Exploring the Potential of Generative AI for Augmenting Choice-Based Preference Elicitation in Recommender Systems

Benedikt Loepp
Fraunhofer IMS
Duisburg, Germany
benedikt@loepp.eu

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

The recent boost in generative artificial intelligence has also reached the field of recommender systems. However, as is often the case, much of the work focuses on the algorithms, overlooking the crucial aspect of improving the systems from a user perspective. In this initial research, we explore the potential of large language models to achieve improvements in preference elicitation. The interactive choice-based method we are augmenting has previously demonstrated significant improvements in a number of aspects related to the user experience. Through an exploratory user study, we show that the item set comparisons presented by this method can be successfully accompanied by independently generated textual summaries, thereby improving the user experience even further.

CCS CONCEPTS

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

KEYWORDS

Recommender systems, Preference elicitation, Large language models, User experience.

ACM Reference Format:
Benedikt Loepp and Jürgen Ziegler. 2024. Exploring the Potential of Generative AI for Augmenting Choice-Based Preference Elicitation in Recommender Systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 1–4, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3631700.3664873

1 INTRODUCTION

Generative artificial intelligence (GAI) techniques have recently received tremendous attention due to their impressive results in a variety of application scenarios, especially in the creation of digital content such as text, images, and video [7]. The popularity of GAI has also spread to the field of recommender systems (RSs). These systems aim to automatically present users with relevant content from the typically large number of alternatives available in today's web. In the context of RSs, large language models (LLMs) are the most popular type of GAI, leveraged for a variety of purposes [28, 30, 31, 46, 50]. However, while the importance of the user experience has been increasingly accepted in the RS community over the last decade [3, 21, 23, 25, 33], the potential of LLMs to improve aspects such as interactive control or system transparency has been largely ignored so far. In contrast, as the overview of the state of the art in the next section will show, existing work has primarily focused on improving the underlying algorithms.

The most important requirement for an algorithm to generate accurate recommendations is that the system is able to adequately capture the user's preferences. The main advantage of collaborative filtering (CF), the most popular method in RSs, is that it only requires his or her feedback on a number of items. This is true for traditional memory-based techniques [12], but also for embedding-based approaches such as matrix factorization [27] or neural networks [51]. The feedback can either be expressed explicitly in the form of item ratings (stars, thumbs up), or it can be gathered implicitly from past interactions with the system, such as clicks on items or dwell time [19]. While implicit feedback data are usually more readily available and often lead to more accurate recommendations [19, 38], explicit feedback data still has its place in many cases [39], as illustrated by the recent introduction of the "double thumbs up" at Netflix. However, the next section will also show that rating-based preference elicitation suffers from several drawbacks. Accordingly, improving this aspect remains to be one of the most important goals of user-oriented RS research [3, 21, 33].

In this paper, we build on one of the most successful alternatives, choice-based preference elicitation. We explore whether GAI can further improve the corresponding interactive dialogs by augmenting them with LLM-generated textual summaries. After reviewing the literature and describing our approach, we present an initial user study ($N = 27$). The results show the potential of LLMs to improve the user experience of RSs without further ado.

2 STATE OF THE ART

In this section, we discuss existing choice-based preference elicitation methods for RSs and their drawbacks. We also illustrate that RS research is an area where the potential of GAI is far from being fully realized.

2.1 Choice-based preference elicitation

Many studies have shown that item ratings tend to be noisy, inaccurate, and unstable over time [4, 20]. Moreover, assigning absolute numbers to single items is cognitively demanding and may be unreliable. Thus, it is not surprising that—inspired by consumer behavior in physical environments, where purchase decisions are typically made after comparing a number of products—expressing preferences in relative terms has been proposed as a more appropriate way to elicit user preferences. In particular, choosing items from pairwise comparisons has been found to be easier than rating

UMAP Adjunct '24, July 1–4, 2024, Cagliari, Italy

each item separately and to improve the perceived recommendation quality [20, 41].

Several authors have proposed to implement such choice-based methods as interactive dialogs that show items sampled directly from the embedding space derived by state-of-the-art CF algorithms [15, 34, 43]. However, all these approaches still fall into the trap of requiring users to specify *item* preferences—be it for a single item or multiple items: If users do not know the items presented by the RS, it makes no sense to ask them which one(s) they prefer. Of course, it is usually allowed to skip dialog steps if one is not familiar with (some of) the items. However, this leads to less information being provided to the system, and thus, lower recommendation quality. This is also the case when users have to make a choice purely based on some selected item information. This typically includes the item title, an image, and, at best, some metadata, but has shown to be insufficient for RS users in many cases [32]. On the other hand, several attempts have been made to ensure a certain level of familiarity with the items displayed, e.g., by considering a popularity threshold [34]. Nevertheless, there may be users, e.g., with low domain knowledge, who know only a limited number of items. At the same time, items may be omitted by the system even though for certain users they would provide the most information about the respective preferences. In contrast to pairwise comparisons, the approaches in which *sets* of items need to be compared (e.g., [8, 34, 43]) increase the likelihood that users are familiar with at least some items or can deduce more information from the context. However, especially when items are automatically sampled from latent embedding spaces (e.g., in [15, 34, 43]), it may still happen that users do not understand the semantics of the comparisons or perceive the item characteristics in the sets as inconsistent or even contradictory.

2.2 Generative AI in recommender systems

As shown in recent surveys (e.g., [28, 30, 31, 46, 50]), GAI is mainly used in RSs in the form of LLMs, most often as a generic means to perform item ranking tasks (e.g., in [10, 11, 17, 48]). For instance, Harte et al. suggest exploiting the semantically rich embeddings learned by existing models or fine-tuning the models with dataset-specific information in order to generate next-item recommendations. Their results show significant improvements over methods designed specifically for sequential recommendation tasks. Other works that suggest the use of LLMs instead of or in addition to specific RS methods include, e.g., [6, 35, 45, 47]. In addition, Cui et al. present a framework for replacing the individual algorithms that are used for the different tasks and domains in industrial settings with a single foundation model [9].

Only a few authors have integrated LLMs into RSs for purposes other than improving recommendation performance. Agrawal et al. use generative models to annotate movies with "micro-genres" [2]. In this way, they expect to learn more accurate item representations and to better organize the items in the user interface. Silva et al. present a form-based user interface to inquire about the user's needs [28]. This information is then used in a prompt sent to the *ChatGPT-API*, asking the model to generate both recommendations and explanations tailored to the specified requirements. However, these efforts are still at the case study or proof of concept stage. Other authors have focused on the natural language capabilities of LLMs. Zhou and Joachims describe a survey study with a mock-up RS in which they compared model- and human-generated movie reviews [52]. In some cases, participants perceived no difference, but in other cases, the LLM-generated texts were superior. This highlights the potential of GAI for providing post-hoc explanations. Acharya et al. use a LLM to generate detailed item descriptions simply from the features included in the *MovieLens* dataset, e.g., cast and directors [1]. According to several metrics, these descriptions were of similar quality to information they scraped from the web. This illustrates the ability of LLMs to step in when the available data are insufficient. Mysore et al. generate narrative queries from past user-item feedback data to obtain synthetic training data [36].

Finally, several attempts have been made to leverage LLMs for conversational RSs. Friedman et al. present a framework and a demonstrator that can represent users through transparent natural language profiles [13]. Once constructed over multiple sessions from the user's interaction with the system, such a profile can be ingested by the underlying model to improve personalization and to generate textual justifications for each item displayed. Other work in this area includes, e.g., [14, 18, 49]. In summary, however, the literature review shows that there are still very few approaches that consider the use of LLMs as a practical means to improve the user experience of RSs. Other GAI techniques, such as generative adversarial networks or diffusion models, are used even more rarely, e.g., to analyze visual features to improve the recommendations [5, 22], or to adapt product images and videos about the recommended content to the user's taste [22, 44]. At the same time, however, the few existing works illustrate the advantages of this new technology. Consequently, it seems promising to exploit generative models also to augment RSs for purposes for which the use of GAI has not yet been explored, such as making preference elicitation more intelligible, accurate, and intuitive.

3 AUGMENTING CHOICE-BASED PREFERENCE ELICITATION WITH LLM-GENERATED TEXTS

As illustrated in the previous section, choice-based methods have shown their superiority over traditional ratings for eliciting preferences in RSs. However, these methods still rely heavily on the user's familiarity with the items selected by the system. Moreover, the semantics of the comparisons may not be immediately clear, and, in the case of item sets, the item composition may lead to confusion. To overcome these problems, we propose to augment the interactive dialogs presented by these methods with natural language texts generated by a LLM.

3.1 Background

We build on the method proposed by Loepp et al. [34], which has the advantage of sampling items automatically from a latent embedding space. A matrix factorization algorithm [26] is used to derive the matrices $\mathbf{P} \in \mathbb{R}^{|U| \times k}$ and $\mathbf{Q} \in \mathbb{R}^{|I| \times k}$ from the original user-item matrix $\mathbf{R} \in \mathbb{R}^{|U| \times |I|}$ that contains the feedback of the users $u \in U$ for the items $i \in I$. These two matrices represent the relationships between users or items and the k latent factors learned by the algorithm. When learned one after the other, the factors are

inherently ordered by their importance. Consequently, the matrix $\mathbf{Q}$ allows determining representative items for the $n \ll k$ most relevant dimensions of the embedding space. In an iterative dialog, these items are presented to the user in two sets for each factor, with titles, images, and metadata. The sets are composed based on three criteria, so that they eventually contain four items, each of which is *popular*, has a *low/high score* for the respective factor, and is very *specific*, i.e., tends to be neutral with respect to all other factors. At each step, the user can then indicate which set he or she prefers, or skip the comparison. From these decisions, an artificial user-factor vector p_u is constructed and finally used to generate recommendations in the usual way via dot multiplications.

3.2 Workflow

To augment the comparisons with explanatory texts, we propose the workflow illustrated in Figure 1. The first three steps reflect the steps of the original choice-based preference elicitation method as described above. Next, however, the information about the representatives from the two sets for one of the n factors is used to prompt a LLM (see details below). The responses are presented along with the item sets, which are displayed as in the original approach, to help users understand the underlying semantics and make sense of the comparison if they are not well familiar with individual items. In addition, as shown by the gray dotted lines, a diffusion model can be prompted to generate an image for each set that characterizes the items it contains. After several prompt-response cycles (one for each of the n factors) in which the user chooses one of the two sets, recommendations are presented as described above.

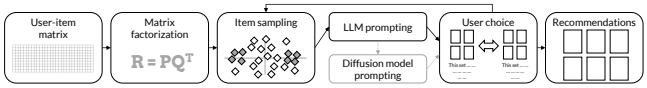

Figure 1: The proposed workflow.

3.3 Prompting

We use the *ChatGPT-API* to generate the textual descriptions for the sets of representatives, which aim to summarize the item commonalities and the meaning of the respective factor. We prompt the generative model *GPT-3.5-text-davinci-003* as shown in Figure 2 (left). In an iterative qualitative evaluation of the responses, this prompt turned out to produce the best results for this particular model at this time. At the beginning, (1) we provide the LLM with contextual information. Then, (2) we ask it to use domain-specific terminology (here movies). While we also tried to provide item titles separately for each set, we found that the generated summaries are easier to distinguish when (3) the LLM is informed of all titles for a factor at once, (4) but is asked for individual responses. Finally, (5-8) we further constrain the output for readability.

As mentioned in the previous section, we also experimented with additional prompts to a diffusion model to abstract from the individual items in a set by creating a single image that characterizes all of the contained representatives. Specifically, we used the responses generated by the LLM as described above to prompt *DALL-E 2*, again using the *ChatGPT-API*, as follows: "An epic shot of a climactic film scene, according to the description: LLM_Response". However, both in a qualitative evaluation of the images generated by this and other prompts, and in the user study (see Section 4), the results were not convincing (e.g., images were too unspecific, visually confusing, or thematically disconnected from the items). Since we did not pursue this direction further, we omit more details for the sake of space.

3.4 Prototype

We implemented our approach in a web-based prototype system using content-boosted matrix factorization [33], the *MovieLens 25M* and *Tag Genome* datasets, and additional metadata crawled from *The Movie Database* (TMDB). Figure 2 (right) shows a single step of the resulting interactive preference elicitation dialog. It is easy to see that the items selected as representatives possess different characteristics, which are well summarized by the LLM-generated descriptions below. The buttons at the bottom allow the user to navigate through the dialog and to settle for one of the two sets.

4 USER STUDY

To investigate the effectiveness of our augmentation approach and to understand whether it is an improvement over the original method from a user perspective, we conducted an exploratory user study with the prototypical web-based RS presented in the previous section. Through personal contacts and social media, we recruited 27 participants, aged 18–48 ($M = 28.30$, $SD = 7.99$), 17 female and 10 male. Fifteen were students, the rest were employed (7), self-employed (2), or responded otherwise. Students from a specific program were rewarded with study credits. Domain knowledge was rather high ($M = 3.24$, $SD = 1.06$). The study was approved by the department's ethics committee.

4.1 Study design

We designed the study as a lab experiment with a within-subjects design. In a randomized order, each participant was first assigned to one of the following two conditions, and then to the other:

- GEN Interactive preference elicitation dialog with five steps, each showing two sets of four representative items, with textual summaries generated as described in the previous section (see Figure 2 for a screenshot).
- TAG The same interface as in the experimental condition, but with a tag cloud derived as described below for each set instead of the textual summaries generated by the LLM (a screenshot is provided in the supp. material).

We assumed that a within-subjects comparison with an interface showing only the items would naturally lead to inferior results. Therefore, we decided to provide participants also in the control condition with additional information to make the comparison with GEN more fair. For this purpose, we created the tag clouds mentioned above, using the three most popular tags of each contained item according to the underlying metadata dataset. After going through these two conditions, participants were confronted with the IMG condition, which was not part of the comparison. The interface was again the same, but instead of items, two images generated by a diffusion model were shown along with the textual summaries to characterize and contrast the two sets (a screenshot is provided in the supp. material).

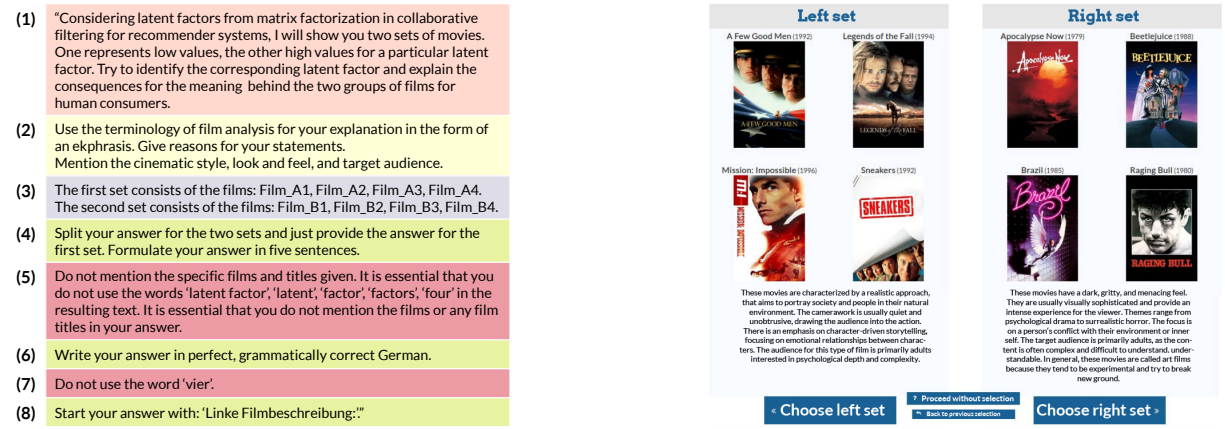

Figure 2: The left side shows a prompt to obtain textual summaries for the two sets of items created by the choice-based method to represent the characteristics of one dimension of the underlying embedding space. The right side shows the prototype system. Note that the interface, including the summaries, is translated from German for the convenience of the reader.

4.2 Procedure and task

After a short introduction, participants filled out the first part of the questionnaire with personal data. Then, they had to interact with the prototype system twice, based on the two conditions GEN and TAG. They were asked to select a total of two movies for an evening with friends, one in each of the two interfaces. After completing the respective preference elicitation phase and selecting a movie, participants were briefly redirected to the questionnaire to rate the quality of the process and the resulting recommendations. At the very end, they were confronted with the IMG condition (only five static screenshots were shown, i.e., no recommendations were generated) and the final part of the questionnaire, asking about this condition and the intention to use one of the methods again. A supervisor was present at all times.

4.3 Questionnaire

The questionnaire shown before, between, and after the tasks was administered online using *SosciSurvey*. For the constructs shown in Table 1, we mainly used items from established RS evaluation frameworks [24, 40]. In addition, we used a few self-generated items and items from [29]. To assess the general user experience, we used the short UEQ [42]. We also collected demographics and asked participants about their domain knowledge using self-generated items. All items had 5-point Likert response scales, except for the UEQ (7-point bipolar). We also assessed typical decision-making traits using the short maximization scale (MAX) [37] and the decision styles scale (DSS) [16], and asked for qualitative feedback. Due to space limitations, we omit the corresponding results in the remainder of this section.

4.4 Results

Table 1 shows the questionnaire results for the comparison of the two main conditions in terms of the subjective assessment of the system and the recommendations. We used paired t-tests to examine the differences and applied Benjamini-Hochberg correction to account for multiple comparisons (FDR of 0.05).

Table 1: t-test results for the comparison of the main conditions. Higher values indicate better results (difficulty and effort are reversed accordingly), with best values highlighted in bold. * indicates significance using a FDR of 0.05. d is Cohen's effect size.

	GEN		TAG				
Construct	*M*	*SD*	*M*	*SD*	*T*	*p*	*d*
Perceived recommendation quality [24]	**3.83**	1.02	3.61	0.84	0.876	0.526	0.17
Perceived recommendation diversity [24]	**3.59**	1.01	3.30	1.24	1.217	0.526	0.23
Choice satisfaction [24]	**4.15**	1.06	3.96	0.98	0.708	0.526	0.14
Choice difficulty [24]	**3.59**	1.22	3.19	1.27	1.203	0.446	0.23
Perceived usage effort [24]	**4.06**	0.76	3.91	0.86	0.750	0.526	0.14
Perceived effectiveness and fun [24]	**4.00**	1.00	3.59	1.04	1.954	0.134	0.38
Context compatibility	**3.78**	0.75	3.19	0.96	3.049	0.026*	0.59
Control [29]	**3.15**	0.90	2.73	0.97	2.011	0.134	0.39
Understandibility [29]	**3.72**	0.92	3.53	0.82	0.978	0.526	0.19
Overall satisfaction [40]	**3.85**	0.99	3.81	1.00	0.137	0.892	0.03
Pragmatic quality [42]	**6.06**	0.81	5.19	1.57	3.060	0.026*	0.59
Hedonic quality [42]	**5.39**	1.13	4.62	1.46	2.660	0.046*	0.51
Overall quality [42]	**5.72**	0.89	4.91	1.44	3.000	0.026	0.58

In all comparisons, GEN outperformed TAG, with small to medium effect sizes. This was true for both the process (e.g., control and understandability) and the final outcome (e.g., recommendation quality and diversity). Remarkably, the only difference was whether natural language texts or tag clouds were shown in the preference elicitation dialog, i.e., the representative items and especially the final recommendations were always determined by the system in the same way. Nevertheless, we found significant differences in terms of user experience. Moreover, participants in the GEN condition found the system to be significantly more compatible with their current context.

For the IMG condition, we obtained the following results: Participants liked the support of the method (M=3.85, SD=1.13), the compatibility with the current context (M = 3.74, SD = 0.90), and the amount of information provided (M = 3.53, SD = 1.17). In the qualitative feedback, 13 participants had a positive opinion of the

generated images, 8 a negative. The rest of the statements were neutral or gave no insight into the specific method.

Finally, regarding their intention to use one of the methods again, participants responded as follows: They would very much like to use the interface with the LLM-generated descriptions again (GEN: $M=3.96$, $SD=1.06$). In contrast, their intention to return to the tag cloud interface was much lower (TAG: $M=3.26$, $SD=1.23$). The condition with images from the diffusion model was in between the other two (IMG: $M=3.59$, $SD=1.34$). While these results support the superiority of GEN over TAG as indicated by the results for the specific constructs above, a one-factorial repeated-measures analysis of variance did not indicate considerable differences, $F(2,52)=1.81$, $p=.173$, $\eta_p^2=0.07$.

5 DISCUSSION AND CONCLUSIONS

With this paper, we have made a first attempt to improve preference elicitation in RSs by exploiting the capabilities of GAI. The results of our exploratory user study show that a relatively simple workflow and rather basic prompt engineering can already improve the subjective assessment of the system and the recommendations in a number of relevant dimensions. Admittedly, our study sample was small and consisted mainly of students. Hence, it is not surprising that we did not find significant effects for most of the dependent variables. Nevertheless, for a proof-of-concept evaluation, the results seem very promising, especially considering that the tags shown in the control condition also provided descriptions of the item set characteristics. The score for context compatibility suggests that participants perceived the system to be more aligned with their tasks simply due to the presentation of additional natural language texts as they progressed through the interactive dialog. We also observed significant improvements in user experience, both in terms of pragmatic and hedonic qualities. However, further research is needed to understand why one method outperformed the other, especially considering potential differences in cognitive load.

Given that the descriptive texts can now be generated on the fly by prompting a LLM in the background, our results highlight the potential of GAI for improving RSs not only from an algorithmic perspective, but also in terms of user-oriented aspects such as control and transparency. It should thus be easy to augment RSs with post-hoc explanations of recommendations or relationships between items and user model. This will no longer require the development of specialized methods that can only be used in certain scenarios, but, as demonstrated in this paper, can be achieved with a foundation model independent of the underlying recommendation approach. It should be noted, however, that the results depend on model, prompt engineering, and training data. Regardless of the ability of LLMs to generalize, it thus remains to be investigated whether the texts are still meaningful when something changes, less knowledge about the items is present in the data, or representatives are shown for which little external data are available.

In summary, our contribution is certainly only a first step towards a user-oriented integration of LLMs. However, we hope that it will help focus future work on the use of GAI in RSs more on the user experience than has been the case so far, where the introduction of new technologies has usually led to exclusively algorithm-oriented research efforts.

ACKNOWLEDGMENTS

The authors thank Robert Sczech, who implemented the proposed method and conducted this study as part of his Bachelor's thesis. This work was carried out while the first author was at the University of Duisburg-Essen.

REFERENCES

[1] Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. LLM Based Generation of Item-Description for Recommendation System. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 1204–1207.

[2] Saurabh Agrawal, John Trenkle, and Jaya Kawale. 2023. Beyond Labels: Leveraging Deep Learning and LLMs for Content Metadata. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA.

[3] Oscar Luis Alvarado Rodriguez, Veronika Vanden Abeele, David Geerts, and Katrien Verbert. 2019. "I Really Don't Know What 'Thumbs Up' Means": Algorithmic Experience in Movie Recommender Algorithms. In *Human-Computer Interaction — INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Lecture Notes in Computer Science, Vol. 11748. Springer, Berlin, Germany, 521–541.

[4] Xavier Amatriain, Josep M. Pujol, Nava Tintarev, and Nuria Oliver. 2009. Rate it Again: Increasing Recommendation Accuracy by User Re-Rating. In *RecSys '09: Proceedings of the 3rd ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 173–180.

[5] Betul Ay and Galip Aydin. 2021. Visual Similarity-Based Fashion Recommendation System. In *Generative Adversarial Networks for Image-to-Image Translation*, Arun Solanki, Anand Nayyar, and Mohd Naved (Eds.). Academic Press, 185–203.

[6] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 1007–1014.

[7] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv:2303.04226 [cs.AI]

[8] Shuo Chang, F. Maxwell Harper, and Loren G. Terveen. 2015. Using Groups of Items for Preference Elicitation in Recommender Systems. In *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 1258–1269.

[9] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. arXiv:2205.08084 [cs.IR]

[10] 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 *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 1126–1132.

[11] Dario Di Palma. 2023. Retrieval-Augmented Recommender System: Enhancing Recommender Systems with Large Language Models. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 1369–1373.

[12] Michael D. Ekstrand, John Riedl, and Joseph A. Konstan. 2011. Collaborative Filtering Recommender Systems. *Foundations & Trends in Human-Computer Interaction* 4, 2 (2011), 175–243.

[13] Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, Brian Chu, Zexi Chen, and Manoj Tiwari. 2023. Leveraging Large Language Models in Conversational Recommender Systems. arXiv:2305.07961 [cs.IR]

[14] 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]

[15] Mark P. Graus and Martijn C. Willemsen. 2015. Improving the User Experience During Cold Start Through Choice-Based Preference Elicitation. In *RecSys '15: Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 273–276.

[16] Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The Development and Validation of the Rational and Intuitive Decision Styles Scale. *Journal of Personality Assessment* 98, 5 (2016), 523–535.

[17] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large Language Models are Zero-Shot Rankers for Recommender Systems. arXiv:2305.08845 [cs.IR]

[18] 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]

[19] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and

Implicit User Feedback. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (2014), 8:1–8:26.

[20] Nicolas Jones, Armelle Brun, and Anne Boyer. 2011. Comparisons Instead of Ratings: Towards More Stable Preferences. In *WI-IAT '11: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, Washington, DC, USA, 451–456.

[21] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders – Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 10:1–10:46.

[22] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. (2017). arXiv:1711.02231 [cs.CV]

[23] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. *Recommender Systems Handbook*. Springer US, Boston, MA, USA, Chapter Evaluating Recommender Systems with User Experiments, 309–352.

[24] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa. 2011. A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 321–324.

[25] Joseph Konstan and Loren Terveen. 2021. Human-Centered Recommender Systems: Origins, Advances, Challenges, and Opportunities. *AI Magazine* 42, 3 (2021), 31–42.

[26] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.

[27] Yehuda Koren, Steffen Rendle, and Robert Bell. 2022. *Recommender Systems Handbook*. Springer US, New York, NY, Chapter Advances in Collaborative Filtering, 91–142.

[28] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions. arXiv:2309.01157 [cs.IR]

[29] Yu Liang and Martijn C. Willemsen. 2021. Interactive Music Genre Exploration with Visualization and Mood Control. In *IUI '21: Proceedings of the 26th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 175–185.

[30] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2023. How Can Recommender Systems Benefit from Large Language Models: A Survey. arXiv:2306.05817 [cs.IR]

[31] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems. arXiv:2302.03735 [cs.IR]

[32] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2018. Impact of Item Consumption on Assessment of Recommendations in User Studies. In *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 49–53.

[33] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive Recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies* 121 (2019), 21–41.

[34] Benedikt Loepp, Tim Hussein, and Jürgen Ziegler. 2014. Choice-Based Preference Elicitation for Collaborative Filtering Recommender Systems. In *CHI '14: Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3085–3094.

[35] Hanjia Lyu, Song Jiang, Hanqing Zeng, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, Yinglong Xia, and Jiebo Luo. 2023. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. arXiv:2307.15780 [cs.CL]

[36] Sheshera Mysore, Andrew Mccallum, and Hamed Zamani. 2023. Large Language Model Augmented Narrative Driven Recommendations. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 777–783.

[37] Gergana Y. Nenkov, Maureen Morrin, Andrew Ward, Barry Schwartz, and John Hulland. 2008. A Short Form of the Maximization Scale: Factor Structure, Reliability and Validity Studies. *Judgment and Decision Making* 3 (2008), 371–388.

[38] Denis Parra and Xavier Amatriain. 2011. Walk the Talk: Analyzing the Relation Between Implicit and Explicit Feedback for Preference Elicitation. In *UMAP '11: Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization*. Springer, Berlin, Germany, 255–268.

[39] Ladislav Peska and Stepan Balcar. 2022. The Effect of Feedback Granularity on Recommender Systems Performance. In *RecSys '22: Proceedings of the 16th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 586–591.

[40] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 157–164.

[41] Lior Rokach and Slava Kisilevich. 2012. Initial Profile Generation in Recommender Systems Using Pairwise Comparison. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews* 42, 6 (2012), 1854–1859.

[42] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (2017), 103–108.

[43] Taavi T. Taijala, Martijn C. Willemsen, and Joseph A. Konstan. 2018. MovieExplorer: Building an Interactive Exploration Tool from Ratings and Latent Taste Spaces. In *SAC '18: Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, New York, NY, USA, 1383–1392.

[44] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative Recommendation: Towards Next-generation Recommender Paradigm. arXiv:2304.03516 [cs.IR]

[45] 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. 2024. Enhancing Recommender Systems with Large Language Model Reasoning Graphs. arXiv:2308.10835 [cs.IR]

[46] 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. arXiv:2305.19860 [cs.IR]

[47] 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:2401.04997 [cs.IR]

[48] Bin Yin, Junjie Xie, Yu Qin, Zixiang Ding, Zhichao Feng, Xiang Li, and Wei Lin. 2023. Heterogeneous Knowledge Fusion: A Novel Approach for Personalized Recommendation via LLM. In *RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 599–601.

[49] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as Instruction Following: A Large Language Model Empowered Recommendation Approach. arXiv:2305.07001 [cs.IR]

[50] Qian Zhang, Jie Lu, and Yaochu Jin. 2021. Artificial Intelligence in Recommender Systems. *Complex & Intelligent Systems* 7, 1 (2021), 439–457.

[51] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *Comput. Surveys* 52, 1 (2019), 5:1–5:38.

[52] Joyce Zhou and Thorsten Joachims. 2023. GPT as a Baseline for Recommendation Explanation Texts. arXiv:2309.08817 [cs.AI]