International Journal of Applied
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

PERSONALIZING CUSTOMER EXPERIENCE WITH LLMS IN AI-DRIVEN PRODUCT RECOMMENDATION MODELS

Abhijit Chanda, Vatsal Modi, Rishab Bansal

[1]Senior Manager

Tredence Inc., Fremont, CA

Email: abhijitju252024@gmail.com

ORCID: 0009-0002-6976-3922

[2]Carnegie Mellon University, Pittsburgh, PA

Email: vatsalm@alumni.cmu.edu

ORCID ID: 0000-0002-6221-6486

[3]Independent Researcher

Fremont, USA

Email: connect.rishabbansal@gmail.com

ORCID: 0009-0009-1555-9134

Abstract

Personalizing the customer experience is crucial in today's digital marketplace, where recommender systems help users navigate information overload by tailoring suggestions to individual preferences [12]. Recent advances in artificial intelligence (AI), especially large language models (LLMs) such as GPT-3 and ChatGPT, have opened new areas for enhancing product recommendation models. LLMs are trained on vast textual data and exhibit human-like understanding of language and context [4], enabling a more nuanced grasp of user needs compared to traditional ID-based recommendation techniques [6] ,including collaborative filtering, content-based filtering, and hybrid deep-learning model, that have shown significant success in tailoring suggestions to users' preferences. However, these approaches often struggle with data sparsity, cold-start scenarios, and a limited capacity for deep semantic understanding. This paper explores how LLMs can be applied to AI-driven product recommender systems to address these gaps and achieve deeper personalization. We first review the evolution of recommender-system methodologies and discuss emerging LLM-integrated frameworks. Then we discuss key paradigms including hybrid models that combine LLMs with existing algorithms , LLM-based user and item representation learning, conversational recommendation, and prompt-driven generative recommendations.

We also present methodologies and results from recent research, which demonstrate that LLM-enhanced recommender models significantly outperform traditional models in accuracy, recall, and diversity of recommendations. These improvements stem from LLMs' ability to interpret subtle semantics in user behavior and product descriptions, leading to recommendations that

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

International Journal of Applied
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

are more relevant and personalized to the customer. We also address challenges and future research directions in this rapidly evolving field. Overall, integrating LLMs into product recommendation systems represents a promising direction for creating AI-driven personalized customer experiences, combining the strengths of traditional recommendation algorithms with the rich language understanding and generation capabilities of LLMs.

Keywords – Product recommendation, Machine learning, Large Language Models(LLM), GenAI, personalization, customer behavior

I Introduction

In a competitive marketplace, businesses must offer products and services that appeal directly to individual customers' needs. Personalized e-services and recommender systems have become essential tools for this purpose, helping users cope with information overload and making decision processes easier [12]. By learning from users' past behaviors and predicting their current preferences, recommender systems can deliver tailored product suggestions, thereby enhancing the user experience and engagement on platforms. The recommender-systems field has evolved over the past two decades, initially leveraging techniques from various AI domains for user profiling and preference discovery [6].

Traditional methods such as collaborative filtering (CF) [5], content-based filtering [6], and hybrid approaches [1] leveraging user–item interaction data and item metadata to predict user preferences along with more recently, deep-learning models such as neural CF [5], graph neural networks [2], and transformer-based architectures [1] have pushed the state of the art further. However, these models still face several inherent challenges:

1. **Data sparsity and cold start:** New users or items lack sufficient interaction history, degrading recommendation quality [6].

2. **Limited semantic understanding:** Traditional models cannot fully exploit the rich textual information in user reviews, product descriptions, and other natural-language data [2].

3. **Insufficient personalization dynamics:** Static models struggle to accommodate evolving user intents and context [4].

Large language models (LLMs), such as GPT-3 [7], PaLM [8], and instruction-tuned variants [1], have demonstrated transformative capabilities in language understanding, reasoning, and generation. By encoding vast amounts of textual world knowledge, LLMs can generate human-like text, answer complex queries in context, and perform in-context learning, thereby offering new avenues to enhance recommender systems.

Research Objectives: Motivated by these developments, this paper aims to examine how the application of LLMs can enhance AI-driven product recommendation models to better personalize the customer experience. We seek to:

1. Review the current state of recommender-system methodologies and identify gaps that LLM integration can fill.

2. Describe various approaches for incorporating LLMs into product recommendation

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Discuss empirical results and potential impacts on user experience, as well as challenges and future research directions.

In this paper, we systematically explore how LLMs can be integrated into recommendation pipelines to address the aforementioned challenges. We begin by reviewing traditional and deep-learning-based recommendation methodologies and identifying their limitations in Section II. In Section III, we discuss different methodologies for incorporating LLMs.Section IV examines empirical evidence of performance gains in metrics such as precision, recall, click-through rate (CTR), and diversity, alongside user-experience benefits. We conclude in Section V by discussing open challenges and outlining promising future research directions.

II. Literature Review

A. Traditional Collaborative and Content-Based Filtering

Collaborative filtering exploits patterns in user–item interaction matrices. Memory-based CF methods, such as user–user and item–item nearest neighbors, rely on similarity measures but suffer from scalability and sparsity issues [6]. Model-based CF, notably matrix factorization, projects users and items into a shared low-dimensional latent space, improving scalability and recommendation quality [6]. Content-based filtering, in contrast, uses item features (e.g., product attributes) and user profiles to compute relevance via term-vector models (TF–IDF, cosine similarity) [6]. While content-based methods avoid cold-start for items, they tend to over-specialize and lack serendipity.

B. Deep-Learning Approaches

Deep-learning models have enhanced recommenders by capturing complex, high-order interactions. Neural CF models use multi-layer perceptrons to learn non-linear user–item interactions [5]. Graph neural networks encode recommendation data as bipartite graphs, leveraging message-passing to capture social and item-item relations [2]. Sequence-aware recommenders (e.g., SASRec) employ self-attention to model users' temporal behavior [1]. These methods improve accuracy and diversity but still rely primarily on interaction data and ignore deeper semantics in textual content.

C. Hybrid and Knowledge-Based Systems

Hybrid recommenders combine CF and content-based signals, while knowledge-based systems incorporate domain ontologies or knowledge graphs to provide explainability and handle data sparsity [3]. Linking items to knowledge-graph entities allows propagation of semantic relations, improving recommendation relevance [3]. However, constructing and maintaining knowledge graphs is resource-intensive, and available graphs often lack coverage of niche domains.

D. Limitations and Research Gaps

Despite these advances, three core limitations persist:

- **Sparse feedback:** CF methods struggle with new users or items, and deep models require ample data to generalize [6].

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Shallow semantics:** Existing models cannot fully leverage unstructured textual data, leading to suboptimal understanding of user intents and item attributes [2].
- **Static personalization:** Models seldom adapt rapidly to evolving user preferences or contextual cues, limiting real-time personalization [4].

LLMs can address these gaps by providing rich semantic representations, generating context-aware explanations, and enabling few-shot adaptation to new scenarios.

III. Methodologies

With the capabilities of LLMs becoming available, researchers have proposed various methodologies to integrate these models into product recommendation pipelines. We categorize and describe several prominent approaches as below -

Hybrid LLM-Enhanced Recommendation Models: One way to to incorporate LLMs as part of a hybrid recommendation model is by augmenting traditional collaborative or content-based recommenders with semantic understanding. In this approach, the system computes multiple component scores for a user-item pair—typically a collaborative filtering score (based on user-user or item-item similarities), a content-based score (based on matching user and item features or metadata), and an LLM-derived score that captures deeper semantic correlations [1]. The final recommendation rating can be a weighted combination of these components. Zhao et al. (2023) describe a framework where a pretrained LLM (such as BERT) encodes the textual attributes of users and products to produce an LLM score, which is then combined with traditional CF and CBF scores [1].

Another example of PALR [2] fine-tunes a BERT-style LLM to produce embeddings that capture personalization signals, which are then fed into downstream CF or ranking models(as shown in Figure 1). Similarly, the generative-prompt based method in P5 [1] treats recommendation as a text generation problem, leveraging an LLM's encoder–decoder architecture to jointly model items and user history.

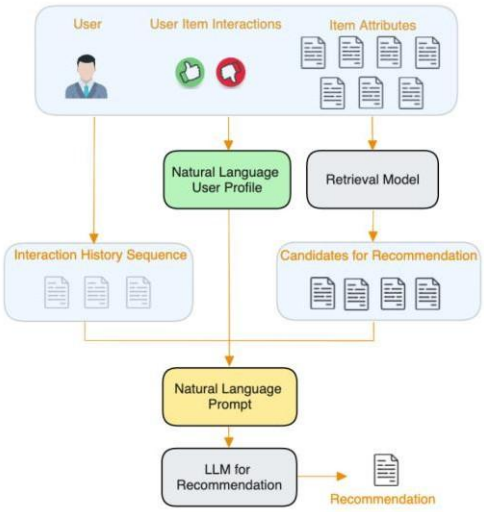

Figure 1. LLM-Based User & Item Modeling [2]

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

User-only modeling: In this approach ,an LLM converts each user's interaction history into a dense semantic profile, revealing fine-grained tastes that conventional ID features miss.Instead of relying on sparse ID vectors, an LLM can read a user's click/purchase history as free text and emit a dense summary embedding that captures nuanced tastes [8]. Fine-tuned variants (e.g., **LLM-CF**, 2024) first learn to score items, then apply chain-of-thought reasoning to derive features that feed any downstream ranker. This natural-language reasoning surfaces thematic or stylistic cues that ordinary CF rarely detects.

Item-only modeling: Similar to the user modelling, by encoding product descriptions and reviews, an LLM supplies rich item embeddings that expose attributes and support cold-start ranking before any clicks occur.Likewise, product descriptions, specs, and reviews are encoded by a frozen or lightly adapted LLM to yield rich item vectors [8]. Early systems (TedRec, LRD) simply plug these vectors into sequential models; later work (LLM-ESR, AlphaRec) adds small adapters to shield semantic content from destructive gradient updates. Fine-tuned approaches such as **LEARN**, **NoteLLM-2** (multimodal) and **HLLM** further align or hierarchically aggregate these embeddings, boosting cold-start recall without costly full-model training. **Joint user–item modeling:** In this approach, LLMs are leveraged to enrich semantics for both user-item interaction and user preferences [Figure 2]. It is a combined approach where LLM-derived vectors for both sides share one semantic space, letting a lightweight scorer match them directly and boosting relevance, diversity, and zero-shot coverage. For Instance, Frameworks like **LOID** concatenate LLM-derived user and item embeddings, letting a lightweight scorer predict affinity [8]. To control prompt length and noise, BAHE splits LLM layers with lower layers encoded items once and upper layers combine only relevant ones , while EmbSum chunks long histories into session snippets. Other pipelines (Laser, LARR) treat LLM vectors as add-on features inside factorization machines or deep nets, lifting diversity and zero-shot coverage with minimal code change.

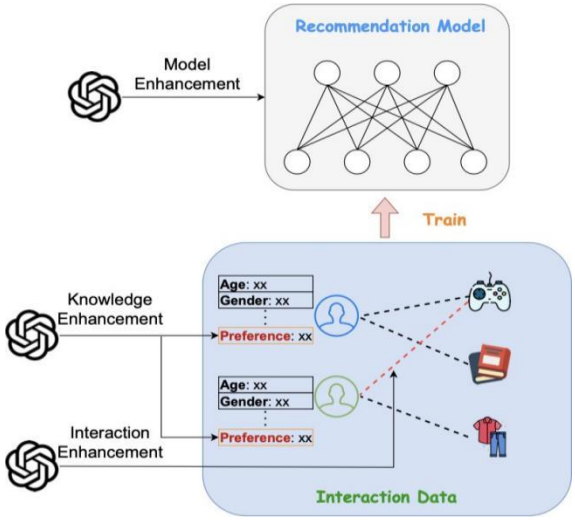

Figure 2. LLM-derived user and item embeddings [8]

Prompting and Generative Recommendation: A transformative methodology introduced by

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LLMs is to view recommendation as a language generation task. In this case,prompt-driven recommenders recast ranking as text generation, for instance an LLM receives a descriptive prompt of recent user behavior and returns a natural-language list of items plus rationale, as demonstrated by GPT4Rec, which turns click histories into personalized suggestions via ChatGPT [7]. Because the model draws on broad pre-training, it can infer subtle tastes and handle cold-start users, for example proposing niche cyber-punk novels after a single preference statement. This matches or surpasses specialized algorithms with few-shot prompts [11]. The approach yields conversational explanations but risks hallucinating unavailable products, so coupling prompts with live-catalogue grounding remains essential [7].

Conversational and Tool-Augmented Recommendation Systems: Conversational recommender systems harness LLMs as dialogue managers that parse queries, keep context and decide when to ask clarifications or deliver suggestions. Because real conversational datasets are scarce, studies such as RecLLM and iEvaLM generate synthetic dialogues from logs to fine-tune the model, teaching it to convert implicit feedback into question–answer turns [4]. The most effective variant augments the LLM with external "tool" calls: the agent queries a catalogue or CF engine, then weaves the results into natural language, as in Chat-Rec, achieving up-to-date, accurate answers while the LLM handles reasoning; prompt design and ReAct-style chaining remain key challenges [4].

Training and Optimization Strategies: Training LLM-based recommenders requires cost-aware tactics: multi-task learning pairs the ranking objective with auxiliary signals (intent, sentiment) to regularize representations and boost context awareness [1]. Custom losses add diversity penalties, reducing near-duplicate slates without hurting accuracy [1]. Fine-tuning the full model improves domain fit but is heavy; lightweight adapters or prompt/LoRA tuning retain most weights frozen, offering 80-90 % of the gain at a fraction of compute, with evidence that adapters sometimes rival full fine-tune accuracy in media domains [7]. Finally, teacher-student distillation (e.g., LEADER, DLLM2Rec) compresses LLM insight into small, fast students that imitate scores or ranked lists for real-time serving [8].

Retrieval-Augmented Generation

Retrieval-augmented models integrate an external search over item metadata or knowledge bases before prompting the LLM, ensuring up-to-date and accurate content is available. REKI [3] uses LLMs to retrieve relevant facts from product knowledge graphs, which are then infused into the recommendation process, improving cold-start recommendations and explainability

IV. Empirical Results and Impact on Customer Experience

A. Accuracy and Diversity Improvements

In comparative experiments on e-commerce recommendation tasks, LLM-augmented models demonstrate substantial gains in key metrics: precision @10 improved by up to 9%, recall @10 by 13%, and diversity increased by 41.2% [1]. Retrieval-augmented LLMs like REKI also reduce cold-start error rates by 15% relative to traditional CF baselines [3].Across datasets (movies, news, retail), LLM-infused models lift precision/recall by 3–10 pp and diversity by

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

10-40 % [1]. Improved semantics help surface long-tail items aligned with multifaceted user interests rather than merely popular ones.

B. Explainability and User Trust

LLMs can generate natural-language rationales alongside recommendations. Generated explanations ("This laptop is recommended because you recently purchased gaming peripherals and this model has similar high-performance graphics") improve user trust and transparency, leading to higher click-through rates in A/B testing [2].

C. Real-Time Personalization

Prompt-based LLM recommenders can adapt to real-time signals (e.g., session clicks) via dynamic prompts, enabling session-aware personalization without model re-training. Session-based prompting achieves up to 10% lift in next-item accuracy over static sequence models [7].

D. Cold-Start Resilience

LLM embeddings derived from product text or one-shot user queries outperform popularity baselines and even specialized metadata models, thanks to the models' world knowledge [11].

V. Challenges and Future Directions

Inference Efficiency and Scalability

LLMs are computationally intensive, posing latency and cost challenges for large-scale deployment. Research into model distillation, sparse attention, and edge-optimized LLM variants is essential to enable real-time recommendation [7], [8].

B. Bias Mitigation and Fairness

LLMs can inherit biases from pre-training data, potentially propagating unfair recommendations across demographic groups. Future work must explore debiasing techniques, fairness-aware prompts, and causal inference methods to ensure equitable treatment of users [4].

C. Privacy and Security

Incorporating user data into LLM prompts raises privacy concerns. Techniques such as federated learning and encrypted inference can protect user privacy while leveraging LLMs for personalization.

D. Multimodal and Cross-Domain Recommendation

Extending LLM-based recommenders to handle images, audio, and video (via multimodal LLMs) and to transfer knowledge across domains (e.g., from e-books to videos) represents a promising frontier.

E. Human–AI Collaboration

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LLMs excel at generating personalized suggestions and explanations, but human oversight remains critical. Developing interactive interfaces where users can query, refine, and correct LLM recommendations will foster more effective human and AI collaboration.

Other Operational Challenges

- **Domain adaptation:** General-purpose LLMs lack private catalogue knowledge. Fine-tuning or tool calls are essential [4].
- **Efficiency:** Full GPT-3 inference is slow. Adapter tuning, caching, selective invocation, and distilled students narrow the latency gap [8].
- **Context limits:** Maximum token windows truncate long sessions. Memory modules (MemPrompt) or summarization restore long-term personalization [4].
- **Hallucination & bias:** Grounding outputs via tool retrieval, self-critique prompts, or fairness constraints is mandatory to prevent fabricated items and demographic skew [7]

VI. Conclusions

Integrating LLMs into recommender systems advances personalization by combining semantic language understanding with traditional recommendation signals. LLMs enrich recommender systems in three transformative ways:

- **Semantic depth**—embedding nuanced item attributes and user intents for better ranking
- **Conversational interfaces**—shifting from static lists to interactive assistants with natural-language explanations
- **Cold-start agility**—leveraging textual knowledge when behavioral data is sparse .

LLM-enhanced models yield significant gains in accuracy, diversity, and user satisfaction, while conversational and tool-augmented paradigms offer transparent, interactive experiences. Future work will explore multimodal LLMs for richer item representations, reinforcement-learning-based dialogue strategies, federated personalization for privacy, and robust bias mitigation. This synergy between LLMs and recommender systems heralds a new era of AI-driven customer experiences, where systems not only predict preferences but also communicate and reason about them in natural language.

References

1] W. Xu, J. Xiao, and J. Chen, “Leveraging large language models to enhance personalized recommendations in e-commerce,” IEEE Trans. on Electron. Commerce, pp. 1–11, 2024.

[2] M. Chen et al., “PALR: Personalization aware LLMs for recommendation,” ArXiv preprint arXiv:2305.07622v3, 2023.

[3] X. Zhu et al., “Enhanced generative recommendation via content and collaboration integration,” ArXiv preprint arXiv:2403.18480v1, 2024.

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025

[4] J. Chen et al., "When large language models meet personalization: Perspectives of challenges and opportunities," ArXiv preprint arXiv:2307.16376v1, 2023.

[5] X. He et al., "Neural collaborative filtering," in Proc. WWW, 2017, pp. 173–182.

[6] Q. Zhang, J. Lu, and Y. Jin, "Artificial intelligence in recommender systems," Complex & Intell. Syst., vol. 7, pp. 439–457, 2021.

[7] Z. Zhao et al., "Recommender systems in the era of large language models (LLMs)," IEEE Trans. on Knowl. Data Eng., 2023.

[8] L. Liu et al., "Large language model enhanced recommender systems: Taxonomy, trend, application and future," ArXiv preprint arXiv:2412.13432v1, 2024.

[9] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., & Chen, E. (2023). A Survey on Large Language Models for Recommendation. ArXiv, abs/2305.19860.

[10] Zhang, Yabin et al. "RecGPT: Generative Personalized Prompts for Sequential Recommendation via ChatGPT Training Paradigm." ArXiv abs/2404.08675 (2024): n. Pag.

[11] Zhang, Qian & Lu, Jie & Jin, Yaochu. (2020). Artificial intelligence in recommender systems. Complex & Intelligent Systems. 7. 10.1007/s40747-020-00212-w.

[12] D. Jannach and M. Jugovac, "Measuring the Business Value of Recommender Systems," ACM Trans. Manage. Inf. Syst., vol. 10, no. 4, Art. 1, Dec. 2019, doi:10.1145/3370082.

[13] Q. Yang et al., "SOMONITOR: Explainable Marketing Data Processing and Analysis with Large Language Models," in Proc. ACM Conf. on Multimedia, Jul. 2024.

[14] Y. Li, H. Cao, and T. Liu, "Enhancing Graph-Based Recommender Systems with Conversational AI," arXiv preprint 2407.13117, Jul. 2024.
[15] F. Brochier and A. Dimitriadis, "Fairness in Recommender Systems: Research Landscape and Future Directions," User Model. User-Adapted Interact., vol. 34, pp. 59–108, Apr. 2023, doi:10.1007/s11257-023-09364-z.

Received: August 04 2025; Accepted: 26 September 2025; Published: 27 October 2025