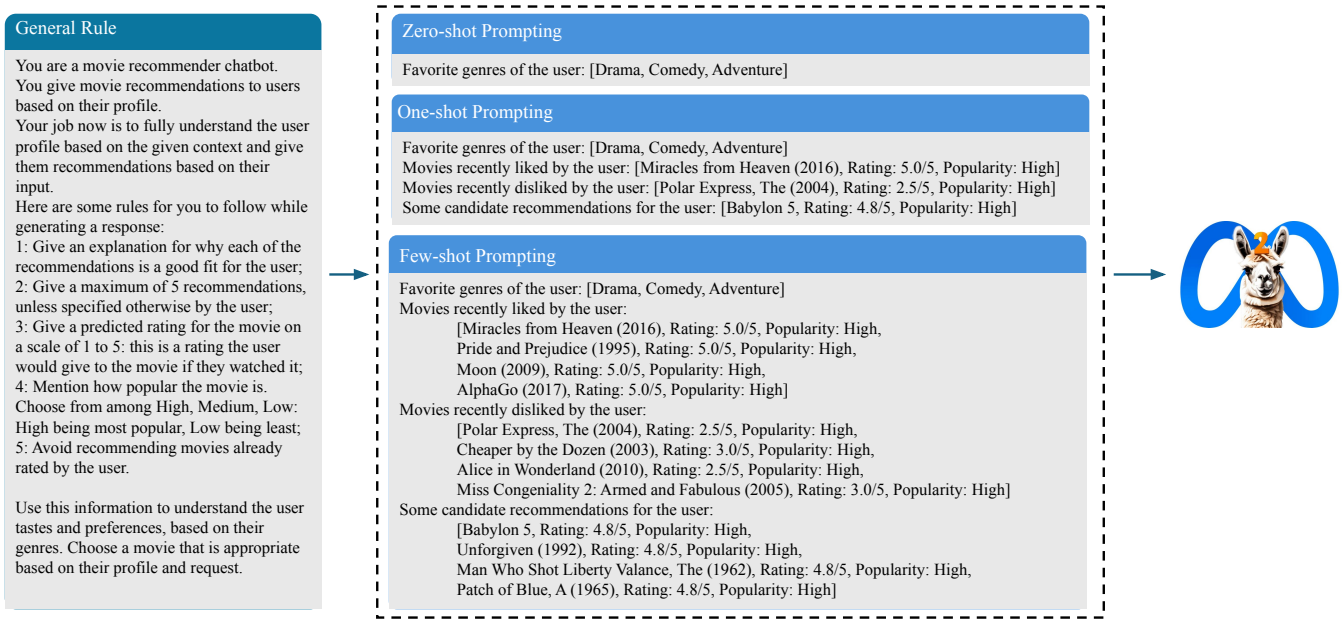

Figure 2: Diagram depicting the general rule and randomly selected prompting technique of the LLM for engaging with users at the 1st phrase of the user study.

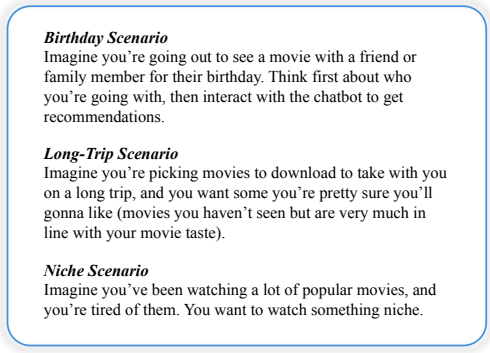

Figure 3: Three recommendation scenarios.

of user response time is 17.62 minutes for completing evaluation of all three scenarios.

4.1 User-assessed LLMRec Quality

Based on user survey responses comparing their MovieRec to LLMRec experiences, we categorized responses into three groups: *prefer MovieRec*, *About the Same*, and *prefer LLMRec*. In comparison to MovieRec, LLMRec stood out for its explainability, but users generally found MovieRec to offer more novel, diverse, accurate, and trustworthy recommendations across all three scenarios (Fig. 5).

To gain a deeper understanding of individual user experiences, we conducted a qualitative analysis of the free-response data (N=137) using the classic grounded theory method (GMT) [8]. In this approach, three researchers who were trained in GMT independently coded the user responses. To ensure accuracy, we cross-checked 10% of each other's coding at random. Subsequently, all researchers met to do thematic analysis [12], assigning each code to one or more topic clusters and iterating this process until all ambiguities were resolved. Ultimately, we identified three major themes.

4.1.1 Defect of Recommendation Quality. The first major theme we identified revolves around the attributes where LLMRec fell short. Many users complained about a lack of personalization (N=39) and novelty (N=56) in the recommendations they received. As participant P50 remarked, *"...I received repetitive recommendations for movies I had already watched. While it correctly grasped the concept of gore and violent films, it predominantly suggested love stories for me and my wife, which lacked the variety I anticipated."* Another significant issue users noted was low diversity (N=10) and high-popularity (N=18). P376 summarized this by saying, *"It pretty much always stuck to only the most popular movies of all time."*

4.1.2 Interactivity, Explainability, and Context. In this theme, users shared their positive interaction experience with LLMRec (N=49). P153 praised, *"I like that it is interactive and takes many of my already established preferences into consideration when making recommendations."* Building on this interactivity, users like P368 also enjoyed getting explanations on why they got certain recommendations, *"I like that it explains why I might like the films–this is far better than the traditional MovieRec interface."* In addition to explanations, users found LLMRec supported context-oriented recommendation well, accommodating mood, interests of others, or even quickly extracting preferences from new users. As P407 remarked,*"I like the additional option to tell the chatbot some information about what I want to watch which I can't do if I only rate movies. Additionally it can be helpful to get recommendations faster, if you're new to recommender services and haven't rated that much yet."*

4.1.3 Control, Trust, and Transparency. Despite generally positive interactions, users expressed concerns about the reliability of LLMRec (N=42). One issue was the difficulty in user control. As P169

Scenario Questions

Metric	Survey Question
Enjoyment	I like this interaction scenario.
Satisfaction	The recommended movies satisfy my needs.

Strongly Disagree <---------->Strongly Agree

Personalization	Which gives you more recommendations matching your interest?
Diversity	Which gives you more recommendations with higher variety?
Novelty	Which gives you more recommendations you have never heard about?
Serendipity	Which gives you more unexpected but enjoyable recommendations?
Trustworthiness	Which gives you more trustworthy recommendations?
Explainability	Which gives you more explainable recommendations?

Definitely MovieRec <---------->Definitely LLMRec

Summary Questions

Metric	Survey Question
Responsiveness	LLMRec was responsive to my instructions.
Understandability	LLMRec understood my goals.
Usefulness	LLMRec helped me get useful movie recommendations.
Future Use	I would interact with LLMRec as a feature in the system in the future.

Strongly Disagree <---------->Strongly Agree

Can you share some thoughts on what you like and dislike about getting recommendations from LLMRec?

Free Response

Figure 4: Survey Question as evaluation metrics.

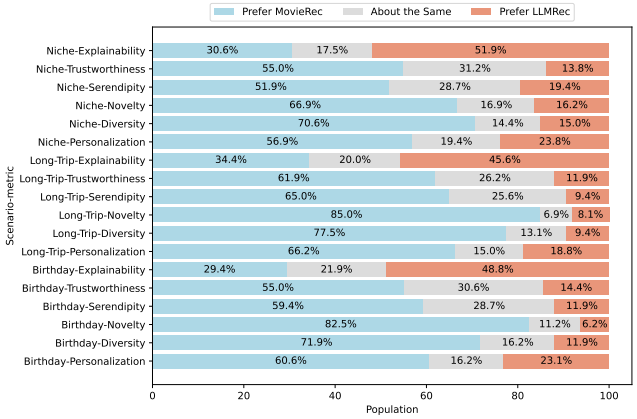

Figure 5: User perception of recommendation quality from LLMRec compared to the classic MovieRec experience. All LLMRec-MovieRec pairs are tested with paired t-test and shows p <= 0.05 statistical significance except for the Birthday-Explainability and Long-Trip-Explainability.

shared, *"...In the road trip scenario, it got fixated on horror movies and started only suggesting those, even though I dislike them very much."* Similar to previous empirical study on LLMs [21], we also observed hallucination in recommendation scenarios, which deeply undermines user trust. P47 commented, *"I am also wary about the chatbot suggestions because I know it can often hallucinate and not understand me clearly etc. I feel like I can't rely fully on chatbot, compared to relying on my own judgement and judgement of other users that in regular movie selection experience."* This sense of disappointment prompted users to call for more transparency in LLM recommendations. P23 noted, *"It was also a black box, I like knowing that the algorithm is SVD based (for example), as opposed to whatever the collaborative intelligence model has scraped from the internet."* With data from the MovieRec-LLMRec comparison and qualitative analysis, we answer ***RQ1***:

RQ1: *How do users perceive the LLM recommender compared to a classic recommender experience?*

Our findings suggest that while users found LLM recommenders superior to classic systems in terms of explaining recommended movies, they also noted a lack of algorithmic transparency in the recommendation process. Users appreciated the LLM's smooth interaction flow and its ability to adapt to customized recommendation contexts. However, the recommendations were often non-personalized and skewed toward popular, homogeneous content, which diminished user trust and reduced the perceived reliability of the LLM recommender.

4.2 Effect of Prompt, Scenario, and Rating

In the second stage of our analysis, we used a between-subject prompt-wise test and a within-subject scenario-wise test to examine differences across various response groups. We applied one-way ANOVA and pairwise Tukey's HSD [1] tests to assess statistical significance. Surprisingly, we did not find significant differences in evaluation metrics between different prompts (see Appendix Table A1). However, our ANOVA tests revealed group-wise statistical differences in the means for *Enjoyment, Satisfaction, Novelty*, and *Serendipity* scores (Fig. 6 and Table 1). Notably, users found the *Niche* scenario more enjoyable and satisfying compared to the *Trip* scenario. They also felt they received more novel and serendipitous recommendations in the *Niche* scenario than in the other two scenarios.

	F-val	P-val	B-T	B-N	T-N
Enjoyment	3.244	.040	-	-	.363 (.031)
Satisfaction	4.239	.015	-	-	.419 (.011)
Novelty	8.461	.000	-	.410 (.002)	.438 (.001)
Serendipity	3.476	.032	-	-	.319 (.027)

Table 1: AVOVA and pairwise Tukey's HSD test results for different scenario metrics. Pairwise stats include the mean difference and p-val in bracket. N: Niche, T: Trip, B: Birthday.

The absence of differences across prompting techniques led us to investigate other potential factors affecting user perceptions beyond LLM attributes. By querying the MovieRec database, we categorized users into three groups based on their total number of ratings, as outlined in Table 2. This allowed us to determine if a user's historical movie consumption correlated with variations in recommendation and chatbot quality metrics (Fig. 6 and Table 3).

Using ANOVA, we found statistically significant differences among these groups for several metrics, including recommendation *Enjoyment, Satisfaction, Personalization, Novelty, Serendipity*, and chatbot *Responsiveness, Understandability*, and *Usefulness*. Light

	percentile	# ratings	# users
light-rater	<=25%	<=406	40
medium-rater	25 -75%	406-1522	80
heavy-rater	>=75%	>=1522	40

Table 2: User rating buckets

	F-val	P-val	L-M	M-H	L-H
Enjoyment	7.360	.001	-.429(.007)	-	.600(.001)
Satisfaction	8.195	.000	-	-.417(.011)	-.658(.000)
Personalization	6.976	.001	-	-.350(.037)	-.609(.001)
Novelty	7.718	.001	-.396(.003)	-	.508(.001)
Serendipity	9.087	.000	-	-.363(.009)	.592(.000)
Responsiveness	5.290	006	-	-.538(.047)	-.825(.005)
Understandability	4.626	.011	-	-	-.750(.010)
Usefulness	5.349	.006	-	-.650(.018)	-.825(.008)

Table 3: AVOVA and pairwise Tukey's HSD test results for different rater metrics. Pairwise stats include the mean difference and p-val in bracket. L: Light, M: Medium, H: Heavy.

raters found LLMRec more enjoyable and novel compared to medium and heavy raters. Heavy raters, however, were less satisfied with LLMRec, reporting it was less personalized, less responsive to their requests, and less likely to generate unexpected recommendations. They also indicated that LLMRec did not understand their needs as well as it did for those who had rated fewer movies in the past.

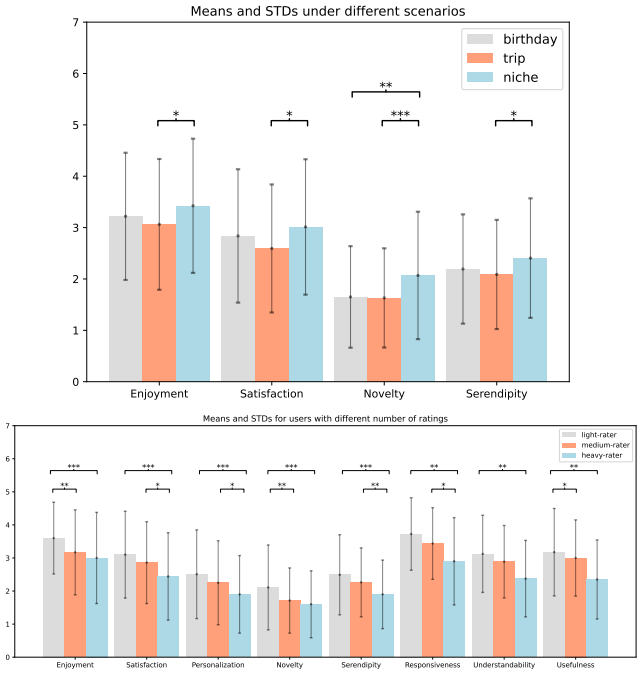

Figure 6: Within-subject and between-subject test results. Top: Differences of recommendation quality from different scenarios. Bottom: Differences of recommendation quality for users based on their historic rating count. Asterisk (*) indicates p-val between conditions with Tukey's HSD test: $p < .1$ (*); $p < .05$ (**); $p < .01$ (***).

We then answer RQ2 with our findings:

***RQ2**: How do different prompts, scenarios, or users' native consumption factors influence perceived LLM recommendation qualities?*

We found no significant differences in user perceptions based on varying personalization prompts. Among the three scenarios, users derived more enjoyment and felt their needs were better met when asking for niche recommendations. This scenario also provided users with more novel and unexpected movie suggestions. Additionally, the level of users' past movie consumption proved to be a critical factor influencing their satisfaction with LLM recommendations and the chatbot's overall performance. Users who had rated more movies tended to find the recommendations less satisfactory and personalized, which likely contributed to a more negative view of the LLM recommender's responsiveness and overall usefulness.

4.3 Useful Conversation Strategies

In our third analysis stage, we examined user conversation data with LLMRec to identify relationships and patterns tied to different types of user experiences. Our initial step involved running an Ordinal Linear Regression (OLR) to evaluate the correlation between the number of sentences users exchanged with LLMRec and their satisfaction and understandability score. The results indicated significant negative relationships for both Satisfaction ($coef$ = -0.079, p = 0.041) and Understandability ($coef$ = -0.118, p = 0.002), implying that users who engaged in shorter dialogue feel the LLMRec had a better overall goal understanding. Also, those who engaged in longer conversations with LLMRec felt were less likely to be satisfied with their recommendations.

4.3.1 Pattern Coding. Building on this initial finding, we conducted a qualitative analysis to identify conversation patterns associated with user responses to the *Satisfaction* question. We categorized responses into positive (i.e., "Agree" and "Strongly agree") and negative (i.e., "Disagree" and "Strongly disagree"). Using the GMT and thematic analysis, we identified 10 key patterns that correlated with both positive and negative user conversations.

Figure 7 illustrates that when users provided specific information such as previously watched movies, expressed their preferences (likes or dislikes), or specified genres, their satisfaction with recommendations tended to be higher. In contrast, users who offered minimal context and expected LLMRec to infer all their preferences on its own, treating it like a search engine, generally had lower satisfaction. Based on these insights, we refined the conversation patterns and defined five discrete tags to characterize conversation attributes, as shown in Table 4. Some example conversation classified under these tags can be found in Appendix Table A2.

4.3.2 Regression Model. We further analyzed all 417 user conversation queries across different scenarios. A researcher manually labeled these conversations as "Applicable" or "Not Applicable" under each of the five tags. To ensure validation, two other researchers independently labeled 20% of the conversation data, checking for inter-rater reliability with Cohen's Kappa and Krippendorff's Alpha coefficients [19], both yielding values of 0.652, indicating substantial agreement among raters.

We built Ordered Logistic Regression (OLR) models to predict user survey evaluation metrics based on five conversation tags, as summarized in Table 5. A key finding was that the main factor

Tag	Description
Dialogue	When users have a conversation by sharing details and giving feedback on how they feel about the recommendations. It should also be like talking to humans instead of a search engine, not always dumping questions or steering the response to what they want.
Context	When users clearly specify their tastes/requirements/preferences and what they're looking for. Specifically, they need to come up with something on their own, not just reuse the scenario context we provided in the survey.
Steering	When users try to make the recommendations converge to their choice without clarifying their objective. They may have multiple rounds of conversation and iteratively try to nudge the chatbot to get closer to their goal.
Testing	When users try to see how the system works and usually test it negatively. Some examples include asking questions outside of movie recommendations, or trying to decode the system prompts the bot was based on, or using inappropriate language.
Retry	When users re-generate a response to their previous query. This can be identified by repeated queries in conversation history.

Table 4: Main tags associated with user conversation with LLMRec.

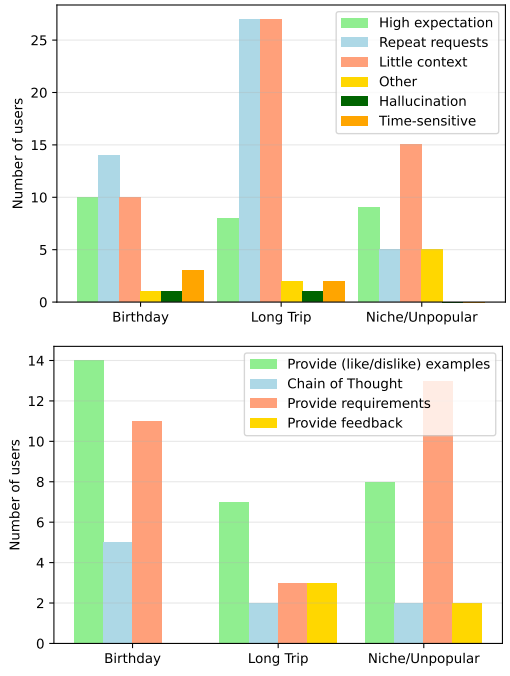

Figure 7: User conversation patterns based on qualitative coding. Top: the distribution of patterns in negative conversation. Bottom: the distribution of patterns in positive conversation.

contributing to user satisfaction with recommendations was the context provided by users. As shown in Fig. 7, users who were dissatisfied with LLMRec typically shared less context compared to those who were satisfied and provided specific examples and requirements. The regression results also suggest that providing context helps LLMs generate more personalized and diverse recommendations. Additionally, context positively influences LLMRec's responsiveness, understandability, and general helpfulness. Meanwhile, *Dialogue* had a negative impact on user-perceived understandability. This indicates that even when users communicated with the bot in plain, conversational language, they often felt that the bot didn't fully understand their objectives.

Users who tried to test LLMRec generally felt less satisfied with their interactions. However, since their goal was not to get movie recommendations, the negative impact of the *Testing* label on recommendation or general interaction metrics is not necessary of concern. Additionally, there was a positive relationship between the *Retry* label and recommendation explainability, indicating that repeating the same queries might improve understanding of why specific recommendations were made. However, Fig. 7 shows that repeating the same request is among the top patterns linked with negative interaction experiences, while using a chain-of-thought strategy could lead to higher satisfaction with the recommendations. With all the findings, we answer ***RQ3***:

RQ3*: What are some effective interaction strategies users can employ with the LLM to attain more satisfactory recommendations?*

Users need to provide adequate context and details about their preferences or requirements while asking LLM for recommendations. Providing examples of what they like or dislike is a good strategy to make LLM actually learn and expand upon. While having a human-like dialogue might not make the LLM understand user goals better, simply dumping queries and treating the LLM-Rec as a search engine without follow-up feedback and tuning is also not the optimal way to get good recommendations. Moreover, repeating the same queries might temporarily improve the explainability of recommendations, but that could not directly contribute to their ultimate satisfaction goals. In fact, some highly-satisfied users provided feedback or executed chain-of-thought strategy in their conversation. In addition, users need to be reminded about the ability boundary of LLM recommenders, that they should focus their conversation in the specific application domain, but not challenge the bot with irrelevant topics.

5 DISCUSSION

In this section, we summarize our novel findings and future design implications of LLM recommenders from user experience perspectives in two folds: 1) How can RecSys researchers better utilize and improve current LLM modeling, prompting, or information retrieval strategies to better accommodate user needs; and 2) How can users be better informed or educated to achieve a more satisfactory interaction experience from LLM recommenders, without requiring specific technical background or understanding of terminology?

5.1 Implications for LLM-Rec designers

Our findings showed that few-shot prompts didn't significantly improve recommendation quality compared to zero-shot or one-shot approaches. This aligns with Dai et al., who noted that recommendation quality doesn't always improve with more examples [13]. To address the lack of personalization and novelty in our findings, retrieval-augmented generation (RAG) has been proposed, allowing

	Satisfaction	Personalization	Diversity	Explainability	Responsiveness	Understandability	Helpfulness	Future Use
Dialogue	-.375(.086)	.031(.890)	-.051(.821)	.040(.854)	-.269(.240)	**-.427(.050)**	-.098(.665)	-.064(.775)
Context	**.464(.029)**	**.584(.007)**	**.748(.001)**	.210(.317)	**.564(.011)**	**.465(.027)**	**.513(.017)**	.188(.382)
Steering	.056(.793)	-.127(.560)	.003(.991)	.078(.715)	-.124(.575)	-.281(.188)	.162(.448)	.186(.382)
Testing	**-.938(.008)**	**-1.061(.010)**	-.816(.056)	**-.778(.028)**	.301(.404)	-.645(.057)	-.456(.208)	**-1.026(.003)**
Retry	-.021(.960)	.428(.299)	.598(.119)	**.845(.041)**	-.471(.212)	-.011(.977)	.201(.608)	.620(.118)

Table 5: OLR Analysis of conversation label features to user survey response metrics. Each column of metrics is run with one single OLR model. Value in each cell is formatted as coef (p-val).

for user and item context to be integrated into LLM-based recommendations [27]. Research by Di Palma has looked into retrieval-augmented RecSys with offline datasets [14], finding it effective for re-ranking more diverse, though potentially less relevant, recommendations [9]. However, success with RAG requires careful selection of embedding and retrieval models, with field tests needed to validate effectiveness.

LLMs' native recommendation capabilities are often limited by their training data, which tends to be biased toward popular content [61]. However, LLMs offer two distinct advantages over traditional recommendation systems: 1) they excel at explaining recommendations using personal context, as noted by Acharya et al. [2], Wang et al. [53], and Silva et al. [48]; 2) they can learn in-context, adapting to users' real-time interests during recommendation [7, 45]. We suggest that future practitioners leverage these benefits by integrating LLMs into recommendation systems to improve user experience. Additionally, a hybrid approach can help reduce LLM-generated content hallucination by anchoring recommendations in grounded, genuine item databases.

The qualitative analysis of survey responses revealed several areas for improving LLM-based recommenders. Users suggested incorporating beyond-text recommendations, like movie posters or trailers, for a more intuitive experience. This could be achieved with visual-enhanced chatbots or by exploring multimodal recommendation generation [59]. Users also wanted LLMRec to improve short-term memory and retain context throughout a conversation session, as seen in MemGPT [40]. Additionally, users requested LLMRec to be more proactive in seeking context and feedback during recommendations. Researchers might experiment with different LLM personas [23, 37] to determine the optimal level of proactivity and user-customizable recommendations.

Another issue raised was the balance between AI-generated content censorship and recommendation quality. Users proposed confidence thresholds and disclaimers to address LLM hallucinations and misinformation [16, 35, 36]. However, some users, like P241, expressed concerns about overly strict content guardrails: *"...the guardrails were a bit too strict. I wanted to ask the system for a more naughty movie, but was prevented as it would not recommend anything with violence or sex. I can see why this is a problem in the US, but it is also a problem if the chatbot is unable to recommend high-quality movies just because they have a violent or sexual theme."* This highlights a research gap for future studies to explore the right balance of AI ethics in LLM recommenders.

5.2 Learnings for End Users

The success of LLM recommenders cannot be achieved without well-informed users with proper goals and interaction flows. The first learning we summarize from the study is to remind users to provide adequate context in order to get more personalized and specific recommendations that can better satisfy their needs, echoed by what Vaithilingam et al. and Skjuve et al. identified before [49, 52]. As we observed from the lack of difference in prompting techniques, it is challenging for designers to include the most relevant context to system prompt since users' needs are constantly changing. Future studies can explore different ways to encourage more context from users to benefit in-context learning [7, 45]. Some solutions can be tuning a proactive LLM that asks users for more contextual information. However, such model needs to be carefully tuned to respect the boundary between useful context and sensitive private information, and also keep conversation length in mind because users are impatient. Another way can be introducing some query examples for users to mimic or follow before they start their interaction.

Another important implication is to inform users about LLM recommender's capability and set reasonable expectations, similar to what Zamfirescu-Pereira et al. found before [60]. As reflected in Fig. 7, one of the top patterns associated with conversations of unsatisfied users it having too high expectation, such as P54 shared, *"While the chatbot claims to know my profile, it clearly does not. It claimed that I recently disliked the Matrix, that is not true, I liked it, and it was a long time ago. But was right about the other movies."* This user case implies that while we need to set more strict rules for LLM to avoid such hallucination, we also need to help users construct the right expectation and understand there should be a room for LLM to make mistakes, which is also the case to encourage them to correct any wrong information and allow LLMs to learn. For example, users can be educated about how to utilize chain-of-thought interaction strategy to elicit step-wise reasons of making certain recommendations and iteratively improve the recommendations they get.

Finally, we also find future opportunity for developing built-in mechanism in LLM to detect users' off-the-track behavior and advise them to keep their conversation within the certain recommendation domain. Results from Fig. 7 and Table 5 both suggest that having irrelevant topics to recommendation or testing and challenging the LLM recommender contribute to negative user experience. To avoid unintentional drifting from the recommendation theme, users should be gently reminded of such behavior, either by LLM directly from conversation or from some extra UI alert.

5.3 Limitations

Our study has several limitations. First, we were unable to test with more advanced pre-trained LLMs due to hardware constraints. Second, our prompts only incorporated recent user ratings, without exploring a broader range of user context combinations. Third, we only applied qualitative coding to user conversation data, without utilizing other NLP techniques for analysis. Addressing these

limitations could be valuable for future research to enhance understanding in this domain.

6 CONCLUSION

In this study, we conducted a field experiment with LLM-based movie recommenders using zero-shot, one-shot, and few-shot personalized prompts to evaluate user interaction across three recommendation scenarios. By analyzing real-user conversations and semi-structured survey responses, we observed the following: 1) Users valued LLMs for their superior explainability and interactivity but found their generated recommendations not as good compared to those from classic recommender systems; 2) Recommendations for unpopular movies were more enjoyable and better met users' needs than those for personalized or ask-for-others scenarios. Additionally, the number of movies a user has watched significantly influenced their perception of LLM recommenders; 3) Providing personal or example-based context can lead to more personalized and satisfactory recommendations from LLMs. We believe these findings offer a useful starting point for future research aiming to enhance personalized interaction experiences with LLM-based recommender applications.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design* 3, 1 (2010), 1–5.

[2] Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1204–1207.

[3] Luiz Agner, Barbara Necyk, and Adriano Renzi. 2020. Recommendation systems and machine learning: Mapping the user experience. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 3–17.

[4] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.

[5] Junseong Bang, Byung-Tak Lee, and Pangun Park. 2023. Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI. In *2023 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 109–113.

[6] Lenz Belzner, Thomas Gabor, and Martin Wirsing. 2023. Large language model assisted software engineering: prospects, challenges, and a case study. In *International Conference on Bridging the Gap between AI and Reality*. Springer, 355–374.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Antony Bryant. 2013. The grounded theory method. In *Reviewing qualitative research in the social sciences*. Routledge, 108–124.

[9] Diego Carraro and Derek Bridge. 2024. Enhancing Recommendation Diversity by Re-ranking with Large Language Models. *arXiv preprint arXiv:2401.11506* (2024).

[10] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems* 47, 1 (2023), 33.

[11] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[12] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.

[13] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1126–1132.

[14] Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1369–1373.

[15] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. 2014. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*. 161–168.

[16] Emilio Ferrara. 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* (2024), 1–21.

[17] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). *arXiv preprint arXiv:2203.13366* (2022).

[18] Mateo Gutierrez Granada, Dina Zilbershtein, Daan Odijk, and Francesco Barile. 2023. VideolandGPT: A User Study on a Conversational Recommender System. *arXiv preprint arXiv:2309.03645* (2023).

[19] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[20] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian Mcauley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (, Birmingham, United Kingdom,) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 720–730. https://doi.org/10.1145/3583780.3614949

[21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[22] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505* (2023).

[23] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547* (2023).

[24] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[25] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22 (2012), 441–504.

[26] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22 (2012), 101–123.

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[28] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized Prompt Learning for Explainable Recommendation. *ACM Trans. Inf. Syst.* 41, 4, Article 103 (mar 2023), 26 pages. https://doi.org/10.1145/3580488

[29] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Prompt-Based Generative News Recommendation (PGNR): Accuracy and Controllability. In *European Conference on Information Retrieval*. Springer, 66–79.

[30] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 374–382.

[31] Junling Liu, Chaoyong Liu, Renjie Lv, Kangdi Zhou, and Yan Bin Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. *ArXiv* abs/2304.10149 (2023). https://api.semanticscholar.org/CorpusID:258236609

[32] Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. arXiv:2304.10149 [cs.IR]

[33] Weizi Liu and Yanyun Wang. 2024. Evaluating Trust in Recommender Systems: A User Study on the Impacts of Explanations, Agency Attribution, and Product Types. *International Journal of Human–Computer Interaction* (2024), 1–13.

[34] Robert L Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353* (2021).

[35] Jonathan Luckett. 2023. Regulating Generative AI: A Pathway to Ethical and Responsible Implementation. *Journal of Computing Sciences in Colleges* 39, 3 (2023), 47–65.

[36] Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ digital medicine* 6, 1 (2023), 120.

[37] Ishani Mondal, S Shwetha, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the Humans and For the Humans: Harnessing LLMs for Generating Persona-Aware Slides from Documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2664–2684.

[38] Fedelucio Narducci, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2018. Improving the user experience with a conversational recommender system. In *AI* IA 2018–Advances in Artificial Intelligence: XVIIth International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20–23,*

2018, Proceedings 17. Springer, 528–538.
[39] Tien Tran Tu Quynh Nguyen. 2016. *Enhancing user experience with recommender systems beyond prediction accuracies*. Ph. D. Dissertation. University of Minnesota.
[40] Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560* (2023).
[41] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.
[42] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22 (2012), 317–355.
[43] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).
[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[45] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633* (2021).
[46] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*. 890–896.
[47] Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-shot recommendation as language modeling. In *European Conference on Information Retrieval*. Springer, 223–230.
[48] Ítallo Silva, Leandro Marinho, Alan Said, and Martijn C Willemsen. 2024. Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 597–608.
[49] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of ChatGPT: Findings from a questionnaire study of early users. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–10.
[50] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.
[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[52] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.
[53] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding User Experience in Large Language Model Interactions. *arXiv preprint arXiv:2401.08329* (2024).
[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[55] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.
[56] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. *arXiv preprint arXiv:2401.04997* (2024).
[57] Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. 2023. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487* (2023).
[58] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* 36 (2024).
[59] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.
[60] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
[61] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.
[62] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language Models as Recommender Systems: Evaluations and Limitations. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*.
[63] Zizhuo Zhang and Bang Wang. 2023. Prompt Learning for News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM. https://doi.org/10.1145/3539618.3591752

APPENDICES

	F-val	P-val	F-O	F-Z	O-Z
Enjoyment	.208	.812	.086(.825)	.071(.867)	-.015(.994)
Satisfaction	1.213	.298	-.197(.375)	-.191(.372)	.007(.999)
Personalization	1.030	.358	.095(.791)	-.113(.699)	-.208(.327)
Diversity	2.325	.099	-.235(.161)	-.232(.146)	.003(.100)
Novelty	2.930	.054	-.266(.081)	-.238(.112)	.028(.973)
Serendipity	2.002	.136	-.057(.894)	-.232(.132)	-.176(.342)
Trustworthiness	.124	.883	-.057(.893)	-.048(.916)	.009(.997)
Explainability	.008	.993	-.008(.999)	-.018(.992)	-.001(.998)
Responsiveness	1.020	.363	-.161(.772)	-.321(.329)	-.161(.772)
Understandability	.278	.758	-.161(.765)	-.125(.839)	.036(.987)
Usefulness	.928	.397	-.140(.837)	-.321(.365)	-.182(.742)
Future-Use	1.107	.333	-.095(.935)	-.375(.325)	-.280(.559)

Table A1: AVOVA and pairwise Tukey's HSD test results for different prompt metrics. Pairwise stats include the mean difference and p-val in bracket. F: Few-shot, O: One-shot, Z: Zero-shot.

Example Raw User Conversation with LLMRec	Dialogue	Context	Steering	Testing	Retry
"Can you recommend a few movies for me?"	NA	NA	NA	NA	NA
"Can you recommend a movie for a date. it has to be something i like but she has to like it too. shes a bit more normal then me', 'I dont want a rom com', 'something i havent seen', ':those are too weird for a first date with someone whos taste i dont know. also ive seen 3 of those. try again. Ideally it would be a movie that maves the viewer feel sophisticated', '2 and 3 are good recs thanks.'	A	A	A	NA	NA
'Can you recommend some movie for me', 'i have watched inception', 'i have already watched the matrix', 'Can you recommend some movie for me, 'but i have already watched inception', 'Can you recommend some movie for me', 'but i have already watched inception, the shawshank redemption and the lords of the rings'	NA	NA	A	NA	A
"Ok, that seems useless and tedious, but let's try this. Recommend 10 movies", "I've seen all of those. Recommend 10 more movies and don't recommend those again. Also, I hated The Godfather (1972) and The Big Lebowski (1998)", "I've seen all of those. Recommend 10 more movies I haven't seen. I liked all of them except The Lobster (2015)", 'Who directed The Lighthouse (2019)?', 'List all movies directed by Robert Eggers', "That isn't correct. Robert Eggers also directed The Northman and The Witch", "I've watched Eternal Sunshine of the Spotless Mind (2004), The Lighthouse (2019), The Florida Project (2017), The Killing of a Sacred Deer (2017), The Witch (2015), The Invitation (2015), The Love Witch (2016), and The Northman (2022). Recommend 10 more movies I haven't seen. Don't recommend horror films.", 'Stop', "Recommend 10 movies I haven't seen that my boyfriend might also like", 'I asked for 10 and you give me 5', "I can't find any information about The Prisoner (1977)"	A	A	A	A	A
'can you recommend some movies that are perfect for a road trip? preferably something with action, "recomend me a movie that I haven't seen", 'something like "shutter island"', "can you recommend me a good oscar worthy movie preferably something I haven't seen", 'can you give me list', "can you recommend me a list of good Thiller-worthy movies preferably something I haven't seen"	NA	A	A	NA	NA

Table A2: Example user conversations and their associated taggings. A denotes Applicable, NA denotes Not Applicable.

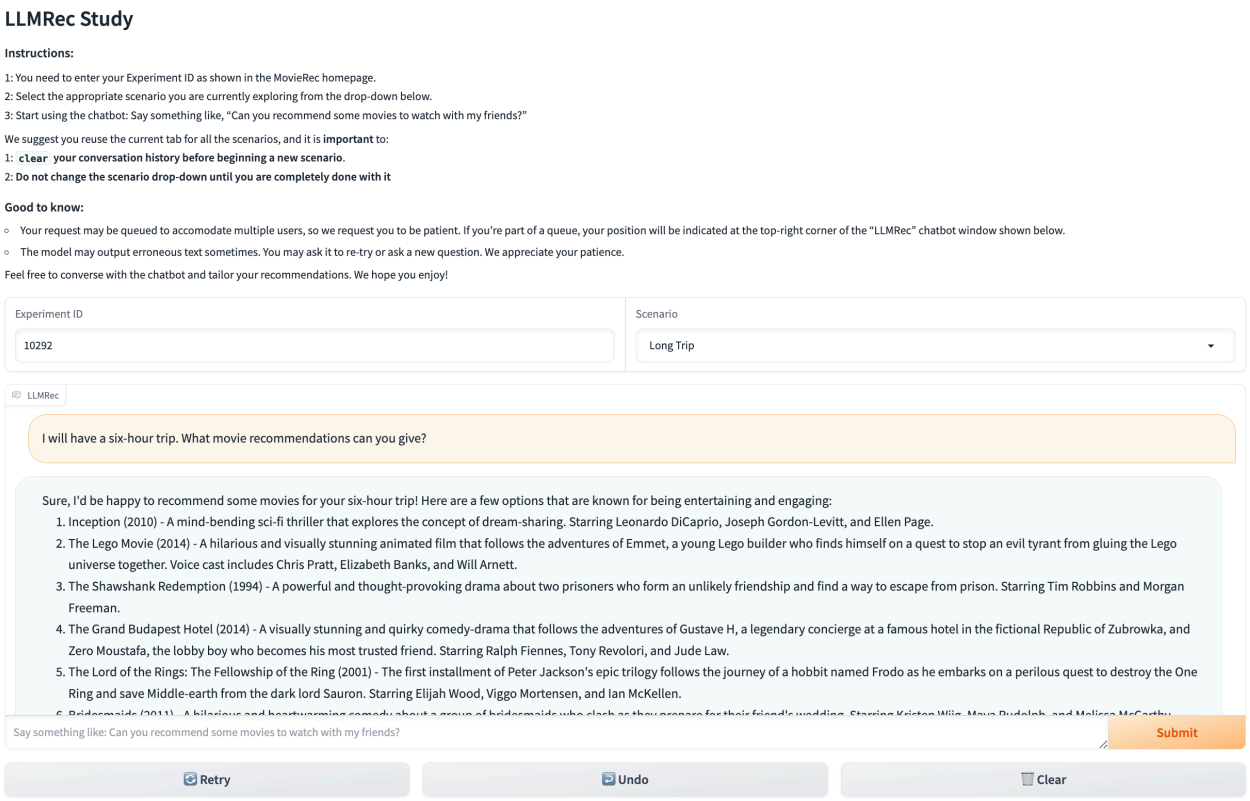

Figure A1: Example Chatbot Recommendation Interface.