# SAER: Scalable Assessment of E-commerce Recommendations using Large Language Models

Xiang Liu
xanlu@amazon.com
Amazon
Vancouver, Canada

Manoj Srivatsav Regulagedda
manojsrv@amazon.com
Amazon
Vancouver, Canada

Sapan Patel
sapanp@amazon.com
Amazon
Seattle, USA

Walter Wong
walterwg@amazon.com
Amazon
Toronto, Canada

## Abstract

Selecting which recommendation algorithm variant to advance to online experimentation is a critical decision in industry practice. Manual evaluation is subjective and time-consuming, while offline metrics such as nDCG often fail to correlate with real-world customer preferences. We present SAER, a two-stage framework (pointwise filtering and pairwise comparison) that uses Large Language Models as judges to evaluate e-commerce recommendation quality. For pointwise evaluation, SAER achieves higher agreement with human consensus ($\kappa_w = 0.58$) than human annotators achieve with each other ($\kappa_w = 0.38$). For pairwise evaluation, SAER mitigates position bias (88.0% consistency) and demonstrates strong adversarial robustness (91.0% detection rate). In a large-scale online case study serving over 10M impressions, the algorithm SAER preferred offline also achieved a statistically significant click-through rate (CTR) lift ($p < 0.0001$), providing encouraging directional evidence, though this single observation cannot establish a general predictive relationship. SAER produces reproducible, explainable signals in approximately 30 minutes for $28, positioning it as a scalable pre-screening layer for recommendation development.

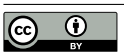
## CCS Concepts

• **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; *Personalization*; • **Applied computing** → **Online shopping**; • **Computing methodologies** → **Natural language processing**; Neural networks; • **General and reference** → Experimentation.

## Keywords

recommendation systems, large language models, LLM-as-a-judge, evaluation metrics, e-commerce, prompt engineering, A/B testing

**ACM Reference Format:**
Xiang Liu, Manoj Srivatsav Regulagedda, Sapan Patel, and Walter Wong. 2026. SAER: Scalable Assessment of E-commerce Recommendations using Large Language Models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3808462




*SIGIR '26, Melbourne, VIC, Australia*

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808462

## 1 Introduction

Modern e-commerce recommendation systems employ numerous algorithms across different customer touchpoints. During algorithm development, teams continuously explore variations: which customer segments to target, which item pools to draw from, what filtering criteria to apply, which touchpoint to surface, what time to show the strategy, and which model parameters to tune. At each decision point, the natural question arises: should we bring this variant to online experimentation? The standard approach relies on a combination of offline metrics and developer sign-off.

**Unreliable offline metrics.** Relying solely on offline metrics is often misleading due to the offline-online evaluation gap [2, 23]. Marginal offline accuracy gains frequently fail to reflect real-world performance, and metrics like nDCG can even contradict live user satisfaction [23]. Furthermore, standard offline setups suffer from exposure biases [21] and reproducibility issues [1, 9]. Even beyond-accuracy measures (diversity, novelty) [28, 30] fall short of capturing holistic human judgment [12]. Since algorithmic accuracy influences behavior only through subjective perception [15], teams struggle to know if offline precision gains will improve customer experience or merely reflect overfitting to logged biases [3].

**Subjective qualitative inspection.** To bridge this gap, teams use expert qualitative inspection, introducing significant subjectivity. Evaluators apply different internal quality bars—one might prioritize category diversity, another brand relevance—leading to conflicting conclusions even on nearly identical outputs. Furthermore, human judges frequently exhibit a "tie-bias," defaulting to neutral judgments for cognitively demanding comparisons. This produces inconsistent, irreproducible signals tied to the evaluator rather than algorithmic quality. Consequently, teams launch costly online experiments without a confident prior expectation of which algorithm should win [7, 8].

To address these challenges, we present SAER (Scalable Assessment of E-commerce Recommendations), a two-stage LLM-as-a-Judge framework that provides reproducible, explainable pre-screening of algorithms prior to online experimentation. SAER operates sequentially: pointwise evaluation identifies specific list-level quality issues with actionable reasoning (e.g., flagging a redundant

near-term repurchase of a multi-pack), while pairwise comparison identifies the likely superior algorithm. While recent literature demonstrates that LLMs can successfully mimic human annotators offline, less is known about whether these offline judgments carry practical relevance for live user behavior. SAER yields three contributions: (i) we demonstrate that SAER effectively mitigates manual evaluation subjectivity, capturing human consensus more reliably than individual expert annotators; (ii) we provide preliminary evidence that SAER's offline preferences can align with live user behavior, with the algorithm SAER preferred also achieving a statistically significant client-level click-through rate (CTR) lift in a large-scale online experiment; and (iii) we establish that domain-aware prompt engineering (incorporating short-term user intent and e-commerce domain constraints like consumable pacing) is critical for robustness, enabling the system to catch subtle algorithmic failures such as off-category injections and mismatched consumable pacing.

## 2 Related Work

**LLMs as evaluators.** Zheng et al. [32] established the LLM-as-a-Judge paradigm, and subsequent works validated it across search [10, 26], translation [16], and persona-conditioned evaluation [4], alongside studies on hallucination limits [11, 14]. Within recommender systems, LLMs can generate persuasive explanations [25] and are increasingly adopted for broad evaluation tasks [31].

**LLM-as-Judge for recommendations.** Recent works explore LLMs as automated recommendation evaluators. Fabbri et al. [5] and Penha et al. [20] adapted LLM judges for podcast and movie recommendations, achieving high agreement with offline human labels using profile-aware prompts and Cranfield-style IR techniques. While these works successfully simulate offline human annotators, SAER advances the paradigm in two critical dimensions. First, it addresses unique e-commerce constraints (e.g., consumable pacing, durable goods redundancy), which require distinct reasoning from media consumption. Second, whereas prior works validate strictly against offline annotations, SAER provides preliminary evidence of alignment with real-world user behavior through a large-scale online case study, though further validation is needed.

## 3 The SAER Framework

### 3.1 System Architecture

**Data preparation.** For each customer, we compile six months of purchase history, including product titles, brands, categories, and ratings. We additionally derive search keywords, recent browsing categories, and implicit preferences from recent customer activities. This rich context enables the judge to accurately assess contextual appropriateness and short-term intent.

**LLM judge configuration.** We utilize Claude 4.5 Sonnet via Amazon Bedrock. To ensure deterministic and reproducible assessments, the model temperature is set to 0.0. The framework is deployed as an offline batch process, allowing for high-throughput evaluation of recommendation datasets without impacting production latency. The LLM is configured to produce a structured response containing a qualitative rationale, a categorical quality label (Good/Partial/Poor Match), and a list of specific flagged items for diagnostic review.

### 3.2 Prompt Engineering

Crafting an effective evaluation prompt required careful design and significant iterative refinement to align the LLM's reasoning with expert human judgment. We found that simply providing a generic scoring rubric was insufficient for capturing the complex realities of e-commerce behavior. Instead, we developed highly category-specific, task-oriented few-shot examples that teach the LLM to evaluate recommendations against real-world shopping constraints.

The prompt instructs the judge to identify nuanced catalog and timing issues, such as minimum variation redundancy, consumable pacing, and durable goods redundancy. The context also extends beyond historical purchases to incorporate short-term user intent (recent searches, clicks, and browsed categories). This is critical for evaluating unfinished shopping tasks; for example, if recent clicks indicate active exploration of a durable good despite a prior purchase, the LLM learns to suspend redundancy penalties, recognizing the customer is still comparing alternatives.

**Prompt structure:** 5 few-shot examples (good match, minimum variation redundancy, durable goods redundancy, consumable pacing, off-category) → Customer context (6-month purchase history, recent searches, recent clicks, browsed categories) → Recommendation list (10 items) → Evaluation guidelines → Output: REASONING, CATEGORY (Good/Partial/Poor Match), FLAGGED ITEMS.

### 3.3 Evaluation Modes

**Pointwise evaluation** assesses each recommendation list independently using three named categories: Good Match (7+ relevant items, diverse coverage, no quality issues), Partial Match (4–6 relevant items or minor issues), and Poor Match (fewer than 4 relevant items or severe issues). Named categories with explicit semantic boundaries prevent the score clustering common in LLM evaluators. LLMs evaluating on numerical scales tend to collapse outputs onto specific integers, which compresses variance and artificially inflates tie rates [18].

**Pairwise evaluation** compares full recommendation lists from two algorithms for the same customer. To mitigate position bias [29, 32], each comparison is evaluated twice with reversed presentation order using neutral labels ("Set 1", "Set 2"). When both orderings agree, that algorithm wins. When they disagree, the result is a position-dependent tie. An explicit tie example in the prompt calibrates the LLM's threshold for declaring ties.

### 3.4 Experimental Setup

**Recommendation algorithms.** We evaluate two fundamental approaches in the Beauty & Personal Care category. Co-occurrence filtering recommends items frequently purchased together by other customers, capturing behavioral co-purchase patterns. Embedding-based filtering retrieves semantically similar items using sentence transformers and approximate nearest neighbor search.

**Evaluation datasets.** We construct two datasets for evaluation. The standard dataset is a stratified random sample of 200 customers spanning diverse engagement levels. For each customer, we generate top-10 recommendation lists from both algorithms, producing 400 pointwise lists and 200 pairwise comparisons. To test the judge's robustness against "rubber-stamping," we develop an adversarial

dataset containing 100 blind pairwise comparisons. Each case pairs a customer's genuine recommendation list (50 sourced from the co-occurrence algorithm and 50 from the embedding algorithm) against a random list taken from a different customer.

**Human benchmark.** To establish a rigorous baseline, a stratified random sample of 200 customer recommendation scenarios is independently evaluated by two experienced recommendation developers. Crucially, to ensure a fair comparison, the human annotators evaluate the lists using the exact same guidelines and quality criteria provided in the LLM prompt. Both annotators complete the pointwise and pairwise evaluations for all 200 standard cases, as well as the 100 adversarial cases. This allows us to capture expert consensus, calculate inter-annotator agreement metrics, and establish a human baseline for adversarial robustness.

**Adjudication rules.** To establish consensus ground truth from human annotators, we apply deterministic adjudication rules. For pointwise evaluation, we use strict adjudication, taking the harsher label whenever annotators disagree (e.g., a Good and Poor match defaults to Poor). For pairwise comparison, we use conservative adjudication, defaulting to a Tie whenever annotators prefer different algorithms. Similarly, to mitigate LLM position bias in pairwise evaluations, SAER evaluates each pair twice with swapped orderings; if the LLM's preference changes based on position, the verdict is recorded as a Tie.

**Reliability metrics.** To quantify the reliability of the assessments, we measure inter-rater agreement using Cohen's Quadratic Weighted Kappa ($\kappa_w$). Unlike simple percentage agreement, $\kappa_w$ accounts for agreement occurring by chance and applies heavier penalties to extreme disagreements on our ordinal scales (e.g., a Good/Poor mismatch is penalized more severely than a Good/Partial mismatch).

## 4 Results

### 4.1 Offline Evaluation

**Pointwise assessment.** As shown in Table 1, co-occurrence produces more polarized results, yielding higher rates of both Good and Poor Matches compared to the more conservative embedding-based filtering. Co-occurrence's higher Good Match rate reflects cases where behavioral co-purchase patterns surface genuinely useful complementary products. SAER's reasoning frequently highlights cross-category relevance. For example, it correctly identifies a thermal heat protectant spray as highly relevant for a user who recently purchased a curling iron, whereas embedding often just retrieves more curling irons.

Both human annotators confirm SAER's directional finding (Table 1). While Annotator A was systematically stricter, resulting in a low inter-annotator agreement ($\kappa_w = 0.38$), the strictly adjudicated SAER–human agreement reached $\kappa_w = 0.58$ (82.0% agreement). This confirms SAER's automated judgments align with human consensus substantially better than the two annotators agree with each other. We note that SAER's Partial Match rates are higher than both annotators', suggesting some residual central tendency despite named categories; future work could introduce finer-grained sub-categories to improve discrimination at the distribution tails.

**Pairwise comparison.** Table 2 demonstrates SAER's clear preference for co-occurrence over embedding. The LLM maintained its verdict regardless of presentation order in 88.0% of cases, indicating low position bias, and retained a strong adjudicated lead (51.5% vs. 33.0%).

**Table 1: Pointwise quality assessment comparing SAER and two human annotators (N=200 lists per algorithm).**

| Algorithm | Rating | SAER | Human A | Human B |
|---|---|---|---|---|
| Co-occurrence | Good Match | 12.5% | 13.5% | 19.5% |
| | Partial Match | 82.0% | 74.5% | 68.0% |
| | Poor Match | 5.5% | 12.0% | 12.5% |
| Embedding | Good Match | 1.5% | 7.5% | 9.0% |
| | Partial Match | 88.0% | 75.0% | 73.0% |
| | Poor Match | 10.5% | 17.5% | 18.0% |
| *Agreement Metrics* | | | | |
| Human–Human | | $\kappa_w = 0.38$ (70.0% agree) | | |
| SAER–Human (strict adjudication) | | $\kappa_w = 0.58$ (82.0% agree) | | |

**Table 2: Pairwise algorithm comparison between SAER and human annotator consensus (N=200 pairwise comparisons).**

| Evaluation Scope | Co-occ. Win | Embed. Win | Tie |
|---|---|---|---|
| SAER Raw (N=400 calls) | 57.0% | 38.5% | 4.5% |
| SAER Adjudicated (N=200) | 51.5% | 33.0% | 15.5% |
| Human Merged (N=200) | 26.0% | 16.0% | 58.0% |
| *Consistency & Agreement Metrics* | | | |
| SAER Position Consistency: 88.0% (176/200) | | | |
| Human-Human Raw Agreement: 44.5%, $\kappa_w = 0.13$ | | | |
| SAER-Human Merged Agreement: 49.0%, $\kappa_w = 0.43$ | | | |

Human annotators showed a similar directional preference, favoring co-occurrence 1.6× more often than embedding when they agreed. Although conservatively assigning ties to human disagreements inflated their tie rate, the SAER-human merged agreement reached $\kappa_w = 0.43$, confirming alignment on the ordinal scale. Notably, when humans agreed on a directional preference, SAER concurred 82.1% of the time.

**Adversarial robustness and prompt ablation.** To validate that SAER evaluates recommendation content rather than surface features, we construct 100 blind pairwise comparisons. Each pairs a customer's genuine recommendation list (50 from co-occurrence, 50 from embedding) against a random list taken from a different customer. The evaluator receives the identical pairwise prompt used in pairwise comparison (with no indication of adversarial intent) and must judge purely on relevance to the purchase history. Position is randomized to control for presentation bias.

Using the SAER prompt (Table 3), the LLM correctly identifies the original recommendation as better in 91.0% of cases, with only 4.0% incorrect preferences and 5.0% ties. To quantify the contribution of prompt engineering, we test two ablation variants on the same 100 adversarial pairs. A generic prompt that provides only purchase history and asks "which is better?" achieves 74.0% detection. An over-specified prompt with a detailed five-criterion rubric reaches 88.0%. The SAER prompt's 91.0% represents a 17-point improvement over generic and a 3-point gain over over-specified,

**Table 3: Adversarial robustness and prompt ablation for blind pairwise detection (N=100 pairs).**

| Evaluator Variant | Correct | Wrong | Tie |
|---|---|---|---|
| Generic Prompt (v1) | 74.0% | 18.0% | 8.0% |
| Over-specified Prompt (v2) | 88.0% | 8.0% | 4.0% |
| SAER Refined Prompt (v3) | 91.0% | 4.0% | 5.0% |
| Human Annotator A | 78.0% | 21.0% | 1.0% |
| Human Annotator B | 73.0% | 18.0% | 9.0% |

demonstrating that calibrated evaluation criteria (neither too sparse nor too verbose) yield the best discrimination.

Human annotators achieve 78.0% (Annotator A) and 73.0% (Annotator B) detection, bracketing the generic prompt but falling below both the over-specified and SAER variants. This confirms that the task is genuinely difficult for humans, who must mentally cross-reference two 10-item lists against a purchase history, and that structured LLM prompting provides a reliable, scalable alternative.

## 4.2 Online Experiment Case Study

To examine whether SAER's offline judgments carry practical relevance, we conducted a controlled online A/B test [17] in the Beauty & Personal Care category, serving over 10M client impressions across a 4-week period with 50/50 customer-level traffic allocation. Co-occurrence achieved a 0.619% client CTR (95% CI: [0.617%, 0.622%]) compared to embedding's 0.523% (95% CI: [0.521%, 0.526%]), a relative lift of +18.4% (95% CI: [+17.6%, +19.1%], $p < 0.0001$). This is directionally consistent with SAER's adjudicated pairwise preference (51.5% for co-occurrence vs. 33.0% for embedding) and with human annotators' consensus preference (26.0% vs. 16.0%). All three evaluation signals (SAER, human annotators, and live user engagement) point in the same direction, though we emphasize this is a single-experiment observation and cannot establish a general predictive relationship. Multiple confounds could explain this alignment: for instance, both SAER and users may independently favor the co-occurrence algorithm for different reasons, such as SAER rewarding textual cross-category relevance while users respond to visual variety or familiarity.

SAER produced this directional signal in approximately 30 minutes for $28, whereas the online experiment required 4 weeks of live traffic. For comparison, equivalent crowdsourced evaluation costs $108–$153 [19, 22], which significantly understates the true expense, as internal engineering teams typically perform these qualitative reviews at a much higher hourly rate.

The primary value of SAER is not to act as a strict gatekeeper to block weak algorithms, since falsely rejecting a viable variant incurs an opportunity cost. Instead, SAER provides a scalable evaluation layer capable of integrating complex textual metadata and short-term behavioral intents that standard offline metrics fail to capture. By continuously updating the judging criteria to better reflect live traffic patterns, engineering teams can iteratively tighten the alignment between offline pre-screening and real customer behavior. This positions A/B testing not as a blind empirical check, but as a guided workflow informed by a reproducible prior expectation.

# 5 Discussion and Conclusion

**Consistency and tunability.** Human evaluators do not represent ultimate ground truth; live behavior does. While SAER exceeds individual human consistency [6], its primary value lies in explicit tunability. Quality standards encoded in prompts are shareable, versionable, and can be iteratively aligned with live A/B test outcomes. Furthermore, as foundation models advance, baseline reasoning improves without requiring infrastructure changes or new data annotation.

**Position bias and uncertainty.** While our position swap limits inconsistency to 12.0%, positional bias remains a structural artifact [29]. Our approach controls this effectively but relies on deterministic point-estimates. Future work should explore uncertainty-aware calibration to handle marginal comparisons more gracefully. Recent advances in conformal prediction and confidence-driven evaluation [4, 24, 27] could further mitigate residual order effects without strict position swapping.

**Semantic similarity bias.** SAER exhibits a semantic similarity bias, rating lists higher than humans do when items are relevant but lack category diversity (e.g., SAER assigns significantly fewer 'Poor Match' labels to the embedding algorithm than humans do). This indicates LLMs weight item-level relevance over list-level diversity, motivating the need for complementary diversity metrics.

**Limitations and future work.** Our evaluation spans a single product category and contrasts fundamentally different algorithmic paradigms; future studies should evaluate variants with narrower gaps (e.g., hyperparameter tuning). Furthermore, SAER relies entirely on textual metadata, whereas human shoppers heavily weight visual signals. To address this, we plan to extend the framework using multimodal foundation models to process product images. Finally, we aim to deploy SAER in a continuous shadow mode to longitudinally track offline preferences against live A/B test outcomes, helping close the offline-online evaluation gap [13].

**Conclusion.** SAER demonstrates that LLM-based evaluation can serve as a practical pre-screening tool for recommendation development. A single large-scale online case study provides encouraging directional evidence that LLM judgments can align with live user behavior, producing this signal at a fraction of the time and cost of A/B testing. However, further validation across distinct algorithm pairs and product categories is needed to establish broader generality. By providing a reproducible, explainable directional signal, SAER enables recommendation teams to iterate with confidence rather than launching experiments without informed expectations.

## Acknowledgements

We sincerely thank Money Dhaliwal for his valuable contribution to the work. We are also grateful for the support and guidance from Ankur Datta, Nirav Desai, and Sam Heyworth, as well as the broader Everyday Essentials leadership team. Finally, we thank the reviewers for their insightful comments.

## Presenter Biography

**Xiang Liu** is a Senior Machine Learning Engineer with the Everyday Essentials Science team at Amazon. He specializes in the architecture of scalable customer understanding frameworks and recommendation systems. His work focuses on human-centered

AI, particularly the application of Large Language Models (LLMs) to consumer behavior modeling and personalization. Xiang is dedicated to developing sustainable, trustworthy systems that ensure machine learning accessibility and reliability at a massive real-world scale.

## References


[1] Joeran Beel, Corinna Breitinger, Stefan Langer, Andreas Lommatzsch, and Bela Gipp. 2016. Towards reproducibility in recommender-systems research. *User Modeling and User-Adapted Interaction* 26, 1 (2016), 69–101. doi:10.1007/s11257-016-9174-x

[2] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (Hong Kong, China) *(RepSys '13)*. Association for Computing Machinery, New York, NY, USA, 7–14. doi:10.1145/2532508.2532511

[3] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 456–464. doi:10.1145/3289600.3290999

[4] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM Be a Personalized Judge? arXiv:2406.11657 [cs.CL] https://arxiv.org/abs/2406.11657

[5] Francesco Fabbri, Gustavo Penha, Edoardo D'Amico, Alice Wang, Marco De Nadai, Jackie Doremus, Paul Gigioli, Andreas Damianou, Oskar Stål, and Mounia Lalmas. 2025. Evaluating Podcast Recommendations with Profile-Aware LLM-as-a-Judge. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. ACM, 1181–1186. doi:10.1145/3705328.3759305

[6] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2305016120 doi:10.1073/pnas.2305016120

[7] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018)*. ACM, 198–206. doi:10.1145/3159652.3159687

[8] Meisam Hejazinia, Kyler Eastman, Shuqin Ye, Abbas Amirabadi, and Ravi Divvela. 2019. Accelerated learning from recommender systems using multi-armed bandit. arXiv:1908.06158 [cs.IR] https://arxiv.org/abs/1908.06158

[9] Balázs Hidasi and Ádám Tibor Czapp. 2023. Widespread Flaws in Offline Evaluation of Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 848–855. doi:10.1145/3604915.3608839

[10] Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, and Ana Peleteiro Ramallo. 2024. Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation. arXiv:2409.11860 [cs.IR] https://arxiv.org/abs/2409.11860

[11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (Jan. 2025), 1–55. doi:10.1145/3703155

[12] Aryan Jadon and Avinash Patil. 2024. A Comprehensive Survey of Evaluation Techniques for Recommendation Systems. arXiv:2312.16015 [cs.IR] https://arxiv.org/abs/2312.16015

[13] Olivier Jeunen. 2019. Revisiting offline evaluation for implicit-feedback recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 596–600. doi:10.1145/3298689.3347069

[14] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[15] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504. doi:10.1007/s11257-011-9118-4

[16] Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. arXiv:2302.14520 [cs.CL] https://arxiv.org/abs/2302.14520

[17] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press. doi:10.1017/9781108653985

[18] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634 [cs.CL] https://arxiv.org/abs/2303.16634

[19] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. doi:10.1016/j.jesp.2017.01.006 Recommends $8/hr as an ethical baseline for MTurk studies.

[20] Gustavo Penha, Aleksandr V. Petrov, Claudia Hauff, Enrico Palumbo, Ali Vardasbi, Edoardo D'Amico, Francesco Fabbri, Alice Wang, Praveen Chandar, Henrik Lindstrom, Hugues Bouchard, and Mounia Lalmas. 2025. Do LLM-judges Align with Human Relevance in Cranfield-style Recommender Evaluation? arXiv:2511.23312 [cs.IR] https://arxiv.org/abs/2511.23312

[21] Bruno L. Pereira, Alan Said, and Rodrygo L. T. Santos. 2025. On the Reliability of Sampling Strategies in Offline Recommender Evaluation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 360–369. doi:10.1145/3705328.3748086

[22] Prolific Team. 2025. Fair Pay on Prolific. https://researcher-help.prolific.com/en/article/2273bd. Prolific requires minimum wage compliance and recommends ˜$12/hr (USD) as a fair baseline.

[23] Marco Rossetti, Fabio Stella, and Markus Zanker. 2016. Contrasting Offline and Online Results when Evaluating Recommendation Algorithms. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 31–34. doi:10.1145/2959100.2959176

[24] Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2025. Analyzing Uncertainty of LLM-as-a-Judge: Interval Evaluations with Conformal Prediction. arXiv:2509.18658 [cs.CL] https://arxiv.org/abs/2509.18658

[25] Ítallo Silva, Leandro Marinho, Alan Said, and Martijn C. Willemsen. 2024. Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 597–608. doi:10.1145/3640543.3645171

[26] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. arXiv:2309.10621 [cs.IR] https://arxiv.org/abs/2309.10621

[27] Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang, and Lizi Liao. 2025. Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution. arXiv:2508.06225 [cs.AI] https://arxiv.org/abs/2508.06225

[28] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, Chicago, IL, USA, 109–116. doi:10.1145/2043932.2043955

[29] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL] https://arxiv.org/abs/2305.17926

[30] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. 2024. Fairness and Diversity in Recommender Systems: A Survey. arXiv:2307.04644 [cs.IR] https://arxiv.org/abs/2307.04644

[31] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (2024), 6889–6907. doi:10.1109/TKDE.2024.3392335

[32] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36. https://arxiv.org/abs/2306.05685