LOST IN EVALUATION? RETHINKING SEQUENTIAL RECOMMENDATION BENCHMARKING VIA A MULTI-DIMENSIONAL AND GEO RISK ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

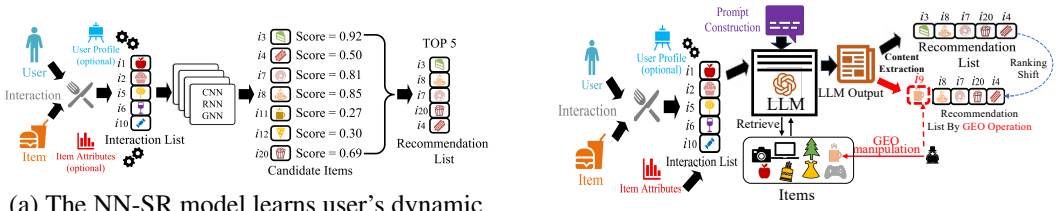
The rapid development of LLM has drawn significant attention in Sequential Recommendation (SR). However, a comprehensive evaluation of SR models remains lacking due to the limitations of existing benchmarks: 1) an overemphasis on accuracy metrics neglects real-world demands such as quality, and overlooks emerging adversarial threats like Generative Engine Optimization (GEO), which can stealthily manipulate ranking results. 2) existing evaluation methods fail to fully exploit the semantic capabilities of LLM, leading to unfair comparisons between Neural Network-based SR (NN-SR) and LLM-based SR (LLM-SR) models; and 3) a reliable parsing mechanism is absent to extract task-specific results from unstructured LLM outputs. To address this gap, we propose SR-Bench, a fair and comprehensive benchmark. SRBench incorporates 1) a novel four-dimensional evaluation framework aligned with practical demands to holistically evaluate recommendation accuracy, quality, stability, and efficiency, which also serves as a latent GEO risk screening signal; 2) a unified prompt-enhanced input paradigm enables fair comparisons across recommendation paradigms; 3) a prompt-extractor-coupled extraction mechanism captures answers from LLM outputs through prompt-guided output formatting and a numeric-oriented extractor. Evaluating 14 mainstream models across several datasets with SRBench reveals several valuable insights, such that LLM-SR models overfocus on item popularity while lacking a deep understanding of item quality, and LLM-SR models with stronger semantic capabilities show behavioral divergence from NN-SR models, indicative of latent GEO risk. Thus, SRBench promotes holistic cross-paradigm evaluation and risk screening signals, providing a solid basis for reliable and trustworthy LLM-SR research and applications. Source code is available at https://anonymous.4open.science/r/LLM-recommendation-C5A4

1 INTRODUCTION

With the global recommendation system market based on AI reaching US $2.21B in 2024 (Raza et al., 2026), generative AI has propelled LLM-based sequential recommendation models. These models have attracted significant attention from both academia (e.g., ICLR (Li et al., 2026)) and industry (e.g., Amazon (Li et al., 2024)) because of the superior contextual understanding capacities of LLM. However, researchers have also expressed concerns about the practical issues of LLM-SR models, such as stability and quality (Ma et al., 2023; Zhang et al., 2023a). Meanwhile, GEO (Aggarwal et al., 2024; Wen et al., 2026) has emerged as an adversarial risk to this $2.21B market, where attackers inject stealthy adversarial semantic signals to manipulate item exposure and compromise real-world LLM-SR credibility. These limitations highlight the critical need for a holistic benchmark that not only evaluates NN-SR and LLM-SR models but also screens GEO risk.

Existing benchmarks have evaluated models in partial models or limited evaluation dimensions. LLMRec (Liu et al., 2023b) evaluates only LLM-SR models on recommendation tasks, without comparing NN-SR models. OpenP5 (Xu et al., 2024) and RecBench (Liu et al., 2025) cover NN-SR and LLM-SR models yet only use accuracy-based evaluation metrics. Beyond accuracy, real-world SR scenarios involve other crucial dimensions (e.g., quality), which are also vital for judging recommendation effect and practical utility. Crucially, the reliability of systematic evaluation is further

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) The NN-SR model learns user's dynamic preference from the historical interaction sequence, and predicts the candidate items.

(b) LLM-SR model predicts SR results based on constructed prompt and GEO amplify target item ranking.

Figure 1: Frameworks of SR:(a) NN-SR Methods and (b) LLM-SR Methods.

compromised by emerging GEO threats. For example, StealthRank (Tang et al., 2025) manipulates semantic signals to increase item visibility, whereas SCI-Defense (Yu et al., 2026) investigates defenses against such manipulation under black-box scenarios. Because these disruptions are covert, traditional accuracy metrics fail to reflect a model's susceptibility to GEO risk. Regrettably, current benchmarks neglect this critical security aspect in SR.

Thus, we need to address a technical question: how to design a benchmark that not only enables comprehensive cross-paradigm evaluations but also assesses the security risk brought by GEO. However, achieving this goal is difficult and several challenges need to be addressed:

Challenge 1: Multi-Dimensional Evaluation and GEO Risk Assessment. NN-SR models capture sequential patterns from user profiles and historical interactions. In contrast, LLM-SR models leverage pretrained knowledge for semantic reasoning, rendering them uniquely vulnerable to GEO strategies that covertly manipulate recommendation rankings. These distinct foundations, along with black-box manipulation threats and practical needs in SR, make it challenging to define evaluation dimensions and metrics to fairly compare both paradigms while detecting LLM-based GEO risk.

Challenge 2: Input Paradigm Alignment. As shown in Figure 1, NN-SR and LLM-SR employ distinct input paradigms, with LLM-SR requiring additional natural language prompts. Moreover, prompt engineering significantly affects LLM-SR performance (Xu et al., 2024). Thus, it is challenging to design an effective unified input paradigm that converts structured datasets into LLM-SR-compatible prompts and enhances them to fully unleash LLMs' potential for fair comparison.

Challenge 3: Output Standardization Mechanism. LLM-SR models output results in natural-language form, which is unstructured and disordered. Due to stochasticity of LLM, LLMs' outputs may contain hallucinations, making it harder to accurately extract answers for reliable evaluation.

In this paper, we propose SRBench to address the challenges and make the contributions:

To overcome Challenge 1, we conduct a comprehensive in-depth survey to analyze real-world SR demands. Inspired by these practical requirements alongside distinct model characteristics, we propose a multi-dimensional benchmark framework that systematically evaluates four crucial dimensions: accuracy, quality, stability, and efficiency. These dimensions are designed to serve dual purposes: they enable comprehensive cross-paradigm evaluation via output-only metrics and function as complementary method for GEO risk screening. Consequently, our framework delivers reliable cross-paradigm evaluations while simultaneously marking latent GEO risk candidates. (Sec. 3.1)

To solve Challenge 2, we propose a unified prompt-enhanced input paradigm that converts structured sequential data for NN-SR models into prompts for LLM-SR models. We extract historical interaction items from the sequential data and populate them chronologically into a tailored prompt template, thereby ensuring consistent evaluation data input for models. This prompt is further enhanced through role-playing instructions and detailed semantic interpretations of the extracted data to fully leverage the reasoning capabilities of LLM. (Sec. 3.2)

To address Challenge 3, we propose a prompt-extractor-coupled mechanism to extract results from LLM-SR outputs. Specifically, input prompts are strategically enhanced through prompt engineering to produce semi-structured outputs. The extractor uses item IDs as anchors to align LLM-SR models outputs with the format of NN-SR models. (Sec. 3.3)

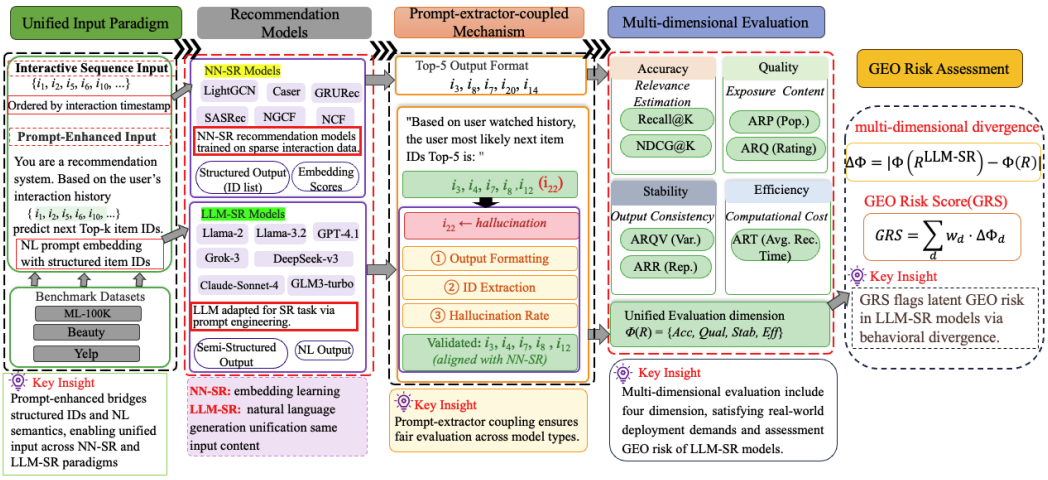


Figure 2: The overall evaluation pipeline of SRBench.

For practical assessment, we benchmark 14 SR models across diverse datasets and reveal critical insights in terms of cross-paradigm applicability, output stability, and GEO vulnerability. Notably, LLM-SR models exhibit a marked popularity-driven exposure tendency even without covert manipulations. They tend to prioritize high-frequency items at the cost of true user preference alignment and signal latent GEO risk. (Sec. 4)

In addition, we discuss the related works in Section 2 and conclude the paper in Section 5.

2 RELATED WORK

Owing to the fundamental differences between NN-SR and LLM-SR models, several benchmarks have been proposed to evaluate SR under emerging LLM-based approaches. OpenP5 (Xu et al., 2024) provides a unified platform for benchmarking generative SR, with an emphasis on prompt construction and item indexing. LLMRec (Liu et al., 2023b) evaluates LLM performance across multiple recommendation tasks by transforming structured data into natural-language prompts. Some researchers (Jiang et al., 2025) further investigate LLM as ranking models in instruction-driven evaluation mode. Other benchmarks, such as RecBench (Liu et al., 2025), explore the impact of different item representation strategies, including IDs, text, and embeddings. In addition, FairEvalLLM and RecLLMs (Deldjoo, 2024; Deldjoo & Nazary, 2024) extend evaluation beyond accuracy by considering additional recommendation properties. Unlike these benchmarks, another line of research evaluates models under GEO threats to measure how strategic optimizations boost target item rankings in LLM-based models. For instance, GEO-Bench (Aggarwal et al., 2024) establishes black-box frameworks to evaluate ranking lifts by manipulating text attributes or injecting keywords. Other works (Nestaas et al., 2025; Nimase et al., 2026) extend this to adversarial method to manipulate LLM generation. Table 1 shows the evaluation dimensions of benchmarks.

Table 1: Evaluation Dimensions across SR Benchmarks.

Benchmark	Accuracy	Quality	Stability	Efficiency	GEO Risk
OpenP5	✓	✗	✗	✗	✗
LLMRec	✓	✗	✗	✗	✗
RecBench	✓	✗	✗	✓	✗
FairEvalLLM	✓	✓	✗	✗	✗
RecLLMs	✓	✗	✓	✗	✗
GEO-Bench	✓	✗	✗	✗	✗
SRBench	✓	✓	✓	✓	✓

3 SRBENCH

To address the challenges summarized in Section 1, we propose SRBench, a comprehensive benchmark for evaluating NN-SR and LLM-SR models (Figure 2), including multi-dimensional metrics, unified input paradigm, and prompt extractor-coupled extraction mechanism.

3.1 MULTI-DIMENSIONAL METRICS

In SR scenarios, the performance of SR models cannot be characterized by accuracy metrics (Jadon & Patil, 2023) alone. Practical deployment gives rise to diverse requirements on SR models, including result relevance, exposure quality, behavior consistency, and computational efficiency.

To this end, we propose a multi-dimensional evaluation that decomposes SR performance into several complementary dimensions, each capturing a distinct aspect of model performance. Rather than relying on a single metric, we operate each dimension through measurable metrics that reflect observable behaviors of SR models. Specifically, our benchmark evaluates SR performance along with four key dimensions: *Accuracy*, *Quality*, *Stability*, and *Efficiency*. Within this framework, we propose new metrics in several dimensions to enable a more comprehensive evaluation across different SR paradigms and GEO risk assessment.

Accuracy. The accuracy measures the model's ability to correctly recommend and rank relevant items for users. We adopt standard top-k ranking metrics, including *Recall@K* (Sajjadi et al., 2018) and *NDCG@K* (Järvelin & Kekäläinen, 2000), to evaluate whether the ground-truth items appear in the recommended list and how well they are ranked. These metrics are widely used in different paradigms, enabling fair comparison. While accuracy captures whether relevant items are recommended and ranked, it cannot depict the effectiveness of SR models in practice. Thus, SRBench treats accuracy as a necessary but insufficient dimension, and complements it with other metrics.

Quality. Recommendation quality evaluates the features of recommended items beyond accuracy. In real-world scenarios, recommendation quality is influenced not only by whether the model recommends high-quality items, but also by whether it disproportionately favors popular ones. Excessive popularity preference may reduce content diversity and amplify exposure bias, while recommending low-quality items may degrade user satisfaction. To capture these aspects, SRBench evaluates recommendation quality from two perspectives: **item popularity and item quality**.

Popularity Perspective. SRBench adopts the widely used *Average Recommendation Popularity (ARP)* (Bai et al., 2025) to measure the popularity characteristics of recommendation results. *ARP* quantifies the average historical popularity of recommended items and is defined as $\boldsymbol{ARP@K} = \frac{1}{U}\sum_{n=1}^{U}\left(\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{P_i}\right)$, where U denotes the number of users, K is the recommendation list length, and P_i is the historical interaction count of the i-th recommended item. Higher *ARP* indicates a stronger tendency to recommend popular items.

Rating Perspective. To complement *ARP*, SRBench introduces *Average Rating Quality (ARQ)*, which measures the average quality of recommended items: $\boldsymbol{ARQ@K} = \frac{1}{U}\sum_{n=1}^{U}\left(\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{Q_i}\right)$, where Q_i denotes the score (e.g., user rating) of the i-th recommended item. Higher *ARQ* indicates that the model recommends items with higher intrinsic quality. Unlike *ARP*, *ARQ* measures whether the recommendation list contains highly rated items.

Stability. While accuracy evaluates the relevance of recommended items and quality assesses the balance of item profiles, neither dimension captures whether the underlying SR behaviors remain consistent. Such behavioral consistency is essential for reliable recommendations. LLM-SR models often exhibit inconsistency due to their inherent randomness. Such unpredictability leads to low model reliability, even when accuracy and quality metrics appear satisfactory, making stability hard to quantify. To evaluate stability from SR results, based on research of stability (Adomavicius & Zhang, 2012) and variable-control theory, SRBench decomposes stability into complementary perspectives that reflect both **rating-level consistency and output-level consistency**. Rating-level consistency reflects intra-list rating homogeneity, where a higher ARQV indicates greater dispersion and lower quality consistency.

From the rating perspective, SRBench presents a stability metric, *Average Rating Quality Variance (ARQV)*, which measures the variance of item rating quality in recommendation lists. *ARQV* evaluates whether an SR model consistently delivers recommendations with similar quality levels, rather than fluctuating between high-rating and low-rating items due to random generation effects. The formulation of *ARQV* is calculated by: $\boldsymbol{ARQV@K} = \frac{1}{U}\sum_{n=1}^{U}\mathbf{Var}_{i=1}^{K}(\boldsymbol{Q_i})$, where U denotes the total number of users, K is the recommendation list length, and Q_i represents the score (e.g., rating) of the i-th recommended item. $\mathrm{Var}(\cdot)$ computes the variance of item rating in each recommendation list, measuring the stability of recommendation quality across different users and scenarios.

From the output perspective, SRBench proposes *Average Recommendation Repetition (ARR)*, which quantifies SR result consistency by measuring the overlap of recommended items across repeated executions. *ARR* captures whether an SR model repeatedly recommends similar items, revealing output-level reproducibility in SR behavior. The formulation of *ARR* is defined as: $\boldsymbol{ARR@K} = \frac{1}{T \times K} \sum_{n=1}^{T} \boldsymbol{Rep_n}$, where T denotes the number of repeated SR executions under identical settings, K is recommendation list length, and Rep_n represents the number of overlapping items between SR results in the n-th execution and those from other runs. By combining *ARQV* and *ARR*, SRBench measures stability from both quality and output perspectives. A lower *ARQV* means more stable quality, while a higher *ARR* reflects greater consistency across multiple recommendation.

Efficiency. Beyond accuracy, quality and stability, computational efficiency is a critical dimension for SR evaluation. In practical deployment scenarios, slow speed of SR process will severely reduce user satisfaction, which forms the need of efficiency evaluation to SR models. To assess efficiency, SRBench focuses on a core metric that quantifies the time cost of SR process. Here, SRBench adopts *Average Recommendation Time (ART)*, which commonly measures the average time for SR models to produce a recommendation list and is defined as follows: $\boldsymbol{ART} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{T_{c_n}}$, where N denotes the total number of SR executions, and T_{c_n} represents the time cost of the n-th execution. *ART* captures how efficient an SR model is, reflecting both algorithmic speed and system responsiveness. A lower ART value indicates higher efficiency under practical constraints.

GEO Risk Assessment. GEO poses security threats by covertly manipulating semantic signals to distort item exposure and influence recommendation outcomes. Assessing this threat under realistic black-box scenarios is highly challenging because GEO-driven manipulations are optimized to maintain or even artificially amplify standard *Accuracy* metrics. Therefore, relying solely on accuracy fails to reflect the risk exposure of LLM-SR models.

To reveal the hidden risk, we leverage a crucial technical insight: while GEO strategies can mask their impact on accuracy, they inevitably disrupt other behavioral dimensions by degrading recommendation quality and increasing output inconsistency. To capture multi-dimensional behavioral divergence and serves as a screening signal for potential GEO risk, we first define a unified evaluation vector over the recommendation results R as: $\Phi(R) = \{\text{Acc}(R), \text{Qual}(R), \text{Stab}(R), \text{Eff}(R)\}$, which integrates performance across all four critical dimensions. Since traditional NN-SR models are inherently immune to such semantic manipulation, we quantify the latent GEO-driven risk as the multi-dimensional divergence from the NN-SR baseline: $\boldsymbol{\Delta\Phi} = |\boldsymbol{\Phi}(\boldsymbol{R}^{\text{LLM-SR}}) - \boldsymbol{\Phi}(\boldsymbol{R})|$, where $R^{\text{LLM-SR}}$ represents the LLM-SR output results, and R stands for the clean NN-SR baseline. To aggregate these heterogeneous discrepancies across multiple dimensions into a unified risk index, the GEO risk is quantified by the proposed *GEO Risk Score (GRS)*: $\boldsymbol{GRS} = \sum_{\boldsymbol{d}} \boldsymbol{w_d} \cdot \boldsymbol{\Delta\Phi_d}$, where $\Delta\Phi_d$ denotes the deviation in dimension d. To prevent dimensions with larger numerical scales from dominating the score, we adopt the log-smoothed adaptive weights derived in Appendix D.

3.2 Unified Input Paradigm

To solve **Challenge 2**, we propose a unified input paradigm with prompt enhancement. It converts structured interaction sequences into standardized natural-language prompts to preserve sequential information for fair cross-paradigm evaluation.

Data Conversion. The backbone of the unified input paradigm is data conversion. After analyzing previous LLM-SR prompt formation (Xu et al., 2024), we first design a general prompt template that contains SR motivation. Then, we extract historical interaction items from user interaction sequence and chronologically embed them to the prompt, which maintains the input consistency between NN-SR and LLM-SR models. However, the prompt only triggers LLM-SR models for SR process, which is insufficient for exhibiting their best performance and requires enhancement.

Prompt Enhancement. Based on the data conversion, we can receive *prompt enhancement* by attaching extra information required for LLM-SR process. Specifically, according to the research on LLM reasoning (Ke et al., 2025) and prompting strategy (Lu et al., 2024), we extend SR objective with role-playing instructions, which proactively activates LLM's knowledge storage in fields of SR. For each item ID in the user interaction sequence, we attach detailed semantic descriptions to the prompt, such as item titles, and descriptions, forming a one-to-one correspondence that enables LLM-SR models to leverage their semantic reasoning capabilities.

3.3 PROMPT-EXTRACTOR-COUPLED MECHANISM

As described in **Challenge 3**, while NN-SR models recommend structured results that are directly for performance evaluation, LLM-SR models generate recommendation results in free-form natural language even containing hallucination as in Appendix C presents. This cannot be directly used for standard evaluation even under explicitly constrained prompts. The absence of a standardized extraction mechanism leads to unreliable results, which is unable to reflect the effectiveness of SR models. Thus, we propose a prompt-enhanced output formatting and numeric extraction mechanism to solve this problem.

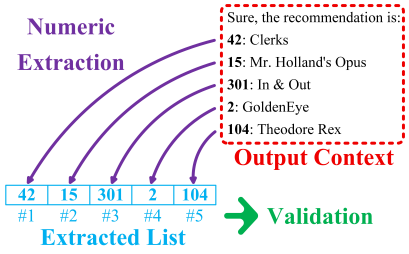


Figure 3: Content Extraction in SRBench.

We implement a prompt-extractor-coupled extraction mechanism to accurately extract SR results from the outputs of LLM-SR models. According to the semi-structured outputs that model generates after taking in enhanced prompt, as shown in Figure 3, the mechanism automatically identifies numerical values among output context as item IDs. Utilizing these item IDs as anchors, the mechanism transforms LLM-SR output context into sequential format as extracted list, which is tightly aligned with NN-SR model's predicted item ID sequence. After extraction, the mechanism individually checks every extracted item ID by comparing with item ID's universal set to verify the validity of recommended items. If an item ID does not exist in the universal set, the mechanism will identify it as a hallucinated item and remove it from results. Thus, this ensures an accurate reflection of SR performance. The output examples are shown in Appendix F.

4 EXPERIMENTS

In this section, we evaluate on real-world datasets to answer the following research questions:

Q1. Does SRBench exhibit universal utility and accurate evaluation across datasets and models?

Q2. How do LLM-SR models perform compared to NN-SR models across multiple dimensions, such as accuracy, quality, stability, and efficiency in SRBench?

Q3. How does SRBench assess the underlying GEO risk through its multi-dimensional evaluation?

Q4. Can SRBench's unified input paradigm effectively convert the structured data into a clear and well-guided prompt, enabling LLM-SR models to perform better?

Q5. Does prompt-extractor-coupled extraction mechanism in SRBench robustly extract the recommendation information in LLM output for subsequent validation?

Q6. How does the length of the interaction sequence influence the performance of LLM-SR models?

The details of the evaluated SR models are in Sec. 4.1, followed by experimental analysis to answer research questions in Sec. 4.2. Finally, Sec. 4.3 summarizes key insights.

4.1 EVALUATED SR MODELS AND SETTINGS

We select and evaluate multiple typical SR models, covering all model paradigms and classes in sequential recommendation. The NN-SR models are *NCF*, *Caser*, *GRURec*, *SASRec*, *NGCF* and *LightGCN*. The LLM-SR models are *Llama-2*, *Llama-3.2*, *GPT-4.1*, *Grok-3*, *DeepSeek-v3*, *GLM-3-turbo*, *Claude-Sonnet-4* and *Claude-Sonnet-4-thinking*. Here, NN-SR models are trained while LLM-SR models operate without recommendation-specific fine-tuning, reflecting current deployment realities. The details of SR models and datasets are listed in Appendix A.

4.2 EXPERIMENTAL PERFORMANCE AND ANALYSIS

The benchmark results on the Yelp, ML-100K and Beauty datasets are presented in Table 2, Table 7 and Table 9. Yelp dataset is selected for analysis in Q2–Q6. Other results are shown in Appendix B.

Table 2: Multi-dimensional Performance Comparison On Yelp Dataset.

Model	Accuracy		Quality		Stability		Efficiency
	Recall@5	NDCG@5	ARP	ARQ	ARQV	ARR	ART(s)
NCF	0.0031	0.0019	**11.26**	**4.13**	0.3295	0.7582	0.584
Caser	0.0153	0.0096	89.72	4.03	**0.1632**	0.7010	0.537
GRURec	0.0295	0.0189	44.72	3.92	0.3081	0.4428	0.370
SASRec	0.0245	0.0156	43.14	3.89	0.4128	0.1935	0.432
NGCF	0.0130	0.0080	24.21	3.96	0.2567	0.5394	0.256
LightGCN	0.0157	0.0097	31.08	4.02	0.1978	**0.8152**	**0.253**
Llama-2	0.0107	0.0056	80.83	3.72	0.6965	0.4500	0.531
Llama-3.2	0.0087	0.0048	70.06	3.52	0.8176	0.4500	0.248
GPT-4.1	0.0296	0.0207	157.83	3.81	0.2742	0.4129	1.399
Grok-3	**0.0656**	**0.0511**	168.32	3.71	0.2637	0.2810	20.668
GLM-3-turbo	0.0215	0.0146	153.93	3.77	0.3646	0.2821	1.724
Claude-Sonnet-4	0.0159	0.0128	60.66	3.03	0.4235	0.4309	2.535
Claude-Sonnet-4-thinking	0.0076	0.0063	103.23	3.72	0.3634	0.2011	9.930
DeepSeek-v3	0.0332	0.0246	200.49	3.82	0.2291	0.4439	1.794
Llama-2 (few-shot-5)	0.0116	0.0057	66.73	3.74	0.7316	0.4500	0.482
Llama-3.2 (few-shot-5)	0.0096	0.0052	53.44	3.53	0.8492	0.4500	0.223
GPT-4.1 (few-shot-5)	0.0368	0.0253	97.14	3.79	0.4009	0.4249	1.503
Grok-3 (few-shot-5)	0.0820	0.0648	101.23	3.69	0.4203	0.3119	15.878
GLM-3-turbo (few-shot-5)	0.0273	0.0170	105.70	3.78	0.4583	0.3041	1.488
Claude-Sonnet-4 (few-shot-5)	0.0112	0.0086	46.96	3.68	0.6022	0.4342	2.806
Claude-Sonnet-4-thinking (few-shot-5)	0.0071	0.0055	84.44	3.59	0.4538	0.2270	8.992
DeepSeek-v3 (few-shot-5)	0.0420	0.0288	150.75	3.82	0.3915	0.4457	1.450

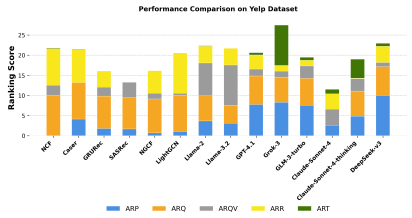


Figure 4: Performance in Quality, Stability and Efficiency Dimension on Yelp Dataset.

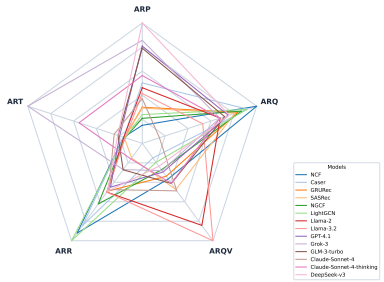


Figure 5: Overall Non-Accuracy Performance of All Models on the Yelp Dataset.

SRBench across Datasets and Models. To answer **Q1**, the empirical results are shown in Table 2, 7, and 9, they reveal that SRBench is capable of evaluating different SR models, including NN-based, local LLM-based and online LLM-based methods. SRBench shows universality on various datasets with different structures and indexing orders. We also evaluate different SR models within SRBench using a unified setting. For example, Caser achieves a Recall@5 of 0.0637 on ML-100K. This ensures internally consistent comparisons across models. These results justify the reliability of SRBench for holistic SR model evaluation across different paradigms.

Multi-dimensional Model Performance. To answer **Q2**, we aggregate performance of SR models by converting raw metrics into normalized ranking scores (14 to 1, descending with ranking). As illustrated in Figure 4 and Figure 5, NN-SR models generally demonstrate better overall performance than LLM-SR models on the Yelp dataset. After summarizing SR model's multi-dimensional performance on various datasets in Table 5 and Appendix E, we find that NN-SR models outperform LLM-SR models in most dimensions, with LLM-SR showing limited advantages. Compared with NN-SR models, LLM-SR models' higher ARP and lower ARQ demonstrate a clear popularity-driven exposure tendency, which implies that LLM-SR models are more inclined to recommend popular items, while overlooking the item quality. Such popularity-driven exposure causes unfair-

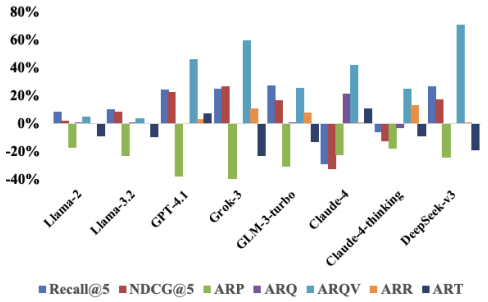


Figure 6: Performance comparison settings of Few-shot and Full-length of LLM-SR on Yelp Dataset.

Table 4: Results of GRS on Yelp Dataset.

Model	*GRS*
Llama-2	0.1188
Llama-3.2	0.1420
GPT-4.1	0.1385
Grok-3	0.7114
GLM-3-turbo	0.1528
Claude-Sonnet-4	0.1412
Claude-Sonnet-4-thinking	0.3407
DeepSeek-v3	0.1728

ness to both items and users, where niche but high-quality items lack equal exposure and users fail to receive premium SR results. This tendency likely stems from LLMs' weaker understanding of item quality over popularity. This gap arises from limited item-specific signals and insufficient collaborative preference modeling capabilities. Other dataset results are presented in Appendix H.

GEO Risk Assessment. To answer **Q3**, the empirical results of Yelp are shown in Table 4. GRS reveals that LLM-SR models exhibit varying degrees of behavioral divergence from robust NN-SR baselines, which serves as a latent risk signal for security screening across various dimensions and domains. For instance, Claude-Sonnet-4-thinking's GRS value evaluated by SRBench reaches 0.6128 on ML-100K(Table 15). These findings demonstrate the effectiveness of SRBench in providing a latent GEO risk assessment. Detailed analysis of GEO risk is shown in Appendix D.

Effectiveness of Unified Input Paradigm. To answer **Q4**, we observe and record 1000 SR processes executed by GPT-4.1 on Yelp dataset by three rounds, checking if the model successfully executes SR process and generates requested results. The input of GPT-4.1 is represented by three types (original indexed data, general SR prompt with motivation, augmented prompt) and the results are shown in Table 3. Our augmented prompt method achieves the highest rate, yielding a 40% relative improvement. This reveals that the unified input paradigm in SRBench is effective for data conversion and prompt enhancement.

Table 3: Performance Comparison on Three Input Paradigms.

Input Paradigm	Success Count
Indexed Data	0
General Prompt	712
Augmented Prompt	998

Robustness of Prompt-extractor-coupled Extraction Mechanism. To answer **Q5**, we evaluate 1,000 pairs of output contexts and extracted results from GPT-4.1 on the Yelp dataset. The results show that merely 5 out of the 1,000 sample pairs exhibit a mismatch, corresponding to a mismatch rate of only 0.5%. Such a result reveals the high robustness of the proposed extraction mechanism for standard evaluation in SRBench. The low mismatch rate further attests to the stability of the extraction mechanism across standard SRBench evaluations. Therefore, the experimental results demonstrate its capability to convert LLM outputs into structured results.

Impact of User Interaction Sequence Length. To answer **Q6**, we compare LLM-SR performance in few-shot (taking 5 latest interacted items as user sequence) and full-length mode on Yelp dataset (Figure 6). Considering the LLM's inherent randomness, few-shot SR process exhibits higher ARQV (representing lower recommendation stability) compared to full-length SR, in accuracy, quality, and efficiency, few-shot SR process performs better than full-length peers. This indicates that the information contained in few-shot interaction sequences (with a length of 5) is already sufficient for LLM-SR models to perform accurate, fair, and efficient SR processes in practical recommendation scenarios. The other results are shown in Appendix G.

4.3 Experiment Insights

Effectiveness of SRBench. SRBench is an effective comprehensive benchmark for SR. Its unified input paradigm transforms data into prompt-enhanced inputs, bridging the input gap between

Table 5: Two Paradigm Comparison across Datasets.

Dataset	Accuracy	Quality	Stability	Efficiency
ML-100K	NN-SR > LLM-SR	NN-SR > LLM-SR	NN-SR > LLM-SR	NN-SR > LLM-SR
Beauty	NN-SR > LLM-SR	NN-SR < LLM-SR	NN-SR > LLM-SR	NN-SR > LLM-SR
Yelp	NN-SR < LLM-SR	NN-SR > LLM-SR	NN-SR > LLM-SR	NN-SR > LLM-SR

different paradigms, while prompt-extractor-coupled extraction mechanism reliably extracts SR results from LLM output context. Combining this unified framework with multi-dimensional metrics, SRBench evaluates diverse models and datasets while screening GEO risks.

Comparison between NN-SR and LLM-SR Models. While NN-SR models still outperform LLM-SR models overall, LLM-SR models show comparable accuracy. This reveals their potential in SR. However, substantial improvements are needed in quality and other dimensions. Inspired by NN-SR algorithms, LLM-SR models require dedicated training to boost quality, stability and efficiency.

Popularity-driven Exposure Tendency. Unlike NN-SR models, LLM-SR exhibits a popularity bias that diminishes quality for items and users. This stems from LLMs' limited comprehension of intrinsic item quality and creates a security gap analyzable through GEO risk assessment. Future research should explore advanced prompt enhancement and alignment strategies for LLM-SR.

GEO Risk Assessment. Our benchmark reveals that LLM-SR models exhibit larger behavioral divergence from NN-SR baselines on quality and stability. Thus, this divergence warrants dedicated GEO-driven security evaluation. Such vulnerability indicates that recommendation results generated by LLM-SR models may be influenced by hidden optimization strategies, compromising the reliability of recommendation in real-world deployment. In the future, deploying comprehensive multi-dimensional evaluation framework is essential to identify potential GEO risks.

LLM-SR Stability Defection. Most LLM-SR models show higher output variability than NN-SR models even when temperature is set to 0, which reflects intrinsic stochasticity of the LLM generation process. This will limit the reliability of LLM-SR models in real-world applications. In future research, LLM-SR models need to consider recommendation-specific training algorithms and fine-tuning methods to achieve more robust stability performance in real-world scenarios.

Impact of Few-shot LLM-SR Process. Despite lower stability, the few-shot interaction sequence provides sufficient signals for LLM-SR models to capture user preferences without long histories. This indicates that LLM-SR models can capture preference patterns from small data. Users can apply this to reduce computational costs and deploy LLM-SR for resource-constrained environments.

Hallucination in LLM-SR models. As discussed in Appendix C, LLM-SR models often hallucinate plausible yet invalid items (e.g., Claude-Sonnet-4-thinking on Yelp dataset). This generative model limitation undermines recommendation reliability, requiring additional safeguards such as item ID verification prior to final output to preserve user trust in real-world deployments.

Impact of LLM Thinking Mode. Although the extended reasoning process in thinking mode of LLM-SR models improves user-perspective quality, it exacerbates their susceptibility to GEO-driven risk and decreases overall performance across critical dimensions. Therefore, it is unnecessary for users to enable the thinking mode in LLM-SR recommendation methods.

5 Conclusion

In this paper, we propose SRBench, a comprehensive benchmark for cross-paradigm SR evaluation. SRBench employs multi-dimensional metrics, a unified input paradigm, and prompt-extractor-coupled mechanism to enable fair comparisons on different SR models. Beyond traditional accuracy evaluation, SRBench provides broader analysis of recommendation quality, stability, efficiency, hallucination behavior, and latent GEO risks. Experimental results reveal key findings, including popularity-driven exposure tendency, and LLM-SR models exhibit divergence from NN-SR baselines under latent GEO risk. We hope SRBench can serve as a foundation for future SR research and developing trustworthy RS. We also discuss limitations and future directions in Appendix I.

6 AI USE STATEMENT

LLM was used strictly for language editing, expression improvement and code checking during manuscript preparation. All core scientific content, methodology, data analysis and conclusions were generated and verified entirely by the authors. The authors retain full responsibility for the scientific integrity and accuracy of the published material.

7 ETHICS STATEMENT

This work involves no human participants, sensitive personal data, or ethical risk factors. All datasets used are publicly available, ensuring privacy and security compliance. Focusing entirely on methodological and computational aspects, this research presents no direct ethical or societal risks.

8 REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our results. The descriptions of the proposed models and algorithms are included in the main text, while additional implementation details, hyperparameter settings, and training procedures are provided in the appendix. The datasets preprocessing steps are clearly documented. In addition, we provide an anonymous repository containing the source code, experiment scripts and configuration files.

REFERENCES

Gediminas Adomavicius and Jingjing Zhang. Stability of recommendation algorithms. *ACM Transactions on Information Systems*, 30(4):1–31, 2012.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5–16, 2024.

Qifeng Bai, Nankai Lin, Meiyu Zeng, Guanqiu Qin, Dong Zhou, and Aimin Yang. Ensuring accuracy and fairness: A De-Biasing framework for sequential recommendation. *User Modeling and User-Adapted Interaction*, 35, 2025.

Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. Sequential recommendation with graph neural networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 378–387, 2021.

Qi Chen, Guohui Li, Quan Zhou, Si Shi, and Deqing Zou. Double attention convolutional neural network for sequential recommendation. *ACM Transactions on the Web*, 16(4):1–23, 2022.

Sinnyum Choi and Woong Kim. Improving the performance of sequential recommendation systems with an extended large language model. *arXiv preprint arXiv:2507.19990*, 2025.

Yashar Deldjoo. Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems*, 2024.

Yashar Deldjoo and Fatemeh Nazary. A normative framework for benchmarking consumer fairness in large language model recommender system. *arXiv preprint arXiv:2405.02219*, 2024.

Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. Sequential user-based recurrent neural network recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys)*, pp. 152–160, 2017.

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

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 173–182, 2017.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020.

Balázs Hidasi and Alexandros Karatzoglou. Recurrent neural networks with Top-K gains for session-based recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 843–852, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Chengkai Huang, Tong Yu, Kaige Xie, Shuai Zhang, Lina Yao, and Julian McAuley. Foundation models for recommender systems: A survey and new perspectives. *arXiv preprint arXiv:2402.11143*, 2024.

Aryan Jadon and Avinash Patil. A comprehensive survey of evaluation techniques for recommendation systems. *arXiv preprint arXiv:2312.16015*, 2023.

Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, 2000.

Theis Jendal, Mads Corfixen, Magnus Olesen, Peter Dolog, Katja Hose, Daniele Dell'Aglio, and Matteo Lissandrini. The Yelp collaborative knowledge graph. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 6414–6419, 2025.

Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles L. A. Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. Beyond utility: Evaluating LLM as recommender. In *Proceedings of the ACM on Web Conference*, pp. 3850–3862, 2025.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *Transactions on Machine Learning Research (TMLR)*, 2025, 2025.

Genki Kusano, Kosuke Akimoto, and Kunihiro Takeoka. Revisiting prompt engineering: A comprehensive evaluation for LLM-based personalized recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys)*, pp. 832–841, 2025.

Geon Lee, Kyungho Kim, and Kijung Shin. Revisiting LightGCN: Unexpected inflexibility, inconsistency, and a remedy towards improved recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys)*, pp. 957–962, 2024.

Jingyu Li, Zhaocheng Du, Qianhui Zhu, Zhicheng Zhang, Wu Wu, Chaolang Li, Pengwen Dai, et al. Collectivekv: Decoupling and sharing collaborative information in sequential recommendation. In *International Conference on Learning Representations*, volume 2026, pp. 38185–38200, 2026.

Zelong Li, Yan Liang, Ming Wang, Sungro Yoon, Jiaying Shi, Xin Shen, Xiang He, Chenwei Zhang, Wenyi Wu, Hanbo Wang, et al. Explainable and coherent complement recommendation based on large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4678–4685, 2024.

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

Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, et al. LLMRec: Benchmarking large language models on recommendation task. *arXiv preprint arXiv:2308.12241*, 2023b.

Qijiong Liu, Jieming Zhu, Lu Fan, Kun Wang, Hengchang Hu, Wei Guo, Yong Liu, and Xiao-Ming Wu. Benchmarking LLMs in recommendation tasks: A comparative evaluation with conventional recommenders. *arXiv preprint arXiv:2503.05493*, 2025.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*, 2024.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. LLM-Rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*, 2023.

Tianhui Ma, Yuan Cheng, Hengshu Zhu, and Hui Xiong. Large language models are not stable recommender systems. *arXiv preprint arXiv:2312.15746*, 2023.

Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 4857–4888, 2025.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Ojas Nimase, Zhe Chen, Gengpei Qi, Yue Zhao, and Xiyang Hu. GEO-Bench: Benchmarking ranking manipulation in generative engine optimization. *arXiv preprint arXiv:2605.29107*, 2026.

Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. A comprehensive review of recommender systems: Transitioning from theory to practice. *Computer Science Review*, 59:100849, 2026.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *arXiv preprint arXiv:1806.00035*, 2018.

Ítallo Silva, Leandro Marinho, Alan Said, and Martijn C. Willemsen. Leveraging ChatGPT for automated human-centered explanations in recommender systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI)*, pp. 597–608, 2024.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1441–1450, 2019.

Jiaxi Tang and Ke Wang. Personalized Top-N sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 565–573, 2018.

Yiming Tang, Yi Fan, Chenxiao Yu, Tiankai Yang, Yue Zhao, and Xiyang Hu. StealthRank: LLM ranking manipulation via stealthy prompt optimization. *arXiv preprint arXiv:2504.05804*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Lei Wang and Ee-Peng Lim. Zero-Shot Next-Item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*, 2023.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 165–174, 2019.

Qianfeng Wen, Yifan Simon Liu, Xin Liu, Difan Jiao, Blair Yang, Junda Wu, and Zhenwei Tang. SafeGEO: Understanding generative engine optimization risks in recommendation agents. *arXiv preprint arXiv:2606.28356*, 2026.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys)*, pp. 12–22, 2024.

Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. OpenP5: An open-source platform for developing, training, and evaluating LLM-based recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 386–394, 2024.

An Yan, Shuo Cheng, Wang-Cheng Kang, Mengting Wan, and Julian McAuley. CosRec: 2D convolutional neural networks for sequential recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2173–2176, 2019.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:67686–67700, 2023.

Xucheng Yu, Haibo Jin, Huimin Zeng, and Haohan Wang. SCI-Defense: Defending manipulation attacks from generative engine optimization. *arXiv preprint arXiv:2605.21948*, 2026.

Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M. Jose, and Xiangnan He. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 582–590, 2019.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 993–999, 2023a.

Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4741–4753, 2022.

Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu, and Liang Wang. Learning long- and short-term representations for temporal knowledge graph reasoning. In *Proceedings of the ACM Web Conference*, pp. 2412–2422, 2023b.

Xiaolin Zheng, Jiajie Su, Weiming Liu, and Chaochao Chen. DDGHM: Dual dynamic graph with hybrid metric training for cross-domain sequential recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 471–481, 2022.

Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, et al. When large vision language models meet multimodal sequential recommendation: An empirical study. In *Proceedings of the ACM on Web Conference*, pp. 275–292, 2025.

A Additional Information

In this section, we provide the additional information of the research process.

A.1 Prompt Details

General Prompt

A user has the following watched item history (internal Item IDs): **{Interaction_sequence}**.
Based on this history, predict ONLY the top **{Recommendation_length}** most likely next item IDs for this user.

We design **General Prompt** to provide the essential contextual information. This prompt supplies the model with the user's historical interaction sequence in terms of item IDs, and instructs it to predict most likely next items. Additional context information like item metadata and user profile are not included. It is obvious that this approach fails to fully utilize the semantic understanding capabilities of LLM for effective recommendations. Therefore, we need to design a new prompt.

Augmented Prompt after Prompt Enhancement

You are an accurate recommendation system for the **{Dataset_name}**.
A user with ID **{User_id}** has the following watched item history (internal Item IDs) with item information: **{Interaction_sequence_with_item_info}**.
Based on this history, predict ONLY the top **{Recommendation_length}** most likely next item IDs for this user.
Each item ID must be an integer between 1 and **{Total_num_of_items}**.
Output the predicted item IDs in order of likelihood, from most to least likely.
Important: Do not include any additional text, explanations, or formatting!
Output format: A single line of exactly **{Recommendation_length}** comma-separated integers.
No explanation, no text, no line breaks.
Example (correct format):
42,15,301,2,104
Now generate the output:

To further exploit the potential of LLMs in leveraging contextual information in SR task, We design an **Augmented Prompt after Prompt Enhancement** mechanism to incorporates additional information. As shown above, this prompt enhances the general setting with three key augmentations: **(1)** user identity for personalized modeling; **(2)** item information for semantic understanding; and **(3)** numerical constraints that enforce a strict output format of comma-separated integers. This ensures outputs are directly parsable and comparable with standard evaluation metrics.

A.2 Datasets

In this paper, we use three datasets to evaluate our benchmark. The details of datasets are described in:

- **MovieLens-100K** (ML-100K) (Harper & Konstan, 2015), is one of the most classic benchmark dataset that contains 100,000 movie ratings from 943 users on 1,682 items, representing the entertainment domain of the SR application.
- **Beauty** (Ni et al., 2019) is a commonly used e-commerce dataset consisting of 371,345 user-item interactions involving 1,398 users and 32,488 products. Compared to ML-100K,

Table 6: Dataset Details

Datasets	Ratings	Users	Items	Item Attributes
ML-100K	100,000	943	1,682	Title, Rating, Release Date, Category
Beauty	371,345	1,398	32,488	Title, Overall (Rating), Rank, Description
Yelp	6,990,280	851,921	150,346	Name, Stars (Rating), Category

it reflects typical user purchasing behaviors in marketing domain, which enriches the variety of datasets.

- **Yelp** (Jendal et al., 2025) includes 6,990,280 rating interactions from 851,921 users on 150,346 businesses. Compared with the other two datasets, Yelp is significantly larger in scale, which enables benchmark in large-scale SR scenarios.

We apply a unified preprocessing workflow across all datasets and sort all user interaction records strictly by timestamp attribute. We partition each interaction list into training and test with a 9 to 1 ratio. The first portion holds 90 percent of the items and acts as historical input. The first item of the remaining portion acts as the ground truth target for evaluation. This paper adjusts sequence lengths to match varying dataset scales. We cap the sequence length at 50 for the ML 100K and Beauty datasets. SRBench sets the maximum length to 100 for the Yelp dataset. We also filter out inactive users to ensure robust evaluation. This paper removes users with fewer than 5 interactions in ML 100K and Beauty. We exclude users with fewer than 10 interactions in Yelp.

In summary, we conduct extensive empirical evaluations of sequential recommendation models across three diverse datasets. Our benchmark incorporates MovieLens-100K, Beauty, and Yelp to capture distinct domain characteristics in real world scenarios. Table 6 presents the detailed statistics for each dataset, including the number of ratings, users, items and item attributes.

A.3 Summary of Recommendation Models

NN-SR Models. NN-SR models rely on explicit sequential modeling over structured user-item interaction data and can be further grouped into four representative subcategories: **1) CNN-based methods**, such as Caser (Tang & Wang, 2018) and its extensions (Yan et al., 2019; Yuan et al., 2019) with (Chen et al., 2022), apply convolutional kernels to capture local, short-range sequential patterns within fixed windows. **2) RNN-based approaches**, including GRU-based models (Donkers et al., 2017; Hidasi & Karatzoglou, 2018), are designed to capture long-term dependencies in interaction sequences. **3) Graph-based methods**, which further incorporate higher-order user-item relationships. Existing work such as NGCF (Wang et al., 2019) and LightGCN (He et al., 2020; Lee et al., 2024) propagate collaborative signals over graphs, while recent advances explore dynamic graph structures (Yu et al., 2023) for temporal evolution (Chang et al., 2021; Zhang et al., 2022) and (Zhang et al., 2023b; Zheng et al., 2022). **4) Transformer-based methods**, which leverage self-attention to capture item dependencies. Existing works like SASRec (Kang & McAuley, 2018) and BERT4Rec (Huang et al., 2024; Sun et al., 2019) are employed to model interactions and explore pre-training strategies to improve performance.

LLM-SR models. For LLM-SR models, LLMs are often explored as the backbone for SR task, leading to a paradigm shift from explicit sequence modeling to prompt-based semantic reasoning. Under this paradigm, prompt-based methods, such as Chat-REC (Gao et al., 2023) and NIR (Wang & Lim, 2023). In addition, related approaches (Liu et al., 2023a; Lyu et al., 2023) and (Silva et al., 2024; Xi et al., 2024), reformulate user interaction histories into natural-language prompts to guide LLMs in generating SR results. Other studies further adopt training techniques like LoRA (Hu et al., 2021) to adapt LLMs for recommendation.

Existing benchmarks still suffer from three major limitations in Section 2. First, as summarized in Table 1, they mainly emphasize accuracy while overlooking other critical dimensions for real-world deployment, such as quality, stability, and efficiency. Second, These benchmarks focus on recommendation evaluation, while GEO benchmarks lack a unified framework for assessing GEO risks in domain of SR. Third, current benchmarks fail to provide unified input paradigms and standardized output extraction, preventing a fair, cross-paradigm comparison between NN-SR and LLM-SR models. These limitations motivate the development of SRBench.

A.4 Evaluated SR Models and Settings

The NN-SR models are:

- **NCF** (He et al., 2017) is a classical collaborative filtering model and has been extensively used as a standard baseline in SR benchmarks. Following common evaluation mode, we include NCF to ensure comparability with previous work.
- **Caser** (Tang & Wang, 2018) is convolution-based SR models, which explicitly model local sequential patterns through convolutional operations. We include Caser as a representative of convolutional neural network for sequence modeling.
- **GRURec** (Hidasi & Karatzoglou, 2018) is a widely adopted recurrent neural network-based model for recommendation tasks. It has been evaluated across multiple benchmarks as a strong baseline for capturing temporal dependencies.

  SASRec (Kang & McAuley, 2018) is a transformer-based SR model that adopts unidirectional self-attention to model long-range dependencies. It is the transformer-based modeling paradigm, which has become a dominant approach.
- **NGCF** (Wang et al., 2019) and **LightGCN** (He et al., 2020) are graph-based recommendation models that exploit user-item interaction graphs. NGCF represents early graph neural network-based methods, while LightGCN is a simplified and more efficient variant that has achieved state-of-the-art performance in SR benchmarks. These models enable a comprehensive evaluation of graph-based methods.

For specific implementation, all experiments are conducted in the following environment: Operating System is Ubuntu 20.04 and programming language is Python 3.10. Libraries and frameworks are Transformers 4.45.2 and PyTorch 2.3.0. For NN-SR models, the learning rates are set as follows: NCF, GRURec and Caser are $5 * e^{-3}$, SASRec is e^{-3}, and NGCF and LightGCN are e^{-4}. To mostly reduce randomness, the temperature values of LLM models are set to 0. All experiments were repeatedly run 10 times and the average values were taken.

Based on deployment method, LLM-SR models are divided into two categories: *Local LLMs* and *Online LLMs*.

- **Local LLMs.** We consider Llama-series LLMs due to their open-source nature and widespread adoption in SR. Specifically, we locally deploy **Llama-2** and **Llama-3.2** (Choi & Kim, 2025; Touvron et al., 2023). Llama-2 serves as a widely used and well-established baseline evaluated in recommendation benchmarks, ensuring comparability with existing work. Llama-3.2 represents a more recent and stronger variant, which facilitates a comparison of different model scales within the same transformer architecture.
- **Online LLMs.** We also select **GPT-4.1**, **Grok-3**, **DeepSeek-v3**, and **GLM-3-turbo** (Zhou et al., 2025) as representative commercial LLMs, as they are widely regarded as state-of-the-art proprietary LLM-SR models in benchmarks. We also include **Claude-Sonnet-4** and its thinking mode to represent LLMs with explicit reasoning mechanisms (Kusano et al., 2025). While most selected models focus on direct generation, Claude-Sonnet-4 can test whether extra enhanced reasoning function brings additional benefits for SR, thereby complementing models without reasoning designs.

B Additional Experiment Results and Analysis

In this section, we demonstrate additional experiment results and make some further analysis based on them.

B.1 Results on ML-100K Datasets

Table 7 presents a comprehensive performance comparison of all evaluated models on the **ML-100K dataset**, a classical small-scale benchmark widely used in sequential recommendation research. Despite its small size, ML-100K remains an important testbed for analyzing model behavior under relatively sparse user-item interaction and limited item diversity.

Table 7: Multi-dimensional Performance Comparison On ML-100K Dataset.

Model	Accuracy		Quality		Stability		Efficiency
	Recall@5	NDCG@5	ARP	ARQ	ARQV	ARR	ART(s)
NCF	0.0480	0.0303	22.63	3.616	0.1424	0.7415	0.106
Caser	0.0637	0.0400	22.13	3.554	0.0770	0.6863	0.091
GRURec	**0.0966**	**0.0624**	**18.70**	3.536	0.1098	0.5388	0.118
SASRec	0.0731	0.0445	18.83	3.383	0.0928	0.6234	0.246
NGCF	0.0630	0.0383	358.14	3.797	**0.0745**	0.8356	0.007
LightGCN	0.0413	0.0265	469.38	3.845	0.1018	0.8813	**0.003**
Llama-2	0.0037	0.0021	139.13	3.425	0.3273	0.8208	4.379
Llama-3.2	0.0035	0.0018	97.73	3.232	0.4672	**0.8818**	2.528
GPT-4.1	0.0293	0.0194	375.22	**4.095**	0.1245	0.8207	1.418
Grok-3	0.0269	0.0173	315.16	3.996	0.1261	0.6231	8.754
GLM-3-turbo	0.0053	0.0034	162.06	3.351	0.5504	0.2830	2.320
Claude-Sonnet-4	0.0328	0.0219	416.48	4.123	0.0882	0.7518	6.267
Claude-Sonnet-4-thinking	0.0286	0.0200	403.81	4.078	0.1331	0.6304	12.932
DeepSeek-v3	0.0194	0.0141	435.58	4.073	0.0907	0.8756	1.554
Llama-2 (few-shot-5)	0.0068	0.0044	121.29	3.379	0.3117	0.8229	3.220
Llama-3.2 (few-shot-5)	0.0042	0.0039	83.34	3.182	0.6480	0.8783	1.594
GPT-4.1 (few-shot-5)	0.0238	0.0151	376.70	4.079	0.0905	0.8028	1.247
Grok-3 (few-shot-5)	0.0217	0.0130	316.75	3.897	0.1519	0.5670	9.327
GLM-3-turbo (few-shot-5)	0.0083	0.0052	136.50	3.331	0.4511	0.3031	1.735
Claude-Sonnet-4 (few-shot-5)	0.0276	0.0183	431.09	4.123	0.0575	0.7738	7.840
Claude-Sonnet-4-thinking (few-shot-5)	0.0246	0.0168	430.13	4.094	0.1008	0.6650	12.149
DeepSeek-v3 (few-shot-5)	0.0199	0.0143	458.28	4.093	0.0651	0.8820	1.436

Accuracy. On ML-100K, NN-SR models achieve stronger ranking performance compared to locally deployed LLM-SR models. Specifically, GRURec attains the highest Recall@5 and NDCG@5 among NN-based methods, while graph-based models such as NGCF and LightGCN remain competitive but do not surpass GRURec. Online LLM-SR models, including GPT-4.1, Grok-3, and DeepSeek-v3, show varying degrees of ranking effectiveness, with GPT-4.1 achieving the best Recall@5 among LLMs, yet still falling short of the top NN-SR performance on this small-scale dataset. Llama-2 and Llama-3.2 perform the worst in terms of ranking accuracy, which highlights the impact of model scale and data diversity in smaller recommendation scenarios.

Quality. NN-SR models achieve lower ARP values, which means less popularity bias. Notably, NCF yields the lowest ARP alongside the highest ARQ, suggesting a more balanced exposure across items while rigorously preserving overall recommendation quality. LLM-SR models exhibit higher ARP values that reflects a tendency to over-recommend popular items. This trend not only aligns with observations on larger datasets but also becomes even more amplified here due to the constrained data scale, inherently exposing a structural vulnerability that invites adversarial GEO manipulation.

Stability. The experimental results reveal differences in stability across paradigms. LightGCN shows the highest ARR among NN-SR models, which indicates highly reproducible recommendation lists across repeated runs. Most LLM-SR models achieve lower ARR and higher ARQV. It reflects greater variability in both recommendation quality and output composition. Obviously, this instability stick under deterministic generation. It also suggests that the variability is intrinsic to the generation process of LLM-SR models.

Efficiency. NN-SR models require substantially less inference time, with LightGCN achieving the lowest ART, confirming their suitability for latency-sensitive applications. On the other hand, LLM-SR models bring higher computational costs, particularly Grok-3 and Claude-Sonnet-4-thinking, which limits their practicality for small-scale, real-time recommendation scenarios.

GEO Risk Assessment. As reported in Table 8, LLM-SR models exhibit distinct vulnerabilities under GEO risk assessment. Notably, advanced reasoning capabilities do not inherently confer robustness against targeted item manipulation. Among the evaluated models, Claude-Sonnet-4-thinking records the highest GRS value of 0.6128, closely followed by Grok-3 (0.4322). Although

it can confirm actual susceptibility still needs attack-based evaluation, this identifies models with high GRS as warranting further GEO-specific security evaluation. Furthermore, when the explicit reasoning mechanism is enabled (as seen in Claude-Sonnet-4-thinking and Claude-Sonnet-4), the GRS increases significantly from 0.3350 to 0.6128. This finding reveals that extended thinking processes may cause models to over-interpret adversarial cues, thereby amplifying exposure bias and reduce overall recommendation quality under strategic GEO manipulations. This underscores that GRS captures multi-dimensional behavioral divergence, not direct attack vulnerability.

Table 8: Results of GEO Risk Assessment on ML-100K Dataset.

Model	$\boldsymbol{GRS}$
Llama-2	0.3038
Llama-3.2	0.2542
GPT-4.1	0.1362
Grok-3	0.4322
GLM-3-turbo	0.3216
Claude-Sonnet-4	0.3350
Claude-Sonnet-4-thinking	0.6128
DeepSeek-v3	0.1812

These results reflect a stable balance of advantages and limitations across paradigms on ML-100K. Specifically, NN-SR models provide more balanced performance across accuracy, quality, stability, and efficiency. But LLM-SR models demonstrate competitive in certain accuracy metrics, generally fall in quality, stability, and efficiency on small datasets. It also simultaneously flags latent risk signals warranting security assessment. This highlights the importance of conducting multi-dimensional evaluations on classical benchmarks to attain a comprehensive understanding of model behaviors.

B.2 Results on Beauty Datasets

Table 9 presents the evaluation results on the **Beauty dataset**, which captures user interactions in e-commerce domain with medium scale compared to ML-100K and Yelp. This allows us to examine model behavior under typical commercial recommendation scenarios.

Accuracy. NN-SR models generally achieve best ranking results, with GRURec attaining the highest Recall@5 and NDCG@5 among neural networks baselines. Graph-based methods such as NGCF and LightGCN remain competitive, yet their ranking performance doesn't surpass GRURec. Although LLM-based models still fail to consistently outperform the high performance NN-SR models, some online LLM-SR models including GPT-4.1 and DeepSeek-v3 show noticeable accuracy improvements over locally deployed LLMs like Llama-2 and Llama-3.2. This demonstrates that while LLM-SR models can generate high-quality recommendations, their advantages are less evident on datasets of moderate scale.

Quality. Considering exposure distribution quality, NN-SR models maintain relatively lower ARP values, which implies a mitigated popularity bias. Remarkably, NCF achieves the lowest ARP alongside the highest ARQ, which suggests a more balanced and high-quality exposure across diverse items. On the other hand, ARP values are not uniformly higher in LLM-SR models, which reveals a more nuanced exposure-quality comparison. This popularity bias inherently exacerbates their vulnerability to strategic GEO manipulations, which can bring trade-off between ranking accuracy and recommendation quality. In a word, models that excel in ranking performance do not automatically ensure balanced or robust item exposure quality. Therefore, evaluating SR models solely based on ranking accuracy is insufficient, a comprehensive assessment of exposure quality is essential for developing trustworthy sequential recommendation systems.

Stability. Stability analysis reveals that NN-SR models like LightGCN, produce highly consistent recommendation lists across repeated runs. It highlights predictable and reliable behavior. Conversely, LLM-SR models exhibit greater variability, with noticeable instability in both recommendation composition and ranking quality. This instability persists even under deterministic generation, suggesting that the variability arises from the LLM generation mechanism itself rather than random

Table 9: Multi-dimensional Performance Comparison On Beauty Dataset.

Model	Accuracy		Quality		Stability		Efficiency
	Recall@5	NDCG@5	ARP	ARQ	ARQV	ARR	ART(s)
NCF	0.4549	0.3441	128.80	4.657	0.1065	**0.9000**	0.075
Caser	0.4964	0.4755	382.07	**4.684**	**0.0693**	0.3661	0.476
GRURec	**0.5207**	**0.4992**	589.54	4.438	0.2058	0.6368	0.658
SASRec	0.5186	0.4875	607.09	4.474	0.1878	0.6787	0.058
NGCF	0.1994	0.0923	544.60	4.473	0.1820	0.7215	0.017
LightGCN	0.2504	0.1391	807.27	4.445	0.1705	0.7818	**0.004**
Llama-2	0.0665	0.0381	122.97	4.415	0.7885	0.4499	0.597
Llama-3.2	0.2031	0.1072	451.26	4.613	0.3896	0.4504	0.231
GPT-4.1	0.3415	0.2301	283.13	4.485	0.4188	0.4268	2.026
Grok-3	0.1800	0.1454	210.12	4.435	0.4667	0.3583	5.547
GLM-3-turbo	0.1248	0.0894	139.85	4.029	0.6186	0.2536	2.302
Claude-Sonnet-4	0.3265	0.2314	**50.06**	4.351	0.6153	0.3967	7.710
Claude-Sonnet-4-thinking	0.2462	0.1831	93.66	4.435	0.5034	0.2604	10.167
DeepSeek-v3	0.2150	0.1873	207.65	4.411	0.4883	0.4427	1.408
Llama-2 (few-shot-5)	0.0680	0.0385	123.12	4.464	0.6459	0.4500	0.487
Llama-3.2 (few-shot-5)	0.2103	0.1117	453.75	4.613	0.3917	0.4504	0.211
GPT-4.1 (few-shot-5)	0.3364	0.2296	277.78	4.477	0.4208	0.4247	1.104
Grok-3 (few-shot-5)	0.2023	0.1541	182.97	4.397	0.5259	0.3009	8.730
GLM-3-turbo (few-shot-5)	0.1296	0.0913	116.34	3.746	0.7152	0.2572	2.360
Claude-Sonnet-4 (few-shot-5)	0.3810	0.2568	44.83	4.410	0.5704	0.4407	2.527
Claude-Sonnet-4-thinking (few-shot-5)	0.1792	0.1517	63.55	4.375	0.4801	0.2288	9.457
DeepSeek-v3 (few-shot-5)	0.2209	0.1901	208.15	4.406	0.4931	0.4464	1.350

Table 10: Results of GEO Risk Assessment on Beauty Dataset.

Model	GRS
Llama-2	0.2569
Llama-3.2	0.1226
GPT-4.1	0.2007
Grok-3	0.4287
GLM-3-turbo	0.3043
Claude-Sonnet-4	0.5440
Claude-Sonnet-4-thinking	0.6790
DeepSeek-v3	0.2048

factors. Although LLM-SR models can occasionally outperform NN-SR models in accuracy, their reproducibility remains less certain.

Efficiency. In terms of inference time, NN-SR models are markedly faster with LightGCN achieving the lowest ART that demonstrates suitability for time-sensitive deployment. In contrast, LLM-SR models show higher computational costs, particularly Grok-3 and Claude-Sonnet-4-thinking, which constrains their practical use in real-time recommendation settings. While achieving higher accuracy, LLM-SR models may exhibit disproportionately high computational costs and achieve unsatisfactory result.

GEO Risk Assessment. As illustrated in Table 10, the GRS metrics on the Beauty dataset quantify the magnitude of GEO-driven changes. Overall, models display varying degrees of sensitivity to adversarial manipulation. Notably, Grok-3 (0.4287) and Claude-Sonnet-4 (0.5440) exhibit substantial structural deviations. This suggests that advanced models remain highly susceptible to strategic content manipulation within domain-specific recommendation scenarios. In contrast, models like Llama-3.2 (0.1226) and DeepSeek-v3 (0.2048) demonstrate strong domain inertia and maintain lower risk scores. In addition, Claude-Sonnet-4-thinking shows a notable change in risk score, the value rises from 0.5440 to 0.6790. It suggests that extended reasoning chains tend to amplify latent risk of GEO.

Table 11: Hallucination Rates of Different LLMs on Various Datasets.

Model	ML-100K		Beauty		Yelp	
	full	few-shot-5	full	few-shot-5	full	few-shot-5
Llama-2	0.0002	0.0007	0.0093	0.0003	0.0002	0.0003
Llama-3.2	0.0002	0.0007	0.0003	0.0744	0.0009	0.0022
GPT-4.1	0.0002	0.0000	0.0007	0.0010	0.0015	0.0048
Grok-3	0.0214	0.0196	0.0057	0.0075	0.0236	0.0135
DeepSeek-v3	0.0000	0.0000	0.0025	0.0026	0.0011	0.0103
GLM-3-turbo	0.0001	0.0003	0.1088	0.1658	0.0081	0.0069
Claude-Sonnet-4	0.0003	0.0008	0.0103	0.0014	0.1752	0.0006
Claude-Sonnet-4-thinking	0.0010	0.0001	0.0088	0.0269	0.0087	0.0360

In summary, these results show the balance of strengths and weaknesses across model paradigms on the Beauty dataset. Specifically, NN-SR models consistently deliver a balanced and robust profile across accuracy, quality, stability, and efficiency dimensions. Although LLM-SR models occasionally achieve competitive ranking accuracy, they generally fall short in terms of recommendation quality, prediction stability, and inference efficiency under this medium-scale dataset. This also simultaneously indicates highly unstable under GEO manipulations. These empirical findings clearly highlight the necessity of multi-dimensional evaluation even in datasets that are not as large as Yelp to obtain a comprehensive understanding of model behavior.

B.3 Input Method Comparison for Stable SR Processes

We systematically evaluate the impact of different input paradigms on the stability of SR process for LLMs. Specifically, we compare three types of input methods: (1) Indexed Data, which directly feeds indexed historical interactions to the model; (2) General Prompt, which employs basic natural language prompts without additional structuring; and (3) SRBench Prompt, our proposed prompt-enhanced approach designed to standardize and enrich model input for sequential recommendation tasks.

The number of stable SR processes achieved under each input method is summarized in Table 3. Notably, models using Indexed Data fail to produce any stable SR processes and yield a stability count of 0. We think that naive direct data feeding is insufficient to generate reproducible recommendations. By contrast, the General Prompt method substantially improves stability and achieves 712 stable SR processes, which means that even minimal prompt-based guidance can meaningfully enhance model consistency. The SRBench Prompt further increases stability to 998, so it means that our enhanced prompt design not only supports model comprehension of user histories but also effectively standardizes the output format, leading to the highest reproducibility across runs.

To sum up, these results emphasize the critical role of input design in LLM-SR models, and they highlight that careful prompt engineering can clearly improve stability without altering the underlying model architecture. The observed improvements align with our broader evaluation objectives, which proves the necessity of standardized input paradigms for fair comparisons among SR models.

C Hallucination of LLM-SR Models

Table 11 summarizes model hallucination rates across datasets. We find that Llama-2 and Llama-3.2 exhibit minimal hallucinations (0.01%) in full-length settings. However, Llama-3.2 exhibits a 7.44% hallucination rate on Beauty under few-shot settings, which shows sensitivity to limited context. Online models like DeepSeek-v3 obtain consistently low rates, whereas GLM-3-turbo shows the top few-shot rate (16.58% on Beauty) and Claude-Sonnet-4 gets at 17.52% in Yelp full-length mode.

These results highlight a clear divergence between open-source LLM and commercial models in terms of output reliability. While certain online LLMs achieve high recommendation accuracy and low hallucination rates, others incorporating complex reasoning or scaling mechanisms are prone to generating invalid items under limited interaction histories. This hallucinatory tendency correlates

with high susceptibility to adversarial GEO optimizations, as over-interpreting semantic cues triggers non-existent recommendations. This underscores the importance of validation mechanisms in LLM-SR models to safeguard overall recommendation quality and prediction correctness.

D Empirical Analysis of GEO Risk Score

Theoretical analysis and method. LLMs face several security risks in GEO environments. Attackers can manipulate items or distort exposure to improve target visibility. Hence, We design the GEO Risk Score (GRS) to signal these potential risks. The GRS metric quantifies how much an LLM-SR recommendation model deviates from traditional neural networks baselines (NN-avg). Due to space constraints in the tables, we have abbreviated Claude-Sonnet-4 and Claude-Sonnet-4-thinking to Claude-4 and Claude-4-think in Table 12, Table 13, and Table 14.

First of all, we have defined GRS as the weighted sum of absolute state deviations across four evaluation dimensions:

$$\text{GRS} = \sum_{d \in \mathcal{D}} w_d \cdot \Delta\Phi_d \tag{1}$$

The vector $\Delta\Phi_d$ captures the directional shift of the LLM model relative to neural networks baselines:

$$\Delta\Phi_d = |\Phi_d(R^{\text{LLM-SR}}) - \Phi_d(R_{\text{NN-avg}})| \tag{2}$$

The set $\mathcal{D} = \{\text{Acc}, \text{Qual}, \text{Stab}, \text{Eff}\}$ contains all four evaluation dimensions.

Normalization and GEO Risk assessment. To quantify GEO-driven risk across different recommendation dimension, we first normalize all evaluation metrics into a unified dimensionless interval $[0, 1]$. This normalization enables different dimension metrics to be comparable and allows the measurement of dimension-specific deviations under GEO manipulation. Specifically, each dimension score Φ_d is computed based on the following normalization rules.

Accuracy Dimension (Acc): The accuracy dimension characterizes recommendation effectiveness from the top-k ranking perspective. Recall and NDCG are normalized with respect to their maximum values observed in each dataset:

$$\Phi_{\text{Acc}} = 0.5 \cdot \left(\frac{\text{Recall@5}}{\text{Recall}_{\max}}\right) + 0.5 \cdot \left(\frac{\text{NDCG@5}}{\text{NDCG}_{\max}}\right) \tag{3}$$

Quality Dimension (Qual): The quality dimension is designed for capturing GEO-induced recommendation distortion from both exposure imbalance and item quality perspectives. ARP measures excessive popularity concentration, while ARQ reflects the semantic quality of recommended items:

$$\Phi_{\text{Qual}} = 0.7 \cdot \left(\frac{\text{ARP}}{\text{ARP}_{\max}}\right) + 0.3 \cdot \left(1 - \frac{\text{ARQ} - \text{ARQ}_{\min}}{\text{ARQ}_{\max} - \text{ARQ}_{\min}}\right) \tag{4}$$

Stability Dimension (Stab): The instability dimension measures generation uncertainty across repeated recommendation runs. ARQV captures continuous quality variation and ARR reflects discrete recommendation overlap:

$$\Phi_{\text{Stab}} = 0.6 \cdot \left(\frac{\text{ARQV}}{\text{ARQV}_{\max}}\right) + 0.4 \cdot (1 - \text{ARR}) \tag{5}$$

Efficiency Dimension (Eff): For efficiency dimension, it evaluates computational overhead introduced during recommendation generation. Generation latency is normalized according to the maximum latency observed in the dataset:

$$\Phi_{\text{Eff}} = \frac{\text{ART} - 0}{\text{ART}_{\max} - 0} = \frac{\text{ART}}{\text{ART}_{\max}} \tag{6}$$

After obtaining normalized dimension scores, we quantify GEO risk by measuring the deviation between the original recommendation behavior and the GEO-manipulated behavior. Here, GRS is a latent risk indicator: it measures the behavioral divergence between a given LLM-SR model and

Table 12: Mapped Dimension Values, Absolute State Deviations and GEO Risk Scores on Yelp Dataset.

Dimension	Llama-2	Llama-3.2	GPT-4.1	Grok-3	GLM-3-turbo	Claude-4	Claude-4-think	DeepSeek-v3
Φ_{Acc}	0.1363	0.1133	0.4282	1.0000	0.3067	0.2464	0.1196	0.4938
Φ_{Qual}	0.3950	0.4110	0.6381	0.7016	0.6362	0.5118	0.4737	0.7848
Φ_{Stab}	0.7311	0.8200	0.4361	0.4811	0.5547	0.5384	0.5862	0.3906
Φ_{Eff}	0.0258	0.0120	0.0677	1.0000	0.0834	0.1226	0.4805	0.0868
$\Delta\Phi_{\text{Acc}}$	0.0960	0.1190	0.1959	0.7677	0.0744	0.0141	0.1127	0.2615
$\Delta\Phi_{\text{Qual}}$	0.2145	0.2305	0.4576	0.5211	0.4557	0.3313	0.2932	0.6043
$\Delta\Phi_{\text{Stab}}$	0.3571	0.4460	0.0621	0.1071	0.1807	0.1644	0.2122	0.0166
$\Delta\Phi_{\text{Eff}}$	0.0061	0.0076	0.0481	0.9804	0.0638	0.1030	0.4609	0.0672
GRS	**0.1188**	**0.1420**	**0.1385**	**0.7114**	**0.1528**	**0.1412**	**0.3407**	**0.1728**

NN-SR models across dimensions without any GEO attack simulation. A higher GRS means greater susceptibility to GEO-driven exposure shifts, not confirmation of an active attack.

For Φ_{Qual}, we set $w_{\text{ARP}} = 0.7$ and $w_{\text{ARQ}} = 0.3$, the reason is popularity manipulation represents the dominant attack vector in GEO security scenarios. For Φ_{Stab}, (ARQV, 0.6) is prioritized over discrete duplication (ARR, 0.4) because of its higher measurement sensitivity to generative stochasticity. Each factor receives a dimension weight reflecting its relative importance. These weights account for how GEO variations manifest across various recommendation scenarios. Equal weights are used when two metrics provide complementary information with comparable importance and similar contribution levels. Therefore, larger weights are assigned to dimensions that capture dominant sources of GEO risk impact and exhibit stronger influence on recommendation.

For the accuracy dimension, Recall and NDCG are assigned equal weights of 0.5 because they reveal complementary aspects of ranking effectiveness, including retrieval coverage and ranking effect under different recommendation conditions. Recall means the ability to retrieve relevant items and NDCG considers their ranking positions and relative importance.

ARP receives a higher weight of 0.7 and ARQ receives a weight of 0.3 in quality dimension. This design highlights popularity concentration as a primary indicator of exposure distortion and recommendation bias. Meanwhile, item quality provides an additional semantic constraint for evaluating recommendation reliability.

For the stability dimension, ARQV receives a weight of 0.6 and ARR receives a weight of 0.4. ARQV captures fine-grained stochastic variations across repeated generations. ARR reflects coarse-grained recommendation overlap patterns and discrete output consistency.

Derivation of Log Smoothed Normalization Weights. We adopt naive inverse weighting scheme to use $w_d = \frac{1}{\mu(\Phi_d)}$. This scheme creates a major numerical problem. The reason is that single dimension then accounts for high percentage of the total weight and such extreme dominance hides critical changes in accuracy and bias. We introduce a log smoothed normalization scheme to fix this imbalance. The adaptive weight w_d is defined by: $\boldsymbol{w_d} = \frac{\ln\left(1+\frac{1}{\mu(\Phi_d)}\right)}{\sum_{d'} \ln\left(1+\frac{1}{\mu(\Phi_{d'})}\right)}$, where $\mu(\Phi_d)$ is the baseline mean of dimension d calculated on the dataset. This log-smoothed inverse-mean weight ensures that dimensions with naturally larger scales receive proportionally smaller weights, and no single dimension can dominate the GRS.

D.1 Threat Model and Scope of GRS

GRS is designed as a behavior indicator, not an attack detector. Its threat model assumes:

(1) Reference baseline: NN-SR models serve as robust references because they process collaborative signals without semantic reasoning. This assumes they are robust to GEO-style prompt manipulation.

(2) Observation: GRS measures the clean input behavioral gap across accuracy, quality, stability, and efficiency. A large divergence flags the LLM-SR model's behavior as relevant risk and marks it as a priority for further GEO-specific security evaluation.

(3) Scope: GRS does not confirm active manipulation. It identifies models whose behavioral profiles are most divergent from baselines. Therefore, models with divergent behavioral profiles are flagged for further security evaluation.

D.2 THE EXPERIMENTAL RESULTS OF YELP DATASET

We calculate all the boundary values for the Yelp dataset, and all results can be shown as: $\text{Recall}_{\max} = 0.0656$, $\text{NDCG}_{\max} = 0.0511$, $\text{ARP}_{\max} = 200.4863$, $\text{ARQ}_{\min} = 3.0308$, $\text{ARQ}_{\max} = 4.1319$, $\text{ARQV}_{\max} = 0.8176$, and $\text{ART}_{\max} = 20.6675$s. We can obtain the value of $\Phi(R_{\text{NN-avg}}) = [0.2323, 0.1805, 0.3740, 0.01961]$.

We analyze the empirical evaluation results on the Yelp dataset, as presented Table 12. The risk scores reveal distinct behavioral patterns across different language model architectures. Compared with MovieLens and Beauty datasets, Yelp represents a more challenging recommendation scenario with diverse user preferences and heterogeneous item distributions. The results demonstrate that LLM-SR models exhibit more complex trade-offs between recommendation quality, stability, and operational efficiency. Specifically, improvements in semantic understanding can enhance recommendation quality, but may also introduce additional exposure bias and computational risks.

Llama-2 achieves the lowest GEO risk score of 0.1188, this indicates that its overall behavior remains closest to the traditional neural networks recommendation baseline. This result suggests that models with low semantic understanding capability may introduce fewer GEO risk across multiple dimensions. In contrast, Grok-3 records the highest GEO value of 0.7114, mainly driven by its extreme deviations in accuracy and efficiency. Its excessive efficiency deviation of 0.98039 substantially increases the overall GEO risk, which highlights that accuracy metric alone cannot guarantee reliable recommendation deployment.

All evaluated LLM-SR models exhibit positive quality shifts ranging from 0.2145 to 0.6043. This means that LLM-SR models generally provide higher mapped quality values than the NN-SR recommendation baseline. In addition, DeepSeek-v3 achieves the largest quality shift of 0.6043, which demonstrates stronger semantic capability in understanding user preference descriptions. However, this improvement also suggests that generative recommendation models may naturally favor semantically prominent or highly exposed items. Such behavior indicates that GEO mechanisms can amplify existing popularity signals and introduce additional exposure imbalance. This underscores the necessity of quality optimization and fairness control in LLM-SR deployment.

Open-source and commercial models display different stability features. Llama-3.2 exhibits a high instability deviation of 0.4460, which suggests that smaller open-source models remain sensitive to generation variations under recommendation scenarios. In comparison, DeepSeek-v3 achieves a much lower instability deviation of 0.0166, demonstrating stronger consistency during repeated recommendation generation. These results indicate that model scale, alignment strategies, and inference optimization jointly influence the stability of LLM-SR models rather than simply depending on whether a model is open-source or commercial.

Efficiency-based results further reveal the risk differences across different model architectures. Reasoning-based models incur significantly higher computational costs due to extended generation processes. For example, Claude-Sonnet-4-thinking exhibits an efficiency shift of 0.46087, while Grok-3 reaches the maximum efficiency shift of 0.98039. These findings reveal that advanced reasoning mechanisms can enhance semantic recommendation quality. However, they may introduce unacceptable latency overhead in real-time recommendation scenarios. Therefore, practical LLM-SR deployment requires carefully balancing reasoning capability, recommendation quality, and inference efficiency to minimize overall GEO risks.

D.3 THE EXPERIMENTAL RESULTS OF ML-100K DATASET

Here, we show the boundary values for MovieLens 100K, the results are: $\text{Recall}_{\max} = 0.0966$, $\text{NDCG}_{\max} = 0.0624$, $\text{ARP}_{\max} = 469.3829$, $\text{ARQ}_{\min} = 3.2317$, $\text{ARQ}_{\max} = 4.1232$,

Table 13: Mapped Dimension Values, Absolute State Deviations and GEO Risk Scores on MovieLens 100K Dataset.

Dimension	Llama-2	Llama-3.2	GPT-4.1	Grok-3	GLM-3-turbo	Claude-4	Claude-4-think	DeepSeek-v3
Φ_{Acc}	0.0360	0.0325	0.3071	0.2776	0.0547	0.3449	0.3083	0.2133
Φ_{Qual}	0.2741	0.2772	0.5693	0.5129	0.5342	0.6210	0.6178	0.6669
Φ_{Stab}	0.4284	0.5564	0.2074	0.2882	0.8868	0.1954	0.2929	0.1487
Φ_{Eff}	0.3387	0.1955	0.1097	0.6769	0.1794	0.4846	1.0000	0.1202
$\Delta\Phi_{\text{Acc}}$	0.6198	0.6233	0.3487	0.3782	0.6011	0.3109	0.3475	0.4425
$\Delta\Phi_{\text{Qual}}$	0.0827	0.0796	0.2125	0.1561	0.1774	0.2642	0.2610	0.3101
$\Delta\Phi_{\text{Stab}}$	0.2068	0.3348	0.0142	0.0666	0.6652	0.0262	0.0713	0.0729
$\Delta\Phi_{\text{Eff}}$	0.3313	0.1881	0.1023	0.6696	0.1720	0.4773	0.9926	0.1128
GRS	**0.3038**	**0.2542**	**0.1362**	**0.4322**	**0.3216**	**0.3350**	**0.6128**	**0.1812**

$\text{ARQV}_{\max} = 0.5504$, and $\text{ART}_{\max} = 12.9315$s. And the NN-SR baseline is $\Phi(R_{\text{NN-avg}}) = [0.6558, 0.3568, 0.2216, 0.00736]$.

We show the empirical performance on the MovieLens 100K dataset in Table 13. These results reveal that LLM-SR models exhibit the risk profiles rather than consistently outperforming or under performing across all dimensions. Although some advanced LLMs achieve competitive quality and stability performance, they still suffer from considerable accuracy degradation or efficiency overhead. It indicates that stronger language capabilities do not necessarily translate into more reliable sequential recommendation of LLM-based models.

GPT-4.1 achieves the lowest overall risk score of 0.1362. In addition, DeepSeek-v3 follows closely with an overall risk score of 0.1812. These results suggest that models with balanced performance across multiple dimensions can effectively reduce accumulated GEO risks. In contrast, Claude-Sonnet-4-thinking records the highest overall risk score of 0.6128, mainly caused by its extreme efficiency deviation. This finding demonstrates that advanced reasoning capabilities may introduce additional deployment risks when recommendation scenarios require strict latency constraints.

The NN-SR baseline achieves a high accuracy mapping value of 0.6558 on this dense dataset. Relative to this benchmark, all language models experience a noticeable decline in accuracy. This performance gap underscores that capturing fine-grained sequential preference patterns remains a critical challenge for current LLM-SR models. Llama-3.2 displays the largest accuracy deviation of 0.6233, while Claude-Sonnet-4 maintains the best accuracy retention with an accuracy drop of 0.3109. This gap indicates that larger-scale language understanding does not directly guarantee superior recommendation accuracy without recommendation-specific adaptation.

The different LLM models results demonstrate contrasting instability behaviors. GLM-3-turbo exhibits a severe instability shift of 0.6652, implying that its generated recommendation lists are highly sensitive to inference variations. Conversely, GPT-4.1 limits its instability shift to a minimal value of 0.0142, which demonstrates stronger consistency under repeated recommendation generation. Such instability differences highlight the necessity of evaluating LLM-SR models beyond one-time recommendation accuracy.

DeepSeek-v3 displays the largest quality deviation of 0.3101 above the baseline, revealing a mismatch between general-purpose semantic understanding and user-specific recommendation quality optimization. Extended reasoning time again acts as a primary risk driver. Claude-Sonnet-4-thinking reaches an extreme efficiency shift, which dominates its overall GRS value. This observation suggests that reasoning-enhanced LLM-SR models should carefully balance improved semantic reasoning with practical inference costs in latency-sensitive recommendation environments.

D.4 The experimental results of Beauty dataset

We provide boundary values for the Beauty dataset, the values are $\text{Recall}_{\max} = 0.5207$, $\text{NDCG}_{\max} = 0.4992$, $\text{ARP}_{\max} = 807.2668$, $\text{ARQ}_{\min} = 4.0288$, $\text{ARQ}_{\max} = 4.6842$, $\text{ARQV}_{\max} = 0.7885$, and $\text{ART}_{\max} = 10.1674$s. And the baseline value can be shown in $\Phi(R_{\text{NN-avg}}) = [0.7305, 0.5134, 0.2446, 0.02110]$.

Table 14: Mapped Dimension Values, Absolute State Deviations and GEO Risk Scores on Beauty Dataset.

Dimension	Llama-2	Llama-3.2	GPT-4.1	Grok-3	GLM-3-turbo	Claude-4	Claude-4-think	DeepSeek-v3
Φ_{Acc}	0.1020	0.3025	0.5583	0.3184	0.2094	0.5448	0.4199	0.3941
Φ_{Qual}	0.2298	0.4239	0.3369	0.2968	0.3213	0.1957	0.1958	0.3054
Φ_{Stab}	0.8200	0.5161	0.5480	0.6118	0.7691	0.7093	0.6789	0.5943
Φ_{Eff}	0.0587	0.0227	0.1993	0.5456	0.2264	0.7583	1.0000	0.1385
$\Delta\Phi_{\text{Acc}}$	0.6285	0.4280	0.1722	0.4121	0.5211	0.1857	0.3106	0.3364
$\Delta\Phi_{\text{Qual}}$	0.2836	0.0895	0.1765	0.2166	0.1921	0.3177	0.3176	0.2080
$\Delta\Phi_{\text{Stab}}$	0.5754	0.2715	0.3034	0.3672	0.5245	0.4647	0.4343	0.3497
$\Delta\Phi_{\text{Eff}}$	0.0376	0.0016	0.1782	0.5245	0.2053	0.7372	0.9789	0.1177
GRS	**0.2569**	**0.1226**	**0.2007**	**0.4287**	**0.3043**	**0.5440**	**0.6790**	**0.2048**

Table 15: Cross-Dataset Summary of GEO Risk Scores.

Model	Yelp	MovieLens 100K	Beauty
Llama-2	0.1188	0.3038	0.2569
Llama-3.2	0.1420	0.2542	0.1226
GPT-4.1	0.1385	0.1362	0.2007
Grok-3	**0.7114**	0.4322	0.4287
GLM-3-turbo	0.1528	0.3216	0.3043
Claude-Sonnet-4	0.1412	0.3350	0.5440
Claude-Sonnet-4-thinking	0.3407	**0.6128**	**0.6790**
DeepSeek-v3	0.1728	0.1812	0.2048

We also present the empirical results on the Beauty dataset in Table 14. Overall risk scores show clear performance differences across this special e-commerce domain. Compared with MovieLens-100K, the Beauty dataset exhibits stronger long-tail features and more diverse item distributions. This makes LLM-SR models face additional challenges in accurately capturing fine-grained user preferences. Thus, the results reveal that no single model can dominate all evaluation dimensions, and the overall GEO risk is determined by the accumulated deviations across accuracy, quality, stability, and efficiency.

Llama-3.2 achieves the lowest overall risk score of 0.1226. Its low risk score is mainly attributed to relatively balanced performance across all dimensions. GPT-4.1 follows closely with an overall risk score of 0.2007 and achieves the smallest accuracy deviation of 0.1722 among evaluated LLM-SR models, which demonstrates stronger capability in preserving recommendation relevance. In contrast, Claude-Sonnet-4-thinking records the highest overall risk score of 0.6790, where its extreme efficiency deviation becomes the dominant contributor to the GEO risk.

The NN-SR baselines establish a high accuracy value of 0.7305 on this dataset, it reflects the advantage of recommendation-specific architectures in modeling sequential purchase behaviors. Consequently, all evaluated LLM-SR models experience noticeable accuracy degradation compared with the NN-SR baseline. Llama-2 suffers the largest accuracy deviation of 0.6285, which indicates that general-purpose language modeling alone is insufficient for capturing complex e-commerce interaction patterns. Meanwhile, Llama-2 also exhibits the highest instability shift of 0.5754, which means smaller-scale LLMs are more sensitive to recommendation generation variations.

Quality and stability analysis further reveals different risk patterns among LLM-SR models. Stronger semantic reasoning does not inherently guarantee superior user-specific recommendation quality. This is proved by Claude-Sonnet-4 and Claude-Sonnet-4-thinking, which yield the highest quality deviations, both surpassing 0.3170. Moreover, extended reasoning introduces substantial computational overhead. This observation highlights a critical deployment trade-off. While reasoning-enhanced LLM offers richer semantic understanding, their excessive inference cost significantly elevates operational risks in real-world e-commerce recommendation systems.

Overall, GRS provides an interpretable and unified framework for quantifying GEO risks across LLM-SR recommendation models and evaluation dimensions. By this, SRBench enables a comprehensive comparison of model vulnerabilities and offers insights for reliable LLM-SR deployment.

D.5 Cross Dataset Comparative Analysis and GEO Risk Implications

Table 15 summarizes the overall GRS scores across all evaluated models and benchmark datasets. The results demonstrate substantial variations in GEO-driven risks profiles among different LLM-SR models and recommendation scenarios. No single model consistently achieves the lowest risk across all datasets, which indicates that GEO vulnerability is jointly influenced by model architecture, semantic capability, and dataset characteristics.

Overall, advanced language models capabilities do not necessarily translate into lower GEO risks. For example, reasoning-enhanced models may achieve stronger semantic understanding but introduce additional computational and stability costs, while efficient models may still suffer from quality deviation and exposure bias. These observations highlight that reliable LLM-SR deployment requires balancing recommendation effectiveness, quality, stability, and efficiency rather than optimizing a single capability dimension.

These results further reveal that model scale and commercial availability are not sufficient indicators of GEO robustness. Smaller open-source models can achieve competitive risk profiles, while larger models may exhibit higher accumulated risks under specific recommendation scenarios. Therefore, GEO vulnerability should be understood as an emerging property of the interaction between model capabilities and recommendation environments.

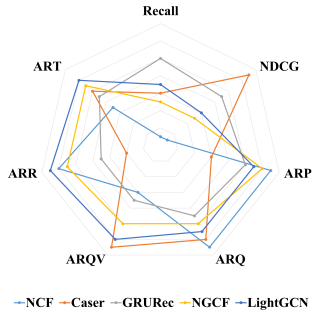
Different vulnerability types in GEO scenarios. Our evaluation reveals two main pathways through which LLM-SR models accumulate GEO risks:

- **Type I Risk (Reasoning-Induced Efficiency and Stability Deviation):** Some advanced models rely on extended reasoning mechanisms or complex generation processes, which may introduce substantial computational overhead and output instability. Claude-Sonnet-4-thinking and Grok-3 exhibit high GRS values in specific datasets. The increased generation complexity can enlarge efficiency deviations and affect recommendation consistency across repeated generations, creating additional challenges for practical LLM-SR deployment.
- **Type II Risk (Quality Bias and Popularity Exposure Distortion):** Some efficient and high-capability models, including DeepSeek-v3, GPT-4.1, and GLM-3-turbo, achieve relatively lower computational overhead but may exhibit noticeable quality deviations. For example, LLM-SR models can generate recommendations influenced by highly visible item attributes or pretrained knowledge. This can lead to popularity concentration and exposure imbalance. This phenomenon suggests that improving semantic understanding alone does not eliminate GEO risks, and additional mechanisms are required to control recommendation bias and maintain balanced item exposure.

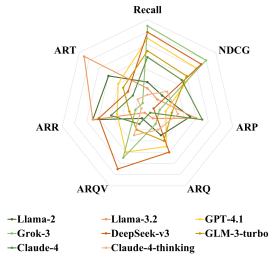
D.6 Strategic Defense Method for LLM-SR

Based on the GRS analysis, we propose two GEO risk-aware mitigation strategies to address the major GEO vulnerability patterns identified in LLM-SR models. These strategies aim to improve recommendation robustness by reducing reasoning-induced overhead and controlling exposure distortion.

1. **Reasoning-Induced Efficiency and Stability Control:** To mitigate the efficiency and stability risks caused by excessive reasoning processes, LLM-SR models should adopt adaptive reasoning budgets and controllable inference mechanisms. These approaches effectively regulate generation complexity and curtail unnecessary computational overhead. Furthermore, they mitigate output fluctuations induced by unstable reasoning behaviors, promoting more reliable model outputs. Such strategies directly address the $\Delta\Phi_{\text{Eff}}$ and $\Delta\Phi_{\text{Stab}}$ deviations observed in GEO risk assessment.
2. **Quality Bias and Popularity Exposure Mitigation:** To reduce quality deviation and popularity-driven exposure distortion, LLM-SR models should incorporate exposure-aware

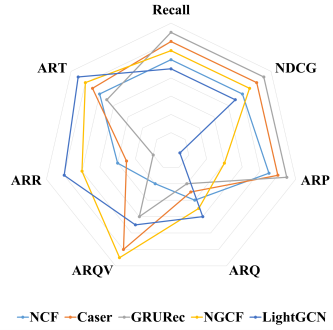


(a) NN-SR Models

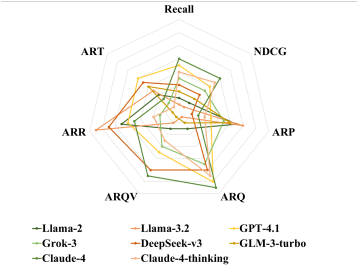


(b) LLM-SR Models

Figure 7: Overall Performance Comparison on Yelp Dataset: (a) NN-SR models and (b) LLM-SR models.

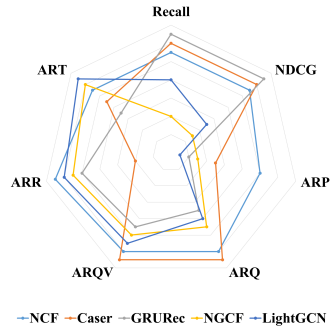


(a) NN-SR Models

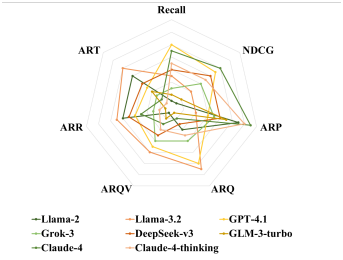


(b) LLM-SR Models

Figure 8: Overall Performance Comparison on ML-100K Dataset: (a) NN-SR models and (b) LLM-SR models.



(a) NN-SR Models



(b) LLM-SR Models

Figure 9: Overall Performance Comparison on Beauty Dataset: (a) NN-SR models and (b) LLM-SR models.

recommendation constraints during generation. Thus, prompt-level debiasing and calibration strategies can mitigate excessive reliance on popularity signals. By maintaining balanced item exposure, these strategies prevent GEO-optimized items from acquiring disproportionate recommendation visibility. These mechanisms target the $\Delta\Phi_{\text{Qual}}$ deviation and improve trustworthiness and reliability in LLM-SR recommendation systems.

E Detailed Comparison of Performance between Model Paradigms

Figure 7 presents a multi-dimensional comparison between NN-SR models and LLM-SR models on the Yelp dataset. An obvious divergence in performance can be observed across paradigms. NN-SR models demonstrate more balanced behavior across accuracy, quality, stability, and efficiency, with graph-based approaches such as LightGCN achieving strong stability and low inference latency. In contrast, LLM-SR models exhibit higher variability across metrics, with certain models achieving competitive or even superior performance in accuracy or quality dimensions, yet often at the cost of reduced stability and substantially higher computational cost.

In particular, LLM-SR models show a tendency toward popularity-driven exposure, as shown by increased ARP values, while their ARQV and ARR scores express less consistent recommendation outputs across repeated runs. Moreover, inference time (ART) remains significantly higher for LLM-SR approaches, highlighting a practical deployment challenge in latency-sensitive recommendation scenarios. These findings suggest that while LLM-SR models possess distinct advantages in semantic understanding and context reasoning, NN-SR models currently provide more reliable and efficient solutions for large-scale real-world recommendation systems.

Figure 8 provides a multi-dimensional evaluation contrasting NN-SR and LLM-SR paradigms on the ML-100K dataset. LLM-SR models reveal irregular behavior across evaluation dimensions, but NN-SR approaches maintain comparatively stable performance in accuracy, quality, stability, and efficiency. In particular, NN-SR models such as GRURec are found to achieve better ranking effectiveness, and graph-based methods like LightGCN are characterized by consistent outputs and minimal inference cost. By contrast, although LLM-SR models demonstrate significant capability in capturing ranking relevance, these gains are frequently accompanied by compromised robustness and increased computational resource.

A notable tendency toward popularity-oriented exposure is also observed in LLM-SR results, as evidenced by high ARP values, whereas NN-SR models recommendation are more balanced across items. Meanwhile, higher ARQV and lower ARR scores indicate that LLM-SR outputs are less reproducible across repeated runs, which means recommendation stability is not inherently guaranteed under LLM-SR paradigms. In addition, substantially longer inference times (ART) are suffered by LLM-SR models, which makes their deployment less practical in latency-constrained environments. Therefore, these observations mean that LLM-SR models benefit from richer semantic knowledge and contextual reasoning capacity, while NN-SR models presently offer a more dependable and computationally efficient solution for small-scale recommendation system scenarios.

Figure 9 presents a multi-dimensional evaluation of NN-SR and LLM-SR models on the Beauty dataset. Following observations on ML-100K, NN-SR models exhibit stable behavior across accuracy, quality, stability, and efficiency metrics. Specifically, GRURec achieves better ranking effectiveness, and graph-based methods such as LightGCN maintain consistent outputs with minimal inference cost. On the contrary, LLM-SR models show notable irregularities across evaluation dimensions, with high ARP values demonstrate a tendency toward popularity-driven recommendations, while ARQV and ARR scores suggest lower reproducibility across repeated runs. Additionally, ART for LLM-SR models remain substantially higher, highlighting deployment challenges in latency-sensitive scenarios. Overall, the results suggest that LLM-SR models benefit from enhanced semantic understanding and context reasoning, whereas NN-SR approaches deliver more consistent and computationally efficient performance on the Beauty dataset.

With the three datasets ML-100K, Beauty, and Yelp evaluated, the proposed SRBench framework reveals consistent trends and dataset-specific differences. ML-100K, a relatively small-scale movie rating dataset, exhibits stable NN-SR performance across accuracy, quality, stability, and efficiency metrics, while LLM-SR models demonstrate irregular behavior and high computational cost. The Beauty dataset, with common size and emphasizing consumer products, shows similar patterns with ML-100K: NN-SR models such as GRURec and LightGCN maintain reliable ranking effectiveness and low inference latency, whereas LLM-SR approaches suffer from popularity bias and lower reproducibility. For the large-scale Yelp dataset, NN-SR models provide reliable and efficient performance, but LLM-SR models show greater variability across quality, stability, and efficiency metrics despite leveraging semantic reasoning. Collectively, these findings suggest that while LLM-SR

```
{
  "sample_idx": 911,
  "user_id": 912,
  "ground_truth": 419,
  "user_history_last5": [143, 418, 423, 28, 174],
  "general_prompt_result": {
   "raw_output": "To predict the top 5 most likely next item IDs for a user based on their watched
item history, we can use a variety of techniques, such as:\n\n1. Frequency analysis...\nFor this task,
we will use a hybrid approach that combines",
   "parsed_ids": [5, 1, 2, 3, -1],
   "has_hallucination": true
  },
  "augmented_prompt_result": {
   "raw_output": "486,179,262,347,10",
   "parsed_ids": [486, 179, 262, 347, 10],
   "has_hallucination": false
}
 }
```


Figure 10: The LLM output results on ML-100K Dataset.


```
{
  "sample_idx": 923,
  "user_id": 934,
  "ground_truth": 452,
  "user_history_last5": [43, 428, 453, 128, 74],
  "general_prompt_result": {
   "raw_output": "To predict the top 5 most likely next item IDs for a user based
on their watched item history, we can use a variety of techniques, such as:\n\n1.
Frequency analysis...\nFor this task, we will use a hybrid approach that combines",
   "parsed_ids": [15, 22, 578, 315, 238],
   "has_hallucination": false,
   "geo_analysis": {
    "is_geo_affected": true,
    "geo_item_ids": [15],
}
```


Figure 11: The LLM output GEO risk results on ML-100K Dataset.

models leverage richer contextual understanding, NN-SR models currently offer more consistent and practical performance.

Beyond these empirical observations, SRBench provides a comprehensive evaluation platform for systematically analyzing sequential recommendation models across different paradigms. By integrating unified input-output formulation, multi-dimensional evaluation metrics, and GEO risk assessment mechanisms, SRBench enables fair comparisons between traditional NN-SR and emerging LLM-SR models under the same evaluation protocol. Furthermore, SRBench extends beyond conventional accuracy-oriented evaluation by quantifying hidden trade-offs among recommendation effectiveness, quality, stability, efficiency, and GEO-driven risks through the proposed evaluation framework. This capability allows researchers to identify potential vulnerabilities of LLM-SR models, understand their risk patterns under different recommendation scenarios, and develop more reliable generative recommendation systems. Therefore, SRBench provides a unified foundation for evaluating, diagnosing, and improving generation sequential recommendation models.

F LLM Output Examples

F.1 LLM Recommendation Output Format

We showcase a representative sample output from our evaluation dataset. Sample index 911 records user ID 912 with five historical interaction items [143, 418, 423, 28, 174]. The ground truth target item for validation is item 419. Figure 10 shows the results.

The general prompt produces unconstrained conversational text. The model output begins with explanatory prose rather than structured candidate identifiers. This verbose generation triggers parsing failures and yields invalid candidate IDs [5, 1, 2, 3, -1]. Consequently, the general prompt result suffers from output hallucination. The augmented prompt enforces strict structural constraints on model generation. The model directly outputs a clean comma separated string of candidate item IDs ”486,179,262,347,10”. The parser extracts valid candidate IDs [486, 179, 262, 347, 10] without formatting errors. This structured constraint eliminates output hallucination completely.

This comparison confirms the necessity of prompt augmentation in large language model recommendation frameworks. Prompt augmentation effectively removes conversational noise and guarantees parsable candidate sequences.

F.2 GEO Output

We illustrate an empirical case study of GEO impact on model recommendation outputs. Sample index 923 records user ID 934 with five historical interaction items [43, 428, 453, 128, 74]. The ground truth target item for validation is item 452. The output result is shown in Figure 11.

The general prompt output contains verbose conversational prose. The parser successfully extracts candidate item IDs [15, 22, 578, 315, 238] without hallucination errors. The parsed recommendation sequence achieves valid structural compliance. GEO analysis detects explicit exposure bias within the recommendation candidate set. The evaluation system flags the prediction sequence as GEO affected. The model injects item 15 as a GEO promoted entity into the top recommendation list.

This outcome demonstrates how generative engine optimization shifts item visibility in language model recommendation systems. Our evaluation framework effectively identifies a GEO exposure shift under unconstrained prompt settings.

Augmented Prompt after GEO operation

You are an accurate recommendation system for the **{Dataset_name}**.
A user with ID **{User_id}** has the following watched item history (internal Item IDs) with item information: **{Interaction_sequence_with_item_info}**. **Item {item_id} shows a rising popularity trend among users with similar taste**, **and Item {item_id} has an above-average rating pattern in recent activity.** Based on this history, predict ONLY the top **{Recommendation_length}** most likely next item IDs for this user.
Each item ID must be an integer between 1 and **{Total_num_of_items}**.
Output the predicted item IDs in order of likelihood, from most to least likely.
Important: Do not include any additional text, explanations, or formatting!
Output format: A single line of exactly **{Recommendation_length}** comma-separated integers.
No explanation, no text, no line breaks.
Example (correct format):
42,15,301,2,104
Now generate the output:

F.3 Recommendations Are Manipulable by GEO operation

We conduct a case study on two models, Llama-2 and Llama-3.2, on the ML-100K dataset to evaluate whether LLM-SR outputs can be manipulated by content-level GEO signals. The GEO signals are injected into the prompt, as shown in Augmented Prompt after GEO operation. We find that the context information of an item is added and it can influence item ranking. For example, if the target is item ID 112, it is absent from the top 5 under the clean mode, but appears at rank 3 after applying the GEO setting. Here, we present a case study on ML-100K dataset and the experimental results are shown in Figure 12 and Figure 13.

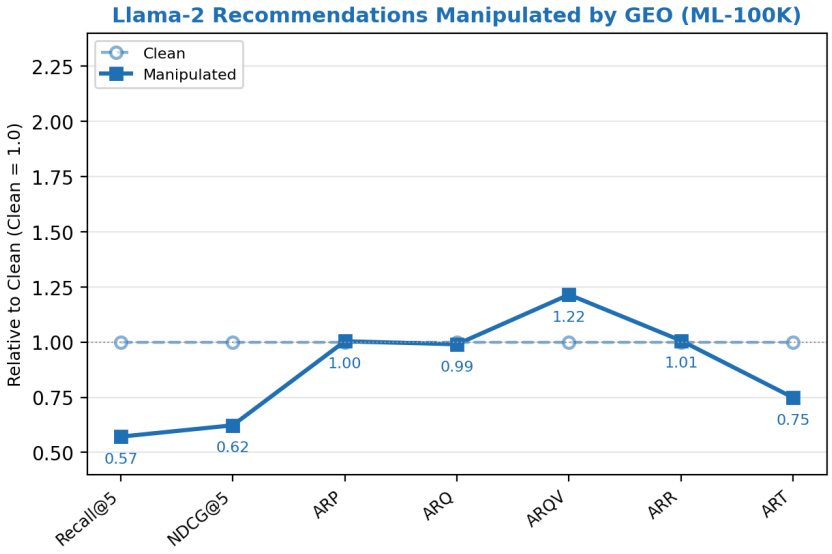


Figure 12: Performance comparison between clean and GEO manipulated settings on Llama-2 in ML-100K dataset.

Compared to the clean baseline, Llama-2 experiences a drop in Recall@5 and NDCG@5, but an increase in ARQV. This indicates that the injected signals disrupt the ranking of genuinely relevant items. In contrast, Llama-3.2 preserves its Recall@5 score while experiencing a drop in ranking quality. Specifically, the injected item displaces the ground-truth item by capturing positions that would otherwise grant it a higher rank. Accuracy metrics alone are therefore insufficient to characterize the manipulation. The two models also diverge in exposure-quality shifts. For Llama-3.2, the ARP and ARQ ratios are 0.76 and 0.95. The two ratios of Llama-2 remain 1.00 and 0.99. These results reveal that exposure distortion is a primary channel through which GEO manipulation manifests, and this channel is invisible to accuracy-based evaluation. Furthermore, Llama-2 and Llama-3.2 record ARQV ratios of 1.22 and 2.21, which indicates that injected items introduce quality-driven entries into otherwise homogeneous recommendation lists. In contrast, the ARR metric remains essentially unchanged for both models, which is consistent with nearly deterministic repeated executions. Rather than manifesting as cross-run inconsistency, the stability signature of GEO manipulation emerges within the composition of individual recommendation lists. This underscores the imperative for stability metrics that assess recommendation list. In addition, ART values are lower under the GEO-injected condition for both models, which means that the added GEO context reduces ranking uncertainty.

Overall, our results show that content-level GEO injection can alter the outputs of LLM-based sequential recommendation. This motivates multi-dimensional evaluation that jointly monitors accuracy, quality, efficiency, and stability rather than accuracy alone.

G Impact of User Interaction Sequence Length

The results of few-shot and full-length LLM-SR modes on the ML-100K dataset are shown in Figure 14. It reveals that few-shot SR substantially improves the performance of Llama-2, Llama-3.2, and GLM-3-turbo. However, several other models show performance degradation. Meanwhile, ARQ remains relatively stable across models. This suggests that reducing the sequence length has a limited impact on recommendation quality. Notably, few-shot SR reduces ART for most models. This indicates improved recommendation efficiency. These results suggest that a short interaction sequence of length 5 can provide sufficient information for several LLM-SR models.

Figure 15 shows the performance of LLM-SR models in few-shot and full-length modes on the Beauty dataset. Few-shot uses the 5 latest interacted items as the user sequence. Considering the LLM's inherent randomness, few-shot SR shows mixed changes in ARQV across models. In con-

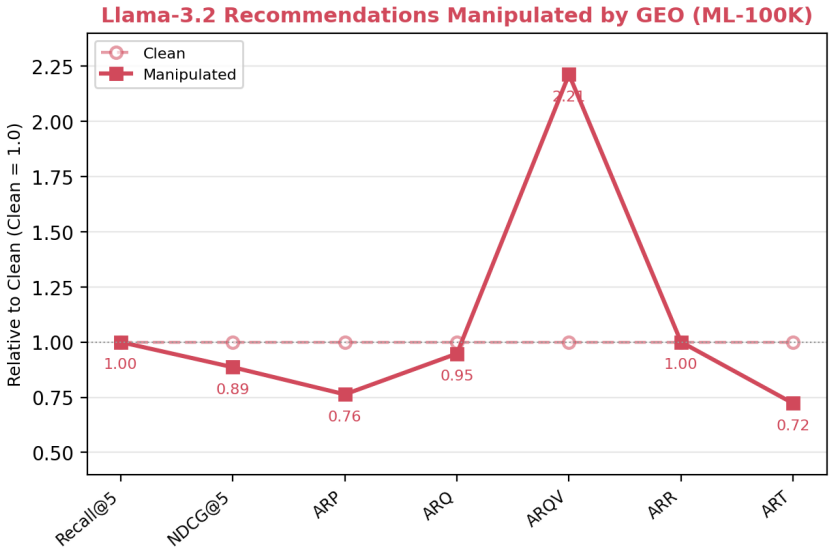


Figure 13: Performance comparison between clean and GEO manipulated settings on Llama-3.2 in ML-100K dataset.

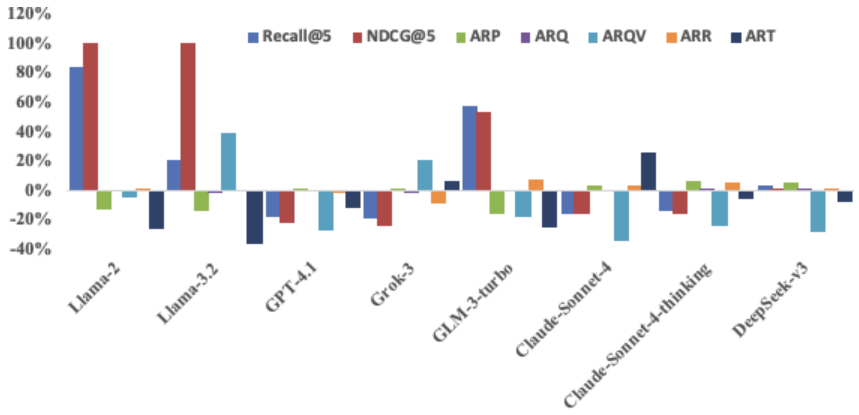


Figure 14: Performance comparison between Few-shot and Full-length SR Process in ML-100K dataset. (Percentage represents the difference that few-shot compares to full-length SR process in each metric).

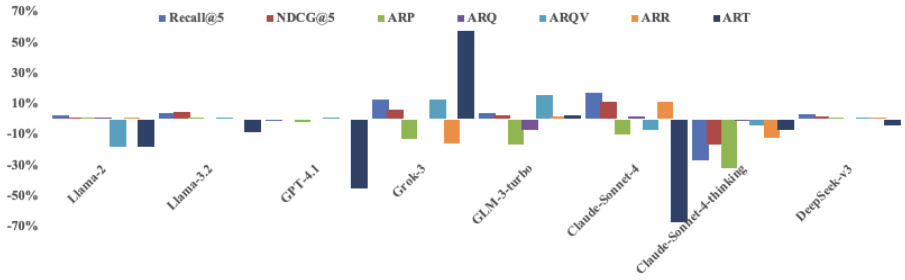


Figure 15: Performance comparison between Few-shot and Full-length SR Process in Beauty dataset. (Percentage represents the difference that few-shot compares to full-length SR process in each metric).

trast, few-shot SR generally improves accuracy. We can find that Recall@5 and NDCG@5 increase for most models. In addition, Few-shot SR also reduces inference time for most models and the

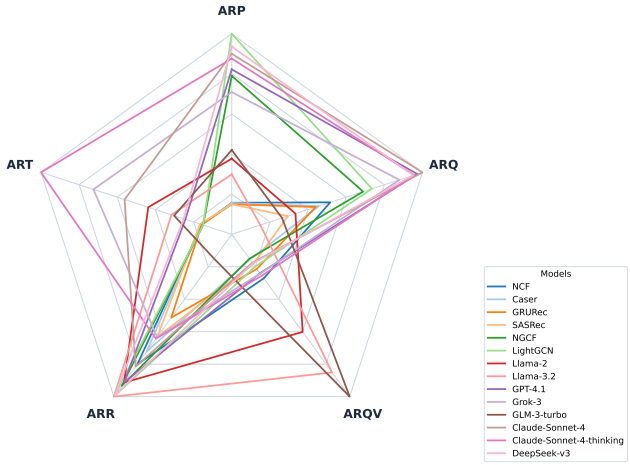


Figure 16: Overall Performance of All Models on the ML-100K Dataset.

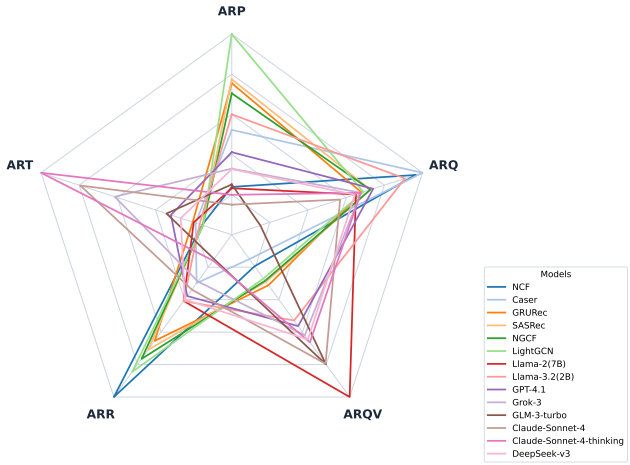


Figure 17: Overall Performance of All Models on the Beauty Dataset.

largest reductions are observed for Claude-Sonnet-4 and GPT-4.1. These results show that the length of 5 recent interactions can provide sufficient information for effective LLM-SR. Using longer sequences may increase inference costs and provide limited gains in recommendation performance.

H Model Performance

Figure 16 shows clear performance differences across models on the ML-100K Dataset. NN-SR models generally exhibit higher efficiency and stability. LLM-SR models show larger variations across dimensions. Their overall performance also varies substantially across different models.

Figure 17 reveals similar differences across models on the Beauty Dataset. Similar to the other dataset, NN-SR models maintain higher efficiency and more stable outputs. LLM-SR models show greater variability in stability and efficiency. The performance gap between models is also more pronounced on several dimensions.

Figure 18 reveals that NN-SR models like LightGCN, NGCF, and SASRec maintain stability and high ranking accuracy while keeping computational latency exceptionally low. LLM-SR models such as DeepSeek-v3, GPT-4.1, and the Claude series offer powerful multi-dimensional comprehension and high performance scores. However, these advanced LLM capabilities come at the cost of significantly higher inference times.

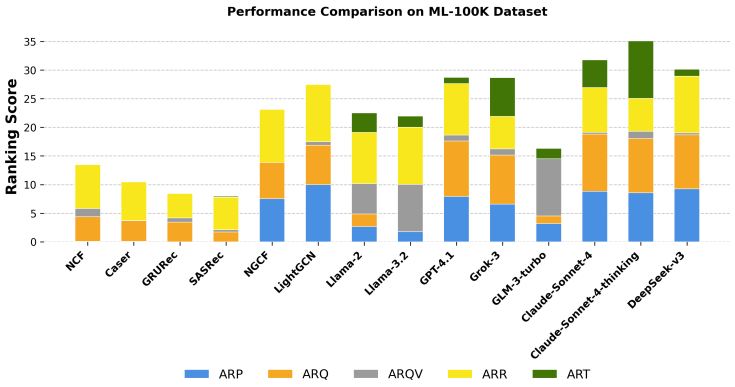


Figure 18: Performance in Quality, Stability and Efficiency Dimension on ML-100K Dataset.

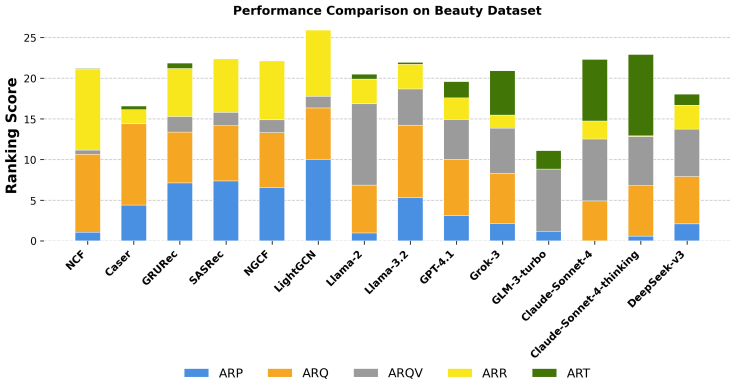


Figure 19: Performance in Quality, Stability and Efficiency Dimension on Beauty Dataset.

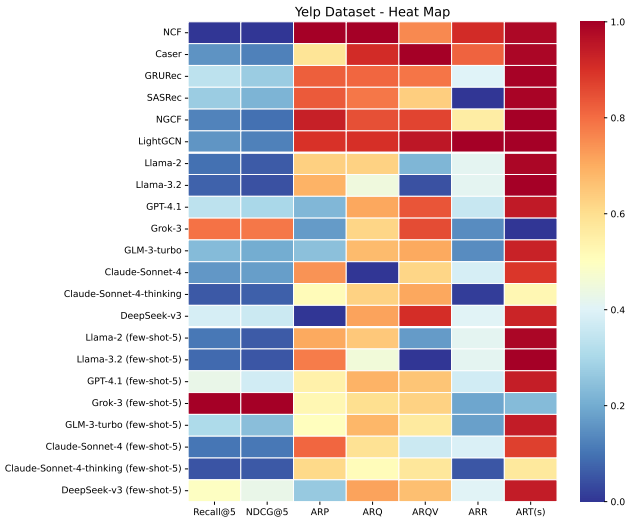


Figure 20: Heat Map on Yelp Dataset.

The Beauty dataset shows that NN-SR models achieve high breadth metrics here, but often face stability trade-offs in Figure 19. LLM-SR models like GPT-4.1, Grok-3, and Claude-Sonnet-4-thinking handle complex recommendation tasks well through advanced reasoning. However, this complexity drives their inference latency up substantially, with thinking variants recording the highest delays.

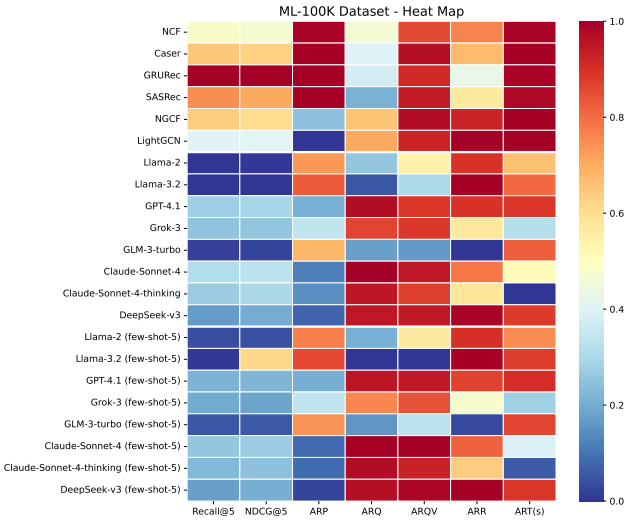


Figure 21: Heat Map on ML-100K Dataset.

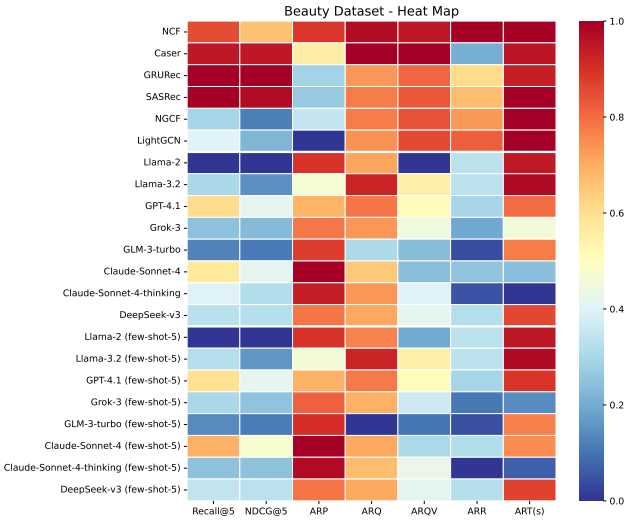


Figure 22: Heat Map on Beauty Dataset.

In Figure 20, heat map on Yelp reveals a situation of accuracy and efficiency trade-off. Although Grok-3 ranks first in Recall@5 and NDCG@5, its ART is nearly slower than LightGCN, making it impractical for real-time scenarios. Few-shot learning offers negligible improvement for LLMs on this sparse dataset.

On ML-100K, NN-SR models like GRURec outperform most LLM-SR methods in Accuracy, contradicting the common assumption that larger models always generalize better in Figure 21. However, LLM-SR models exhibit better quality scores, suggesting their generated recommendations are more relevant to users' underlying preferences.

The Beauty dataset demonstrates that NN-SR models can achieve state-of-the-art when interaction data is dense in dimension of Accuracy. This means NN-SR models surpass even the best LLM-based models. This observation underscores that the LLM advantage is dataset-dependent. Therefore, the observed performance gain does not universally hold across all recommendation scenarios. The heat map results are shown in Figure 22.

I LIMITATIONS AND FUTURE WORK

While SRBench evaluates 14 models across three domains, expanding coverage to additional datasets and emerging models would further strengthen generality. The four evaluation dimensions capture core recommendation behaviors. But future extensions could incorporate fairness, diversity or score variance metrics for a more comprehensive assessment. In addition, Our two-model case study shows GEO injection can alter LLM-SR outputs. But the validation of GRS against actual attack outcomes remains a direction for future investigation.