How Prompting Shapes LLM-Generated Explanations for Recommender Systems: A Multi-Prompt Comparison Across Domains

Hlib Oshchepkov[1], Antonela Tommasel[1,2]

[1]Johannes Kepler University Linz, Austria
[2]ISISTAN, CONICET-UNCPBA, Argentina

Abstract

Large Language Models (LLMs) are increasingly used to generate natural language explanations for recommender systems, offering a flexible alternative to rigid template-based approaches. However, more fluent or engaging explanations do not necessarily imply better user understanding. In this work, we investigate how different prompting strategies shape the quality of LLM-generated recommendation explanations and the trade-offs they introduce. We compare four prompting strategies across three LLMs, three recommender families and two domains. Explanations are evaluated using an LLM-as-a-judge framework across four quality dimensions: persuasiveness, transparency, satisfaction and accuracy, alongside NLP metrics. Results show that all LLM-generated explanations are rated more positively than those by rule-based templates, but different prompting strategies optimize for different qualities. Traditional NLP metrics showed an inverse relationship with LLM-based quality judgments, highlighting their limited usefulness for this task. Overall, the findings position prompt design as a key explanation design choice rather than a purely technical implementation detail.

Keywords

recommender systems, explainability, large language models

1. Introduction

Recommender systems have become integral to modern digital platforms, guiding users through vast catalogs of content by predicting preferences and surfacing relevant items. However, presenting a recommendation without a justification leaves users unable to assess whether a suggested item genuinely matches their interests, limiting trust and engagement [1, 2]. Consequently, *explainability*—the ability to communicate why a particular item was recommended—has become a key research focus [3].

Traditional approaches rely on rule-based templates that populate predefined sentence structures with item metadata or user features. While computationally efficient, these templates are rigid and repetitive, failing to adapt to users [4]. They are especially limited for models based on latent representations, such as matrix factorization, where recommendation logic is not directly interpretable [5].

Large Language Models (LLMs) offer a promising alternative. By conditioning on user history, item metadata, and recommender outputs, they can generate flexible, natural-language explanations tailored to users and items [1, 6]. However, this flexibility introduces a new challenge, explanation quality depends not only on the available information, but also on how the model is instructed to use it. Prompting choices (such as the amount of context, the inclusion of explicit reasoning, or the communicative framing) can shape explanations in different ways, emphasizing different aspects, such as persuasiveness, transparency, or faithfulness. For instance, prompts that encourage a persuasive tone may yield more satisfying explanations, while those that enforce explicit reasoning may improve transparency. Prompting should therefore be viewed as part of the explanation *design*, particularly when fluent narratives may create a stronger sense of personalization than the underlying model provides.

In this work, we study how prompting shapes LLM-generated explanations in recommender systems. To this end, we compare four prompting strategies that vary in the provided user context, reasoning structure and communicative framing, across three recommender families and two content domains.

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden
k12338101@students.jku.at (H. Oshchepkov); antonela.tommasel@jku.at (A. Tommasel)

Explanations are evaluated using an LLM-as-a-judge framework along four dimensions (persuasiveness, transparency, satisfaction, and accuracy), and complemented with traditional NLP metrics. Our study is guided by the following research questions: • **RQ1:** Do LLM-generated explanations outperform rule-based template explanations across standard quality dimensions? • **RQ2:** How do different prompting strategies trade-off persuasiveness, transparency, satisfaction and accuracy? • **RQ3:** Do the findings generalize across generator models, including open-source alternatives to proprietary LLMs?

Our findings show that LLM-generated explanations are consistently rated more positively than rule-based templates, while different prompting strategies emphasize different explanation qualities, highlighting prompt design as a key design choice. We further show that LLM-generated explanations can obscure differences between recommenders by making even non-personalized outputs appear convincing, and that traditional NLP metrics are poorly aligned with explanation quality.

2. Related Work

Explainability in recommender systems. Explainability has long been a central concern in recommender systems. Early work by Herlocker et al. [7] showed that providing explanations improves user acceptance and trust. Since then, a broad range of approaches has been proposed, differing in the information that rely on (e.g., reviews, item features, knowledge graphs), their presentation (e.g., textual, visual), and whether explanations are intrinsic or post hoc [3]. Template-based explanations [4] remain common due to their simplicity, controllability, and low implementation cost, but are limited in their expressiveness. Fixed sentence structures cannot easily adapt to users, context, or communicative goals, and they are particularly restrictive for models with opaque reasoning (e.g., matrix factorization [8] or deep autoencoders). These limitations motivate more flexible explanation mechanisms.

LLMs for recommendation explanation. Recent work explores the use of LLMs in recommender systems, both as part of the recommendation pipeline and for generating explanations. Wu et al. [2] identify explainability as a key capability of LLMs in recommender systems, alongside understanding and conversational interactions. LLMs enable flexible, natural-language justifications that can adapt to users and context. Lubos et al. [1] show through a user study that LLM-generated explanations are generally preferred over more conventional baselines (e.g., feature-based and keyword-based templates) across multiple explanation goals. Luo et al. [5] propose a decoupled framework separating recommendation and explanation, highlighting the role of the language model. Sarkar et al. [6] further demonstrate that reasoning-oriented prompting improves generation quality. While these works establish the premise of LLM-based explanations, they largely treat explanation generation as a single objective and do not systematically examine how different prompting choices shape explanation quality across dimensions.

LLM-Based evaluation of explanations. Evaluating explanation quality remains challenging. Human studies are costly and difficult to scale, while reference-based metrics such as BLEU [9] and ROUGE [10] capture lexical overlap rather than user-perceived usefulness. Recent work explores LLMs as evaluators. Zhang et al. [11] show that strong LLMs can provide reliable multi-dimensional assessments of explanation quality and achieve substantial agreement with human judgments while being considerably more scalable and cost-efficient. More broadly, the LLM-as-a-judge paradigm has been validated across NLP tasks by Zheng et al. [12], demonstrating that strong models can approximate human preferences in comparative evaluation settings.

Prompting as explanation design. Prompt design has been shown to influence LLM outputs across a wide range of tasks. Prior work shows that encouraging intermediate reasoning improves performance on multi-step tasks [13], while assigning specific roles or styles affects the tone and content of generated text. This suggests that prompting is a meaningful design choice that shapes how information is presented. In recommender systems, this is particularly important, as explanations are evaluated not only for correctness but also for how convincing, understandable, and satisfying they are.

Despite this, studies of prompting in recommendation explanation remain limited. Existing works have not yet thoroughly examined prompting as a design space for explanation generation, nor has it explicitly analyzed the trade-offs that different prompt choices may introduce across explanation goals.

3. Study Design

Our experimental design consists of five stages[1]: (1) data selection and preparation, (2) training recommendation models, (3) generating explanations via templates and multiple LLM prompting conditions, (4) evaluating explanation quality using LLM-as-a-judge and traditional NLP metrics, and (5) comparing conditions through statistical analysis.

Dataset selection and preprocessing. We conduct experiments on two datasets from different domains[2]: *Last.fm 1K* (music, implicit feedback) and *MovieLens 100K* (movies, explicit ratings). *Last.fm 1K* contains listening histories from 992 users, treated as implicit feedback (binary interactions). After removing duplicates and applying 10-core filtering, the dataset includes 984 users, 79,239 tracks, and 2,129,372 interactions. *MovieLens 100K* contains 100,000 ratings (1–5) from 943 users on 1,682 movies, along with demographics and metadata; preprocessing leaves it unchanged. Both datasets are split into 80% training and 20% test sets with a fixed seed. All user context used for explanation generation is derived exclusively from the training split to avoid leakage.

Recommendation models. We train three recommendation models that represent fundamentally different approaches, also differing in how easily their outputs can be explained, providing a diverse basis for studying explanation generation. First, a user-based *k-Nearest Neighbors* (kNN) model, a memory-based collaborative filtering approach in which recommendations are derived from the preferences of similar users. We compute cosine similarity on the training interaction matrix and use the top-50 neighbors for ranking, a common default value in the literature. As its recommendations are grounded in the behavior of identifiable similar users, kNN is comparatively intuitive to explain. Second, Alternating Least Squares (ALS), a matrix factorization model based on latent user and item factors [8], implemented with the commonly used implicit-feedback settings ($d = 64$ latent factors, 30 iterations, regularization of 0.1). Unlike kNN, ALS recommendations are based on latent representations (without direct semantic interpretation) and are therefore less directly interpretable. Third, a non-personalized popularity baseline that ranks items by their overall interaction frequency. This setting is particularly informative as explanations cannot rely on user-specific recommendation logic.

Explanation Generation. For each dataset, we randomly sample 50 users and generate top-5 recommendations for each recommender. Each recommendation receives five explanations (one template and four LLM-based), yielding 3,750 explanations per dataset (7,500 total). User sampling allows to manage API costs while ensuring sufficient statistical power for paired comparisons across strategies; with 250 items per condition per dataset, the Wilcoxon signed-rank test achieves high sensitivity.

Template baseline. Following Lubos et al. [1], we define rule-based templates that adapt to both the recommenders and the overlap with the user's training history. When the recommended artist (Last.fm) or genre (MovieLens) matches the user's most frequent past interactions, the template explicitly references this overlap (e.g., *"...because you frequently listen to [artist]"*). Otherwise, it falls back to a generic recommender-specific justification. For example, a kNN template for a user who already listens to the recommended artist reads: *"We recommend [track] by [artist] because you frequently listen to [artist]. Users with similar listening habits also enjoyed this track."*. Templates for ALS reference latent patterns (e.g., *"Our model identified latent taste factors in your history..."*) and popularity templates note the item's broad appeal. While factually grounded, these explanations are limited in flexibility and tone.

LLM-based explanations. We generate explanations with GPT-4.1-mini (OpenAI) and two open-source models run locally via Ollama, Llama 3.1 8B (Meta) and Gemma 2 9B (Google). All models use the same prompting conditions, temperature (0.7), and maximum output length (400 tokens). Each prompt consists of a system message defining the model's role and a user message specifying the recommendation context and generation instruction. Explanations are generated once per item[3].

[1]Prompts and code can be found in: https://github.com/hcai-mms/llm-recommendation-explanations

[2]Although relatively small, these datasets are well suited to our study because they allow controlled comparison of explanation behavior, rather than emphasizing optimization of recommendation performance

[3]While repeated sampling would reduce variance from the stochastic decoding, this was not feasible given the total number of API calls (6,000 per model for generation alone).

Across the four conditions, prompts vary along three design dimensions, the amount of user context provided, whether intermediate reasoning is made explicit, and whether the explanation is framed through a particular communicative style: *Minimal.* The prompt only includes the recommended item and a one-sentence description of the recommendation method. No user history or profile is included. This strategy tests the LLM's ability to generate plausible explanations from minimal information. *Context-rich.* The prompt adds profile user information (age, gender, location or occupation), preferred artists or genres (depending on the data collection), and the most recent (training) interactions. *Chain-of-thought (CoT).* The prompt includes the same context as the context-rich strategy, but instructs the model to analyze the user's preferences before producing the final explanation. The output is structured into `REASONING:` and `EXPLANATION:` sections, with only the latter used for evaluation. This strategy draws on the finding that explicit reasoning improves LLM output quality [13]. *Persona-based.* The prompt includes the same context as the context-rich strategy, but instructs the model to write in the voice of a domain-specific critic writing for a personalized discovery newsletter. This tests whether adopting a communicative persona affects explanation quality. For all conditions, the description of the recommendation logic is adapted to the underlying recommender, referring to similar users for kNN, latent preference patterns for ALS, and overall trends for the popularity baseline. Explanations are generated as short personalized texts (2–4 sentences).

Evaluation framework. We evaluate explanations using an LLM-as-a-judge framework across four quality dimensions, complemented by traditional NLP metrics for comparison.

LLM-as-a-judge. Following Zhang et al. [11], we use an LLM-based evaluation to assess explanation quality across four dimensions (1–5 scale): *persuasiveness* (whether the explanation can convincingly motivate the recommendation), *transparency* (how well it conveys the underlying reasoning), *satisfaction* (perceived usefulness and quality of the explanation) and *accuracy* (whether the explanation is factually consistent with users' known preferences). These dimensions are grounded in established explanation goals in recommender systems [14, 4]. The evaluator LLM (GPT-4.1-mini, temperature 0.0 for deterministic output) receives the user's profile, the recommended item, and the explanation to evaluate, but not which prompting strategy generated it, to prevent the evaluator from relying on prior assumptions about prompting. Then, it returns a JSON object with scores for each dimension. We validate all outputs to ensure valid scores and retry on parsing failures.

Traditional NLP metrics. As a complement, we compute BLEU [9] (1-gram, 2-gram, and 4-gram with smoothing), ROUGE [10] (ROUGE-1, ROUGE-2, ROUGE-L F-measure), and BERTScore (with `roberta-large`) between each LLM-generated explanation and the corresponding template explanation, treating the template as the reference. BLEU and ROUGE measure surface-level n-gram overlap, while BERTScore measures semantic similarity via contextual embeddings. These metrics provide a reference-based comparison to the LLM-based quality judgments.

Statistical testing. We use the Wilcoxon signed-rank test for paired comparisons ($\alpha = 0.001$), evaluating each condition against the template baseline, context-rich against minimal (to isolate the effect of user context), and persona against context-rich (to isolate the effect of framing).

4. Results

Table 1 reports recommendation quality on the test set, evaluated on a random sample of 200 users per dataset. We report NDCG@10 and Precision@10 for ranking quality, as well as Diversity and Novelty [15] to capture the variety and popularity bias of the recommended items. These results are not intended to identify the best-performing recommender, but to contextualize explanation quality across different recommendation behaviors. In particular, the recommenders exhibited distinct performance profiles, with kNN and ALS achieving higher ranking quality, while the popularity baseline produces less personalized recommendations, which creates different conditions for explanation generation.

Table 1
Recommendation performance on Last.fm 1K and MovieLens 100K. Diversity is the average pairwise distance among top-10 items' feature vectors; Novelty is the average inverse popularity of recommended items.

	Last.fm				MovieLens			
Model	NDCG@10	Prec@10	Diversity	Novelty	NDCG@10	Prec@10	Diversity	Novelty
kNN	0.3225	0.3050	0.7652	0.1415	**0.3820**	**0.3180**	0.7206	0.0075
ALS	**0.3847**	**0.3675**	**0.7753**	**0.2665**	0.3485	0.2885	**0.7985**	**0.1060**
Popularity	0.1312	0.1270	0.7446	0.0000	0.1992	0.1810	0.7050	0.0000

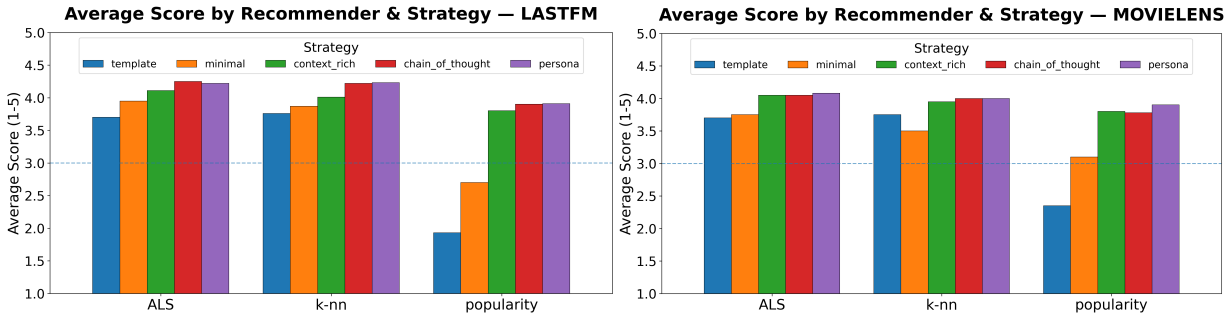

Figure 1: Average score by recommender and strategy for Last.fm (left) and MovieLens (right). The dashed line marks the midpoint (3.0).

4.1. RQ1: LLM vs. Template Explanations

Table 2 reports explanation quality across datasets, recommenders and explanations. All LLM-based explanations achieved significantly higher scores than the template baseline ($p < 0.001$) on the composite average, with one exception, the minimal prompt did not significantly improve accuracy on MovieLens.

The improvement varied by recommender. For kNN and ALS, templates already achieved moderate scores (3.71–3.76), which increased to the 4.2–4.5 range under LLM-based explanations (a gain of approximately 0.5–0.8 points). In contrast, the popularity baseline exhibited a much larger gap. Template explanations scored only 1.93 (Last.fm) and 2.34 (MovieLens), while LLM-based explanations with sufficient context reached 4.03–4.18, corresponding to improvements of over 2 points on a 5-point scale.

This difference was driven primarily by persuasiveness and transparency. Without access to user-specific reasoning, template explanations for popularity recommendations can only state that an item is popular. LLM-based explanations, by contrast, can connect items to users' known preferences, producing explanations that appear personalized even when the underlying recommendation is not.

Figure 1 illustrates these trends by recommender. Across datasets, ALS achieved comparable or slightly higher explanation scores than kNN, despite relying on latent factors without direct semantic interpretation, indicating that explanation quality is largely decoupled from algorithm interpretability. The popularity baseline showed the largest gains, with LLM-based explanations achieving scores comparable to personalized recommenders, suggesting that LLMs can compensate for the lack of personalization by constructing plausible user-aligned narratives. These patterns were consistent across domains, with stable improvements and condition rankings (differences typically <0.2 points).

4.2. RQ2: Effect of Prompting Strategy

Figure 2 shows the mean score for each strategy–dimension pair. A clear hierarchy emerges, as conditions that incorporate user context consistently achieved higher scores that those with limited or no context, with the difference between minimal and context-enriched conditions statistically significant across all dimensions ($p < 0.001$). This suggests that user context seems to be the most influential design factor. Among the context-aware conditions, differences are more nuanced and dimension-dependent.

Persuasiveness and satisfaction. More expressive prompting led to higher persuasiveness and satisfaction scores. For example, persuasiveness reached 4.62 on Last.fm and and 4.5 MovieLens for the persona prompt, followed by context-rich and chain-of-thought. Persona explanations were consistently rated as more engaging and compelling, highlighting the importance of communicative framing.

Table 2
Mean LLM-as-a-judge scores by dataset, recommender, and strategy (GPT-4.1-mini). All LLM strategies significantly outperform the template ($p < 0.001$). Highest scores are in **bold** and second highest are underlined.

		Last.fm					MovieLens				
	Strategy	Pers.	Trans.	Sat.	Acc.	Avg.	Pers.	Trans.	Sat.	Acc.	Avg.
kNN	Template	3.05	3.92	3.05	**5.00**	3.75	3.42	3.76	3.42	4.43	3.76
	Minimal	3.99	3.80	3.99	4.89	4.17	3.75	3.27	3.75	4.26	3.76
	Context-rich	4.38	3.73	4.38	4.52	4.25	4.27	3.75	4.28	4.59	4.22
	Chain-of-thought	4.35	**3.95**	4.35	4.81	4.36	4.10	3.76	4.10	**4.63**	4.15
	Persona	**4.72**	3.74	**4.72**	4.60	**4.45**	**4.56**	3.60	**4.58**	4.61	**4.34**
ALS	Template	3.28	3.31	3.28	**4.98**	3.71	3.33	3.87	3.34	4.32	3.71
	Minimal	2.96	3.71	2.96	4.87	3.62	3.38	3.67	3.39	4.29	3.68
	Context-rich	4.55	3.91	4.55	4.60	4.41	4.44	3.81	4.44	4.53	4.30
	Chain-of-thought	4.40	**3.98**	4.40	4.87	4.42	4.18	3.81	4.20	**4.64**	4.21
	Persona	**4.78**	3.80	**4.78**	4.68	**4.51**	**4.64**	3.66	**4.64**	4.61	**4.39**
Pop.	Template	1.97	1.01	1.97	2.76	1.93	2.08	1.92	2.08	3.27	2.34
	Minimal	2.79	1.96	2.79	2.59	2.53	3.24	2.27	3.21	3.49	3.05
	Context-rich	4.03	3.52	4.05	4.59	4.05	4.06	3.51	4.09	4.46	4.03
	Chain-of-thought	4.08	3.71	4.08	4.75	4.16	4.01	**3.55**	4.03	**4.53**	4.03
	Persona	**4.36**	3.40	**4.36**	4.62	**4.18**	**4.37**	3.45	**4.38**	**4.53**	**4.18**

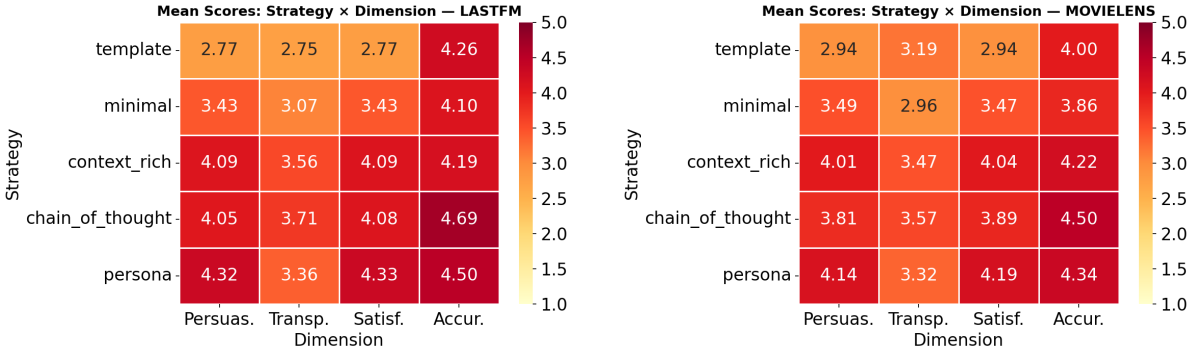

Figure 2: Mean LLM-as-a-judge scores by strategy and dimension for Last.fm (left) and MovieLens (right).

Transparency. Prompting encouraging explicit reasoning achieved the highest transparency scores (3.88 on Last.fm, 3.71 on MovieLens), consistent with the expectation that making intermediate reasoning steps could improve understanding of the recommendation logic. Nonetheless, transparency remained the lowest-scoring dimension overall.

Accuracy. Reasoning-oriented prompting yielded the highest accuracy (4.81 Last.fm, 4.60 MovieLens), suggesting that intermediate reasoning helps ground explanations in user data. More expressive prompting showed slightly lower accuracy, suggesting a trade-off between creativity and grounding.

Communicative framing vs. reasoning. Comparing conditions that share the same input context but differ in framing isolates the effect of prompt design. More expressive framing increased persuasiveness (+0.30 on Last.fm, +0.27 on MovieLens; $p < 0.001$) and satisfaction (+0.29, +0.26; $p < 0.001$), but slightly reduced transparency (−0.07 on Last.fm, $p < 0.001$; −0.12 on MovieLens, $p < 0.001$), while accuracy differences remained small ($p = 0.031$ and $p = 0.018$, for Last.fm and MovieLens, respectively). This highlights a trade-off between engagement and interpretability.

Relationship to NLP metrics. Table 3 reports BLEU, ROUGE and BERTScore using the template explanations as reference. An inverse pattern emerges, minimal prompting achieved the highest scores, while more expressive conditions scored lower, despite being rated higher by the LLM-as-a-judge. For example, Last.fm, minimal achieves ROUGE-1 of 0.419 versus 0.259 for persona. BERTScore shows the same ranking but with a narrower spread (F1 = 0.902 vs 0.869). This reflects that reference-based metrics reward lexical similarity, whereas higher-quality explanations tend to be more diverse. These

Table 3
NLP metrics and BERTScore by strategy, averaged across recommenders (GPT-4.1-mini). Higher values indicate greater similarity to the template. Minimal scores highest despite ranking lowest on LLM-as-a-judge quality.

	Last.fm					MovieLens				
Strategy	BLEU-1	BLEU-4	R-1	R-L	BERT-F1	BLEU-1	BLEU-4	R-1	R-L	BERT-F1
Minimal	0.362	0.157	0.419	0.299	0.902	0.292	0.086	0.370	0.227	0.878
Context-rich	0.241	0.100	0.310	0.199	0.879	0.269	0.047	0.273	0.151	0.859
Chain-of-thought	0.279	0.125	0.322	0.233	0.883	0.296	0.062	0.314	0.195	0.866
Persona	0.208	0.080	0.259	0.173	0.869	0.229	0.036	0.234	0.138	0.848

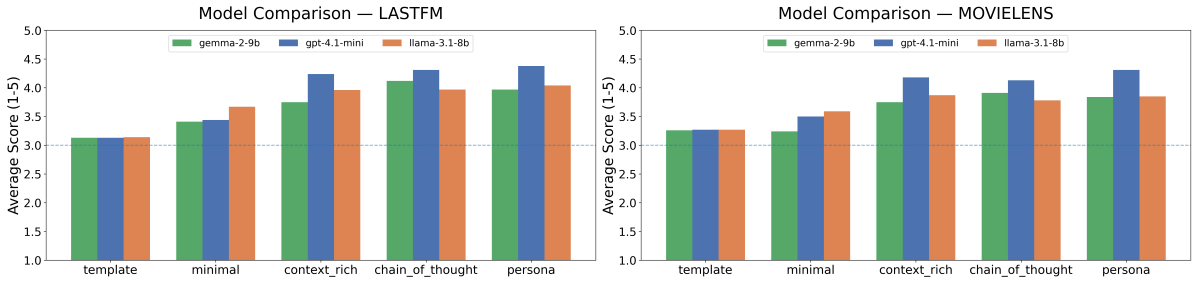

Figure 3: Average LLM-as-a-judge score by generator model and strategy for Last.fm (left) and MovieLens (right). GPT-4.1-mini leads on context-aware strategies, but all models outperform templates.

results reinforce that traditional NLP metrics are poorly aligned with explanation quality, thus resulting of limited relevance for evaluating recommendation explanations [11].

4.3. RQ3: Cross-Model Comparison

To assess whether the findings depend on the choice of LLM, we repeated the explanation generation pipeline using two open-source models. All explanations were evaluated by the same GPT-4.1-mini judge using the same prompt. Figure 3 shows average scores by model and prompting conditions.

GPT-4.1-mini consistently achieved the highest scores across context-aware conditions ($p < 0.001$ in all pairwise comparisons). On Last.fm, average scores ranged from 4.24 (context-rich) to 4.38 (persona), compared to 3.96–4.03 for LLaMa3.1 8b and 3.75–4.12 for Gemma2 9B, with similar gaps observed on MovieLens. Notably, LLaMa3.1 8b achieved slightly higher scores than GPT-4.1-mini in the minimal prompt (+0.23 on Last.fm, $p < 0.001$), suggesting that smaller models could match or exceed proprietary models when little context is provided. Despite lower absolute scores, both open-source models consistently scored higher than the template explanations (above 3.7) across context-aware conditions, suggesting that the benefits of LLM-based explanations generalize across model families. Across LLMs, explanations exhibited high semantic similarity (high BERTScore) but low lexical overlap (low BLEU), suggesting that models convey similar content using different wording. Gemma2 9B was slightly closer to GPT-4.1-mini than LLaMa3.1 8b across these metrics. Importantly, the ranking of prompting conditions was preserved across all models, more expressive and reasoning-oriented conditions achieved higher scores, followed by context-enriched and minimal. This consistency suggests that the relative effects of prompting might be largely model-agnostic.

5. Discussion

Prompt design as an explanation design choice. Results showed that prompt design is not just a technical detail, but a first-order design choice that shapes explanation quality along distinct dimensions. More engaging prompts yielded higher persuasiveness and satisfaction, while prompts that encourage explicit reasoning improved accuracy and transparency. The context-enriched prompt represents the most direct implementation of prior LLM-based explanation approaches [1], performing well across all dimensions. However, the fact that alternative prompting conditions achieved higher scores on specific dimensions indicates that providing more context is necessary, but not sufficient, as how the model is instructed to use the context is also important. This suggest that prompting should be selected based on application goals. For example, systems prioritizing user engagement may favor more expressive framing, whereas systems requiring trust and verifiability may benefit from more structured reasoning.

The transparency gap. Across all conditions, recommenders and datasets, transparency consistently received the lowest scores. This matters because transparency is arguably one of the most critical dimension for explainability as it relates directly to users' understanding of recommendation logic. LLMs appeared to be substantially better at producing persuasive and engaging explanations that at faithfully conveying underlying reasoning. This aligns with Zhang et al. [11], who report lower reliability for transparency judgments. One possible explanation is structural, prompts describe recommenders at a high level (e.g., latent factors), providing limited concrete reasoning for the model to expose.

LLMs as equalizers across recommender types. A key practical finding is that LLM-generated explanations largely equalize perceived quality across fundamentally different recommenders. While template explanations exposed the limitations of non-personalized recommendations, LLM-based explanations recovered high scores even for popularity-based recommendations, reaching levels comparable to personalized models. Similarly, ALS and kNN achieved comparable explanation quality despite differing substantially in interpretability. This suggests that explanation quality is driven more by the available user context than by the internal structure of the recommender. As a result, designers may decouple recommendation accuracy and explainability, using post-hoc LLM explanations to provide high-quality user-facing justifications even for simple or opaque models.

Inadequacy of traditional NLP metrics. The inverse relationship between BLEU/ROUGE scores and LLM-as-a-judge evaluations highlights the limitations of reference-based metrics. Prompts that produced text closer to the template phrasing (e.g., minimal) achieved higher n-gram overlap but lower perceived quality, while more diverse and informative explanations were penalized. BERTScore partially mitigated this effect, but exhibited the same overall trend, suggesting that similarity to a reference (even at embedding level) does not capture explanation quality. These findings reinforce prior work [11] and suggest that evaluation methods should focus on user-perceived qualities rather than textual similarity.

Model size and explanation quality. The cross-model comparison showed that GPT-4.1-mini consistently achieved higher scores than open-source models on context-aware conditions, likely reflecting stronger instruction following capabilities. However, all models preserved the same ranking of prompting conditions and obtained higher scores than the template-based explanations, hinting that the observed effects are not specific to a particular model. This suggests that while model choice influences absolute performance, the relative impact of prompting conditions is robust. For practical deployments, open-source models offer a viable alternative when cost or privacy concerns are important.

6. Conclusions

We presented an evaluation of LLM-generated explanations for recommender systems, comparing multiple prompting conditions across three recommenders, two domains, and three LLMs. LLM-generated explanations consistently achieved higher scores than rule-based templates, with particularly large gains for non-personalized recommendations. Prompting shaped explanation quality along different dimensions, introducing trade-offs between persuasiveness, satisfaction, transparency, and accuracy, with user context as the primary driver.

Several limitations of this study suggest directions for future work. First, we relied on GPT-4.1-mini as the evaluator for all models, introducing potential same-family bias; incorporating alternative evaluators, open-source models or human studies, would strengthen robustness. Second, explanations were generated once per item under stochastic decoding; multiple generations per condition would enable more robust estimates and analysis of variability. Third, the analysis was based on a sample of 50 users per dataset; expanding to larger and more diverse populations would improve generalizability. Fourth, explanations were evaluated in isolation rather than in an interactive system; future work should assess how explanation quality translates to user behavior such as trust and engagement. Fifth, demographic features were included in prompts but not analyzed separately; their contribution to personalization and bias amplification warrants further study. Sixth, comparisons with BLEU and ROUGE relied on template explanations as references, favoring lexically similar outputs; lower overlap thus reflects stylistic divergence rather than lower quality. Alternative evaluation approaches, including human studies and reference-free metrics, should be explored. Finally, our results point to opportunities

for improving explanation generation. Hybrid prompting strategies that combine structured reasoning with more engaging framing may better balance competing dimensions, while incorporating more granular signals (e.g., similar users or latent factors) could help address the transparency gap.

Ethical considerations. Our findings raise important ethical considerations regarding the use of LLM-generated explanations. In particular, the ability of LLMs to produce highly persuasive explanations even for non-personalized or weakly grounded recommendations highlights the risk of *overpersuasion*, where users may be convinced by explanations that do not faithfully reflect the underlying recommendation logic. This may lead to misplaced trust in the system or reduced user autonomy. Overall, while LLMs can improve the perceived quality of explanations, careful design and evaluation are needed to ensure they remain truthful, fair, and aligned with user interests.

Acknowledgments This research was funded in whole or in part by the Austrian Science Fund (FWF): 1255776/COE12.

Declaration on Generative AI During the preparation of this work, the authors used generative AI tools (Claude, Anthropic) to assist with code development, experimental pipeline design, data analysis, and aspects of manuscript drafting. All generated content was reviewed and revised by the authors, who take full responsibility for the final content of this publication.

References

[1] S. Lubos, T. N. T. Tran, A. Felfernig, S. P. Erdeniz, V.-M. Le, LLM-generated explanations for recommender systems, in: Adjunct Proceedings of the 32nd ACM UMAP, 2024, pp. 276–285.

[2] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, E. Chen, A survey on large language models for recommendation, arXiv:2305.19860 (2024).

[3] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Foundations and Trends in Information Retrieval 14 (2020) 1–101.

[4] F. Gedikli, D. Jannach, M. Ge, How should i explain? a comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies 72 (2014) 367–382.

[5] Y. Luo, Q. Liu, Z. Chen, LLMXRec: A two-stage decoupled framework for leveraging LLMs as explainers in recommender systems, in: Proceedings of the 29th DASFAA, 2024.

[6] B. Sarkar, V. Venkataramanan, A. Anand, ReasoningRec: Multi-step reasoning for explainable recommendation, in: Findings of the Association for Computational Linguistics: NAACL, 2025.

[7] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on CSCW, 2000, pp. 241–250.

[8] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), 2008, pp. 263–272.

[9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the ACL, 2002, pp. 311–318.

[10] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[11] X. Zhang, L. Chen, et al., Large language models as evaluators for recommendation explanations, in: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys), 2024.

[12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, in: Advances in NeurIPS, 2023.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in NeurIPS, 2022.

[14] N. Tintarev, J. Masthoff, Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization, User Modeling and User-Adapted Interaction 22 (2012) 399–439.

[15] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM TIIS 7 (2016).