Can Large Language Models Understand Preferences in Personalized Recommendation?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in various tasks, including personalized recommendations. Existing evaluation methods often focus on rating prediction, relying on regression errors between actual and predicted ratings. However, user rating bias and item quality, two influential factors behind rating scores, can obscure personal preferences in user-item pair data. To address this, we introduce PERRECBENCH, disassociating the evaluation from these two factors and assessing recommendation techniques on capturing the personal preferences in a grouped ranking manner. We find that the LLM-based recommendation techniques that are generally good at rating prediction fail to identify users' favored and disfavored items when the user rating bias and item quality are eliminated by grouping users. With PERRECBENCH and 19 LLMs, we find that while larger models generally outperform smaller ones, they still struggle with personalized recommendation. Our findings reveal the superiority of pairwise and listwise ranking approaches over pointwise ranking, PERRECBENCH's low correlation with traditional regression metrics, the importance of user profiles, and the role of pretraining data distributions. We further explore three supervised fine-tuning strategies, finding that merging weights from single-task training is promising but improving LLMs' understanding of user preferences remains an open research problem.

1 Introduction

Personalization tailors system interactions, content, or recommendations to individual users by analyzing their behavior, preferences, and characteristics (Tan and Jiang, 2023; Zhang et al., 2024). It is critical in domains such as content recommendation (Qian et al., 2013; Baek et al., 2023), user simulation (Dejescu et al., 2023), personalized chatbots (Srivastava et al., 2020), user profiling (Gu et al., 2020; Gao et al., 2023), healthcare (Goldenberg et al., 2021), and education (Pratama et al., 2023). Large Language Models (LLMs) excel in diverse natural language tasks, showcasing emergent abilities (Wei et al., 2022; Lu et al., 2023). To align LLM outputs with individual user preferences, personalization has become a key research focus, necessitating benchmarks for evaluation (Li et al., 2024; Sun, 2023, 2024). Personalized recommendation, with its abundant user behavior data and preference signals, is widely adopted as a proxy for assessing LLM personalization (Kang et al., 2023).

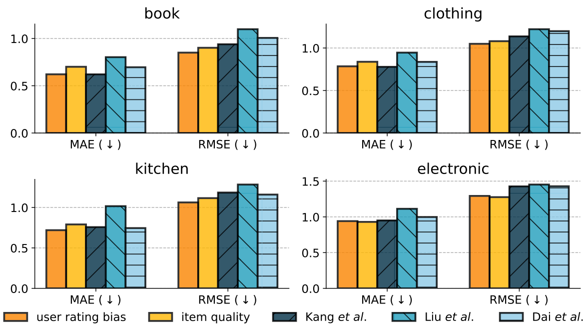

Figure 1: MAE and RMSE performance of user rating bias (average user rating history), item quality (average item rating), and existing LLM-based personalization methods. Simple averages of user rating history and item quality, which do not consider individual preferences, achieve state-of-the-art performance across four shopping domains, questioning the validity of MAE and RMSE for evaluating *personalization*.

Personalized recommendation evaluation can be typically categorized into rating-based and ranking-based paradigms. In rating-based evaluation, models predict a user's rating for an item and calculate regression errors such as MAE and RMSE against actual ratings. However, user rating bias and query item quality are two influential factors behind the rating scores from a user to an item, which might prevent personal preferences in user-item pairs data from being learned. We hypothesize that naive methods that average user rating history (user rating bias) or averaged item rating (item quality) can achieve competitive MAE and RMSE scores. To validate this, we sampled 1,000 user behaviors from the Amazon review dataset (Hou et al., 2024) across books, clothing, kitchen, and electronics domains and compared the performance with existing LLM-based personalized recommendation methods (Kang et al., 2023; Liu et al., 2023a; Dai et al., 2023). Results in Figure 1 show that relying solely on statistics like user bias and item quality can achieve strong regression results without incorporating personalized preferences, a pervasive issue in recommendation evaluation. The ranking-based recommendation involves predicting a user's top-k favorite items based on the user's history, considering a recommendation successful if the predicted item is reviewed and rated highly. However, this approach relies on incomplete signals, as it samples distractors from unreviewed items. These distractors are not inherently poor recommendations, as their exposure to users remains unknown. Effective personalization evaluation should focus on observed signals, distinguishing between low-rated and high-rated preference signals from the user.

To isolate personalization in recommendation evaluation from user rating bias and item quality, we introduce PerRecBench, a benchmark that assesses personalization based on observed user preferences in a grouped ranking framework. Specifically, models rank users within a group by their preferences for a shared query item. To eliminate user rating bias, we define relative rating as the actual rating minus the user's average rating, where a positive relative rating indicates the user prefers the item over other purchased items. To control for item quality, we group users who purchased the same item within a short timeframe, ensuring consistent item quality within each group. Ground truth rankings are derived by ordering users in each group based on their relative ratings for the shared item. PerRecBench evaluates model performance using pointwise, pairwise, and listwise ranking methods to rank users and measure correlations with ground truth rankings. While input is identical across users in a group, outputs are expected to reflect personalized preferences based on individual profiles and histories. By focusing on observed signals and controlling variables on user rating bias and item quality, PerRecBench ensures reliable assessment of personalization.

Using data from Amazon review (Hou et al., 2024), we constructed PerRecBench with 600 user groups, including 200 groups each with 2, 3, and 4 users to represent increasing levels of difficulty. Benchmarking 19 off-the-shelf LLMs revealed generally unsatisfactory performance, with open-source models exceeding 100B parameters approaching the performance of proprietary models. Among these, Claude-3.5-Sonnet performs best overall. While larger LLMs generally outperformed smaller ones, scaling laws did not consistently hold, as increased model size did not always translate to better performance. Moreover, the low correlation between PerRecBench results and MAE/RMSE confirms that personalization is distinct from traditional rating regression tasks. Further analysis highlights the importance of textual user profiles, domain relevance, and shot/retrieval k settings on model performance.

We also investigate three supervised fine-tuning (SFT) strategies to enhance personalization: *single-task training*, *joint training*, and *weight merging*. Single-task training improves task performance and cross-task generalization, while weight merging achieves the best results on PerRecBench. However, developing LLMs with robust personalization capabilities remains an open challenge.

In summary, our contributions include introducing PerRecBench, the first recommendation benchmark specifically designed to evaluate personalization by removing user rating bias and item quality through observed preference signals, and exploring initial strategies to tackle challenges in LLM-based personalized recommendations.

2 PerRecBench

To assess whether LLMs can capture users' personalized preferences rather than relying on rating bias or item quality, we introduce the PerRecBench Benchmark (Figure 2). We first select user groups from diverse shopping domains with varying sizes (§2.1). Next, we evaluate personalization by ranking users by their preferences towards query item using LLM-based ranking methods, including pointwise, pairwise, and listwise approaches (§2.2). Finally, we define evaluation metrics tailored to PerRecBench (§2.3).

2.1 User Group Selection

Let U be the set of all users, and let $\mathcal{H}_u = \{(x_u^t, y_u^t)\}$ denote the historical behavior of user $u \in \mathcal{U}$, where x_u^t is the item purchased at timestamp t, and y_u^t is the corresponding rating. The

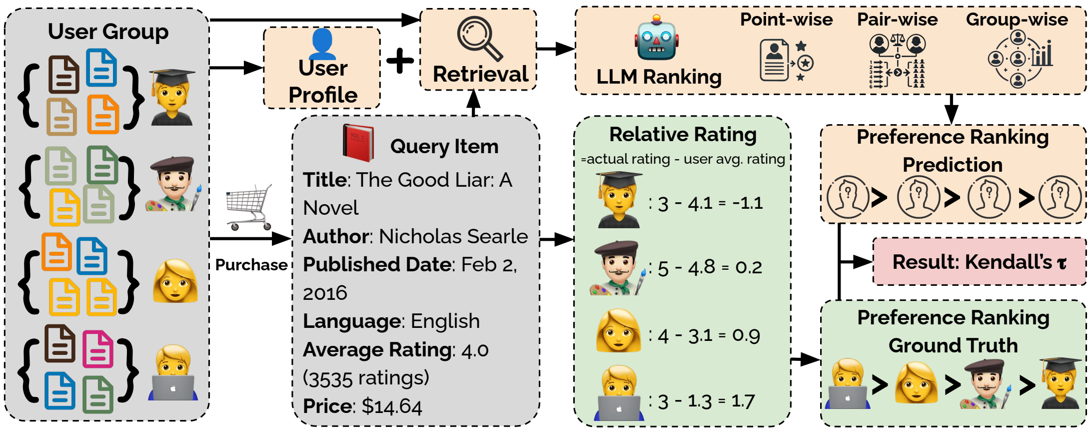

Figure 2: Overview of PERRECBENCH, where the LLM ranks user preferences for a query item using pointwise, pairwise, and listwise prompting. The ground-truth ranking is derived from relative ratings, calculated as the user's actual rating minus their average rating, to mitigate user rating bias. Finally, Kendall's tau is computed between the predicted ranking and the ground-truth ranking to evaluate performance.

goal of user group selection is to select a query item q and the corresponding user subset $\mathcal{U}^* \subseteq \mathcal{U}$ that meet the following criteria:

Temporal Item Co-Purchase. All users in the group must have purchased the query item q within a specific time interval $[t_0, t_0 + \Delta t]$, ensuring consistent item quality. Thus, users share the same query item q but have distinct histories. Formally:

$$\forall u \in \mathcal{U}^* : \exists t \in [t_0, t_0 + \Delta t],\ x_u^t = q.$$

Active Users. Each user must have a sufficient rating history to enable effective personalization. A user is considered active if their history prior to purchasing q exceeds a threshold γ:

$$\forall u \in \mathcal{U}^* : \left|\{(x_u^{t'}, y_u^{t'}) \in \mathcal{H}_u, t' < t(q_u)\}\right| > \gamma,$$

where $t(q_u)$ denotes the timestamp that user u purchased query item q, we set the active level threshold γ to 20 by default.

Significant Relative Rating Difference. To capture meaningful preference signals, users within the group must demonstrate distinct preferences for the query item q. We address user rating bias by introducing relative ratings. Let $y(q_u)$ denote user u's rating for query item q, and define the relative rating as $\tilde{y}(q_u) = y(q_u) - \bar{y}_u$, where $\bar{y}_u$ represents user u's average rating, eliminating the user rating bias. A positive relative rating indicates that the user prefers the query item compared to their average rating, while a negative value suggests the opposite. To ensure distinguishable preferences within the user group, the relative rating difference between any two users must exceed a threshold λ, set to 0.6 by default:

$$\forall u, v \in \mathcal{U}^*, u \neq v : \ |\tilde{y}(q_u) - \tilde{y}(q_v)| > \lambda.$$

Using these criteria, we constructed PERRECBENCH, comprising 663 user groups with five users per group across book, clothing, kitchen, and electronic domains. To create a progressive testbed, we randomly down-sampled 200 user groups with sizes of 2, 3, and 4, representing ascending levels of difficulty.

2.2 LLM-based Ranking Methods

PERRECBENCH evaluates LLMs' ability to rank users' preferences for a shared query item. Formally, the task requires the LLM to predict a ranking r for $u \in \mathcal{U}^*$ and compare it to the ground truth ranking r^* based on user preferences. To comprehensively assess LLMs' personalization capabilities, we adopt multiple ranking approaches, including pointwise rating prediction, pairwise ranking, and listwise ranking, to evaluate their effectiveness in modeling user preferences.

Pointwise Rating Prediction Given a single user u with rating history $\mathcal{H}_u$, we feed the top-k most relevant user's behavior history *w.r.t.* query item q and user profile p_u to the LLM, asking it to predict user's rating s_u for the query item q.

$$s_u = \text{LLM}(\phi_{pt}(q, \mathcal{D}_u^q, p_u)),$$

where ϕ_{pt} is the pointwise rating prediction prompt template, and $p_u = \text{LLM}(\mathcal{H}_u)$ denotes the textual user profile generated by an instruction-tuned LLM.

The retrieved user history $\mathcal{D}_u^q = \mathcal{R}(q_u, \mathcal{H}_u^{<t(q_u)}, k)$ represents the top-k relevant user history prior to the timestamp of q_u, with $\mathcal{R}$ as the retriever. Using the predicted ratings for all users in the group, we compute each user's predicted relative rating $\tilde{s}_u = s_u - \bar{y}_u$, where $\bar{y}_u$ is the user's average rating. Users are then ranked based on their predicted relative rating $r = \text{argsort}(\{\tilde{s}_u, u \in \mathcal{U}^*\})$.

Pairwise Ranking As shown in prior research (Qin et al., 2024; Sun et al., 2023), LLMs can effectively perform text ranking through pairwise comparisons. Similarly, we use pairwise ranking paradigm to rank users based on their preferences on the query item. In pairwise ranking, the fundamental unit is the comparison of user preferences for the same query item. The pairwise comparison function f between user u_i and u_j is:

$$f(u_i, u_j) = \text{LLM}[\phi_{pr}(q, (\mathcal{D}_{u_i}^q, p_{u_i}), (\mathcal{D}_{u_j}^q, p_{u_j}))],$$

where ϕ_{pr} is the prompt template for pairwise user preference comparison. The LLM outputs which user has a stronger preference for the query item. To mitigate the position bias in LLM judgment (Ye et al., 2024; Lu et al., 2022), for each pair of users, we swap the position of user u_i and u_j and only consider preferences differences if the judgments are consistent across both orderings. Using the pairwise comparison function, we rank users with heapsort, which ensures $O(N \log N)$ computational complexity and has been shown effective in LLM-based text ranking (Qin et al., 2024). This process yields the final pairwise ranking r.

Listwise Ranking Previous research (Sun et al., 2023; Ma et al., 2023) has shown that LLMs are effective at listwise text ranking, where they rank the relevance of multiple documents to a query in a single prompt. Similarly, LLMs can rank a group of users within a single prompt input, where each user u is represented by their retrieved rating history $\mathcal{D}_u^q$ and the corresponding user profile p_u. The ranking r is defined as:

$$r = \text{LLM}(\phi_{gp}(q, \{(\mathcal{D}_u^q, p_u), u \in \mathcal{U}^*\})),$$

where ϕ_{gp} is the prompt construction function for listwise ranking.

2.3 Evaluation Metric

To evaluate LLMs' personalization capabilities, we measure the correlation between the predicted user preference ranking and the ground truth ranking. The ground truth ranking r^* is derived from the relative ratings within the selected user group:

$$r^* = \text{argsort}(\{\tilde{y}(q_u), u \in \mathcal{U}^*\}),$$

where $\tilde{y}(q_u)$ is the ground truth relative rating. The evaluation metric, *Personalization Tau Correlation (PTC)*, is computed as Kendall's tau correlation between the predicted ranking r and the ground truth ranking r^* :

$$\text{PTC} = \textit{Kendall-tau}(r, r^*).$$

Overall, we define relative ratings to capture users' preferences for a query item while eliminating user rating bias. All users within a group are ranked based on the same query item, ensuring consistent item quality. User preference signals are clearly observed through their reviewed ratings, derived from differences in relative ratings across users. By using a single query item, PerRecBench expects personalized outputs tailored to each user's history and profile. This evaluation paradigm is specifically designed to assess personalization capabilities, making the personalization signal easy to interpret.

3 Experimental Settings

We evaluate the personalization capabilities of 19 off-the-shelf LLMs, including open-source models: Llama-3.1-8B-it, Llama-3.1-70B-instruct, Meta-Llama-3.1-405B-Instruct (Dubey et al., 2024), Gemma-2-9B-it, Gemma-2-27B-it (Team et al., 2024), Ministral-8B-Instruct-2410, Mistral-Nemo-Instruct-2407, Mixtral-8x22B Instruct v0.1 (Jiang et al., 2024), Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct, Qwen2.5-Coder-32B-Instruct (Qwen et al., 2024), DeepSeek-v3 (Liu et al., 2024a), and proprietary models: Claude-3.5-haiku, Claude-3.5-sonnet, GPT-4o-mini, and GPT-4o (Hurst et al., 2024). For a fair comparison, all models were tested with a temperature of 0.1 using zero-shot prompting by default. For LLM fine-tuning, we applied LoRA (Hu et al., 2021) for efficient fine-tuning with rank 16, training for 2 epochs with a batch size of 32 and a learning rate of 1×10^{-5}. We use BM25 (Trotman et al., 2014) retriever and the number of retrieved history items k was set to 4 by default, and performance is reported under zero-shot settings without further notice.

Table 1: Main results on PerRecBench. Scores range from -1 to 1, with higher values indicating better performance. The best results across different LLMs are highlighted in **bold**, and the second-best results are underlined.

Model	easy			medium			hard			Avg.
	pointwise	pairwise	listwise	pointwise	pairwise	listwise	pointwise	pairwise	listwise	
Open-source LLMs										
Llama3.1-8B-it	-0.27	0.25	0.30	0.03	-0.01	-0.01	-0.03	0.07	0.05	0.04
Gemma-2-9B-it	0.13	0.23	0.25	0.00	0.10	-0.08	0.03	0.05	0.09	0.09
Qwen2.5-7B-it	-0.10	0.17	0.16	0.06	0.05	-0.05	0.02	0.08	0.02	0.05
Ministral-8B-it	-0.04	0.02	0.09	0.00	0.06	-0.01	0.05	0.02	0.01	0.02
Mistral-12B-Nemo-it	-0.11	0.14	0.37	0.00	0.07	0.01	0.05	0.02	-0.10	0.05
Qwen2.5-14B-it	0.09	0.23	0.17	0.05	0.10	0.03	0.08	0.09	0.02	0.10
Gemma-2-27B-it	0.21	0.15	0.15	0.02	0.03	-0.01	0.06	0.07	0.02	0.08
Mixtral-8x22B-it	-0.02	0.25	0.35	0.05	0.11	0.03	0.05	0.09	0.07	0.11
Qwen2.5-32B-it	0.21	0.26	0.35	-0.01	0.09	0.02	0.08	0.07	0.07	0.13
Qwen2.5-Coder-32B-it	0.04	0.26	0.12	0.01	0.06	0.02	0.03	0.06	0.03	0.07
Qwen2.5-72B-it	0.09	0.27	0.24	0.04	0.10	0.06	0.04	0.09	0.03	0.11
Llama-3.1-70B-it	0.09	0.31	0.32	0.00	0.09	0.13	0.09	0.12	0.10	0.14
Mistral-Large-123B-it	0.19	0.31	0.33	0.09	0.09	0.06	0.09	0.08	0.07	0.15
Llama-3.1-405B-it	0.14	0.31	0.38	0.08	0.09	0.05	0.05	0.08	0.11	0.14
DeepSeek-v3-671B	0.23	0.21	0.23	0.13	0.10	0.06	0.14	0.06	0.06	0.14
Proprietary LLMs										
Claude-3.5-Haiku	0.21	0.27	0.32	0.09	0.08	0.10	0.09	0.04	0.02	0.14
Claude-3.5-Sonnet	0.27	0.34	0.31	0.11	0.07	0.14	0.13	0.10	0.12	0.18
GPT-4o-mini	0.20	0.31	0.41	0.13	0.04	0.05	0.07	0.05	0.04	0.15
GPT-4o	0.25	0.29	0.27	0.10	0.07	0.11	0.08	0.07	0.06	0.15

4 Results

Table 1 shows the performance of 19 off-the-shelf LLMs on PerRecBench. We have the observations as follows.

LLMs struggle with personalized recommendation. Across 19 strong LLMs, performance on PerRecBench ranges from 0.02 to 0.18, within Kendall's tau value range of $[-1, 1]$. This indicates a low to moderate correlation between predictions and ground truth rankings. Pointwise, pairwise, and listwise ranking methods all demonstrate limited success, with average Kendall's tau scores ranging from -0.27 to 0.38 across models and methods. Even the best-performing model, Claude-3.5-Sonnet, achieves only 0.18 on average across different group sizes and ranking methods. These results highlight the limited personalized preference understanding ability of current LLMs, underscoring that LLM personalized recommendation remains an open research question.

Scaling law does not always hold for personalization. While the scaling law suggests larger models generally perform better on tasks (Kaplan et al., 2020), our results show that increasing model size does not consistently improve personalization performance. For example, in the Qwen model series, the 7B, 14B, and 32B models perform as expected with scores of 0.05, 0.11, and 0.13, respectively. However, the 72B model performs worse than the 32B model and similarly to the 14B model. Similarly, in the Gemma series, the 27B model performs close to the 9B model. These results challenge the assumption that larger models inherently enhance personalization capabilities.

Pairwise and listwise ranking outperform pointwise. Across all user group sizes and models, the average Kendall's tau scores for pointwise, pairwise, and listwise ranking are 0.19, 0.38, and 0.35, respectively. Pairwise and listwise ranking methods significantly outperform pointwise ranking. We attribute this to the limitations of pointwise ranking, where the model evaluates a single user in isolation, making it difficult to discern subtle preference differences. In contrast, pairwise and listwise methods allow the model to leverage comparative reasoning, capturing nuanced differences by analyzing multiple users within a single prompt.

Strong open-source LLMs rival proprietary models. Open-source models demonstrate competitive performance compared to proprietary counterparts on PerRecBench. For instance, Mistral-Large-123B-it and Llama-3.1-405B-it achieve average Kendall's tau scores of 0.15, slightly outperforming Claude-3.5-Haiku and approaching the performance of the GPT-4o

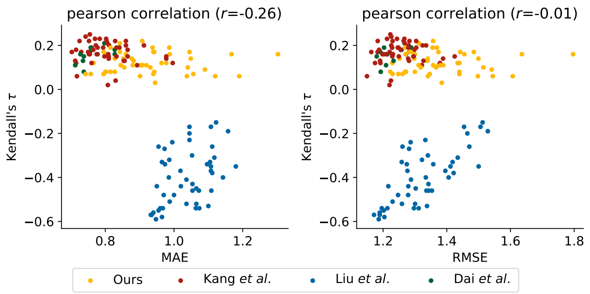

Figure 3: Correlation between Kendall's tau in PERRECBENCH and traditional metrics (MAE and RMSE). The results show weak correlation, further confirming that MAE and RMSE are not reliable indicators of personalization capabilities.

Table 2: Performance of GPT-4O-MINI on PERRECBENCH with different prompting methods, where the best performance across prompting method is in **bold**, the second best is underlined. Incorporating user profiles significantly enhances personalization capabilities, whereas few-shot, self-consistency, and chain-of-thought prompting does not consistently improve performance and may even degrade it.

Prompting	**easy**			**medium**			**hard**			**Avg.**
	pt	pr	ls	pt	pr	ls	pt	pr	ls	
ZERO-SHOT	.20	.31	.41	.13	.04	.05	.07	.05	.04	.15
ZERO-SHOT W/O PROFILE	.14	.24	.27	.11	.06	.02	.05	.08	.05	.11
FEW-SHOT	.15	.31	.37	.13	.07	.03	.09	-.02	.06	.13
FEW-SHOT W/O PROFILE	.16	.20	.29	.10	.07	.03	.08	.03	.08	.11
SELF-CONSISTENCY	.28	.33	.35	.11	.05	.04	.07	.05	.03	.15
CHAIN-OF-THOUGHT	.09	.26	.31	.08	.08	.08	.08	.05	.01	.12

family and CLAUDE-3.5-SONNET. These results suggest that with proper optimization, open-source models can be viable alternatives for personalization tasks, offering performance close to or on par with commercial models.

5 Analysis

Tau's Correlation with MAE and RMSE To validate that PERRECBENCH successfully isolates personalization capabilities in LLM recommendations, while traditional metrics like MAE and RMSE do not, we analyze their correlation with PERRECBENCH performance. Specifically, we adopted prompt templates from Kang et al. (2023); Liu et al. (2023a); Dai et al. (2023) and used the default prompting in PERRECBENCH. Additionally, we generated four prompt variants using GPT-4O based on the original prompts. We varied the number of retrieved history items k in $\{2, 4, 8\}$ and the number of shots in $\{0, 1, 2, 3\}$. For each configuration, we computed MAE, RMSE, and PERRECBENCH performance, setting the decoding temperature to 0 to eliminate randomness. The correlations between Kendall's tau, MAE, and RMSE are visualized in Figure 3, alongside the corresponding Pearson correlation coefficients. The results show that both MAE and RMSE have weak correlations with Kendall's tau in PERRECBENCH. Notably, while Liu et al. (2023a) demonstrates moderate performance on MAE and RMSE, its performance on PERRECBENCH consistently falls below random guessing. This indicates that traditional rating prediction metrics like MAE and RMSE are poor indicators of personalization capabilities.

Comparing Prompting Methods Previous studies (Richardson et al., 2023; Tan et al., 2024b) have shown that incorporating textual user profiles into prompts can enhance model performance. However, conflicting conclusions exist regarding whether few-shot prompting improves LLM personalization capabilities (Kang et al., 2023; Zhiyuli et al., 2023). To address this, we evaluate personalization performance under different prompting methods using GPT-4O-MINI, with results presented in Table 2. Including user profiles p_u in prompts leads to an average performance improvement of 28%. Few-shot prompting, while beneficial for pointwise and listwise methods, reduces performance in pairwise prompting, resulting in relatively stable overall performance. Combining user profiles with few-shot prompting yields worse performance than zero-shot prompting with user profiles. Self-consistency prompting (with 5 sampling times) improved performance in simpler ranking tasks but showed negligible or no gains in more challenging settings. Chain-of-thought prompting, which guides relative rating computation and ranking steps, unexpectedly degraded performance, particularly for easy tasks. These findings suggest that prompting strategies such as few-shot, self-consistency, and chain-of-thought are not universally effective for enhancing personalization performance. This may be due to inconsistencies between user behavior patterns in few-shot demonstrations and those in the query task. These results validate the design choice of profile-augmented zero-shot prompting as the primary method in our main experiments, as it strikes a better balance between simplicity and performance.

Performance across Different Domains Personalized recommendations span various domains, each with unique data distributions that can influence model performance. To examine this, we

Table 3: PerRecBench results of Llama-3.1-8B-it and Mistral-12B-Nemo-it with different supervised fine-tuning strategies, where colors indicate higher, same, and lower performance compared to prompting baseline. While weight merging generally achieves the best performance, it fails to achieve universal improvement across different ranking methods and task difficulties.

Training Method	easy			medium			hard			Avg.
	pointwise	pairwise	listwise	pointwise	pairwise	listwise	pointwise	pairwise	listwise	
Llama-3.1-8B-it										
Prompting	-0.27	0.25	0.30	0.03	-0.01	-0.01	-0.03	0.07	0.05	0.04
Pointwise Only	0.13	0.21	-	0.18	0.00	-	0.11	0.03	-	-
Pairwise Only	-0.04	0.21	-	0.18	0.06	-	0.18	0.09	-	-
Listwise Only	-0.14	0.15	-0.05	0.18	0.03	-0.03	0.19	0.03	0.05	0.04
Multi-task Training	0.12	0.00	0.04	0.16	-0.02	-0.01	0.13	-0.01	0.04	0.05
Weight Merging	0.15	0.22	0.31	0.22	-0.01	-0.03	0.16	0.08	0.06	0.13
Mistral-12B-Nemo-it										
Prompting	-0.11	0.14	0.37	0.00	0.07	0.07	0.05	0.02	-0.10	0.07
Pointwise Only	0.17	0.31	0.20	0.07	-0.01	0.09	0.08	0.02	0.01	0.12
Pairwise Only	0.17	0.40	0.26	0.02	-0.03	0.06	0.05	0.03	0.06	0.12
Listwise Only	0.17	0.30	0.22	0.10	0.08	0.05	0.08	-0.01	-0.02	0.11
Multi-task Training	0.17	0.11	0.12	0.16	0.01	0.02	0.11	0.01	0.03	0.08
Weight Merging	0.19	0.34	0.21	0.08	0.08	0.07	0.06	0.01	0.11	0.13

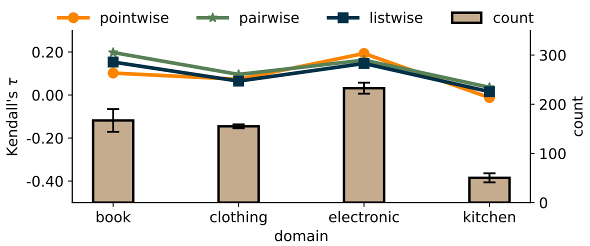

Figure 4: Performance across different domains and average item name count in the pretraining dataset. Query items with higher frequency in pretraining data generally show better performance in PerRecBench.

analyzed the correlation between model performance across different shopping domains and the frequency of item names in the LLM pretraining dataset. Specifically, we selected 50 user groups from the book, clothing, electronic, and kitchen domains, comparing their performance on PerRecBench with the average frequency of item names in the pretraining corpus. Since the training corpora of base LLMs are not publicly accessible, we approximated using the Dolma Corpus (Soldaini et al., 2024), which contains 3.1T tokens. The infini-gram method (Liu et al., 2024b) was used to calculate the average occurrence of item names in the corpus. As shown in Figure 4, query items with higher frequencies in the pretraining dataset generally exhibit better performance in PerRecBench. These findings suggest that LLM personalization capabilities are partially influenced by the domain distribution of their pretraining data, emphasizing the importance of diverse training datasets for improving performance across domains.

6 Enhancing LLM Personalized Recommendation

Results from PerRecBench reveal that current LLMs are not effective personalized recommender systems, as they show limited capability in understanding user preferences when user rating bias and item quality are eliminated. To address this limitation, we explore several supervised fine-tuning (SFT) strategies using held-out user groups. The methods are as follows:

- **Single-task Training** This approach involves training the model exclusively on one ranking method from pointwise, pairwise, or listwise.
- **Multi-task Training** In this method, pointwise, pairwise, and listwise ranking tasks are combined to create a joint training dataset for SFT.
- **Weight Merging** Parameters trained separately on pointwise θ_{pt}, pairwise θ_{pr}, and listwise methods θ_{ls} are merged using a linear combination: $\theta_{fnl} = \alpha \times \theta_{pt} + \beta \times \theta_{pr} + (1 - \alpha - \beta) \times \theta_{ls}$. We set $\alpha = \beta = 1/3$, averaging the weights from the three models.

Results We evaluate these SFT methods using Llama-3.1-8B-it and Mistral-12B-Nemo-it, with results shown in Table 3. Surprisingly, single-task training not only improves performance in the targeted ranking task but also shows moderate improvement in other tasks, suggesting effective task transfer. Conversely, multi-task training often underperforms compared to single-task training, indicating potential negative task transfer. The weight merging method consistently delivers the best results, achieving a notable improvement in average Kendall's tau and performance close to that of 70B parameter models. These findings highlight the importance of positive task transfer for developing LLMs capable of pointwise, pairwise, and listwise preference ranking. Though cannot achieve universal improvement against direct prompting, weight merging emerges as a viable improve strategy.

7 Related Work

Evaluation Metric of Personalized Recommendation Personalized recommendation systems are evaluated using metrics tailored to ranking-based and rating-based tasks. For ranking-based recommendations, models predict an ordered list of items, evaluated using metrics such as Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) and Hit Rate (HR) for ranking quality. Mean Reciprocal Rank (MRR) (Radev et al., 2002) assesses the position of the first relevant item, while Precision@K and Recall@K evaluate the relevance of top-k recommendations. Metrics like Coverage and Diversity capture the range of items recommended and their dissimilarity, respectively. For rating-based tasks, where models predict user ratings for query items. Metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Willmott and Matsuura, 2005) measure regression error, while R-Squared (Nagelkerke et al., 1991) evaluates the fit between predictions and actual ratings. Rating prediction can also be treated as a binary classification task, with metrics like AUC-ROC (Hanley and McNeil, 1982), F1 Score, Precision, Recall, and Log Loss used to assess performance.

LLM-based Personalized Recommendation Existing LLM-based methods for personalized recommendations can be broadly classified into ranking-based and rating-based approaches. Ranking-based methods generate an ordered list of items for users, leveraging user behavior history to predict the top-K items of interest. In-context learning is a popular paradigm, with several works exploring exemplars to improve understanding of user preferences (Liu et al., 2023a; Dai et al., 2023; Zhang et al., 2023a; Liu et al., 2023b; Hou et al., 2023; Du et al., 2023). Zhang et al. (2023b) introduced Chain-of-Thought prompting for top-K recommendations. Fine-tuning LLMs has also been explored to improve representation and domain adaptation. For instance, Chen (2023) and GenRec (Ji et al., 2023) fine-tuned Llama-7B for recommendation tasks, while Zhang et al. (2023b) adapted Flan-T5-XL through instruction tuning. Harte et al. (2023) integrated LLM embeddings and prompts with traditional sequential recommendation approaches. Rating-based methods predict user ratings for specific items, probing LLMs' ability to understand preferences and predict user behavior. Similar to ranking tasks, these methods can involve frozen or fine-tuned LLMs. For frozen LLMs, BookGPT (Zhiyuli et al., 2023) and Dai et al. (2023) used prompt engineering for rating prediction, while KAR (Xi et al., 2023) generated user profiles and item knowledge for use in discriminative recommendation systems. Fine-tuned LLMs include Kang et al. (2023), which adapted LLMs to rating prediction tasks, and TallRec (Bao et al., 2023), which integrated LoRA and instruction tuning. OPPU (Tan et al., 2024b) introduced personalized PEFT for private and accurate rating prediction, while Per-Pcs (Tan et al., 2024a) optimized PEFT for efficient personalization.

Ranking-based evaluations are limited by their reliance on unreviewed items as distractors, as these items may still be suitable recommendations despite lacking user exposure. Rating-based methods are influenced by user rating biases and item quality, often failing to accurately reflect personalization capabilities. To address these limitations, we propose PerRecBench, a benchmark that eliminates biases from ratings and item quality, relying solely on observed user preference signals to evaluate true personalization effectiveness.

8 Conclusion

User rating bias and item quality significantly impact user ratings, often hindering the evaluation of user preferences. To address this, we introduced PerRecBench, a personalized recommendation benchmark that removes the influence of rating bias and item quality, focusing solely on observed user preference signals to evaluate personalization capabilities. Extensive experiments on PerRecBench reveal that current LLMs face substantial challenges in personalization, with performance varying based on prompting methods and training domains. We also investigated supervised fine-tuning strategies, finding weight merging to be the most effective, but enhancing LLMs' personalization capabilities remains an open challenge.

Limitations

We identify two key limitations in PerRecBench. First, the dataset scale is relatively small, comprising 200 user groups for each difficulty level, resulting in a total of 600 groups and 1800 users. This limitation arises from the strict constraints in selecting personalization data. Additionally, the user group selection criteria in PerRecBench may bias the dataset toward more popular items, potentially introducing item-related bias into the evaluation process. Second, the methods explored for improving LLM personalization capabilities are not universally effective in all scenarios, leaving the challenge of enhancing personalized LLM recommendations an open research problem. Future work could explore encoding user histories into personalized PEFT parameters (Tan et al., 2024b), which may offer a promising direction with sufficient computational resources. Furthermore, our experiments employ LoRA for SFT on Llama-3.1-8B-it and Mistral-12B-Nemo-it. While efficient, this approach might impact results compared to full-parameter fine-tuning.

Ethical Considerations

Data Bias The design of PerRecBench relies on observed user preferences, and while efforts are made to eliminate rating bias and item quality effects, biases inherent in the underlying data may still influence the evaluation and personalization capabilities. For instance, popular items may disproportionately appear in user groups, potentially introducing item-related biases. Such biases could skew evaluations and lead to misleading conclusions about LLM personalization performance. Future work should explore methods to ensure diversity and fairness in data selection and mitigate biases in both user and item distributions.

Privacy Personalization inherently requires the use of user-specific data, which may include sensitive or private information. While PerRecBench focuses on observed user preferences and anonymized data, extending this benchmark to real-world applications may involve privacy risks. Care must be taken to ensure that data used for personalization is anonymized, securely stored, and handled in compliance with privacy regulations. Future iterations of PerRecBench could incorporate privacy-preserving techniques, such as differential privacy or personalized parameter-efficient fine-tuning (PEFT), to enhance privacy safeguards.

Accessibility The computational demands of LLM training and evaluation, particularly for benchmarks like PerRecBench, pose challenges for smaller organizations or individual researchers with limited resources. This may exacerbate disparities in access to cutting-edge personalization research and hinder equitable advancements in the field. Efforts should focus on improving the efficiency of benchmarking frameworks and exploring lightweight alternatives to support broader accessibility and inclusivity in AI research.

Fairness in Personalization While PerRecBench aims to evaluate personalization capabilities, care must be taken to ensure that such personalization does not inadvertently reinforce harmful stereotypes or exclude certain user groups. Models evaluated on PerRecBench should be assessed not only for their personalization accuracy but also for fairness and inclusivity, ensuring equitable treatment across diverse user populations.

References

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Sujay Kumar Jauhar, et al. 2023. Knowledge-augmented large language models for personalized contextual query suggestion. *arXiv preprint arXiv:2311.06318*.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *arXiv preprint arXiv:2305.00447*.

Zheng Chen. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. *arXiv preprint arXiv:2305.02182*.

Cosmina Andreea Dejescu, Lucia V Bel, Iulia Melega, Stefana Maria Cristina Muresan, and Liviu Ioan Oana. 2023. Approaches to laparoscopic training in veterinary medicine: A review of personalized simulators. *Animals*, 13(24):3781.

Yingpeng Du, Di Luo, Rui Yan, Hongzhi Liu, Yang Song, Hengshu Zhu, and Jie Zhang. 2023. Enhancing job recommendation through llm-based generative adversarial networks. *arXiv preprint arXiv:2307.10747*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Dmitri Goldenberg, Kostia Kofman, Javier Albert, Sarai Mizrachi, Adam Horowitz, and Irene Teinemaa. 2021. Personalization in practice: Methods and applications. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 1123–1126.

Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. 2020. Hierarchical user profiling for e-commerce recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 223–231.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1096–1102.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2023. Genrec: Large language model for generative recommendation. *arXiv e-prints*, pages arXiv–2307.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *Preprint*, arXiv:2305.06474.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Xinyu Li, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024b. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023a. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023b. A first look at llm-powered generative news recommendation. *arXiv preprint arXiv:2305.06566*.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Nico JD Nagelkerke et al. 1991. A note on a general definition of the coefficient of determination. *biometrika*, 78(3):691–692.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

MJ Pazzani. 2007. Content-based recommendation systems.

Muh Putra Pratama, Rigel Sampelolo, and Hans Lura. 2023. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of Education, Language Teaching and Science*, 5(2):350–357.

Xueming Qian, He Feng, Guoshuai Zhao, and Tao Mei. 2013. Personalized recommendation combining user interest and social circle. *IEEE transactions on knowledge and data engineering*, 26(7):1763–1777.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*. Citeseer.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi. 2020. Personalized chatbot trustworthiness ratings. *IEEE Transactions on Technology and Society*, 1(4):184–192.

Aixin Sun. 2023. Take a fresh look at recommender systems from an evaluation standpoint. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2629–2638.

Aixin Sun. 2024. Beyond collaborative filtering: A relook at task formulation in recommender systems. *ACM SIGWEB Newsletter*, 2024(Spring):1–11.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv preprint arXiv:2406.10471*.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, pages 58–65.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards open-world recommendation with knowledge augmentation from large language models. *arXiv preprint arXiv:2306.10933*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023a. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023b. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Aakas Zhiyuli, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023. Bookgpt: A general framework for book recommendation empowered by large language model. *arXiv preprint arXiv:2305.15673*.

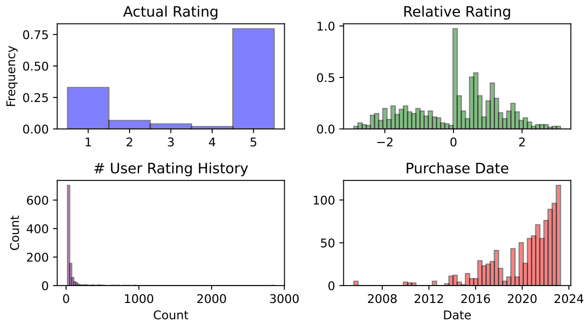

Figure 5: The statistics of PERRECBENCH, including the distribution of actual rating, relative rating, length of user history, and the purchase date.

Table 4: Statistics of PERRECBENCH.

Statistics	Held-out Data	Test Data
# Review	2,438	1,007
# User	2,140	986
# Query Item	466	200
# Rating History	157,480	61,412
# Token	289,164	111,017
Domain	book, clothing, electronic, kitchen	
Time Range	07/16/2022 - 05/16/2023	07/25/2005 - 04/14/2023

A PERRECBENCH Statistics

We present the statistics of PERRECBENCH in Figure 5, which include the distributions of actual ratings, relative ratings, the number of user rating histories, and the purchase dates of query items. The actual ratings are biased toward scores of 1 and 5. The relative rating distribution peaks around 0 and follows an approximately normal distribution. For the distribution of user rating histories, a long-tail pattern is observed, with all selected users having more than 20 rating histories to ensure sufficient data for personalization. The purchase date distribution shows that most user activity in PERRECBENCH occurred around 2022, indicating that PERRECBENCH contains up-to-date data. Additional benchmark statistics are provided in Table 4.

B Performance of Traditional Recommendation Methods

We also experiment with traditional recommendation methods on our PERRECBENCH, including content-based filtering (Pazzani, 2007), matrix factorization (Koren et al., 2009), and neural collaborative filtering (He et al., 2017). Specifically, for content-based filtering, we adopted item-based collaborative filtering, where given a query item, we calculate the similarity between the user's reviewed items using Jaccard similarities and do weighted aggregation on history rating based on the similarities to obtain the predicted rating. For matrix factorization, we only user ID and parent_asin of the item are used as input embeddings. For collaborative filtering, we adopt a two-layer MLP with hidden size of 64 and 32.

Table 5: Performance of traditional recommendation methods on PERRECBENCH: The results indicate that traditional methods struggle with personalization, often producing nearly non-personalized recommendation outputs.

Model	easy	medium	hard
	pointwise	pointwise	pointwise
COLLABORATIVE FILTERING	-0.04	0.04	0.02
MATRIX FACTORIZATION	-0.20	0.04	0.02
NEURAL COLLABORATIVE FILTERING	-0.27	-0.03	-0.08

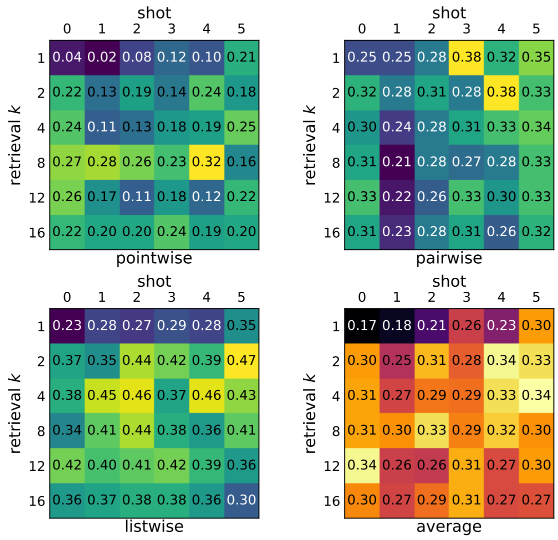

Figure 6: Performance of PERRECBENCH-EASY with different retrieval k and and shot number. By searching the appropriate k and shot combination, we can improve the performance.

Since traditional recommendation methods can only do pointwise rating prediction, we adopt the pointwise ranking setting in PERRECBENCH, predicting each user's actual rating and obtain user relative rating by subtracting the user average rating and do the ranking based on relative rating. Results are presented in Table 5, we find traditional recommendation methods fail to capture the personalized preference in PERRECBENCH and give close to non-personalized predictions. These results indicate that traditional recommendation methods are hard to capture the personalized preference signals.

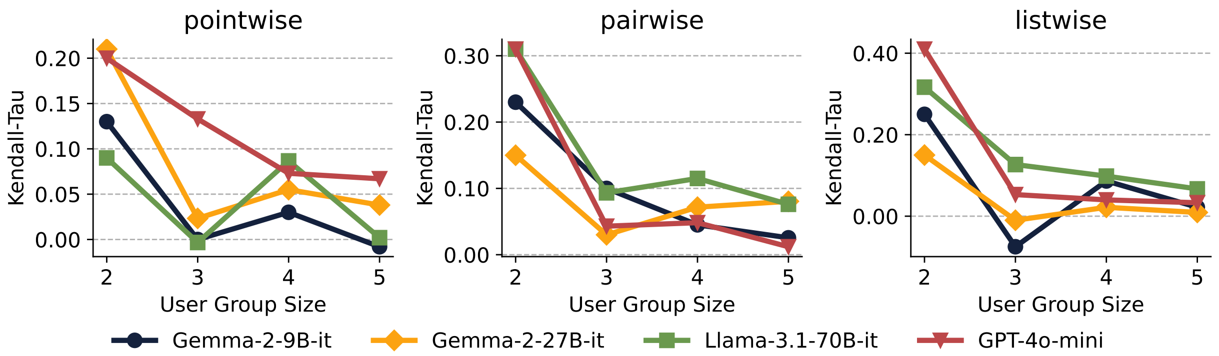

Figure 7: LLM performance using pointwise, pairwise, and listwise ranking methods across different user group sizes. Larger user groups generally result in reduced performance, while pointwise ranking demonstrates stroner robustness to group size variations.

C Analysis (Cont.)

Combination of Shot and k Although the preliminary results in Table 2 indicate that few-shot prompting does not consistently enhance personalization capabilities, its interaction with different retrieval history counts (k) warrants further investigation to understand the role of demonstrations in LLM personalization. To explore this, we vary the shot size (shot $\in \{0, 1, 2, 3, 4, 5\}$) and the retrieval history count ($k \in \{1, 2, 4, 8, 12, 16\}$) and visualize the performance for pointwise, pairwise, and listwise evaluations, as well as the average performance under the easy setting in Figure 6. The results show that while few-shot prompting does not consistently improve performance and can sometimes degrade it, it provides relatively consistent improvements when $k \in \{2, 4\}$ and shot $\in \{4, 5\}$. Furthermore, when $k = 1$, increasing the number of shots results in noticeable improvements; however, this benefit diminishes as k increases. We hypothesize that while both shot examples and retrieval histories provide relevant information, the retrieval history is more directly aligned with the user's personalized preferences. In contrast, patterns in randomly chosen shots may introduce noise, distracting the LLM and negatively affecting predictions.

Performance *w.r.t.* User Group Size From the main results, the average performance for user group sizes of 2, 3, and 4 is 0.201, 0.052, and 0.058, respectively. These findings suggest that groups with two users are less challenging for LLMs, as they only require a single comparison between the two users. However, as the group size increases beyond two, the number of comparisons required for user preference evaluation grows, leading to significant performance degradation. We visualize the performance of pointwise, pairwise, and listwise rankings across different user group sizes in Figure 7. The results show that pointwise ranking is more robust to variations in group size, with performance decreasing only slightly from 0.09 to 0.05 and 0.06. In contrast, pairwise and listwise rankings exhibit significant drops when the group size exceeds two, with pairwise ranking declining from 0.24 to 0.07 and listwise ranking from 0.27 to 0.05. This trend may be attributed to the accumulation of errors in pairwise ranking and the increased complexity of the task in listwise ranking as the group size grows.

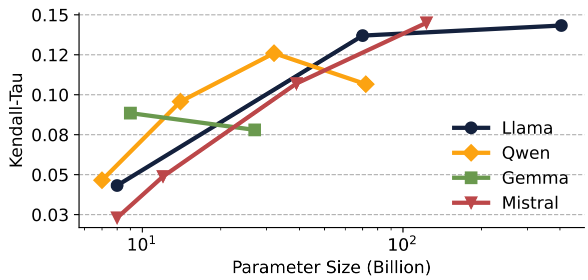

Figure 8: Zero-shot average performance across varying task difficulties and ranking methods for models of different parameter sizes in PerRecBench. The Llama, and Mistral model families demonstrate performance improvements with increased model size, while Gemma and Qwen family cannot guarantee the performance gain with larger model size.

Scaling Law in PerRecBench We analyze the performance of the Llama, Qwen, Gemma, and Mistral model families to investigate the scaling law in PerRecBench, as shown in Figure 8. For the Llama and Mistral families, larger parameter sizes consistently lead to better performance, aligning with the scaling law. However, this trend does not hold uniformly for the Qwen and

Gemma model families. For instance, the Qwen-2.5-72B-it model significantly underperforms the Qwen-2.5-32B-it model, and the Gemma-2-27B-it model falls short compared to the Gemma-2-9B-it model. These findings suggest that while larger models often exhibit stronger performance in PERRECBENCH, simply increasing parameter size does not guarantee performance improvements. This challenges the universal applicability of the scaling law in PERRECBENCH, indicating that factors beyond parameter count may play a critical role in model performance.

D Computational Resources

All experiments are implemented on a server with 3 NVIDIA A6000 GPU and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz with 20 CPU cores. For model fine-tuning, it takes 9 hours, 4 hours, 12 hours, and 25 hours to do pointwise, pairwise, and listwise supervised fine-tuning on a single GPU.

E Scientific Artifacts

PERRECBENCH is built with the help of many existing scientific artifacts, including PyTorch (Paszke et al., 2019), Numpy (Harris et al., 2020), huggingface, vllm (Kwon et al., 2023), and transformers (Wolf et al., 2020). We will make the PERRECBENCH data, source code, and all model output publicly available to facilitate further research.

F Prompt Template

Pointwise Prompt Template

[System]
Act as a personalized product recommender system. Below is a list of user's rating history, shown in [User History]. Your task is to predict the user's rating for the query item, which is described in [Query Item Details]. Analyze each user's preferences for the query item based on their historical ratings to generate the prediction. Output a predicted rating ranging from 1 to 5, where 1 being not recommended and 5 being highly recommended. The final answer should strictly follow this JSON structure: {"predicted_rating": <rating>}

[User History]
`{User Data}`

[Query Item Details]
`{Query Item Information}`

Answer:

Pairwise Prompt Template

[System]
Act as a personalized product recommender system. Below is a pair of users, [User A] and [User B], each with their rating history. Your task is to determine which user is more likely to prefer the query item, based on its details in [Query Item Details]. Analyze each user's preferences for the query item using their historical ratings and output the user more likely to prefer the query item.
[User A]
`{User A Information}`

[User B]
`{User B Information}`

[Query Item Details]
`{Query Item Information}`
Which user prefer the query item more? Output only "[User A]" or "[User B]", do not generate anything else:

Listwise Prompt Template

[System]
Act as a personalized product recommender system. Below is a group of users accompanied with each user's rating history, shown in [Users]. Your task is to rank these users based on their preference of the query item, which is described in [Query Item Details]. Analyze each user's preferences for the query item based on their historical ratings to generate this ranking. Output the list of user indices (e.g., 1 for User1), ranked from highest to lowest preference for the query item. The final output should rank users from most preferred to least preferred for the query item and adhere to the JSON structure shown below: {"predicted_ranking": <user_ranking>}

[Users]
`{All User Information within User Group}`

[Query Item Details]
`{Query Item Information}`

The answer ranks users from most preferred to least preferred for the query item and adhere to the following JSON format, output the list of user indices (e.g., 1 for User1), do not include any additional information: {"predicted_ranking": <user_ranking>}
Answer:

Prompt Template for User Profile Generation

User Behavior history
`{User Behavior List}`

Task Instruction
You are given a list of user behavior history data. Your task is to analyze this data and create a user profile that describes the user's preferences, interests, and patterns of behavior. This profile should be written in a concise and coherent narrative form. Only generate user profile without any additional characters or formatting.

Template for a Single Behavior

\### Item Title
{Item Name}

\### Item Author
{Author Name}

\### User Rating
{User Rating}

Template for a Single User

<|The Start of User Data|>
\### User Profile
{User Profile}

\### User Most Common Rating
{User Most Common Rating}

\### User Average Rating
{User Average Rating}

{Retrieved Top-k User History Behavior}

<|The End of User Data|>

Query Item Template

<|The Start of Query Item Information|>
\### Item Title
{Item Name}

\### Item Author
{Item Author}
<|The End of Query Item Information|>

Pointwise Chain-of-Thought Prompt Template

[System]
To predict a user's rating for a query item, follow these steps: 1. Analyze the user's preference for the query item using their history and profile. 2. Predict the user's rating for the query item: If the item is likely preferred by the user, the predicted rating should be higher than the user's average rating. If the item is unlikely to be preferred by the user, the predicted rating should be lower than the user's average rating. Act as a personalized product recommender system. Below is a list of user's rating history, shown in [User History]. Your task is to predict the user's rating for the query item, which is described in [Query Item Details]. Analyze each user's preferences for the query item based on their historical ratings to generate the prediction. Output a predicted rating ranging from 1 to 5, where 1 being not recommended and 5 being highly recommended. The final answer must strictly follow this JSON structure: {"predicted_rating": <rating>}.

[User History]
{User Data}

[Query Item Details]
{Query Item Information}

Answer: Let's think step by step.

Pairwise Chain-of-Thought Prompt Template

[System]
Act as a personalized product recommender system. Below is a pair of users, [User A] and [User B], each with their rating history. Your task is to determine which user is more likely to prefer the query item, based on its details in [Query Item Details]. Analyze each user's preferences for the query item using their historical ratings and output the user more likely to prefer the query item.

[User A]
{User A Information}

[User B]
{User B Information}

[Query Item Details]
{Query Item Information}

Which user prefer the query item more? First provide a short thinking step, then output your final answer in "[User A]" or "[User B]".
Answer:

Listwise Chain-of-Thought Prompt Template

[System]
To rank user preferences, follow these steps: 1. Predict the rating of the query item for each user. 2. Calculate the relative rating by subtracting each user's average rating from the predicted rating. 3. Rank user preferences based on the relative ratings: users with higher relative ratings should be ranked higher, while those with lower relative ratings should be ranked lower. Act as a personalized product recommender system. Below is a group of users accompanied with each user's rating history, shown in [Users]. Your task is to rank these users based on their preference of the query item, which is described in [Query Item Details]. Analyze each user's preferences for the query item based on their historical ratings to generate this ranking. Output the list of user indices (e.g., 1 for User1), ranked from highest to lowest preference for the query item. The final output should rank users from most preferred to least preferred for the query item and adhere to the JSON structure shown below: {"predicted_ranking": <user_ranking>}

[Users]
{All User Information within User Group}

[Query Item Details]
{Query Item Information}

The answer ranks users from most preferred to least preferred for the query item and adhere to the following JSON format, output the list of user indices (e.g., 1 for User1): {"predicted_ranking": <user_ranking>}
Answer: Let's think step by step.

Table 6: Pointwise ranking output in PerRecBench.

<table>
<tr><td>Input</td><td>[User History]
<|The Start of User Data|>
User Profile
The user appears to have a strong interest in home decor and organization. They have purchased various items such as wall art, a headboard, a folding table, bed sheets, and storage ottomans, suggesting a desire to create a comfortable and stylish living environment. Additionally, their purchase of a vacuum cleaner and coat rack indicates a concern for cleanliness and organization.
The user also seems to have a preference for quality and durability, as evidenced by their purchase of high-thread-count sheets and a stainless steel coffee percolator. They may value products that are long-lasting and well-made.
Furthermore, the user has shown an interest in area rugs, with purchases ranging from small runners to larger rugs, potentially indicating a desire to add warmth and texture to their living spaces.
Overall, the user profile suggests an individual who values a well-organized and aesthetically pleasing home environment, with a focus on quality and functional pieces that contribute to both comfort and style.
User Most Common Rating
5.0
User Average Rating
3.9
Item Title
West Bend 54159 Classic Stainless Steel Electric Coffee Percolator with Heat Resistant Handle and Base Features Detachable Cord, 12-cup, Silver
User Rating
2.0
Item Title
Cloth Napkins Set of 12 Cotton Linen Blend Printed Dinner Napkins Perfect for Parties Dinners Weddings Cocktail Christmas Napkins Cloth 20x20 Blue Floral
User Rating
5.0
Item Title
Weavric Cloth Dinner Napkin Bulk, Set of 12, 20 X 20 Inches Wrinkle-Free Washable Reusable Forest Green Linen Table Napkins with Hemmed Edge for Wedding, Party, Hotel, Restaurant
User Rating
5.0
Item Title
Crown Mark Barlow Bicast Headboard, King
User Rating
5.0
<|The End of User Data|>

[Query Item Details]
<|The Start of Query Item Information|>
Item Title
Keurig K-Slim Coffee Maker, Single Serve K-Cup Pod Coffee Brewer, Multistream Technology, Scarlet Red
<|The End of Query Item Information|>
Answer:</td></tr>
<tr><td>Gemma-2-9B-it</td><td>{"predicted_rating": 4}</td></tr>
<tr><td>Qwen-2.5-14B-it</td><td>{"predicted_rating": 4}</td></tr>
<tr><td>Llama-3.1-70B-it</td><td>{"predicted_rating": 2.0}</td></tr>
<tr><td>GPT-4o-mini</td><td>{"predicted_rating": 4.0}</td></tr>
<tr><td>Claude-3.5-Sonnet</td><td>{"predicted_rating": 3}</td></tr>
<tr><td>Ground Truth</td><td>actual rating: 5.0</td></tr>
</table>

Table 7: Pairwise ranking output in PERRECBENCH.

Input

[User A]
<|The Start of User Data|>
User Profile
The user appears to have a diverse range of interests and preferences based on their purchase history. They seem to be interested in kitchen and household items, with purchases including a foil cutter, cheese slicer, dish soap dispenser, trash can, and cabinet organizers. They also appear to enjoy beer and have purchased beer-related products like a beer dispenser and bar towel.

The user's purchase history suggests an interest in home decor and organization, with items like decorative wall art, a vanity, and storage shelves. They have also purchased alarm clocks, indicating a need for timekeeping devices.

In terms of electronics and appliances, the user has bought a Keurig coffee maker, a gaming chair, and a robot vacuum cleaner, suggesting an interest in convenience and technology.

The user's ratings reveal a preference for high-quality and functional products, with items like the foil cutter, cheese slicer, vanity, and oscillating fan receiving high scores. However, they also seem to be dissatisfied with some purchases, as evidenced by the low scores given to certain items like alarm clocks and towels.

Overall, the user appears to be practical and value-conscious, seeking products that serve specific purposes and offer good quality and functionality. Their interests span across various categories, including kitchen, home organization, decor, technology, and entertainment.

User Most Common Rating
1.0

User Average Rating
2.1

Item Title
Keurig K-Classic Coffee Maker K-Cup Pod, Single Serve, Programmable, 6 to 10 oz. Brew Sizes, Black
User Rating
1.0
Item Title
Leick Favorite Finds Coffee Table
User Rating
1.0

Item Title
Ottomanson CTW1008-16X30 8 Piece Turkish Cotton Towels, 16" X 30"-Set of 6, Brown
User Rating
2.0

Item Title
OKP K8 Robot Vacuum and Mop Combo, 2000Pa Super Suction, Integrated Design of Dust Box Water Tank, Self Charging, Robotic Vacuums for Pet Hair, Blue
User Rating
1.0
<|The End of User Data|>

Table 8: Pairwise ranking output in PERRECBENCH.

Input	[User B] <|The Start of User Data|> ## User Profile The user appears to have a strong interest in home decor and organization. They have purchased various items such as wall art, a headboard, a folding table, bed sheets, and storage ottomans, suggesting a desire to create a comfortable and stylish living environment. Additionally, their purchase of a vacuum cleaner and coat rack indicates a concern for cleanliness and organization. The user also seems to have a preference for quality and durability, as evidenced by their purchase of high-thread-count sheets and a stainless steel coffee percolator. They may value products that are long-lasting and well-made. Furthermore, the user has shown an interest in area rugs, with purchases ranging from small runners to larger rugs, potentially indicating a desire to add warmth and texture to their living spaces. Overall, the user profile suggests an individual who values a well-organized and aesthetically pleasing home environment, with a focus on quality and functional pieces that contribute to both comfort and style. ## User Most Common Rating 5.0 ## User Average Rating 3.9 ### Item Title West Bend 54159 Classic Stainless Steel Electric Coffee Percolator with Heat Resistant Handle and Base Features Detachable Cord, 12-cup, Silver ### User Rating 2.0 ### Item Title Cloth Napkins Set of 12 Cotton Linen Blend Printed Dinner Napkins Perfect for Parties Dinners Weddings Cocktail Christmas Napkins Cloth 20x20 Blue Floral ### User Rating 5.0 ### Item Title Weavric Cloth Dinner Napkin Bulk, Set of 12, 20 X 20 Inches Wrinkle-Free Washable Reusable Forest Green Linen Table Napkins with Hemmed Edge for Wedding, Party, Hotel, Restaurant ### User Rating 5.0 ### Item Title Crown Mark Barlow Bicast Headboard, King ### User Rating 5.0 <|The End of User Data|>

Table 9: Pairwise ranking output in PerRecBench.

<table>
<tr><td>Input</td><td>[Query Item Details]
<|The Start of Query Item Information|>
Item Title
Keurig K-Slim Coffee Maker, Single Serve K-Cup Pod Coffee Brewer, Multistream Technology, Scarlet Red
<|The End of Query Item Information|>

Which user prefer the query item more? Output only "[User A]" or "[User B]", do not generate anything else:</td></tr>
<tr><td>Gemma-2-9B-it</td><td>[User A]</td></tr>
<tr><td>Qwen-2.5-14B-it</td><td>[User B]</td></tr>
<tr><td>Llama-3.1-70B-it</td><td>[User A]</td></tr>
<tr><td>GPT-4o-mini</td><td>[User B]</td></tr>
<tr><td>Claude-3.5-Sonnet</td><td>[User B]</td></tr>
<tr><td>Ground Truth</td><td>[User B]</td></tr>
</table>

Table 10: Listwise ranking output in PerRecBench.

<table>
<tr><td>Input</td><td>[Users]
<|The Start of User1 Data|>
User Profile
Based on the user behavior history data, this user appears to have a strong interest in kitchen and household items. They have purchased a variety of appliances and tools for cooking, baking, and food preparation, such as a stand mixer, sandwich maker, grill, and coffee maker. Additionally, they seem to value convenience and practicality, as evidenced by their purchase of a touchless trash can and a slim, shatterproof pitcher. The user also seems to appreciate furniture and decor items that have a rustic or natural aesthetic, as shown by their purchase of a rustic end table. Overall, this user likely enjoys cooking, entertaining, and creating a comfortable and functional living space.

User Most Common Rating
5.0

User Average Rating
4.5

Item Title
Keurig K-Mini Plus Coffee Maker, Single Serve K-Cup Pod & Keurig K-Cup Pod & Ground Coffee Storage Unit
User1 Rating
5.0

Item Title
Keurig K-Mini Plus Coffee Maker, Single Serve K-Cup Pod & Keurig K-Cup Pod & Ground Coffee Storage Unit
User1 Rating
5.0

Item Title
Signature Design by Ashley - Mestler Rustic Chairside End Table w/ Two Fixed Multi-Colored Shelves, Brown
User1 Rating
5.0

Item Title
Signature Design by Ashley - Mestler Rustic Chairside End Table w/ Two Fixed Multi-Colored Shelves, Brown
User1 Rating
5.0
<|The End of User1 Data|>

<|The Start of User2 Data|>
User Profile
The user appears to have a strong interest in home decor and organization. They have purchased various items such as wall art, a headboard, a folding table, bed sheets, and storage ottomans, suggesting a desire to create a comfortable and stylish living environment. Additionally, their purchase of a vacuum cleaner and coat rack indicates a concern for cleanliness and organization.

The user also seems to have a preference for quality and durability, as evidenced by their purchase of high-thread-count sheets and a stainless steel coffee percolator. They may value products that are long-lasting and well-made.</td></tr>
</table>

Table 11: Listwise ranking output in PerRecBench.

Input

Furthermore, the user has shown an interest in area rugs, with purchases ranging from small runners to larger rugs, potentially indicating a desire to add warmth and texture to their living spaces.

Overall, the user profile suggests an individual who values a well-organized and aesthetically pleasing home environment, with a focus on quality and functional pieces that contribute to both comfort and style.

User Most Common Rating
5.0

User Average Rating
3.9

Item Title
West Bend 54159 Classic Stainless Steel Electric Coffee Percolator with Heat Resistant Handle and Base Features Detachable Cord, 12-cup, Silver
User2 Rating
2.0

Item Title
Cloth Napkins Set of 12 Cotton Linen Blend Printed Dinner Napkins Perfect for Parties Dinners Weddings Cocktail Christmas Napkins Cloth 20x20 Blue Floral
User2 Rating
5.0

Item Title
Weavric Cloth Dinner Napkin Bulk, Set of 12, 20 X 20 Inches Wrinkle-Free Washable Reusable Forest Green Linen Table Napkins with Hemmed Edge for Wedding, Party, Hotel, Restaurant
User2 Rating
5.0

Item Title
Crown Mark Barlow Bicast Headboard, King
User2 Rating
5.0
<|The End of User2 Data|>

<|The Start of User3 Data|>
User Profile
The user appears to have diverse interests spanning personal care, kitchen gadgets, and home organization. They seem to value quality and convenience, as evidenced by their high ratings for items like the vanilla sticks, manual food chopper, herb mincer, and Oster convection toaster oven. However, they also express dissatisfaction with certain products, such as the humidifier, milk frother, can openers, and pepper mill, suggesting a discerning eye for functionality.
The user's interest in kitchen tools and appliances is evident, with a focus on efficient food preparation and storage solutions. The purchase of a high-quality food processor with a spiralizer attachment indicates a potential interest in healthy eating or culinary exploration.
Organization and storage seem to be important to the user, as demonstrated by their purchase of a shoe storage rack and refrigerator liners. Comfort is also a consideration, with the purchase of a heated mattress pad and a leg elevation pillow, although the latter received a low rating.

Table 12: Listwise ranking output in PERRECBENCH.

<table>
<tr><td>Input</td><td>Overall, the user appears to be a practical and discerning consumer who values quality, convenience, and organization in their personal and kitchen-related purchases, while also exploring new culinary avenues and prioritizing comfort at home.

User Most Common Rating
1.0

User Average Rating
2.7

Item Title
Best Cool Mist Humidifier UltraSonic Steam Vaporizer - Whisper Quiet Technology, Moistair Electric with Warm LED Light
User3 Rating
1.0

Item Title
Cusinart Small Pusher
User3 Rating
5.0

Item Title
ZYLISS FastCut Herb Mincer
User3 Rating
5.0

Item Title
FRESH STICKS - Golden Vanilla
User3 Rating
5.0
<|The End of User3 Data|>

<|The Start of User4 Data|>
User Profile
The user appears to have a diverse range of interests and preferences based on their purchase history. They seem to be interested in kitchen and household items, with purchases including a foil cutter, cheese slicer, dish soap dispenser, trash can, and cabinet organizers. They also appear to enjoy beer and have purchased beer-related products like a beer dispenser and bar towel.

The user's purchase history suggests an interest in home decor and organization, with items like decorative wall art, a vanity, and storage shelves. They have also purchased alarm clocks, indicating a need for timekeeping devices.

In terms of electronics and appliances, the user has bought a Keurig coffee maker, a gaming chair, and a robot vacuum cleaner, suggesting an interest in convenience and technology.</td></tr>
</table>

Table 13: Listwise ranking output in PerRecBench.

<table>
<tr><td>Input</td><td>The user’s ratings reveal a preference for high-quality and functional products, with items like the foil cutter, cheese slicer, vanity, and oscillating fan receiving high scores. However, they also seem to be dissatisfied with some purchases, as evidenced by the low scores given to certain items like alarm clocks and towels.

Overall, the user appears to be practical and value-conscious, seeking products that serve specific purposes and offer good quality and functionality. Their interests span across various categories, including kitchen, home organization, decor, technology, and entertainment.

User Most Common Rating
1.0
User Average Rating
2.1
Item Title
Keurig K-Classic Coffee Maker K-Cup Pod, Single Serve, Programmable, 6 to 10 oz. Brew Sizes, Black
User4 Rating
1.0
Item Title
Leick Favorite Finds Coffee Table
User4 Rating
1.0
Item Title
Ottomanson CTW1008-16X30 8 Piece Turkish Cotton Towels, 16Ẍ 30≐Set of 6, Brown
User4 Rating
2.0
Item Title
OKP K8 Robot Vacuum and Mop Combo, 2000Pa Super Suction, Integrated Design of Dust Box Water Tank, Self Charging, Robotic Vacuums for Pet Hair, Blue
User4 Rating
1.0
<|The End of User4 Data|>

[Query Item Details]
<|The Start of Query Item Information|>
Item Title
Keurig K-Slim Coffee Maker, Single Serve K-Cup Pod Coffee Brewer, Multistream Technology, Scarlet Red
<|The End of Query Item Information|>

The answer ranks users from most preferred to least preferred for the query item and adhere to the following JSON format, do not include any additional information: {p̈redicted_ranking:̈ <user_ranking>}
Answer:</td></tr>
<tr><td>Gemma-2-9B-it</td><td>{"predicted_ranking": [1, 2, 4, 3]}</td></tr>
<tr><td>Qwen-2.5-14B-it</td><td>{"predicted_ranking": [1, 2, 4, 3]}</td></tr>
<tr><td>Llama-3.1-70B-it</td><td>{"predicted_ranking": [1, 2, 4, 3]}</td></tr>
<tr><td>GPT-4o-mini</td><td>{"predicted_ranking": [1, 2, 3, 4]}</td></tr>
<tr><td>Claude-3.5-Sonnet</td><td>{"predicted_ranking": [1, 2, 3, 4]}</td></tr>
<tr><td>Ground Truth</td><td>[2, 1, 4, 3]</td></tr>
</table>