

Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems

Itallo Silva

Leandro Balby Marinho

itallo@copin.ufcg.edu.br

lbmarinho@dsc.ufcg.edu.br

Federal University of Campina Grande
Brazil

Alan Said

alansaid@acm.org

University of Gothenburg
Sweden

Martijn Willemsen

m.c.willemsen@tue.nl

TU Eindhoven & JADS

The Netherlands

ABSTRACT

The adoption of recommender systems (RSs) in various domains has become increasingly popular, but concerns have been raised about their lack of transparency and interpretability. While significant advancements have been made in creating explainable RSs, there is still a shortage of automated approaches that can deliver meaningful and contextual human-centered explanations. Numerous researchers have evaluated explanations based on human-generated recommendations and explanations to address this gap. However, such approaches do not scale for real-world systems. Building on recent research that exploits Large Language Models (LLMs) for RSs, we propose leveraging the conversational capabilities of ChatGPT to provide users with personalized, human-like, and meaningful explanations for recommended items. Our paper presents one of the first user studies that measure users' perceptions of ChatGPT-generated explanations while acting as an RS. Regarding recommendations, we assess whether users prefer ChatGPT over random (but popular) recommendations. Concerning explanations, we assess users' perceptions of personalization, effectiveness, and persuasiveness. Our findings reveal that users tend to prefer ChatGPT-generated recommendations over popular ones. Additionally, personalized rather than generic explanations prove to be more effective when the recommended item is unfamiliar.

CCS CONCEPTS

- Information systems → Recommender systems.

KEYWORDS

recommender systems, explanations, large language models

ACM Reference Format:

Itallo Silva, Leandro Balby Marinho, Alan Said, and Martijn Willemsen. 2023. Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Recommender systems (RSs) have gained significant traction in various domains, revolutionizing how users discover and interact with information, products, and services. However, the often highly complex machine learning models used for generating personalized recommendations raise concerns about transparency and interpretability.

The demand for transparency and interpretability in RSs stems from the need to build user trust and confidence. Understanding the reasoning behind recommendations is crucial for users to make informed decisions and for organizations to comply with ethical and regulatory requirements. Consequently, significant efforts have been dedicated to creating explainable RSs [19].

Despite such advancements, there is still a shortage of automated approaches that can deliver meaningful and contextual human-centered explanations. Current methods often fall short in providing explanations that are interpretable, personalized, and capable of addressing individual users' diverse needs and preferences. Thus, there is a growing interest in developing novel techniques that leverage cutting-edge technologies to enhance the quality of explanations provided by RSs.

Several related works have examined the concept of explainable RSs through the lens of human-generated recommendations and explanations [1, 2, 12]. These studies provide insight into how explanations can impact users if they are similar to how people themselves explain things to each other. However, this approach has limitations in terms of scaling for large RSs with millions of users and items. Additionally, it is difficult to ensure the quality of explanations provided by humans [1]. LLM-based services like ChatGPT have demonstrated remarkable natural language processing and generation capabilities, enabling human-like conversational interactions. LLMs have advanced to the point where machine and human-authored texts are indistinguishable for untrained individuals [3]. Leveraging these conversational abilities, we propose using ChatGPT to provide users with personalized, human-like, and meaningful explanations for recommended items.

Research on RSs powered by LLMs has started to flourish [11, 13, 18, 20]. While many current RSs predominantly rely on user behavior data, LLMs extract and incorporate a wealth of knowledge from large-scale web corpora. This allows LLMs to possess knowledge that enriches behavioral data. For example, an RS powered by an LLM, such as ChatGPT, might suggest the best classic movies of all time, utilizing a zero-shot approach, even in the absence of historical data (ratings, clicks, viewing times) regarding the users' movie preferences. By framing user preference data within task prompts, the inherent knowledge within LLMs can be leveraged to produce personalized recommendations. The reasoning abilities of LLMs can discern user preferences from the context supplied in these prompts.

In this paper, we present one of the first user studies for measuring the users' perceptions of explanations provided by ChatGPT acting as an RS. With 94 participants, we evaluate how ChatGPT's

personalized and generic explanations are perceived. We evaluate this for recommendations generated by ChatGPT (based on movies a person liked or disliked) and for a set of random (but popular) recommendations. We also investigate how people evaluate ChatGPT explanations for movies they should avoid (so-called disrecommendations). This research is anchored by the following central research question: "How do users experience and evaluate personalized explanations generated by ChatGPT?"

Our results indicate that users tend to prefer ChatGPT's personalized recommendations over random selections of popular movies. Surprisingly, even when ChatGPT bases its explanations on users' movie preferences, they are not perceived as more personalized than generic ones unless the recommendations are random. This insight also extends to the perceived effectiveness and persuasiveness of the explanations. We further explore why these scenarios occur and investigate the interconnectedness between different explanation aspects like personalization, persuasiveness, and satisfaction.

2 RELATED WORK

Explainable RSs have been a significant research topic in the past [19]. However, the recent emergence of LLMs has introduced new possibilities in this field, particularly regarding personalized explanations. We will now highlight some related works on explainable RSs in general and specifically focus on recent literature that utilizes LLMs for recommendation purposes.

Chang et al. [2] introduced a method for providing personalized natural language explanations for recommendations. They achieved this by employing crowd workers who synthesized explanations from movie reviews. The approach was integrated into the MovieLens website and was found to be more efficient and effective than personalized tag-based explanations. This study emphasizes the benefits of using natural language over other approaches to provide meaningful explanations.

Evaluating explanation quality has become an increasingly important topic in recent years. Balog and Radlinski [1] proposed evaluating seven explanation goals originally introduced by Tintarev and Masthoff [16], namely effectiveness, efficiency, persuasiveness, satisfaction, scrutability, transparency, and trust. These goals align with the dimensions used by Chang et al. [2] and other researchers in the field. Balog and Radlinsky conducted a study where groups of crowd workers developed and evaluated personalized explanations of recommendations. They discovered that the seven goals are not independent and are often highly correlated, even when explanations were created to reflect a specific goal. Surprisingly, non-personalized explanations are equally as good as personalized ones, suggesting that crowd workers may lack the skills to create explanations for specific goals or that users cannot distinguish between them.

Lu et al. [12], in a user study similar to Balog and Radlinski [1], evaluated the effectiveness of machine-generated versus human-generated explanations. The results indicated that users were more satisfied with human-generated explanations. Based on these findings, a new approach inspired by human explanations was proposed, resulting in even higher user satisfaction.

While the evaluation of explanations quality has predominantly been conducted in online settings, Li et al. [10] developed an approach for offline evaluation of explanations analogous to traditional evaluation methods for recommendations. The proposed approach formulates explanations as a ranking task, which can be evaluated using conventional ranking metrics such as nDCG and accuracy metrics such as precision and recall.

The emergence of LLMs like GPT-3 and GPT-4 has opened new avenues in RSs. For example, the works of Harrison et al. [7] and Huang et al. [8] employ GPT-4 to generate accurate and contextually relevant recommendations. Wang et al. [17] merges GPT-3.5 and GPT-4 embeddings into reasoning graphs to improve recommendation quality. Shu et al. [15] and Lyu et al. [13] delve into the personalization aspects of LLM-based RS, emphasizing their ability to provide tailored recommendations based on individual user behavior. Our work is particularly inspired by Gao et al. [6], who introduced Chat-REC, a conversational recommender system framework that converts user profiles and user-item interaction history into prompts and leverages ChatGPT to generate recommendations and explanations through natural language interaction with the user. The evaluation of Chat-REC only measures recommendation quality, where the results demonstrate that ChatGPT outperforms many common recommendation approaches regarding precision/recall and nDCG. The paper also outlines how ChatGPT can summarize the preferences of MovieLens100k users and leverage those summaries in recommendation prompts.

While some of these works highlight how explanations can be generated, such as Chat-REC, where explanations are embedded as part of the conversation with the user, they do not specifically evaluate the effect of personalized recommendations and explanations generated by the underlying LLM. Zhou and Joachims [22] present one of the pioneering studies moving in this direction. Their study compares the effectiveness of ChatGPT-generated text reviews for movies in a mockup recommendation environment to human-written reviews. Through a survey involving 120 participants, they found no significant differences in how participants ranked movies or rated reviews for unfamiliar movies. However, ChatGPT-generated reviews were favored for movies participants had seen before. They also investigated the specific attributes of the review texts that influenced participants' perceptions of quality.

Informed by the aforementioned literature, we perform a user study in which we compare recommendations generated by ChatGPT to random (but popular) recommendations and accompany these recommendations with explanations generated by ChatGPT that are either generic or personalized towards the users' preferences, the latter hereafter referred to as user-based explanations¹. We posit the following research questions to bridge the existing gaps and extend our understanding of the personalized explanations and recommendations generated by ChatGPT:

- RQ1.** How do users value personalized ChatGPT-generated recommendations when compared to random recommendations?
- RQ2.** How do users perceive user-based versus generic ChatGPT-generated explanations in relation to recommendation methods and explanation goals such as effectiveness, personalization, and persuasiveness?

¹To avoid confusion with the personalized recommendations condition

- RQ3.** Do user-based versus generic explanations work differently for familiar or unfamiliar movies?
- RQ4.** How do explanation goals such as personalization, persuasiveness, and effectiveness relate?

RQ1 is grounded on the works of Harrison et al. [7], Huang et al. [8], and Gao et al. [6], which demonstrate the efficacy of GPT models in generating accurate and contextually relevant recommendations. However, these works do not specifically assess if users experience personalization. **RQ1** seeks to fill this gap by measuring the level of personalization users experience in the recommendations. Since our investigation primarily revolves around comparing user-based versus generic explanations, it is key to identify whether users experience personalization from the beginning. This question is not designed to affirm whether ChatGPT is the pre-eminent recommender, but it serves to validate if ChatGPT can deliver a discernibly personalized experience through its recommendations when compared to random selections. **RQ1** thus sets the tone for evaluating user-based versus generic explanations in the subsequent research questions.

For **RQ2**, the inspiration comes from the surprising results of Balog and Radlinski [1], who found non-personalized explanations to be just as good as personalized ones. This contradiction to the commonly held belief underlines the need to explore how users perceive user-based versus generic ChatGPT-generated explanations regarding various aspects such as recommendation method, effectiveness, and persuasiveness.

RQ3 extends the study of Zhou and Joachims [22] to evaluate the effectiveness of user-based versus generic explanations in the context of both familiar and unfamiliar movies, offering a comprehensive insight into the varying effects based on the user's prior knowledge about the recommended items.

RQ4 explores how the different explanation goals and user perceptions relate. This question is inspired by the work of Balog and Radlinski [1], who found strong correlations between explanation goals. However, Balog and Radlinski [1] did not explore these relations in more depth. Similar to earlier work in user-centric evaluation by Knijnenburg et al. [9], we explore to what extent the effectiveness of the explanation and satisfaction of the recommendation will depend on personalization and persuasiveness using path modeling.

To summarize, our outlined research questions are designed to build upon each other, offering a sequential understanding that progresses from establishing the basic effectiveness of ChatGPT in offering personalized recommendations to understanding the depth of its impact on varied contexts and settings involving explanations.

3 EXPERIMENT DESIGN

The user study took place across three separate batches on the Prolific crowdsourcing platform between June 8-15, 2023. Participants received an average hourly reward of £6.08 and were chosen based on their English fluency and a balanced gender distribution. A total of 94 participants concluded the survey, yielding 564 recommendations.

From the Prolific platform, participants were directed to a user interface we developed to carry out our experiment², employing a within-subject design in three stages. First, we requested users to provide a list of six movies, comprising three they enjoyed and three they disliked, to gather their preferences. Subsequently, we used the OpenAI Chat Completions API, which is powered by the GPT3.5-Turbo model, to generate four recommendations and two disrecommendations, along with explanations, based on their stated preferences. Finally, we applied a questionnaire to collect user opinions about each of the provided recommendations and explanations. Below, we present the details for each step.

3.1 Collecting user preferences

We requested each user to provide six movies, equally divided between movies they liked and disliked (see Fig. 1). We integrated the OMDb API³ to assist users in locating the accurate movie title, with the corresponding movie poster displayed for easy verification, limiting the search for movies prior to 2021, due to the knowledge cutoff of GPT3.5-Turbo being September 2021. As soon as the user fills out the questionnaire, the *Next* button is enabled, and when clicked, the user is redirected to a waiting page while the recommendations and explanations are generated in the backend.

3.2 Generating recommendations and explanations

We used the provided preferences to generate recommendations and explanations. Each participant received four recommendations and two disrecommendations. Two of these recommendations and the two disrecommendations were generated by calls to the OpenAI Chat Completions API⁴. The other two recommendations were drawn at random from a pool of 594 movies, curated from the 50 most-rated movies in each genre on IMDb.

By randomly recommending movies, we can control for user familiarity. If users were only recommended movies they already knew, it could skew their perceptions based on prior knowledge or experiences. Random recommendations could introduce movies that users might not be familiar with, allowing the research to isolate the effect of the explanation itself on user decisions rather than the user's prior knowledge of the movie. Inspired by Zhang et al. [21], we also show two movies they should avoid and the reasons why. By showing recommendations and disrecommendations, we provide a broader range of feedback opportunities. For instance, while a recommendation could be based on a user's preference, a disrecommendation provides information about what not to watch. This might be equally valuable for users, as avoiding a potentially bad movie experience can be as beneficial as finding a good one.

The prompts used to generate the recommendations and explanations are shown in Fig. 2. We created a base prompt containing the user's answers, as shown in the upper left part of the figure. This prompt will be concatenated with every prompt that requires personalization. The upper right box shows the prompt used to get the four recommendations generated by the GPT3.5-Turbo model

²The source code is available at <https://github.com/issilva5/llm-recommendations-survey>

³<https://www.omdbapi.com/>

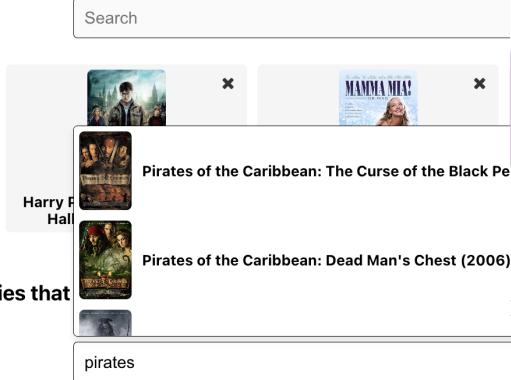
⁴<https://platform.openai.com/docs/guides/gpt/chat-completions-api>

LLM-Based Recommender System

Part I - User Preferences

In this section, we are interested in understanding a little more about your movies preferences.

1. Name three of your favorite movies.*



2. Name three movies that you really disliked (or hated).

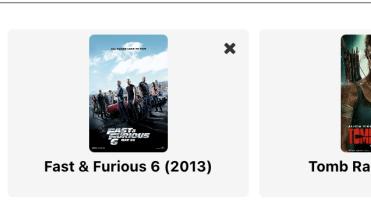


Figure 1: Overview of user preferences elicitation. The user was asked three movies they liked (*Name three of your favorite movies.*) and three movies they disliked (*Name three movies that you really disliked (or hated).*). Here, the user answers the question about the liked movies and searches for *pirates* in the search bar of the disliked movies.

(two positive and two negative). Positive and negative recommendations are requested in the same prompt to improve performance by avoiding an extra request since we did not observe a difference when using two separate prompts.

The lower boxes in Fig. 2 present the prompts used for generating respectively user-based and generic explanations. An explanation is generated for each recommendation and disrecommendation, with an equal distribution of user-based and generic explanations. The only difference between the prompts for generating explanations for recommendations and disrecommendations is the *not* keyword added for the latter case. Each explanation was generated using a single separate request.

In addition to the text presented in Fig. 2, each prompt had some formatting guidelines, such as a description of how the output was to be formatted and an instruction not to recommend any of the movies given as input. In order to access the OpenAI's API, we used the provided python library *openai*⁵ by calling the Chat

⁵<https://github.com/openai/openai-python>

User-preferences base prompt

Given the answers for the following questions about the movie preferences of a person.

Question 1: Name three of your favorite movies (separated by semicolon).

Answer 1: {liked movies}

Question 2: Name three movies that you really disliked or hated (separated by semicolon).

Answer 2: {disliked movies}

Personalized explanation prompt

User-preferences base prompt

Why should someone with these preferences *not* watch the movie: {movie}?

Legend: Auxiliar Personalized Generic

Figure 2: Prompts used for generating recommendations and explanation from the OpenAI GPT3.5-Turbo model

Completion creation interface (*openai.ChatCompletion.create*). We set the temperature to 0.0 for better reproducibility and defined no frequency or presence penalty. A full description of the effect of these parameters can be found in the Chat Completion API Reference⁶.

We randomized the presentation order of all four recommendations and two disrecommendations, with the disrecommendations invariably presented last.

3.3 Evaluating the Explanations

For each recommendation and disrecommendation, we presented the user with a page (see Fig. 3a) containing the poster and name of the (dis)recommended movie, with the explanation next to it, followed by two blocks of questions about the recommendation and explanation (see Fig. 3b)⁷. After fulfilling the evaluation for one particular (dis)recommendation, the user can go to the next one by clicking the next button.

The first set of questions inquires if participants recognized the movie (familiarity) and whether they would enjoy it (satisfaction). Subsequently, we asked three questions related to the explanations' helpfulness, personalization, and convincingness. Note that helpfulness refers to how *effective* the explanation was in helping users make informed decisions, while convincingness refers to how *persuasive* the explanation was in convincing users to consume the item, as defined by Tintarev and Masthoff [16]. Hereafter, we will use the standard terminologies 'effectiveness' and 'persuasiveness' instead of 'usefulness' and 'convincingness,' respectively, in line with prevailing literature. To collect these opinions, we used statements with a 5-point Likert-scale answer option (strongly disagree, disagree, neutral, agree, strongly agree).

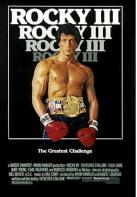
⁶<https://platform.openai.com/docs/api-reference/chat/create>

⁷In practice, the next question block becomes available or unfold after the user completes the previous one.

LLM-Based Recommender System

Part II - Recommendations
In this section, we will present you four recommendations and ask you to evaluate them.

Our recommender system predicts you will like this movie. Below we explain the reason why.

Recommendation	Explanation
	Rocky III is a great movie for someone with your preferences because it combines elements of action, drama, and sports. Like Avengers: Endgame, it has intense and thrilling fight scenes that will keep you on the edge of your seat. If you enjoyed the fantasy world of Harry Potter and the Deathly Hallows: Part 2, you'll appreciate the underdog story of Rocky III and the determination of the main character to overcome obstacles. Mamma Mia! and Rocky III both have memorable soundtracks that will have you singing along. If you didn't enjoy Fast & Furious 6 or Pirates of the Caribbean: The Curse of the Black Pearl, you'll appreciate that Rocky III focuses more on character development and emotional depth rather than relying solely on action sequences. Overall, Rocky III is a must-watch for anyone who enjoys a mix of action, drama, and sports with a compelling story and memorable soundtrack.

About the recommendation

About the explanation

(a) Evaluation page presented to the user

About the recommendation

1. Do you know this movie? * This question is mandatory.

No
 Yes

2. I enjoy this recommendation. * This question is mandatory.

Strongly disagree | Disagree | Neutral | Agree | Strongly agree

About the explanation

3. This explanation helps me to determine how well I will like this movie. * This question is mandatory.

Strongly disagree | Disagree | Neutral | Agree | Strongly agree

4. This explanation resonates well with aspects of movies that I like. * This question is mandatory.

Strongly disagree | Disagree | Neutral | Agree | Strongly agree

5. This explanation is convincing. * This question is mandatory.

Strongly disagree | Disagree | Neutral | Agree | Strongly agree

(b) Questionnaire

Figure 3: Overview of the evaluation step.

4 RESULTS

Our study involves each of the 94 participants providing six sets of responses to the four recommendations and two disrecommendations. In total, our collected data contains responses for 564 recommendations. To analyze our data, we use multilevel linear regression with a random intercept model that accounts for the repeated nature of participant responses. Our results are presented through plots displaying the estimated means from the models. Our initial focus is on the four positive recommendations, followed by a discussion on the two disrecommendations.

4.1 Recommendation Satisfaction

For RQ1, we evaluated how participants determined their potential satisfaction with a movie based on two factors: the personalization of the recommendations and their familiarity with the film. Both recommendation methods generated a substantial number of unfamiliar movies, i.e., users indicated that 25% of the ChatGPT-based recommendations were unfamiliar, whereas 49% of the random recommendations were unfamiliar. Our results show that participants found the movie less enjoyable when the recommendation was random rather than personalized by ChatGPT ($\beta = -.53, p < .001$) and less enjoyable when the movie was unfamiliar rather than known ($\beta = -.78, p < .001$), as shown in Fig. 4a. We did not find any interaction effect between familiarity and recommendation method. These findings suggest that ChatGPT-crafted recommendations are favored over random (but popular) ones, similarly for both familiar and unfamiliar movies.

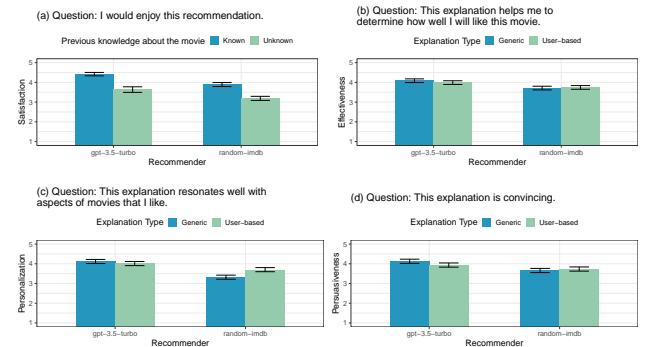


Figure 4: Questionnaire results for the (four) recommendations based on estimated means from the random intercept multilevel regressions. Responses were given on a 5-point, disagree-agree scale, for which 3 means neutral. Error bars are one standard error of the mean.

4.2 User-based vs. Generic Explanations

Regarding RQ2, we compared user-based vs. generic explanations in terms of effectiveness, personalization, and persuasiveness. Our findings suggest that the explanations generated for ChatGPT recommendations are perceived as more effective than the explanations generated for random recommendations ($\beta = .37, p < .001$). However, we did not observe any significant difference in the effectiveness of user-based versus generic explanations ($\beta = -.10, p = 0.37$), as shown in Fig. 4b. Personalization was determined by asking participants if the explanations resonated with their movie preferences. Interestingly, we found that user-based explanations, which explicitly mentioned the movie preferences of the participant (see Fig. 8), were not perceived as significantly more personalized than generic ones. We further discuss this finding in Section 5. Only in the case of random recommendations, which typically score lower on personalization ($\beta = -0.80, p < .001$), user-based explanations were perceived as more personalized than generic ones, as depicted in Fig. 4c and reflected in a significant interaction between the recommendation and explanation type variables in our model.

($\beta = 0.49, p < .01$). Regarding persuasiveness, explanations were less convincing for random recommendations ($\beta = -0.47, p < .001$), with no significant difference between user-based and generic explanations. For ChatGPT-generated recommendations, user-based explanations appeared to be slightly less convincing than generic ones, as shown in Fig. 4d, though not significant ($\beta = -0.27, p = .14$).

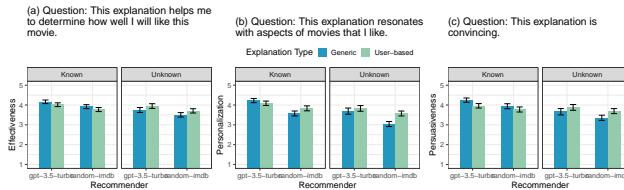


Figure 5: Results regarding how movie familiarity affects explanations' effectiveness, personalization, and persuasiveness.

4.3 Movie Familiarity Analysis

Regarding RQ3 (cf. Fig. 5), we analyzed if perceptions of effectiveness, personalization, and persuasiveness of explanations differed between familiar and unfamiliar movies. In terms of effectiveness, explanations for unknown movies are less effective in general ($\beta = -0.41, p < .01$), but as is clear from the Fig. 5a, this is mostly for generic explanations. User-based explanations are equally effective in all conditions and thus help quite well for unknown movies, as reflected in the positive interaction between explanation type and familiarity ($\beta = 0.35, p < .05$). For personalization, we also observe a main effect of familiarity ($\beta = -0.53, p < .001$), with explanations for unknown movies feeling less personalized in general as can be seen in Fig. 5b. The figure also seems to suggest user-based explanations are less affected by familiarity, but the interaction is not significant ($\beta = 0.274, p = .17$). For persuasiveness, we find again that explanations for unknown movies are less convincing ($\beta = -0.58, p < .05$), but now we find a significant main effect of explanation type ($\beta = -0.28, p < .05$), interacting with familiarity ($\beta = 0.50, p < .05$): as shown in Fig. 5c, the user-based explanations are somewhat less convincing for known movies but more for unknown movies. Overall, we see a consistent pattern in that user-based explanations, compared to generic ones, seem to work better, mostly in cases where movies are unknown. Having less prior knowledge about the movie, a user-based explanation that actually relates the movie to users' preferences is more influential than if the user knows about a movie.

4.4 Path Modeling of Explanation Types and Goals

In RQ4, we asked how our users' different perceptions and experiences relate. We tested a path model in which we try to predict to what extent explanation effectiveness depends on perceptions of personalization and persuasiveness of the explanation, as well as the level of movie satisfaction that users reported. Following the user-centric framework of Knijnenburg et al. [9], we see personalization and persuasiveness as perceptions (Subjective System Aspects:

SSA), whereas satisfaction and effectiveness are experience-type constructs (EXP). Our conditions are the objective system aspects (OSAs) that affect SSAs and EXPs.

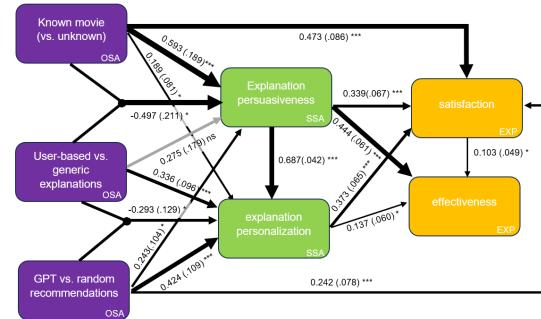


Figure 6: Path model showing how persuasiveness and personalization of the explanations are affected by the conditions and how they subsequently predict satisfaction and effectiveness. OSA=Objective System Aspect, SSA=Subjective System Aspect, EXP=Experience. The thickness of the line represents the strength of the coefficient. Standard errors in brackets, significance: * p<.05, ** p<.01, * p<.001**

In line with the work of Balog and Radlinski [1], we find that persuasiveness, personalization, satisfaction, and effectiveness are correlated, and our path model shows in more depth how they relate. Our analysis finds that effectiveness is predicted by satisfaction, persuasiveness, and personalization, with persuasiveness being the strongest predictor. Satisfaction itself goes up with known movies and GPT-based recommendations (as we already showed in the analysis of RQ1) but is also directly affected by persuasiveness and personalization. Consistent with our analysis of RQ2, we find that personalization is affected by the type of explanation, type of recommendation, and their interaction and a small effect of familiarity of the movie (see again Fig. 4b that shows the same patterns) as well as directly by the level of persuasiveness⁸. Similar to our analysis of RQ3, we find that persuasiveness is affected by explanation type and familiarity (and their interaction) and by familiarity with the movie. Hence, whereas the path model shows the same effects as discussed in the previous sections, it brings insights into how these effects relate. An effective explanation is personalized and persuasive and ideal for a satisfactory movie. We find that persuasiveness is a function of explanation and recommendation type and movie familiarity, and personalization depends on the explanation type, recommendation type, and persuasiveness. Together, this shows when an explanation is effective and how to achieve that.

⁸One could argue that the effect might be in the opposite direction: persuasiveness goes up with personalization, but in our model this led to reduced fit

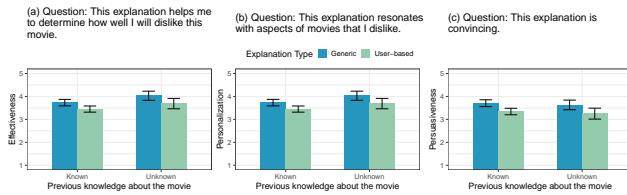


Figure 7: Questionnaire results for the two disrecommendations, based on estimated means from the random intercept multilevel regressions. Responses were given on a 5-point, disagree-agree scale, for which 3 means neutral. Error bars are one standard error of the mean.

4.5 Disrecommendations Analysis

We also analyzed the two disrecommendations. These disrecommendations were always generated by ChatGPT since random recommendations provide no guarantees that we are recommending movies users should avoid. So, in this case, we look only at the effects of explanation type and familiarity.

For disrecommendations, we find that generic explanations seem more effective than user-specific ones ($\beta = -0.32, p < .05$), with no interaction with familiarity, as can be seen in Fig. 7a. For personalization, the effect is smaller and not significant. Concerning persuasiveness, we do find that user-specific explanations are somewhat less convincing ($\beta = -0.36, p < 0.05$). Overall, it seems explanations for disrecommendations do not benefit from user-based explanations. However, further analysis is needed to determine whether this effect is not a consequence of ChatGPT's reluctance to disrecommend, given OpenAI's efforts to avoid harmful and biased discussions⁹.

5 DISCUSSION

Our results highlight several interesting aspects of ChatGPT-based explanations and recommendations. In this section, we discuss the results specifically in terms of user perceptions, popularity bias, and limitations.

5.1 The Anatomy of Explanations and its Influence on the User's Perception

Our findings indicate that generic explanations can feel just as effective, persuasive, and personalized as user-based explanations, as noted by Balog and Radlinski [1]. However, users tend to prefer user-based explanations when it comes to random and/or unfamiliar recommendations. We manually analyzed some of the explanations generated by the GPT model in search of common writing and/or argumentation patterns that could help us understand this finding.

Some examples of user-based vs. generic explanations generated by the GPT3.5-Turbo model for the three possible types of recommendations (GPT-generated recommendation, GPT-generated disrecommendation, and random recommendation) are depicted in Fig. 8. We highlighted four different types of arguments that we observed being used for generating the explanations. The first two, *comparison with liked (light green)* and *disliked (light red)* movies,

⁹<https://openai.com/policies/usage-policies>

GPT Recommendation

Personalized Explanation

The Shawshank Redemption is a classic movie that tells the story of a man who is convicted of murder and sent to prison. It's a movie about hope, friendship, and the human spirit. If you enjoyed the themes of friendship and love in The Shawshank Redemption, then you will definitely appreciate the depth and complexity of the plot in The Shawshank Redemption. On the other hand, if you disliked the superficiality of the plot in The Shawshank Redemption, then you will definitely appreciate the depth and complexity of the plot in The Shawshank Redemption.

Generic Explanation

The Shawshank Redemption is a must-watch movie for anyone who enjoys powerful and emotional tales of hope, friendship, and redemption. The film is beautifully shot and features outstanding performances, including Tim Robbins and Morgan Freeman. The story is gripping and edge-of-your-seat until the very end. If you're looking for a movie that will make you feel, and appreciate the power of the human spirit, then The Shawshank Redemption is definitely worth watching.

GPT Disrecommendation



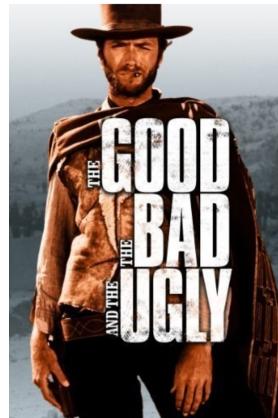
Personalized Explanation

Someone with a preference for action movies like The Fast and the Furious, 3 - Parabellum, and Star Wars: Episode I - The Phantom Menace, Spider-Man: Homecoming, Batman Begins, and Saving Private Ryan, Twilight is a romantic fantasy movie that focuses on the relationship between a vampire and a mortal. While the themes and tone of Twilight may not align with the preferences of someone with a preference for action movies like The Fast and the Furious, Spider-Man: Homecoming, and Batman Begins, therefore, someone with these movie preferences.

Generic Explanation

Twilight is a movie that may not be suitable for everyone. The dialogue, lack of character development, and weak plot. The characters are often criticized for being unrealistic and unconvincing. Twilight has been accused of promoting toxic relationships and perpetuating stereotypes. If you are looking for a well-written and thought-provoking movie, then Twilight is not the choice for you.

Random Recommendation



Personalized Explanation

The Good, the Bad and the Ugly is a classic western movie with iconic characters, memorable soundtrack, and epic showdowns. If you enjoy movies that have a unique style and a distinct visual aesthetic, then this movie is definitely for you. The Good, the Bad and the Ugly, While this movie may not have been your favorite or disliked, it is a timeless classic that is widely regarded as one of the greatest movies ever made.

Generic Explanation

The Good, the Bad and the Ugly is a classic western movie with iconic characters, memorable soundtrack, and epic showdowns. It is known for its stunning cinematography and intense action scenes. If you are a fan of director Sergio Leone's work, you will definitely appreciate this movie. Additionally, the film has influenced many other movies and is a must-see for any film buff.

Colors descriptions

Comparison with liked movies

Comparison with disliked movies

Critic-similar arguments

Movie plot description

Figure 8: User-based vs. Generic Explanations for three recommendations obtained from real users.

are predominant in the user-based explanations. In these types of arguments, the model compares features of the recommended movie with the movies users stated their preferences for. The third type, named *critic-similar arguments (light blue)*, are arguments that resemble specialized-critic opinions and hence may induce an authority bias in users, i.e., a psychological inclination to accept information from authority without critical evaluation. These arguments are featured mostly in generic explanations. Lastly, the fourth type of argument (light yellow) briefly describes the movie plot and can frequently be found in both types of explanations.

The finding that generic explanations were perceived as personalized as user-based explanations for ChatGPT-generated recommendations may be explained, to some extent, by the fact that the generic explanations highlighted movie features that were present in the personalized recommendations, which were tailored to the user's preferred movies. Thus, the user's preference signals might have been unintentionally integrated into the generic explanations. This, together with an authoritative tone, may have made generic explanations as appealing as the user-based ones.

We have found that both user-based and generic explanations share a persuasive tone that is conveyed through comparative arguments, such as "if you do this, then you will get that," and the frequent use of a second-person point of view. This observation may explain why we did not observe significant differences in the persuasiveness of the two types of explanations (see Section 4.2). In Fig. 8, we have highlighted the pronoun "you" in bold to underscore its frequent usage. The high occurrence of this type of sentence may be due to the fact that the prompt used to generate the explanation caused the model to speak to the audience or because the model itself is trained to use this type of discourse when interacting with humans.

5.2 Popularity Bias in GPT-generated Recommendations

While our primary focus in this paper is not to evaluate the quality of the GPT model's recommendations, we observed that popularity bias seems to impact movies with extremely high or low IMDb ratings, causing them to appear more frequently than expected in both recommendations and disrecommendations.

Upon closer examination, we find evidence supporting this observation. There is a positive, albeit weak to moderate, correlation between the frequency of recommendations and the IMDb ratings for the GPT recommendations (Spearman's $\rho = 0.467$, $p < 3e-5$; Kendall's $\tau = 0.383$, $p < 3e-5$). Conversely, we observe a negative, albeit weak, correlation for the disrecommendations (Spearman's $\rho = -0.268$, $p < .05$; Kendall's $\tau = -0.205$, $p < .05$). No significant correlation was found for the random recommendations.

These findings suggest that the recommendations generated by the GPT model may indeed be influenced by popularity bias, but further analysis is needed to explore this bias's implications on the recommendations' overall performance.

5.3 Limitations

While our study has shown significant effects of GPT-generated explanations, the study comes with several limitations. Firstly, the study had a specific focus on the GPT-3.5 model, this was mainly

due to the ease of use provided through its API¹⁰, which has few hardware requirements compared to other LLMs, such as Meta's Llama and its derivatives, since we are not responsible for running the model. This specific focus limits the generalization capabilities of our findings. In future work, we plan to explore the use of other LLMs.

Other limitations arise from the nature of LLMs and ChatGPT specifically and are expected and inherent to this type of technology. When interacting with ChatGPT, the way the prompt is formulated can significantly affect the outcome, even when using its API. For the purpose of our study, we created a standardized set of prompts (see Fig. 2) in order to minimize the effects of differences in output based on irregularities in the prompts given to ChatGPT. Nevertheless, LLMs' (and ChatGPT's specifically) sensitivity to prompt formulation has not been explored in the scope of our work. Another limitation of our study is the arguably limited room for personalization, given that study participants only disclosed six preferences (three liked and three disliked movies). While, in traditional recommendation models, even a modest number of preferences suffice to generate reasonable recommendations [14], we have no insight into the finer details of how our ChatGPT-based recommender creates recommendations. It should, however, be noted that our study focuses on comparing personalized (user-based) and generic explanations. With that in mind, even a relatively modest level of personalization should suffice. In line with the limited insight into the finer details of ChatGPT, another limitation is that of explanation versus justification. In this context, we refer to an explanation as the ability to disclose the algorithmic reason why a certain item is recommended, i.e., an interpretation of the model, whereas by justification, we refer to a motivation as to why a certain item is recommended [5]. Given the recommendation model used in our work, the explanations generated by our system are rather justifications, i.e., human interpretable snippets informing the study participants why a certain item was recommended, than explanations. In the context of our study, the semantic differences between the terms explanation and justification should not have an effect on the outcomes given the study framing and motivations presented to study participants upon taking part in the study.

Lastly, and perhaps most importantly, the number of participants in a study like this may have an effect on the obtained results. However, given that we tested most of our effects within-subject, the data obtained from our sample of 94 participants should have sufficient power as each participant provided four evaluations for the recommendations and two for the disrecommendations. Moreover, also related to data, the crowd worker platform used in our study, Prolific, has been proven to generate higher quality data compared to other crowdsourcing platforms, e.g., Mechanical Turk, Crowdflower [4]. To further ensure the quality of data, we reached out to several of our study participants after having finalized the study to ask about their experiences and reflections on participating in the study. Overall, the participants we contacted were responsive (100% response rate) and content with the experience.

¹⁰<https://platform.openai.com/docs/api-reference/chat/create>

6 CONCLUSION

This paper investigated how users experience and evaluate personalized explanations generated by ChatGPT. Our findings revealed that personalized recommendations from ChatGPT yielded higher user satisfaction than random (but popular) recommendations. This finding expands works based on LLMs-driven RSs that only evaluate the accuracy of recommendations through offline experiments such as Gao et al. [6], Harrison et al. [7]. Interestingly, user-based explanations directly referring to the participant's movie preferences were not perceived as significantly more personalized than generic explanations unless the recommendations were randomly generated. User-based explanations were also not perceived as more effective and persuasive than generic ones, regardless of the recommendation type. This is in line with the findings of Balog and Radlinski [1], who found no significant differences between personalized and non-personalized explanations regarding different explanation goals. We observed that the features of the personalized recommendations may be leaked to the generic explanations to make them feel personalized, even when they do not explicitly mention the participant's movie preferences. We also noticed that ChatGPT has a bias toward producing persuasive explanations. These observations may explain to some extent our findings, although further analysis is needed.

Furthermore, user-based explanations were perceived as somewhat more effective, personalized, and persuasive for unfamiliar movies, perhaps because prior knowledge about movies had less influence on decision-making, leaving more room for explanations to influence users' choices. This finding contradicts the results of Zhou and Joachims [22], who found higher effects for familiar movies. Although our study is not directly comparable to theirs, this difference motivates future works on the impact of movie familiarity on users' perception of what makes a good recommendation/explanation. Regarding disrecommendations, explanations did not seem to benefit from user-based explanations.

We also conducted a path modeling analysis to shed light on explanation types and goal interdependencies. Our analysis revealed that explanation effectiveness is strongly predicted by users' satisfaction, persuasiveness, and personalization perceptions, with persuasiveness exerting the most significant influence. As for personalization, it was influenced by various factors, including the type of explanation, recommendation, and their interaction, as well as persuasiveness. In summary, our path model elucidates the conditions that lead to effective explanations and how to achieve them, providing more in-depth insights into the correlations between explanation goals as observed before by Balog and Radlinski [1].

In future work, we plan to conduct a more comprehensive evaluation, including a larger sample and additional explanation goals and properties to better understand the factors that users perceive most effective in natural language LLM-generated explanations.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

- [1] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) (*SIGIR '20*). Association for Computing Machinery, New York, NY, USA, 329–338. <https://doi.org/10.1145/3397271.3401032>
- [2] Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations (*RecSys '16*). Association for Computing Machinery, New York, NY, USA, 175–182. <https://doi.org/10.1145/2959100.2959153>
- [3] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296. <https://doi.org/10.18653/v1/2021.acl-long.565>
- [4] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (March 2023), 1–17. <https://doi.org/10.1371/journal.pone.0279720> Publisher: Public Library of Science.
- [5] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 0210–0215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- [6] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *arXiv:2303.14524 [cs.IR]*
- [7] Rachel Harrison, Anton Derevenssov, and Anton Bibin. 2023. Zero-Shot Recommendations with Pre-Trained Large Language Models for Multimodal Nudging. *arXiv:2309.01026 [cs.AI]*
- [8] 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:2308.16505 [cs.IR]*
- [9] Bart P. Knijnenburg, Martijn C. Willemse, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22 (oct 2012), 441–504. Issue 4-5. <https://doi.org/10.1007/s11257-011-9118-4>
- [10] Lei Li, Yongfeng Zhang, and Li Chen. 2023. On the Relationship between Explanation and Recommendation: Learning to Rank Explanations for Improved Performance. *ACM Trans. Intell. Syst. Technol.* 14, 2, Article 21 (feb 2023), 24 pages. <https://doi.org/10.1145/3569423>
- [11] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. *arXiv:2304.10149 [cs.IR]*
- [12] Hongyu Lu, Weizhi Ma, Yifan Wang, Min Zhang, Xiang Wang, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2023. User Perception of Recommendation Explanation: Are Your Explanations What Users Need? *ACM Trans. Inf. Syst.* 41, 2, Article 48 (jan 2023), 31 pages. <https://doi.org/10.1145/3565480>
- [13] Hanjia Lyu, Song Jiang, Hanqing Zeng, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiaji Tang, Yinglong Xia, and Jiebo Luo. 2023. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. *arXiv:2307.15780 [cs.CL]*
- [14] István Pilászy and Domonkos Tikk. 2009. Recommending New Movies: Even a Few Ratings Are More Valuable than Metadata. In *Proceedings of the Third ACM Conference on Recommender Systems* (New York, New York, USA) (*RecSys '09*). Association for Computing Machinery, New York, NY, USA, 93–100. <https://doi.org/10.1145/1639714.1639731>
- [15] Yubo Shu, Hansu Gu, Peng Zhang, Haonan Zhang, Tun Lu, Dongsheng Li, and Ning Gu. 2023. RAH! RecSys-Assistant-Human: A Human-Central Recommendation Framework with Large Language Models. *arXiv:2308.09904 [cs.IR]*
- [16] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* 801–810. <https://doi.org/10.1109/ICDEW.2007.4401070>
- [17] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y Zhang, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2023. Enhancing Recommender Systems with Large Language Model Reasoning Graphs. *arXiv:2308.10835 [cs.IR]*
- [18] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2023. A Survey on Large Language Models for Recommendation. *CoRR abs/2305.19860* (2023).
- [19] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (mar 2020), 1–101. <https://doi.org/10.1561/1500000066>
- [20] 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*.

- <https://openreview.net/forum?id=hFx3fY7-m9b>
- [21] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis (*SIGIR '14*). Association for Computing Machinery, New York, NY, USA, 83–92. <https://doi.org/10.1145/2600428.2609579>
- [22] Joyce Zhou and Thorsten Joachims. 2023. GPT as a Baseline for Recommendation Explanation Texts. In *Proceedings of the IntRS'23 Workshop*.