The American Journal of Applied Sciences
ISSN 2689-0992 | Open Access

TYPE Original Research
PAGE NO. 113-120
DOI 10.37547/tajas/Volume07Issue10-13

Check for updates

Leveraging LLMs in recommendation systems

Nisarg B Shah
Product Manager | AI/ML Product Development Seattle, USA

OPEN ACCESS

SUBMITED 25 August 2025
ACCEPTED 28 September 2025
PUBLISHED 30 October 2025
VOLUME Vol.07 Issue 10 2025

CITATION
Nisarg B Shah. (2025). Leveraging LLMs in recommendation systems. The American Journal of Applied Sciences, 7(10), 113–120. https://doi.org/10.37547/tajas/Volume07Issue10-13

Abstract: This paper discusses how large language models are being integrated into recommender systems. The paper attempts to carry out a broad evaluation of the capability of LLM embeddings and generative mechanisms in increasing ranking accuracy as well as decreasing inference latency, ensuring robustness toward the cold-start effect. An application such as this has enormous economic consequences since recommendations drive the central portion of viewing hours on Netflix and purchases on Amazon—over 80% and about 35%, respectively, initiated by recommendation algorithms. The novelty here is that the study applies a unified vector space setup for converting heterogeneous signals—textual descriptions, identifiers, and multilingual metadata—in a comparative analysis of classical and LLM-oriented schemes based on Recall@k, nDCG, and inference-latency metrics. Also included is the hybrid architectures systematization scope that ranges from feature-enrichment pipelines up to fully agent-based solutions. Vocabulary Expansion Techniques, Compressed alongside zero-shot ranking with GPT-4 in the loop, return a dramatic leap in recommendation accuracy at orders of magnitude less latency. Takeaways: LLM helps reduce manual feature engineering, increases ranking accuracy up to 62 per cent, remains stable in multilingual and low data scenarios, and generative and agent components make possible conversational interfaces and multi-step service orchestration. Hybrid solutions offer an optimal trade-off between recommendation quality and computational cost in industrial deployment. This article will be helpful to machine-learning researchers and practitioners, recommender-system developers, and personalization-service architects.

Keywords: large language models, recommender systems, cold start, semantic vectors, generative recommendations

Introduction

Recommender systems have long ceased to be an optional module; they now determine what content users see and which products end up in their shopping carts. Illustrative operational statistics strongly prove how economically important this technology is: more than 80% of viewing hours on Netflix are driven by recommendations (Krysik, 2024), and about 35% of purchases initiated at Amazon are kicked off by the exact mechanism (New America, 2025). Classical approaches provide high accuracy where an extensive history of interactions is available, using collaborative filters, factorization models, and gradient boosters. Pretraining on large data reduces sparsity and enables chat-style preference discovery without manual feature engineering. As a result, personalization becomes more contextual and robust to the cold-start problem. Industrial dynamics follow the same trajectory: major retailers are already testing autonomous shopping agents that accept freely formulated user goals, invoke search filters, and place orders, signaling a shift in focus from pointwise ranking to the orchestration of multiple tools within a unified agent scheme (Choudhary, 2025). These trends indicate a structural transition in the industry from locally optimized models to adaptive, conversational, and universally scalable personalization engines.

Materials and Methodology

The study is drawn from fifteen primary sources under peer review articles and industry reports. The theoretical foundation embraces transformer-based recommendation methods (Shehmir & Kashef, 2025) and hybrid architecture taxonomy with operational dynamics (Liu et al., 2024). Real value-added embeddings by LLM compared with classic feature pipelines are brought in through empirical tests of Compressed Vocabulary Expansion (Zhang et al., 2025) and zero-shot ranking with GPT-4 (Hou et al., 2023), as well as model transferability exploration under multilingual scenarios (Iana et al., 2024). Ethical and practical issues are discussed based on examples from agent-based architecture applications in commercial settings (Choudhary, 2025; McLymore & Bensinger, 2024) as well as user experiments involving conversational assistants.

Methodologically, this work combines several approaches. First, a comparative analysis of classical and LLM-oriented schemes was performed based on ranking-accuracy metrics (Recall@k, nDCG) and inference latency, drawing on data concerning Hit@10 improvements and latency reductions (Liang et al., 2024; Liu et al., 2025). Second, a systematic review of industry case studies identified requirements for the orchestration of multi-step agents and hybrid pipelines (Wang et al., 2023; New America, 2025). Lastly, a plausible content evaluation of client responses and A/B-test outcomes proves the impact of LLM explanations on trust and engagement (Yun & Lim, 2025; Cui et al., 2024). This is a viable approach that results in a sound summary to pave the way for more experiments and thereby derive practical deployment recommendations.

Results and Discussion

The primary value of LLMs for recommendations is that a single model maps heterogeneous signals — texts, identifiers, multilingual descriptions — into dense semantic vectors comparable with one another. A comprehensive review of works from 2018 to 2024 highlights the shift to a unified transformer backbone, which reduces manual feature engineering and improves ranking accuracy (Shehmir & Kashef, 2025). This is validated empirically by results from the Compressed Vocabulary Expansion method: when CoVE-indexed tokens replace the traditional one-hot identifiers, recommendation accuracy increases to 62% and inference latency drops by about 100-fold — critical in online settings (Zhang et al., 2025). A CoVE framework is illustrated in Figure 1.

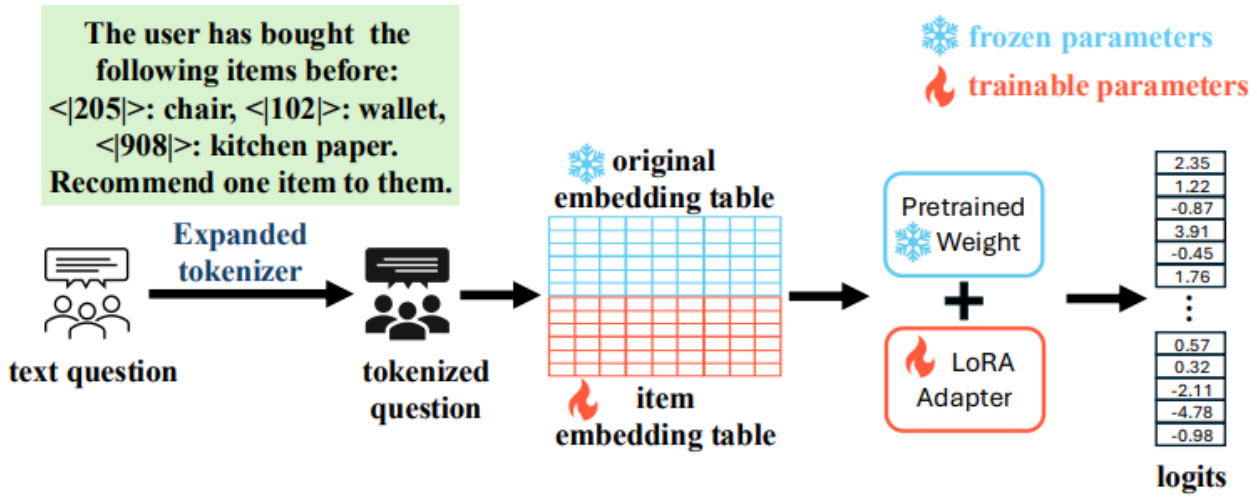

Fig. 1. An overview of our CoVE framework (Zhang et al., 2025)

In multilingual scenarios, the advantages become even more pronounced, since a unified embedding space enables knowledge transfer between languages; for example, in the xMIND suite, a monolingual model lost over one third of its Recall@k when transferred to a new locale, whereas LLM-based embeddings exhibited almost no degradation (Iana et al., 2024).

The second major strength of LLMs lies in their capacity to operate in zero or few-shot settings, where there is virtually no historical data in new markets. Experiments with zero-shot ranking demonstrate that properly crafted prompts enable GPT-4 to compete with domain-specific rankers and to raise nDCG@10 without further training compared to the best classical models (Hou et al., 2023). The TaxRec taxonomic framework amplifies this effect: Recall@10 on the public Movie dataset increases from 0.18 to 0.30, representing a relative improvement achieved in the complete absence of additional training (Liang et al., 2024). When computational cost reduction is required, the knowledge of a large model can be transferred to a lightweight sequential ranker: the DLLM2Rec strategy boosts average Hit@20 across three popular models by 47.97%. It reduces response time from hours to seconds, thereby shortening the product-to-market cycle to mere days (Cui et al., 2024).

Finally, the generative nature of LLMs enables fully conversational interaction, whereby users formulate queries in natural language, refine constraints, and receive explanations of the model's choices. A diary study presented at CHI 2025 involving twelve participants who used custom GPT assistants for music recommendations found that such an interface helps to uncover latent preferences, stimulates exploratory behavior, and deepens understanding of one's tastes (Yun & Lim, 2025).

Integration schemes in which large language models serve as an additional layer atop classical recommendation pipelines form a continuum spanning from lightweight feature augmentation to fully agentic systems. The comprehensive review LLM-ERS notes that hybrid architectures are gradually supplanting attempts to build a single universal model, allowing the combination of the precision of specialized rankers with the semantic flexibility of generative components (Liu et al., 2024). The trend of LLM-enhanced recommender systems is shown in Fig. 2.

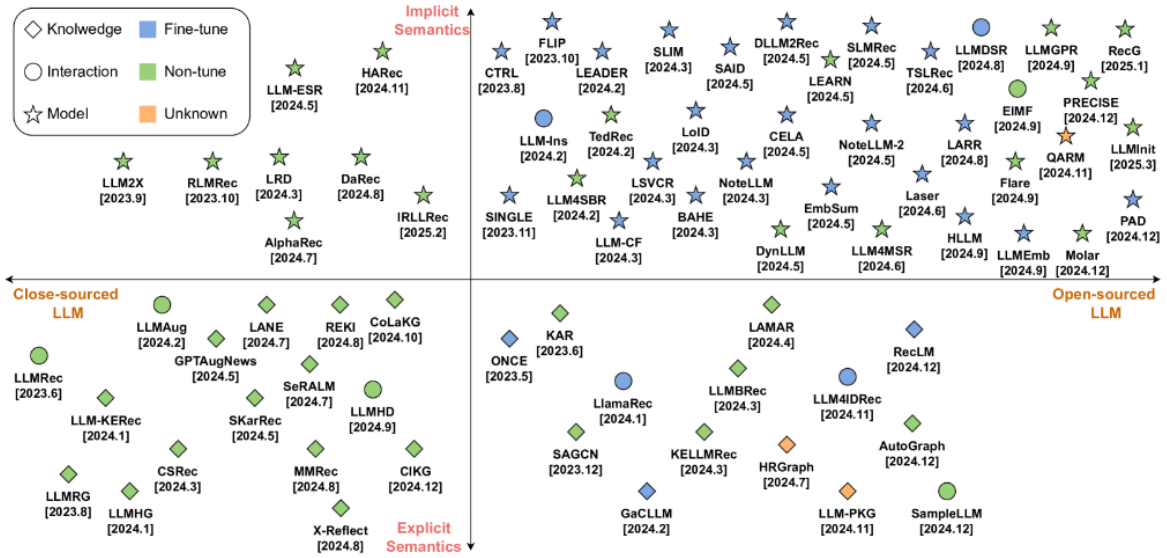

Fig. 2. The trend of LLM-enhanced recommender systems (Liu et al., 2024)

At the beginning of this spectrum are Feature Augmentation schemes in which an LLM generates new attributes for users or items, while a traditional gradient booster performs ranking. Recent work on scaling inference has shown that increasing reasoning depth during feature generation yields an additional 12% gain in NDCG@10 and requires no changes to the online stack, since the computationally intensive stage is executed offline (Liu et al., 2025). In practice, this strategy preserves existing latency metrics while enhancing long-tail coverage, where manual feature engineering typically fails.

The next level consists of LLM rankers, which receive a candidate list and order it through direct text output. With carefully tuned prompt templates, GPT-4 without retraining achieves accuracy comparable to fully trained neural rankers on two public datasets, and in particular outperforms them on the Hit@10 metric for fresh or infrequently accessed items (Hou et al., 2023). Such results explain the industry's growing interest in Top-N scenarios, where an expensive LLM is invoked only for a final shortlist of several dozen items, keeping query costs within reasonable bounds.

The third group of solutions relies on Retrieval-Augmented Generation, connecting to external knowledge stores to address cold-start issues and expand the catalog. The ColdRAG method dynamically constructs an entity graph from textual descriptions of new items, retrieves relevant nodes, and prompts the LLM to rank candidates based on the discovered relationships, as illustrated in Fig. 3; this approach yielded significant improvements in Recall and nDCG across three benchmarks without any model retraining (Yang et al., 2025). Unlike pure generative approaches, this scheme reduces the risk of hallucinations, since each decision is grounded in verifiable facts returned by the retrieval stage.

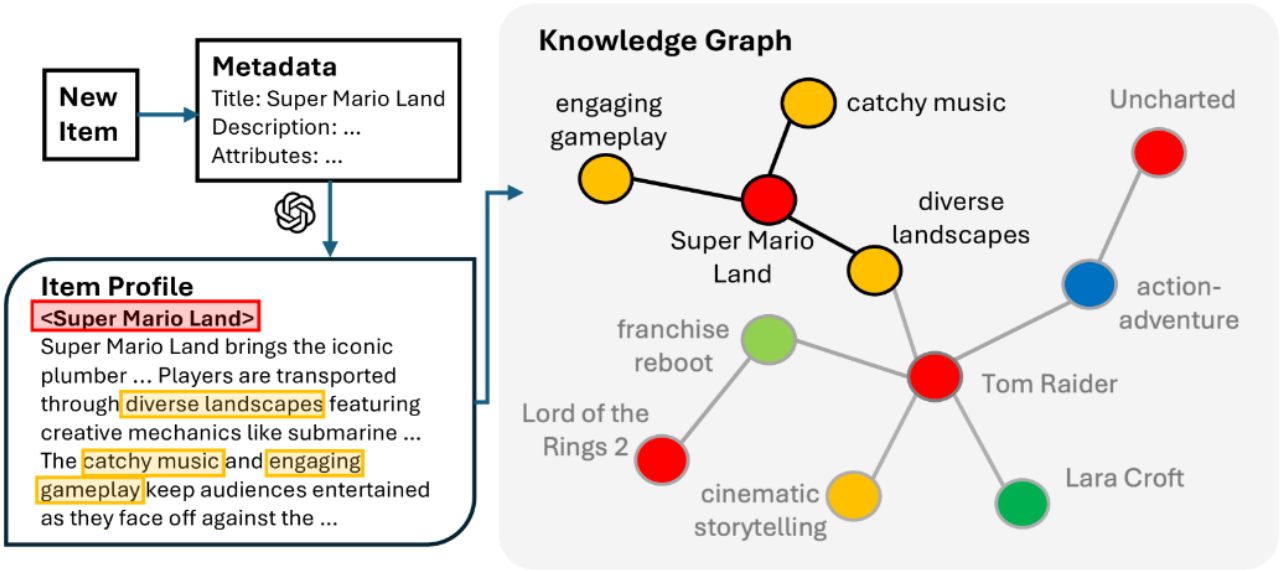

Fig. 3. Illustration of ColdRAG's adaptability to an item cold-start scenario (Yang et al., 2025)

Finally, at the apex lie agentic architectures, in which an LLM orchestrates a chain of tools: it formulates a plan, invokes a search API, filters the results, and explains the final choice. The RecMind study demonstrates that a self-prompting planning algorithm enables the agent to outperform other zero-shot methods consistently and to approach the performance of the trained P5 model when solving a wide range of tasks, including the composition of complex lists based on user constraints (Wang et al., 2023). In parallel, commercial players are deploying similar schemes: in February 2024, Amazon began the phased rollout of the Rufus assistant, capable of answering freely formulated customer questions and translating intents into a ready shopping cart, reflecting a shift from pointwise ranking to multi-step orchestration of services (McLymore & Bensinger, 2024). Thus, the movement from embeddings to agents illustrates the industry's ambition to combine the best aspects of statistical models and linguistic reasoning into a single adaptive personalization platform.

When the integration schemes described earlier transition into real products, their quality is measured not only by ranking accuracy, but also by how deeply they engage users in dialogue, how much trust they inspire, how broad their domain of applicability is, and whether they can complete the path from intent to purchase within a single session.

In chat-based applications for media and shopping, the language model tweaks the user's questions, narrows down the options, and presents the final suggestion in simple words. The push for openness puts reason creation front and center. Organized checks find that there are still not many works that test explanation making in recommendations directly, but even inside fields, the seen data shows higher user trust with clear reasons given. Cross-domain and multilingual scenarios impose special generalization requirements. Under a direct language transfer—moving a model from one language to another—accuracy drops noticeably; this issue is addressed either by adapters or by algebraic operations on embeddings that preserve the user's semantic core when switching contexts. Such an approach is essential for platforms where the same individual consumes news, music, and video content in different languages and on other devices.

An autonomous shopping agent takes the model beyond ranking by entrusting it with planning the entire session. A voice assistant integrated into a primary retail application already answers freely formulated questions about product specifications, use cases, and comparisons, then translates the intent into a ready shopping cart. Thus, conversational interfaces, explanations, multilingualism, and agentic planning form a complementary set of scenarios in which large language models transform recommender systems from hidden mathematical mechanisms into open, interactive, and context-sensitive interlocutors that enhance the value of each user session.

The scenarios considered above convincingly demonstrate that large language models open new horizons for personalization, yet moving to production traffic presents several engineering and ethical challenges. The first challenge relates to response latency and inference cost, since the parameter sizes of the models demand significant computational resources. The industry responds with cascade schemes

offering different quality levels, offline embedding generation, distillation, quantization, and specialized accelerators, to maintain response times within real-time interaction limits without exceeding infrastructure budgets.

The second challenge is inherently of generation: the model might embellish with unreliable facts, use toxic vocabulary, or accidentally leak content that should not be known. Multi-level filters are standard practice now, including safety classifiers at input and output, link verification in the retrieval memory, and manual rules tuned to dynamic moderation signals. Such measures make hallucinations less likely but add to pipeline complexity and require continuous monitoring.

The third aspect concerns quality evaluation. Offline ranking metrics such as AUC or nDCG@k provide a quick a priori assessment, but do not reflect the long-term value to the platform, where user retention and lifetime revenue become key indicators. Consequently, companies increasingly combine offline candidate selection with online experiments and construct multi-objective loss functions that simultaneously optimize instantaneous relevance, conversions, diversity, and strategic business metrics.

Finally, privacy and data-licensing considerations become critical under tightening regulatory requirements. Models require large volumes of user information, yet legal frameworks insist on data minimization and strictly defined processing purposes. Anonymization, federated learning, differential privacy, and provenance control of the training corpus mitigate risks but impose constraints on architecture choice and data sources. Achieving a rational balance between innovation and legal compliance becomes as essential a competitive factor as recommendation accuracy.

Overcoming the challenges described above requires not merely selecting an architecture but devising an operational strategy that maintains responsiveness, ensures quality, and preserves user trust. In practice, transitioning to such a strategy begins with deploying the large language model closer to the point of service, which reduces network latency and enables caching of pre-generated embeddings. Static features are created offline, dynamic features are updated asynchronously, and the heavy generative ranker is invoked only for queries where the gain in quality justifies the cost, thereby keeping infrastructure expenditures under control.

Safety layers shall then be implemented above the main pipeline to sanitize undesirable content further and to tweak ranking after generation. Rules, lightweight neural classifiers, and post-hoc reranking enforce corporate security policies and regulatory constraints as much as possible with minimal degradation of performance. Another post-hoc reranker layer may address factors that are implicitly unaccounted for by the language model, e.g., catalog diversity or strategic profitability.

Even the most carefully calibrated configuration requires a gradual rollout to users. Rather than switching all traffic at once, a phased release is used in which the new version initially handles a small percentage of requests and then gradually expands coverage. In parallel, a comparative experiment with a control group measures not only immediate clicks but also long-term retention, repeat purchases, and user-experience quality. Should any key metric fall below standard, roll back the changes and reassess the hypothesis. This cycle of experimentation makes the LLM deployment a very managed process that allows for novelty and stability to take place.

So, putting big language models into recommender systems provides a way to handle different kinds of signals, reduces manual work on features, improves performance when there is not much data, and helps chat interfaces. Hybrid setups keep the exactness of special rankers and the freedom of making parts; however, to use these setups well, there needs to be a quick setup and thinking about costs in making them real, and many levels of checks against made-up answers, along with a deep look at quality while following privacy rules. That's why here we give the main tips for building a strong, growing, and right-with-ethics personal touch system.

Conclusion

The incorporation of large language models results in the opening of new unified textual, identificational, and multilingual signal processing possibilities with less manual feature engineering efforts and greater algorithmic robustness against cold-start effects that these systems can experience. Heterogeneous input data is enabled online by forming a unified vector space by large language models, thereby delivering a quantum leap in ranking accuracy at still acceptable infrastructure requirements. The generative nature of such models pushes recommender functionality by way of dialoging,

drilling into user preferences, and explaining recommendation logic on the fly.

Hybrid architectures comprising classical rankers and generative components prove in practice the best compromise between the goals of accuracy and computational efficiency. Further quality gains are made under Feature Augmentation schemes without any modifications to the online stack, while LLM rankers applied at the last mile selection among top-N candidates ensure query costs remain within practical limits. Retrieval-Augmented Generation approaches reduce hallucination risk since they are based on external knowledge stores, while agentic architectures allow multi-step service orchestration, translating user intent into a ready order.

Large Language Model-based solutions will only be successful if the engineering and ethical challenges are solved. The problem of latency of response and high inference cost is solved through cascade pipelines, distillation, and quantization; meanwhile, unreliability and unwanted content are fought against through multi-level filters on both input and output, together with offline ranking metrics for quality evaluation, and online experiments that consider long-term value to the platform. They make sure rules about privacy and data licensing are followed by using techniques such as anonymization, federated learning, and differential privacy.

Advices an operational strategy to place the model closer to where it is being consumed, thereby reducing network latency, offline static embeddings generation, asynchronous update for dynamic features, and making generative stages heavy by invoking them only when justified by quality gains. Functional diversification without full retraining of the transformer is enabled through domain adaptation via lightweight adapter layers. Parallel comparative experiments with phased release of new versions, together with monitoring key retention and revenue metrics, allow a controlled deployment process that ensures innovation alongside platform stability.

The mix of LLM meaning ease and the known ways of old recommender models builds a strong base for growing, friendly, and morally right personal touch fixes that can fit quick shifts in user wants and business goals.

References

1. Choudhary, V. (2025). *AI dominance in e-commerce has a new focus: agentic checkout technology*. Retail Brew. https://www.retailbrew.com/stories/2025/07/30/ai-dominance-in-e-commerce-has-a-new-focus-agentic-checkout-technology
2. Cui, Y., Liu, F., Wang, P., Wang, B., Tang, H., Wan, Y., Wang, J., & Chen, J. (2024). Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Model. *ArXiv (Cornell University)*. https://doi.org/10.1145/3640457.3688118
3. Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., & McAuley, J. (2023). Large Language Models are Zero-Shot Rankers for Recommender Systems. *Arxiv*. https://doi.org/10.48550/arxiv.2305.08845
4. Iana, A., Glavaš, G., & Paulheim, H. (2024). MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation. *Arxiv*. https://doi.org/10.48550/arxiv.2403.17876
5. Krysik, A. (2024, June 14). *Netflix Algorithm: How Netflix Uses AI to Improve Personalization*. Stratoflow. https://stratoflow.com/how-netflix-recommendation-algorithm-work/
6. Liang, Y., Yang, L., Wang, C., Xu, X., Yu, P. S., & Shu, K. (2024). *Taxonomy-Guided Zero-Shot Recommendations with LLMs*. Arxiv. https://arxiv.org/abs/2406.14043
7. Liu, Q., Zhao, X., Wang, Y., Wang, Y., Zhang, Z., Sun, Y., Li, X., Wang, M., Jia, P., Chen, C., Huang, W., & Tian, F. (2024). Large Language Model Enhanced Recommender Systems: Taxonomy, Trend, Application and Future. *Arxiv*. https://doi.org/10.48550/arxiv.2412.13432
8. Liu, W., Du, Z., Zhao, H., Zhang, W., Zhao, X., Wang, G., Dong, Z., & Xu, J. (2025). *Inference Computation Scaling for Feature Augmentation in Recommendation Systems*. Arxiv. https://arxiv.org/abs/2502.16040
9. McLymore, A., & Bensinger, G. (2024, February 5). When Amazon's new AI tool answers shoppers' queries, who benefits? *Reuters*. https://www.reuters.com/technology/when-amazons-new-ai-tool-answers-shoppers-queries-who-benefits-2024-02-05/
10. New America. (2025). *Why Am I Seeing This?* New America. https://www.newamerica.org/oti/reports/why-am-i-seeing-this/case-study-amazon/

11. Shehmir, S., & Kashef, R. (2025). LLM4Rec: A Comprehensive Survey on the Integration of Large Language Models in Recommender Systems—Approaches, Applications and Challenges. *Future Internet*, *17*(6), 252. https://doi.org/10.3390/fi17060252
12. Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Huang, X., Lu, Y., & Yang, Y. (2023, August 28). *RecMind: Large Language Model Powered Agent For Recommendation*. Arxiv. https://doi.org/10.48550/arXiv.2308.14296
13. Yang, W., Zhang, W., Liu, Y., Han, Y., Wang, Y., Lee, J., & Yu, P. S. (2025). *Cold-Start Recommendation with Knowledge-Guided Retrieval-Augmented Generation*. Arxiv. https://arxiv.org/abs/2505.20773
14. Yun, S., & Lim, Y. (2025). User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3706598.3713347
15. Zhang, H., Zhang, T., Yin, J., Gal, O., Shrivastava, A., & Braverman, V. (2025). *CoVE: Compressed Vocabulary Expansion Makes Better LLM-based Recommender Systems*. Arxiv. https://arxiv.org/abs/2506.19993