Towards LLM-Based Usability Analysis for Recommender User Interfaces

Sebastian Lubos[1,*], Alexander Felfernig[1], Damian Garber[1], Viet-Man Le[1] and Thi Ngoc Trang Tran[1]

[1]*Graz University of Technology, Inffeldgasse 16b, Graz, 8010, Austria*

Abstract

Usability is a key factor in the effectiveness of recommender systems. However, the analysis of user interfaces is a time-consuming process that requires expertise. Recent advances in *multimodal large language models (LLMs)* offer promising opportunities to automate such evaluations. In this work, we explore the potential of multimodal LLMs to assess the usability of recommender system interfaces by considering a variety of publicly available systems as examples. We take user interface screenshots from multiple of these recommender platforms to cover both *preference elicitation* and *recommendation presentation* scenarios. An LLM is instructed to analyze these interfaces with regard to different usability criteria and provide explanatory feedback. Our evaluation demonstrates how LLMs can support heuristic-style usability assessments at scale to support the improvement of user experience.

Keywords

User Interfaces for Recommender Systems, Usability Analysis, Multimodal Large Language Models

1. Introduction

Recommender systems are a central component of many digital platforms, where they provide personalized item suggestions to help users navigate large sets of options [1]. While the quality of the underlying recommendation algorithm is important [2], the overall effectiveness of a recommender system also depends on how well users can interact with the interface. Usability and user experience play a key role in enabling users to express preferences, interpret recommendations, and make informed choices [3, 4]. Even highly accurate recommendations may fail to deliver value if the interface is difficult to navigate or lacks transparency.

Traditional usability evaluation methods include *usability testing* with real users [5] and *expert inspections* based on heuristic principles [6]. While these methods are effective, they are time-consuming and require expert involvement. General usability guidelines, such as *Nielsen's heuristics* [7] offer structured support, and recommender-specific frameworks further improve contextual relevance [8]. Nevertheless, usability assessments remain resource-intensive and are thus rarely applied across platforms.

To reduce this effort, automated solutions such as rule-based tools and heuristic checkers have been proposed [9, 10]. However, these often cover only limited usability dimensions and struggle with subjective or context-specific issues [11]. More recently, *multimodal large language models (LLMs)* that can process both visual and textual inputs have emerged [12]. Early studies demonstrate their ability to identify usability issues in design mockups [13] and mobile interfaces [14], although expert validation remains necessary. Initial research on the alignment between LLM-based analyses and expert assessments reports promising accuracy in different scenarios [15, 16], but more studies are needed to confirm these results.

While these approaches address general usability, they have not yet been applied to the specific challenges of recommender system interfaces, such as explainability, feedback mechanisms, and preference elicitation workflows, which are central to the user experience in this context. In this work, we explore how a multimodal LLM can help to analyze the usability of ten publicly available recommender interfaces based on explicitly defined criteria. We review the analysis results to highlight the feasibility and benefits of automated usability analysis for recommender interfaces and outline directions for future research.

The paper is organized as follows: Section 2 describes the experimental setup and implementation details. Section 3 presents the results. Section 4 discusses implications and future work. Finally, the paper is concluded in Section 5.

2. Usability Analysis of Recommender Interfaces

A lot of focus in recommender system research is put on the accuracy of algorithms and personalization strategies [2, 17]. However, the quality of recommender user interfaces plays an equally important role in shaping the overall user experience. This experience can be assessed through *usability analysis* that is concerned with the aspect of how effectively users can navigate, interpret, and interact with different parts of the recommender system [3, 4]. These include possibilities to express explicit preferences, understand why items are recommended, review recommended items, and provide feedback.

The following sections outline our considered recommender scenarios, usability criteria, and describe the LLM-based analysis in detail.

2.1. Recommender Scenarios

To explore the LLM-based usability analysis across a diverse set of recommender interfaces, we selected ten publicly accessible platforms from various item domains, which are summarized in Table 1. These systems vary in layout complexity, interaction mechanisms, and types of recommendations. This way, the automated usability analysis could be reviewed for varying contexts to get an impression about its generalizability. Each platform was assessed in two typical usage scenarios: (i) *preference elicitation*, and (ii) *recommendation presentation*.

Platform	Item Domain	URL
Amazon	E-commerce	https://www.amazon.com
Goodreads	Books	https://www.goodreads.com
Google News	News Articles	https://news.google.com
KaptnCook	Recipes	https://www.kaptncook.com
Last.fm	Music	https://www.last.fm
Netflix	Movies & TV Shows	https://www.netflix.com
Pinterest	Visual Content	https://www.pinterest.com
Spotify	Music	https://open.spotify.com
Steam	Video Games	https://store.steampowered.com
YouTube	Videos	https://www.youtube.com

Table 1
Overview of selected recommender platforms for usability analysis.

To have a comparable situation for each platform, we considered a new user scenario of a user interacting with the recommender for the first time. For this purpose, we used a desktop browser in incognito mode to avoid personalization effects and simulate a first-time user experience.[1] For each platform, we captured a representative screenshot[2] for the usage scenarios. Depending on the platform, the *preference elicitation* either showed the default onboarding screens or initial filter options. The

---

[1] A new user account was registered if needed to use the application.
[2] Screenshots were taken at full resolution and included the visible viewport with relevant UI context (e.g., navigation bars, filters, recommendation labels).

recommendation presentation showed either the main homepage (dashboard) or a detail page including recommended items.

To ensure comparability across platforms while accounting for platform-specific interaction patterns, we defined explicit user tasks for each platform. This allowed us to maintain a consistent evaluation structure while respecting the nuances of individual interfaces. The tasks for both scenarios and all platforms are presented in Table 2. They were used to select the screenshots and provide contextual information to the LLM during the usability analysis.

Platform	Preference Elicitation Task	Recommendation Presentation Task
Amazon	Search for "Bluetooth headphones" and interact with product listings to express shopping intent.	Review the related product recommendations shown on the product or search results page.
Goodreads	Rate previously read books as part of the onboarding process to express reading preferences.	Review the initial book recommendations generated based on ratings.
Google News	Select preferred news topics or regions during setup to tailor content delivery.	Review the personalized news feed on the dashboard.
KaptnCook	Indicate disliked ingredients during the initial setup to personalize meal suggestions.	Review the list of recommended recipes based on stated preferences.
Last.fm	Choose a trending artist to indicate music preferences.	Review the list of recommended tracks or artists based on the selected input.
Netflix	Select at least three preferred titles during the onboarding setup to express content preferences.	Review the personalized dashboard with recommended movies and shows.
Pinterest	Select inspirational images reflecting personal interests during the onboarding process.	Review the personalized feed with recommended visual content.
Spotify	Add songs to playlists based on initial recommendations to express musical preferences.	Review the homepage or dashboard with recommended tracks.
Steam	Apply filters (e.g., "Indie" genre) while browsing to express game preferences.	Review the recommended games displayed on the store homepage or Discovery Queue.
YouTube	Search for and watch a video on tennis serve drills to signal viewing preferences.	Review the recommended videos shown after watching the selected content.

Table 2
Standardized user task descriptions for preference elicitation and recommendation presentation.

2.2. Usability Criteria

To analyze the usability of the recommender system interfaces in a structured way, we defined a set of criteria that are shown in Table 3. These are based on general established usability principles [7] and user-centric evaluation metrics for recommender systems [8]. We defined and adapted these criteria to cover general interface qualities and scenario-specific aspects of *preference elicitation* and *recommendation presentation*. Our goal was to assess whether recommender interfaces support users in understanding, influencing, and responding to recommendations.

2.3. LLM-based Usability Analysis

We used the *gemini-2.5-flash* model by Google for the LLM-based usability analysis [18]. This model was designed to process textual and visual inputs, which makes it suitable for our experiments. We instructed the model in Python using the Gemini Developer API.[3] To improve the reproducibility of

---

[3]https://ai.google.dev/gemini-api/docs

Category	Usability Criterion
General (Both Scenarios)	*G1.* Is the layout clear and visually structured?
	G2. Are interactive elements (e.g., buttons, icons) clearly recognizable?
	G3. Is the amount of information per item appropriate and helpful?
	G4. Are interface elements used consistently (e.g., icons, labels, colors)?
Preference Elicitation	*P1.* Can users explicitly express preferences (e.g., ratings, likes, categories)?
	P2. Is there transparency about how input affects recommendations?
	P3. Do users have control and flexibility (e.g., skip, edit, undo inputs)?
Recommendation Presentation	*R1.* Are recommendations clearly labeled as such?
	R2. Are different types of recommendations distinguishable (e.g., "Because you liked...")?
	R3. Are there explanations for why items are recommended?
	R4. Can users interact with recommendations (e.g., rate, hide, save)?

Table 3
Usability criteria used to analyze recommender interfaces.

results, we set the *temperature* to 0.0 for more deterministic LLM output.

For the analysis task, the recommender interface screenshot was provided as context, together with a high-level description of the platform, the considered usage scenario (*preference elicitation* or *recommendation presentation*), and an explicit user task (see Table 2). To define the role and boundaries of the LLM in this scenario, we used the *system prompt* shown in Figure 1.

> You are a usability evaluation assistant specialized in analyzing recommender system interfaces.
> Your role is to assess screenshots of interfaces based on a fixed set of usability criteria.
> You always respond in structured JSON format, using the flat structure defined by the user.
> Be concise, accurate, and avoid assumptions that are not visually grounded.

Figure 1: System prompt for usability analysis.

Figure 2 shows the *user prompt* template to instruct the LLM for the explicit analysis of the provided screenshot. It describes the expected output, which includes a binary evaluation of whether the usability criterion is fulfilled or not, an explanation to justify its decision, and a suggestion on how to improve this usability aspect in case the criterion was not fulfilled. The respective list of usability criteria for each scenario (see Table 3) was provided in the prompt.

> You are a usability expert analyzing a recommender system interface from the perspective of a new user.
> The following screenshot shows the interface of **{platform}**.
> The platform is: **{platform_description}**
> Analyze the interface based on a **{scenario}** scenario.
> The user is performing the following task: **{task_description}**
>
> Evaluate the interface based on the following usability criteria.
> For each criterion, provide:
> 1. Whether the criterion is fulfilled (`true` or `false`)
> 2. A short explanation of your decision
> 3. If the criterion is not fulfilled, suggest a concrete recommendation for improvement
>
> **{result_formatting_instruction}**
>
> The usability criteria to evaluate are:
> - G1. Is the layout clear and visually structured?
> - **{further_criteria}**
>
> Include each of the provided criteria exactly once.
> Respond only with the structured JSON list.

Figure 2: User prompt template for usability analysis. Variables are highlighted in braces ({}).

3. Results

We analyzed 10 platforms across 2 usage scenarios each, and considered 11 different usability criteria. This resulted in 150 individual assessments (respecting partly scenario-specific criteria, see Table 3). The LLM completed the analysis in approximately 216 seconds.

Figure 3 summarizes the fulfillment rates for each usability criterion across all evaluated platforms. The results show that general interface design aspects, such as clear layout (G1), recognizable interactive elements (G2), appropriate information density (G3), and consistent visual styling (G4), are well-supported on most platforms. This suggests that these systems follow established design conventions and offer a solid baseline for usability. This outcome is expected, given that the evaluated platforms are widely used and likely to have invested greatly in interface design and user experience.

In contrast, the recommender-specific criteria show a more diverse picture. Particularly, the presence of explanations (R3) and interactive feedback options (R4) was less frequently fulfilled. The same holds for transparency of underlying algorithms (R2/P2), explicit feedback options (P1), and flexibility of interaction (P3). This suggests that while basic UI design is generally strong, many platforms lack mechanisms to help users understand and influence recommendation behavior. These findings point to a gap in explainability and user control in many "*black-box*" recommender settings.
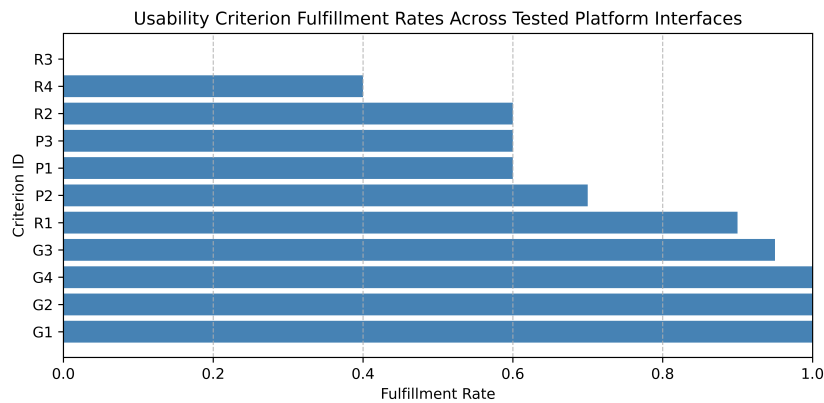

Figure 3: Fulfillment rates for each usability criterion across all platforms. The horizontal bars indicate the proportion of platforms for which each criterion was considered fulfilled by the LLM-based analysis.

Figure 4 shows the average fulfillment rates for each platform, separated by usage scenario. The results indicate that most platforms performed slightly better in the *preference elicitation* scenario, where users explicitly express interests or filters. In contrast, in the *recommendation presentation* scenario, where users review recommended items, more usability issues were revealed. Still, the overall high fulfillment rates are expected as all platforms are widely used.

The quantitative results offer an initial overview of the LLM-based usability analysis. To better understand the practical value of the approach, we examined individual outputs in detail. The following examples show how different criteria were analyzed across platforms.

The analysis revealed only one unfulfilled general usability criterion related to the amount of information shown per item (G3) on the *Goodreads* platform. Figure 5 presents the corresponding LLM explanation and improvement suggestion, alongside a UI sketch that highlights the issue.[4] In this case, the LLM makes a valid point: when book covers are small and titles are unreadable, it is difficult for users to get a first impression of the item without reviewing its details. The suggested improvement is appropriate and addresses this usability gap effectively.

For the recommender-specific criteria, several cases of unfulfilled usability were identified. Figure 6

---

[4]To avoid potential legal issues with publishing proprietary platform screenshots, we redrew the relevant parts as simplified UI sketches for presentation in this paper. Importantly, all analyses were conducted using the original screenshots. The sketches only serve as substitutes for publication and preserve the necessary information for understanding the reported aspects. This procedure was applied to all examples shown. For the original screen design, we refer to the respective platform web pages (see Table 1).
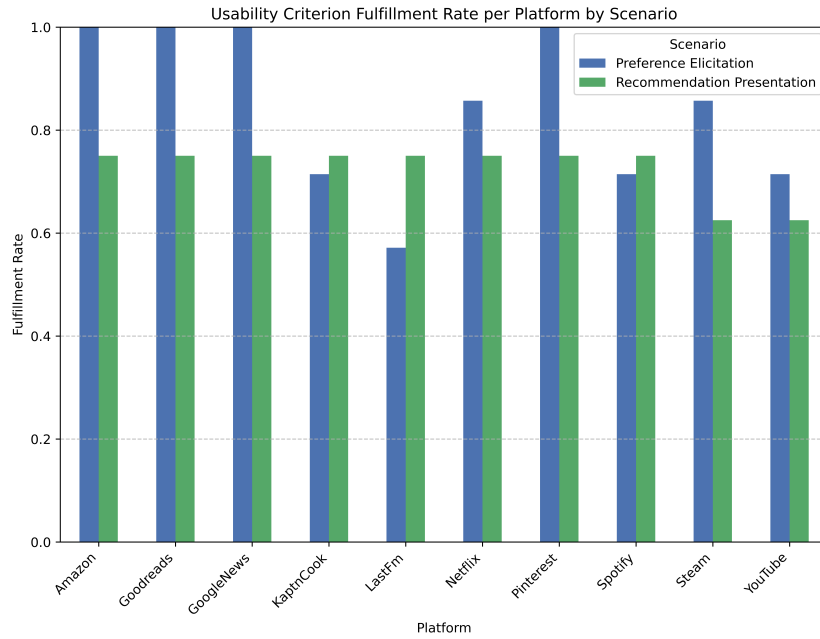
Figure 4: Fulfillment rates of usability criteria per platform and usage scenario.
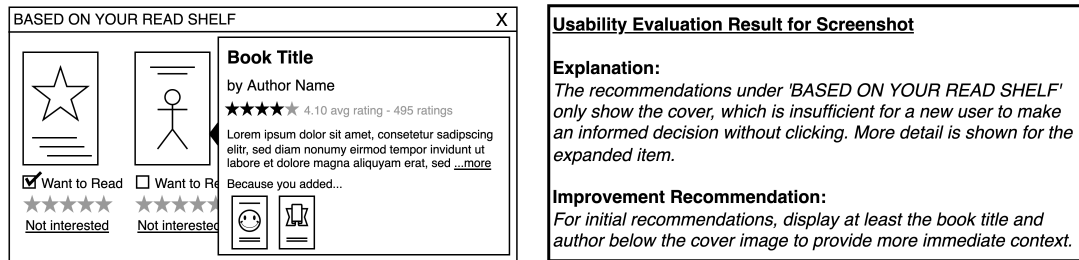

Figure 5: Example usability analysis result for *Goodreads* on the *unfulfilled information per item (G3)* criterion, with corresponding explanations and improvement suggestions. The sketch illustrates the analyzed UI.
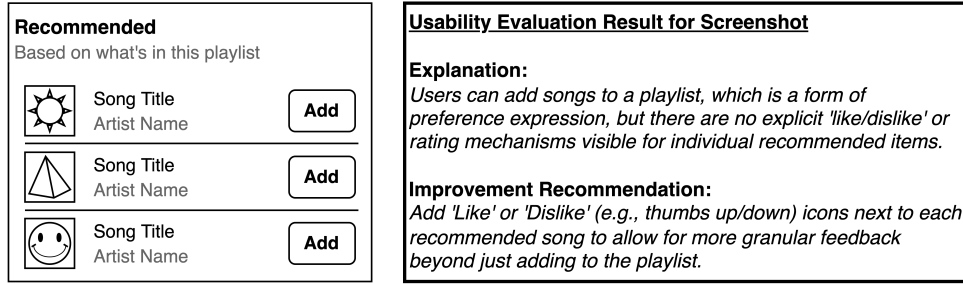
illustrates example results related to criterion P1, which concerns the ability for users to explicitly express preferences. In the playlist creation scenario on *Spotify*, the LLM judged this criterion as unfulfilled, since users can only add recommended songs without any direct feedback mechanism, such as liking or disliking. In contrast, *Amazon* was evaluated as fulfilling the criterion. Users could actively apply clickable filters to indicate item type preferences.

The explanation provided for *Spotify* is particularly nuanced and insightful. While the LLM acknowledges that adding songs to a playlist can be seen as a form of preference expression, it argues that this action alone may not sufficiently satisfy the criterion. This reasoning is persuasive as users might skip a recommendation for various reasons, such as disliking the artist or simply finding the song unsuitable for the current playlist, which is not explicitly communicated to the system. The suggested improvement, to provide more fine-grained feedback options, is both reasonable and actionable.

In the *recommendation presentation* scenario, several criteria were considered unfulfilled. Figure 7 presents example results for criterion R4, which concerns the ability of users to interact with recommended items. For *Google News*, the LLM noted the absence of visible interaction options and suggested improvements. In contrast, *YouTube* was evaluated more positively, as its recommendation cards include an accessible interaction menu ("*three-dot*" menu).

These examples highlight both the usefulness and limitations of our approach. While the LLM was able to identify relevant usability gaps, it also operated solely on static screenshots. As a result, interactive elements that appear only on hover or during user interaction, such as the hidden menu in *Google News*, may remain unrecognized. Nevertheless, the LLM's suggestion remains valid, as for a new or inexperienced user, the lack of visible affordances can be a barrier to effective interaction and may justify more prominent cues.

(a) Unfulfilled P1 criterion (Spotify)

Recommended
Based on what's in this playlist

	Song Title Artist Name	**Add**
	Song Title Artist Name	**Add**
	Song Title Artist Name	**Add**

Usability Evaluation Result for Screenshot

Explanation:
Users can add songs to a playlist, which is a form of preference expression, but there are no explicit 'like/dislike' or rating mechanisms visible for individual recommended items.

Improvement Recommendation:
Add 'Like' or 'Dislike' (e.g., thumbs up/down) icons next to each recommended song to allow for more granular feedback beyond just adding to the playlist.

(b) Fulfilled P1 criterion (Amazon)

All ▾ headphones

🔍 headphones over ear
🔍 headphones noise cancelling
🔍 headphones for kids
🔍 headphones with mic

HEADPHONES BY TYPE

| ▶ | ✳ | ✹ |
| Gaming | Bluetooth | Bud |

Usability Evaluation Result for Screenshot

Explanation:
Users can explicitly express preferences by selecting specific search terms or by choosing a product type (e.g., 'Gaming', 'Bluetooth') from the suggested categories.
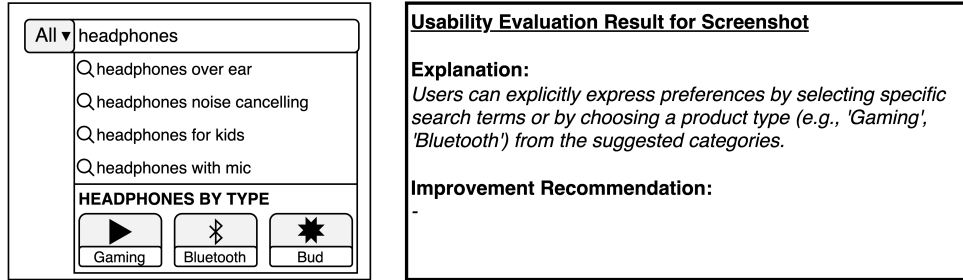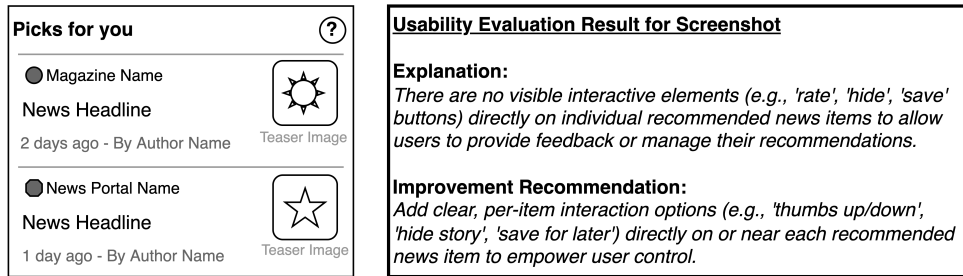
Improvement Recommendation:
-

Figure 6: Example usability analysis result for the *explicit preference expression (P1)* criterion, including explanations and improvement suggestions. The sketches illustrate the analyzed UIs.

(a) Unfulfilled R4 criterion (Google News).

Picks for you   (?)

⬤ Magazine Name

News Headline

2 days ago - By Author Name   Teaser Image

⬤ News Portal Name

News Headline

1 day ago - By Author Name   Teaser Image

Usability Evaluation Result for Screenshot

Explanation:
There are no visible interactive elements (e.g., 'rate', 'hide', 'save' buttons) directly on individual recommended news items to allow users to provide feedback or manage their recommendations.

Improvement Recommendation:
Add clear, per-item interaction options (e.g., 'thumbs up/down', 'hide story', 'save for later') directly on or near each recommended news item to empower user control.

(b) Fulfilled R4 criterion (YouTube).

| ▶ | **Video Name** ⋮
Channel Name
171K views - 5 years ago |
| ▶ | **Video Name** ⋮
Channel Name
1.1M views - 1 years ago |

Usability Evaluation Result for Screenshot

Explanation:
Each recommended video has a three-dot menu icon, which typically allows users to interact with the recommendation (e.g., 'Not interested', 'Save to watch later').

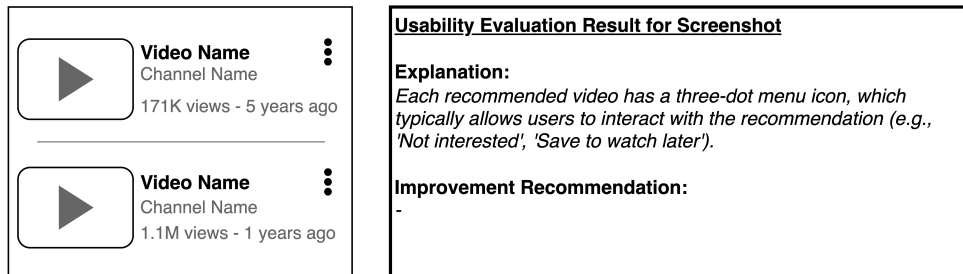**Improvement Recommendation:**
-

Figure 7: Example usability analysis result for the *explicit interactiveness of recommended items (R4)* criterion, including explanations and improvement suggestions. The sketches illustrate the analyzed UIs.

Another notable observation is that the LLM judged none of the platforms as fulfilling criterion R3, which concerns providing explanations for why items are recommended. However, a closer review of the platforms and LLM explanations reveals a more nuanced picture. While many platforms indeed lack explicit explanations, some offer at least high-level contextual hints. For instance, *Spotify* displays recommended playlists with labels such as "*Brand new music from artists you love*", which is a high-level explanation for the recommendation. This suggests that the criterion could benefit from further refinement to distinguish between vague contextual hints and explicit, personalized explanations.

4. Discussion

Our results suggest that LLM-based usability analysis can provide useful, low-effort insights into the strengths and weaknesses of recommender interfaces. The generated explanations and improvement suggestions were generally accurate, understandable, and relevant, which indicates the potential for such tools to support development, especially in early-stage prototyping and iterative design. While LLMs cannot replace expert evaluations, they could act as assistive tools in a human-in-the-loop setting, reducing manual effort and accelerating the identification of usability issues.

Building on these findings, several research challenges emerge:

Prioritization of Issues. The number of identified usability issues of a recommender system can be large and trigger significant effort for developers to evaluate them and prioritize their fixes. The currently used binary fulfillment decisions do not capture the severity of issues, which limits the possibilities for ranking. Using more nuanced assessments using severity ratings could support issue prioritization and make the analysis more actionable by highlighting the most critical usability gaps.

Prompting Design and Context. Prompt design can influence the analysis results. Our current template evaluates multiple criteria simultaneously (see Figure 2), which reduces inference costs but sometimes may overlook issues. Using separate prompts for each criterion could improve effectiveness.

Another opportunity for improvement is the refinement of the context description and clarifying the intended meaning of each criterion. In some cases, the LLM struggled to interpret the criteria consistently, for example, determining what level of detail qualifies as an explanation (R3). Making this clearer by adding more detailed descriptions or examples could lead to more accurate and robust results.

Dynamic UI Behavior. A boundary of our current approach is the reliance on static screenshots, which prevents the model from recognizing dynamic interface elements that appear *on mouse hover*[5] or click. Providing a video of recorded interactions as context could show the complete interface behavior and overcome this limitation. Beyond that, LLM-based agents that directly interact with interfaces could be an even more elaborate approach to also automate the data collection aspect.

Validation Against Expert Assessments. Systematic comparison with expert evaluations is needed to assess the reliability and practical value of LLM-based usability analysis. While early studies report promising alignment with expert judgments on general usability criteria [15, 16], broader comparisons, particularly in recommender-specific contexts, are needed to validate this.

5. Conclusions

In this paper, we explored an LLM-based approach to usability analysis of recommender user interfaces. We applied the method to ten publicly available platforms to assess whether the identified issues were plausible, clearly explained, and accompanied by meaningful improvement suggestions. Our findings demonstrate the potential of multimodal LLMs to support low-effort usability evaluation of recommender interfaces, particularly during early-stage design. Also, the findings highlight different areas for improvement, particularly in handling dynamic interface elements and generating more nuanced, context-aware judgments. Future work will focus on improving prompt strategies and contextual understanding, and validating LLM-generated assessments against expert evaluations.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: *Grammar and spelling check*, *Paraphrase and reword*. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

---

[5]see https://www.w3schools.com/howto/howto_css_display_element_hover.asp for an *on mouse hover* example.

References

[1] F. Ricci, L. Rokach, B. Shapira, Recommender Systems: Techniques, Applications, and Challenges, Springer US, New York, NY, 2022, pp. 1–35. URL: https://doi.org/10.1007/978-1-0716-2197-4_1. doi:10.1007/978-1-0716-2197-4_1.

[2] E. Zangerle, C. Bauer, Evaluating recommender systems: Survey and framework, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3556536. doi:10.1145/3556536.

[3] P. Pu, L. Chen, R. Hu, Evaluating recommender systems from the user's perspective: survey of the state of the art, User Modeling and User-Adapted Interaction 22 (2012) 317–355.

[4] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User modeling and user-adapted interaction 22 (2012) 441–504.

[5] C. Hass, A Practical Guide to Usability Testing, Springer International Publishing, Cham, 2019, pp. 107–124. URL: https://doi.org/10.1007/978-3-319-96906-0_6. doi:10.1007/978-3-319-96906-0_6.

[6] T. Hollingsed, D. G. Novick, Usability inspection methods after 15 years of research and practice, in: Proceedings of the 25th Annual ACM International Conference on Design of Communication, SIGDOC '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 249–255. URL: https://doi.org/10.1145/1297144.1297200. doi:10.1145/1297144.1297200.

[7] J. Nielsen, Enhancing the explanatory power of usability heuristics, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '94, Association for Computing Machinery, New York, NY, USA, 1994, p. 152–158. URL: https://doi.org/10.1145/191666.191729. doi:10.1145/191666.191729.

[8] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 157–164. URL: https://doi.org/10.1145/2043932.2043962. doi:10.1145/2043932.2043962.

[9] A. Namoun, A. Alrehaili, A. Tufail, A review of automated website usability evaluation tools: Research issues and challenges, in: M. M. Soares, E. Rosenzweig, A. Marcus (Eds.), Design, User Experience, and Usability: UX Research and Design, Springer International Publishing, Cham, 2021, pp. 292–311.

[10] J. W. Castro, I. Garnica, L. A. Rojas, Automated tools for usability evaluation: A systematic mapping study, in: G. Meiselwitz (Ed.), Social Computing and Social Media: Design, User Experience and Impact, Springer International Publishing, Cham, 2022, pp. 28–46.

[11] E. Kuric, P. Demcak, M. Krajcovic, J. Lang, Systematic literature review of automation and artificial intelligence in usability issue detection, 2025. URL: https://arxiv.org/abs/2504.01415. arXiv:2504.01415.

[12] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, National Science Review 11 (2024). URL: http://dx.doi.org/10.1093/nsr/nwae403. doi:10.1093/nsr/nwae403.

[13] P. Duan, J. Warner, Y. Li, B. Hartmann, Generating automatic feedback on ui mockups with large language models, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642782. doi:10.1145/3613904.3642782.

[14] A. E. Pourasad, W. Maalej, Does GenAI Make Usability Testing Obsolete? , in: 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE), IEEE Computer Society, Los Alamitos, CA, USA, 2025, pp. 675–675. URL: https://doi.ieeecomputersociety.org/10.1109/ICSE55347.2025.00138. doi:10.1109/ICSE55347.2025.00138.

[15] R. Zhong, D. W. McDonald, G. Hsieh, Synthetic heuristic evaluation: A comparison between ai- and human-powered usability evaluation, 2025. URL: https://arxiv.org/abs/2507.02306. arXiv:2507.02306.

[16] S. Lubos, A. Felfernig, G. Leitner, J. Schwazer, Towards recommending usability improvements with multimodal large language models, 2025. URL: https://arxiv.org/abs/2508.16165. arXiv:2508.16165.

[17] A. Gunawardana, G. Shani, S. Yogev, Evaluating Recommender Systems, Springer US, New York, NY, 2022, pp. 547–601. URL: https://doi.org/10.1007/978-1-0716-2197-4_15. doi:10.1007/978-1-0716-2197-4_15.

[18] G. Team, O. Authors, Gemini: A family of highly capable multimodal models, 2024. URL: https://arxiv.org/abs/2312.11805.