VideolandGPT: A User Study on a Conversational Recommender System

Mateo Gutierrez Granada[1,*,†], Dina Zilbershtein[1,2,*,†], Daan Odijk[1] and Francesco Barile[2]

[1]*RTL Nederland B.V., Hilversum, The Netherlands*
[2]*Maastricht University, Maastricht, The Netherlands*

Abstract

This paper investigates how large language models (LLMs) can enhance recommender systems, with a specific focus on Conversational Recommender Systems that leverage user preferences and personalised candidate selections from existing ranking models. We introduce VideolandGPT, a recommender system for a Video-on-Demand (VOD) platform, Videoland, which uses ChatGPT to select from a predetermined set of contents, considering the additional context indicated by users' interactions with a chat interface. We evaluate ranking metrics, user experience, and fairness of recommendations, comparing a personalised and a non-personalised version of the system, in a between-subject user study. Our results indicate that the personalised version outperforms the non-personalised in terms of accuracy and general user satisfaction, while both versions increase the visibility of items which are not in the top of the recommendation lists. However, both versions present inconsistent behavior in terms of fairness, as the system may generate recommendations which are not available on Videoland.

Keywords

ChatGPT, Conversational Recommender Systems, Video Recommendations, Fairness

1. Introduction

Recommender systems have revolutionized various industries such as e-commerce, media, and online advertising by providing customized experiences based on users' profiles and behaviors. Initially, content filtering was used to match users based on their preferred categories [1], but the development of collaborative filtering techniques such as matrix factorization (MF) has enabled more effective personalization [2, 3]. More recently, the development of attention mechanisms that efficiently connect encoder and decoder via Transformer blocks [4] represented a significant advancement in neural architectures, initially for natural language processing. The emergence of Large Language Models (LLMs), such as BERT [5], and subsequently GPT-3 and chatGPT [6, 7, 8], is a direct result of this breakthrough.

As the Transformer architecture gained popularity in other domains, recommender system scholars also saw potential in the attention mechanism [9, 10], recognizing the utility of sequential information [11, 12, 13]. Breakthroughs in NLP research continued with the addition of new LLMs such PaLM [14] and LLaMA [15]. These advancements have not gone unnoticed by researchers from diverse domains, including Recommender Systems which tried to incorporate LLMs in their toolbox [16, 17]. Overall, LLMs hold great promise for improving the performance and capabilities of recommender systems.

LLMs bring several benefits to recommendation systems, including extensive knowledge, reasoning, natural language processing, and explainability, boosting user engagement and trust. They incorporate context, user preferences, and feedback, and transfer knowledge between domains, making them potent for creating accurate, explainable recommendation systems [18].

However, generating recommendations using LLMs is challenging for quite a few reasons [17]: they are prone to generate incomplete, hallucinatory and biased results [19], along with factually accurate but contextually inconsistent outcomes. Updating the parametric knowledge base and accommodating input token length are also significant challenges. Consequently, modern research often sees LLMs as summarization and reasoning engines rather than knowledge-based solutions for recommender systems, despite efforts to merge these approaches [20].

This paper examines the impact of LLMs on a recommender system that can converse and reason within the users' context, using their preferences and a set of personalised candidates. The study involves users from RTL's Videoland[1], the largest Dutch video-on-demand (VOD) platform. The aim is to investigate through a user study the user experience of personalised recommendations in a conversational context, including situations where users explicitly state their preferences using natural language. We examine whether there is a discernible differ-

Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, September 18–22 2023, Singapore.

[*]Corresponding author.
[†]These authors contributed equally.
Mateo.Gutierrez.Granada@rtl.nl (M. G. Granada); zilbershtein.dina@maastrichtuniversity.nl (D. Zilbershtein); Daan.Odijk@rtl.nl (D. Odijk); f.barile@maastrichtuniversity.nl (F. Barile)

[1] https://www.videoland.com/

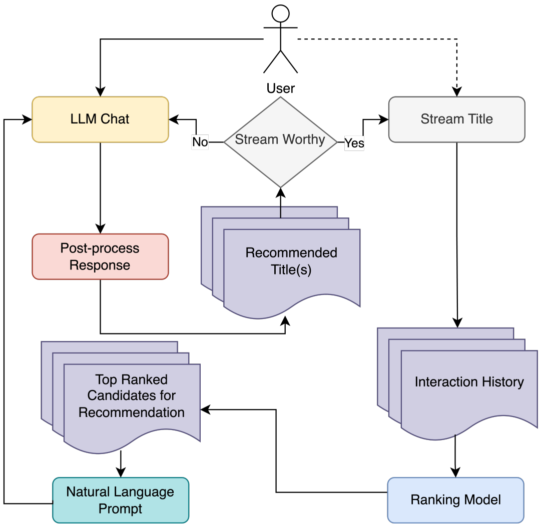


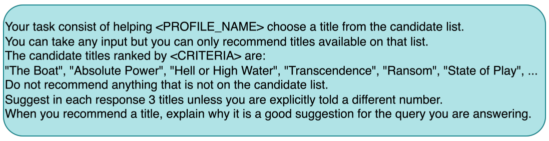
Figure 1: VideolandGPT offers a direct feedback loop with the user, starting with their streaming activity and generating an Interaction History that serves as input to the Ranking Model. The primary filtering is in the Ranking Model, where the top *k* titles are ranked and embedded as Candidates for Recommendation in a Natural Language Prompt. The LLM Chat is instructed to recommend titles from this subset during its conversation with the user. The post-processed response generated by the LLM Chat constitutes the Recommended Title(s), which the user can accept as a successful recommendation and subsequently stream, or continue the conversation with the LLM Chat.

ence in users' personalised and non-personalised LLM recommendations. Furthermore, we aim to determine if users are exposed to titles beyond the top ranking.

In addition to its focus on recommendation accuracy and performance, this study evaluates the safety and fairness of recommendations generated by our proposed Conversational Recommender System (CRS). We analyze if the LLM adheres to fairness definitions proposed by the research community [21]. Adopting the principle of fairness as ”no harm”, it becomes evident that recommending items not accessible on the Videoland platform undermines the platform's interests by encouraging people to find relevant content somewhere else. In this context, our analysis prioritizes aligning our recommender system with Videoland's objectives and avoiding any adverse impact on the platform's operations and goals.

2. VideolandGPT

We evaluate our approach on a prototype conversational recommender system for Videoland, that we detail in this section. We base our prototype on the Ranking Model that we presented in [13]. The architecture used to integrate the Ranking Model with the LLM's knowledge and capabilities, is illustrated in Figure 1.



Your task consist of helping <PROFILE_NAME> choose a title from the candidate list.
You can take any input but you can only recommend titles available on that list.
The candidate titles ranked by <CRITERIA> are:
"The Boat", "Absolute Power", "Hell or High Water", "Transcendence", "Ransom", "State of Play", ...
Do not recommend anything that is not on the candidate list.
Suggest in each response 3 titles unless you are explicitly told a different number.
When you recommend a title, explain why it is a good suggestion for the query you are answering.

Figure 2: This figure provides an example of a personalised prompt that tasks the LLM model with recommending three items to a specific user from a candidate list. The candidate list is personalised and can be dynamically sorted based on varying criteria.

In this architecture, the Ranking Model is considered a critical component of the solution and also a modular building block that can be replaced as needed. In our case, the model is an ensemble comprising a matrix factorization component [3] and a neural component [13], which utilizes the attention mechanism and sequential information in the *Interaction History*. Our Ranking Model retrieves the top 300 titles for each user, reducing the catalog by approximately 90%. We believe this number achieves a balance between relevance and discoverability.

Our Natural Language Prompt is created to give precise instructions to the LLM Chat model to recommend titles from Videoland's candidates. We specify that the model should retrieve three items and provide explanations for each recommendation to improve explainability. An example of the prompt is illustrated in Figure 2, which takes a candidate list of items sorted based on particular criteria and the user profile for which the recommendations are intended. This approach enables flexibility in accommodating various ranking methods.

The LLM Chat is designed to suggest a list of items that best matches the user's query and the candidate list of recommendations. As the conversation progresses, the user can either accept a recommendation or give feedback to the system to refine their discovery preferences. The user can also request new titles, ask for explanations for a particular recommendation, or seek further information related to it. We are testing our prototype with gpt-35-turbo as the LLM.

The post-processing step serves two critical functions. First, it enriches the LLM Chat's response with relevant metadata, such as the title's artwork and a direct link to stream it on Videoland. This additional information enhances the user's experience and makes it easier to access and enjoy recommended titles. Second, the post-processing step acts as a safeguard to remove any recommended title that is not directly aligned with the candidates for recommendation. In our experiment, we intentionally omitted this safeguard to examine the potential impact on platform fairness of not removing any recommended title that is offered by other platforms.

3. User Study

We conduct a small-scale user study to evaluate the performance of the recommender system. We compare two versions of the recommender system: a personalised version, based on users' recommendations and a non-personalised one, based on the most popular titles. The study aims to assess user satisfaction, platform fairness aspects and to answer the main research questions: How can LLMs enhance (our) recommender systems? Can such a system, converse and reason within the user's context, using their preferences and a set of personalised candidates? Is a personalised chat-based recommender system perceived to be more enjoyable and more relevant compared to its non-personalised counterpart?

In a separate study, Radensky et al. [22] examined the impact of confidence signal patterns on user trust and reliance in a music CRS. Their research inspired our evaluation approach, although our study covers broader aspects beyond confidence signals.

Participants The assignment of random groups was done prior to the study. The participants comprised employees within RTL. In total, 27 out of 42 invited participants took part in the study, ages ranging from 26 to 48, being 35% of them women. Participation in the survey was voluntary, and the employees had not previously interacted with VideolandGPT. The sole requirement for participation was that the respondents must have watched at least one title on Videoland within the last 6 months to have personalised recommendations.

Experiment Protocol The experiment's design is presented in Figure 3. All study participants were explicitly requested to engage with the system in English throughout the study. Following this, the respondents were randomly divided into two groups. Participants, unaware of the version they were using, engaged with either a personalized or non-personalized VideolandGPT, the latter featuring top popular titles from Videoland's collection, ensuring unbiased results.

Each participant was assigned a set of five tasks with the specific structures provided for each of them to ensure a more standardized evaluation process. Descriptions of the tasks are provided in Table 1. However, participants were informed they could use their own words during interactions with the system, promoting natural conversation. The study was conducted online over a designated four-day period, offering convenience and flexibility to participants.

Assessment of the system was based on diverse forms of describing users' preferences, which included previously loved titles, topics, current or desired emotions, preferred company for movie-watching, and free-form requests. During the conversations, users had the opportunity to request the system to refine the recommendations twice, resulting in a maximum exposure to 9 items per task. Each task was completed in separate instances of the same version of the recommender, ensuring an isolated examination.

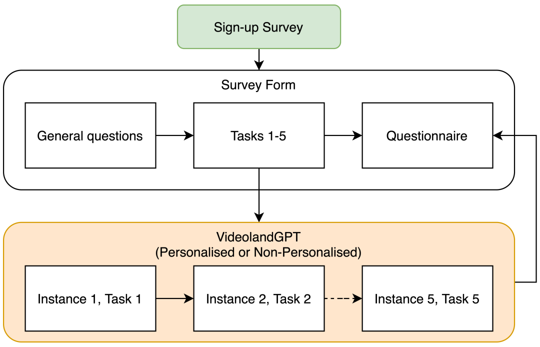


Figure 3: Experiment scheme. Each participant signed up for the survey and subsequently completed the form, which included task descriptions and the questionnaire.

Table 1
Task description with the suggested requests for recommendations.

Task	Suggested initial prompt
Title	Show me the most relevant titles considering that I like <TITLE>.
Topic	Show me the most relevant titles based on my passion for <TOPIC>.
Emotion	Show me the most relevant titles that will make me feel <EMOTION/DESIRE>.
Context	Show me the most relevant titles to watch with <GF/BF/SON/FRIEND> on a <DAY OF THE WEEK and/or EVENING/AFTERNOON/-MORNING>.
Free	<Ask for 3 items to be recommended in any form you would like.>

At the end of each task, respondents specified the title they considered the most relevant recommendation for them or stated that they did not receive a satisfactory recommendation. This feedback was used to understand VideolandGPT's recommendation capabilities, accuracy and fairness to the platform of the recommender.

Furthermore, because the participants were not exposed to VideolandGPT previously and to ensure the experiment's integrity, the order of the tasks was changed every five collected responses. By varying the task order, we sought to avoid any systematic influence on participants' responses, ensuring that the respondents' reactions to the tasks remained impartial and unaffected by the sequence in which they were presented.

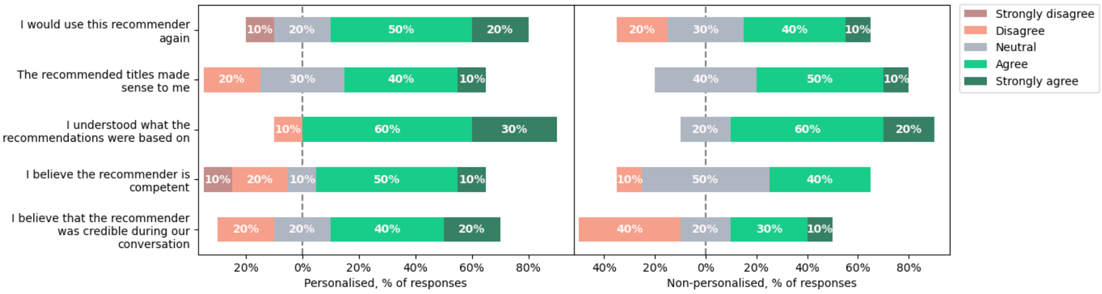


Figure 4: Likert-type questions [23] with the results for both versions of the Recommender System, displaying the users' responses to the evaluation questions. The Likert scale was used to assess users' perceptions and satisfaction.

After completing the tasks, participants were directed to fill out the questionnaire. The results to the Likert questions are presented in Figure 4. Moreover, participants were asked to rank the tasks based on their satisfaction from the conversation with the recommender and were encouraged to provide any additional feedback they had regarding its use. In addition, participants were asked about their native language, to explore any potential correlation between the quality of recommendations and their language background. This question was particularly relevant, as Videoland's collection primarily consists of contents in Dutch (57% of the titles accounting for 63% of the total available minutes).

4. Evaluation

We evaluate this study both quantitatively and qualitatively by analyzing the data collected from the logs of the conversation and the received questionnaire answers. It is important to note that not all responses from the conversations yielded usable data due to various reasons such as incomplete or ambiguous queries. As a result, we obtained 50 valid observations for each version of the recommender (five per respondent, one for each task).

Difference between two versions of the conversational recommender Table 2 presents the metrics used to evaluate both versions. We measured accuracy and relevance of recommendations by allowing participants to interact with 3, 6, or 9 recommended titles (with 8% of sessions interacting with other numbers < 9) in our experiment. We evaluated the recommendations' performance using nDCG@9 and HR@9 metrics, considering all participants regardless of the number of titles they interacted with.

The personalised framework demonstrated a 10% relative improvement over the non-personalised version in all tasks, highlighting the effectiveness of chat-based recommendations in improving user satisfaction and relevance in our research context.

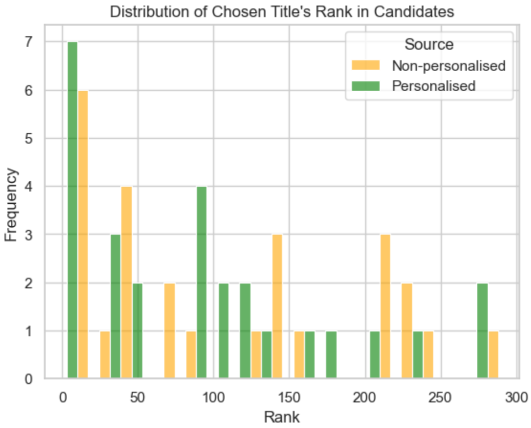


Figure 5: Distribution of the original rank in the candidates list for the titles that were chosen.

To assess the fairness of the recommender system to the platform and its compliance with the rules, we measured the proportion of recommended and chosen titles that were in the candidate list and the chosen titles that were not on the candidate list. Moreover, we consider a measure of efficiency the number of unique titles recommended per user. While the personalised recommender outperformed in relevance metrics, our examination revealed inconsistencies in fairness metrics. For both recommenders, over 22% of tasks had user-selected recommendations that were not available on Videoland, suggesting that the system occasionally generated recommendations beyond the platform's content availability, despite our attempts to control it.

Finally, the results presented in Figure 5 indicate how often users choose titles beyond the top ranking items. Our findings demonstrate that having a large pool of candidates is valuable, as users frequently select titles from across the entire range of recommendations.

Table 2
Experimental results on the two Ranking Models. The best results overall and per task are in boldface.

Task	Ranking Model	nDCG@9	HR@9	Recommended in Candidates	Chosen in Candidates	Chosen but not in Candidates	Unique Titles per User
Overall	**Personalised**	**0.4273**	**0.78**	0.6958	**0.54**	0.24	**24.9**
	Non-personalised	0.3880	0.74	**0.7636**	0.52	**0.22**	26.2
Title	Personalised	0.3537	0.60	0.5750	0.40	0.20	5.6
	Non-personalised	**0.5635**	0.80	0.7188	0.70	0.10	5.1
Topic	Personalised	0.4848	0.80	0.4833	0.30	0.50	**4.3**
	Non-personalised	0.4438	0.70	0.6722	0.30	0.40	5.4
Emotion	Personalised	0.3421	0.70	0.8355	0.60	0.10	6.6
	Non-personalised	0.2185	0.60	**0.9380**	0.60	**0.00**	7.3
Context	Personalised	0.4215	**0.90**	0.9222	**0.80**	0.10	5.9
	Non-personalised	0.4371	**0.90**	0.8611	0.60	0.30	5.2
Free	Personalised	0.5343	**0.90**	0.6633	0.60	0.30	4.6
	Non-personalised	0.2772	0.70	0.6277	0.40	0.30	6.7

Overall experience of using a conversational recommender In the second phase of our evaluation, we analyzed the feedback received from the questionnaire. The metrics substantiated the results, revealing a statistically significant positive correlation (Pearson coefficient of 0.26) between quantitative metrics like nDCG@9 and qualitative metrics like users' task rankings. For instance, the *Title* task was preferred by 30% and 60% of participants for personalised and non-personalised versions, respectively, in their rankings, aligning with corresponding nDCG scores. These findings endorse our metrics' effectiveness in capturing user preferences and judgments. However, it is important to note that this difference, while notable, is not statistically significant due to the relatively small sample size of participants. Consequently, providing an explanation for why the non-personalised version performed better on this task is challenging and requires further investigation.

The Likert questions answers indicate that a comparable proportion of respondents agreed or strongly agreed with three or more statements for both versions of the recommender system (70% for personalised and 60% for non-personalised). However, there is a notable difference: 40% of respondents in the personalised version expressed agreement with all statements, while only 10% did so in the non-personalised version. This suggests that the personalised version garnered a higher percentage of highly satisfied users with its recommendations. Additionally, we can observe that the non-personalised version elicited more neutral responses from the participants. This suggests a more mixed perception of the non-personalised version's recommendations.

The findings from the open-ended questions shed light on user perceptions of the recommender system's experience. Notably, 80% of users perceived the personalised version as enjoyable, even when their specific requests were not entirely met. In contrast, 60% of users found the non-personalised version enjoyable despite similar circumstances. The respondents also mentioned, that this recommender *"could bring added value to the Videoland experience"*. A common feedback from participants who expressed dissatisfaction with their experience was the unavailability of relevant titles on the platform.

5. Discussion and Conclusion

Our study demonstrated that the personalised recommender outperformed the non-personalised version by delivering more relevant recommendations to users. However, it's important to recognise that both versions of the recommender still, in some cases, suggested titles that were not available on the platform, contrary to our initial expectations. This aspect highlights the need for further improvements and considerations in ensuring system consistency. Despite this drawback, the study shed light on the potential of personalised chat-based recommendations to improve user satisfaction and relevance, offering valuable insights for future developments in recommender systems.

Limitations of the study include a primarily Dutch-speaking sample (65% of all of the participants) due to the platform catering to a Dutch-speaking population, limited sample size, and the need to consider privacy and user preferences when implementing conversational recommender systems. Furthermore, if users explicitly share personal details with a conversational recommender system, it could impact their comfort in utilizing the system. Safeguards must be in place to ensure safety and prevent users from exploiting the system.

In conclusion, the study emphasizes the potential of personalised chat-based recommendations to enhance user experience, but further research is required to develop a safer mechanism for LLMs usage, ensuring adherence to rules and understanding potential unfair scenarios.

References

[1] M. Naumov, D. Mudigere, H.-J. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C.-J. Wu, A. G. Azzolini, et al., Deep learning recommendation model for personalization and recommendation systems, arXiv preprint arXiv:1906.00091 (2019).

[2] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37.

[3] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: 2008 Eighth IEEE international conference on data mining, Ieee, 2008, pp. 263–272.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[6] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, OpenAI Technical Report (2018).

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (2019) 9.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[9] T. Donkers, B. Loepp, J. Ziegler, Sequential user-based recurrent neural network recommendations, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, pp. 152–160.

[10] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 197–206.

[11] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450.

[12] Q. Chen, H. Zhao, W. Li, P. Huang, W. Ou, Behavior sequence transformer for e-commerce recommendation in alibaba, in: Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data, 2019, pp. 1–4.

[13] M. Gutierrez Granada, D. Odijk, Recommendations at videoland, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 580–582.

[14] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. Cite arxiv:2302.13971.

[16] J. Li, W. Zhang, T. Wang, G. Xiong, A. Lu, G. G. Medioni, Gpt4rec: A generative framework for personalized recommendation and user interests interpretation, ArXiv abs/2304.03879 (2023).

[17] J. Liu, C. Liu, R. Lv, K. Zhou, Y. B. Zhang, Is chatgpt a good recommender? a preliminary study, ArXiv abs/2304.10149 (2023).

[18] Z. Cui, J. Ma, C. Zhou, J. Zhou, H. Yang, M6-rec: Generative pretrained language models are open-ended recommender systems, ArXiv abs/2205.08084 (2022).

[19] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154.

[20] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, ArXiv abs/2306.08302 (2023).

[21] J. J. Smith, L. Beattie, H. Cramer, Scoping fairness objectives and identifying fairness metrics for recommender systems: The practitioners' perspective, in: Proceedings of the ACM Web Conference 2023, WWW '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3648–3659.

[22] M. Radensky, J. A. Séguin, J. S. Lim, K. Olson, R. Geiger, "i think you might like this": Exploring effects of confidence signal patterns on trust in and reliance on conversational recommender systems, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 792–804.

[23] D. C. Toader, G. D. Boca, R. Toader, M. Macelaru, C. Toader, D. S. Ighian, A. T. G. Rădulescu, The effect of social presence and chatbot errors on trust, Sustainability (2019).