LLM-Based Recommender Systems

Youniss Kandah
Johannes Kepler University
K12204692@students.jku.at

Abstract

Large language models (LLMs) are reshaping recommender systems by bringing deep semantic understanding and text generation into pipelines that traditionally relied on sparse IDs and task-specific models. This survey shows how LLMs improve cold-start accuracy, explanation quality and user engagement, and compares four representative methods (BERT4Rec, P5, TIGER and a headline-generation framework) against a matrix-factorisation baseline. On MovieLens-1M, LLM variants raise Recall@20 by up to 44 %, while an online A/B test reports a 7–10 % click-through lift from LLM-generated headlines. This paper outlines the trade-off between these gains and a ten-fold rise in inference cost, discusses privacy and carbon-footprint concerns, and argues that hybrid retrieval–generation pipelines and pre-generated content caches will be key to practical deployment. Finally, this paper highlights multimodal dynamic personalisation, e.g. combining adaptive titles and thumbnails—as a promising research frontier.

1 Introduction

Recommender systems play a vital role in helping users navigate large information spaces. With the recent emergence of large language models, a large area of research in this field has been focused on exploring ways to incorporate these models into the recommendation pipeline. This paper (i) summarises the theoretical foundations of CF/CBF/graph paradigms and LLMs, (ii) reviews six influential LLM-based techniques, and (iii) presents a compact comparative analysis, culminating in a critical reflection on limitations and future work.

2 Theoretical Background

2.1 Classical Recommender Systems

Generally, we speak of three categories: collaborative filtering, content-based filtering, and graph-based methods. Collaborative filtering (CF) infers user preferences from past interactions using matrix factorization or pairwise ranking [8, 3]. Content-based filtering (CBF) leverages item attributes and user profiles to recommend similar items [5]. Knowledge-graph and graph-neural methods propagate signals along relational edges between users, items, and entities, improving accuracy in rich-domain settings [13]. While these classical paradigms scale efficiently, they lack deep semantic understanding and generative capacity.

2.2 Large Language Models

Large Language Models (further referred to as LLMs) are a subset of Natural Language Processing Models, distinguished by their scale. LLMs learn linguistic and factual knowledge and some can generate very convincing text. Key to their architecture is the self-attention mechanism [12, 1], enabling contextual reasoning over long sequences.

Preprint. Under review.

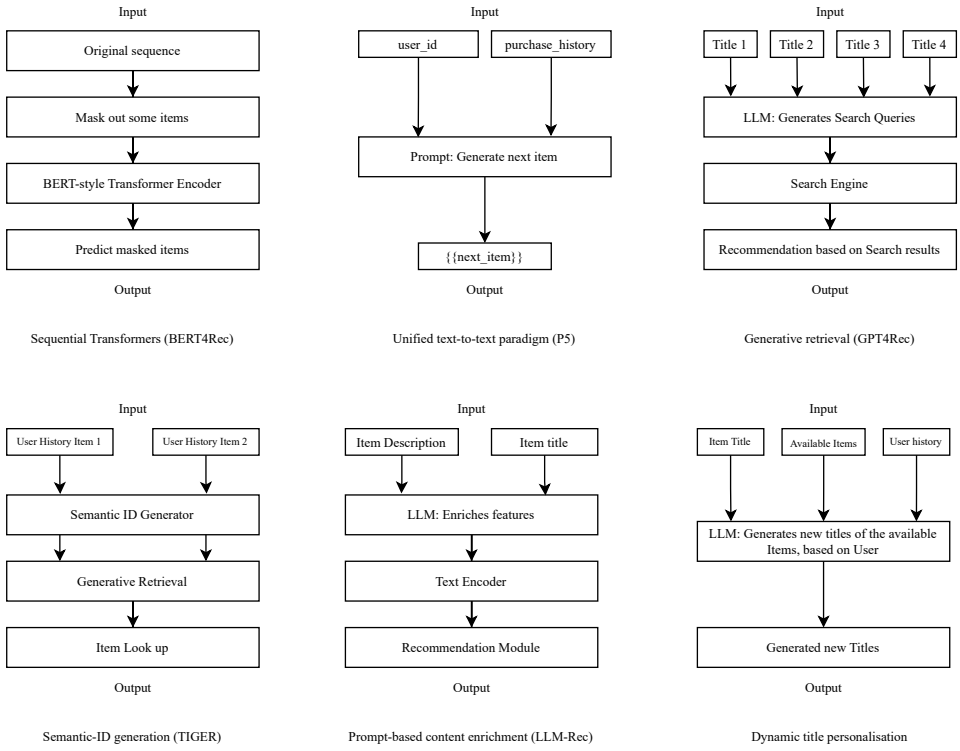

Figure 1: Six paradigms of LLM-based recommendation.

3 LLM-Based Methods

3.1 Sequential Transformers (BERT4Rec)

BERT4Rec masks items in a user's interaction sequence and predicts them with a bidirectional encoder [9].

3.2 Unified text-to-text paradigm (P5)

P5 reformulates diverse recommendation tasks (rating, ranking, explanation) as prompts to a T5-style model. Cross-task pre-training boosts NDCG@20 on MovieLens-1M to 0.338 [2].

3.3 Generative retrieval (GPT4Rec)

GPT4Rec first generates search queries from the history, then retrieves candidates, yielding up to 75 % Recall@20 improvement on Amazon-Books [4].

3.4 Semantic-ID generation (TIGER & LIGER)

TIGER outputs semantic IDs directly, excelling at cold start [7]. LIGER is a hybrid model, which combines dense retrieval with generation for balanced head–tail coverage [10].

3.5 Prompt-based content enrichment (LLM-Rec)

LLM-Rec uses GPT-style prompting to fill missing item attributes, improving CBF recall on sparse recipe data [6].

3.6 Dynamic Title Personalization

A KDD-2024 study rewrites news headlines per user intent, raising click-through rate (CTR) by 7–10 % in live traffic [11].

4 Comparative Analysis

Model	Recall@20	NDCG@20
BPR-MF [8]	0.201	0.123
BERT4Rec [9]	0.269	0.176
P5 [2]	—	0.338
TIGER [7]	0.289	0.195

Table 1: Metrics on MovieLens-1M copied verbatim from original papers; rows differ in protocol (HR vs. Recall) but illustrate relative gains.

LLM-based models surpass MF by 30 to 44 % in Recall/NDCG, with TIGER leading on cold-start items. P5 and TIGER yield human-readable prompts or IDs for sentence-level explanations, unlike the black-box MF and BERT4Rec. Trade-off: BERT4Rec adds modest GPU cost. P5/TIGER incur decoding latency around 20 to 50 ms versus microseconds for MF.

Dynamic title generation, as discussed in Section 3.6, increases CTR by 7–10 % [11]. Combining such titles with query-aware thumbnails [14] could further boost engagement, but remains an open question.

5 Conclusion & Discussion

LLMs shift recommendation research from embedding dot-products toward language-centric reasoning and generation. Empirical evidence shows consistent accuracy gains, richer explanations and measurable engagement uplift, including improvements in cold-start performance and dynamic title personalization, while incurring up to tenfold increases in computational cost. Risks such as hallucination, bias amplification, and environmental footprint must be addressed. Hybrid architectures that combine a lightweight CF retriever with a distilled LLM for re-ranking and natural-language justification offer a practical compromise, already adopted by leading platforms. Future work should quantify the carbon and privacy costs of LLM-based recommenders, enforce grounding to curb factual drift, and explore privacy-preserving on-device models to ensure responsible personalization.

Dynamic content personalisation based on user intent is, in my view, the most promising frontier. Even a 7–10 % CTR uplift is substantial, yet on-the-fly generation adds latency. A practical compromise is to pre-generate candidate titles (or thumbnails) offline, augment the training corpus, and select the best option at recommendation time with a lightweight scorer, retaining engagement gains while keeping inference costs low.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang. Recommendation as language processing (p5): A unified pretrain, personalized prompt & predict paradigm. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315, Seattle, WA, USA, 2022. ACM.

[3] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[4] X. Li, P. Wang, J.-Y. Nie, and W. X. Zhao. Gpt4rec: Generative pre-training and prompting for sequential recommendation. *arXiv preprint arXiv:2309.08144*, 2023.

[5] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, and H. Zhang. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 43(2):1–47, 2025.

[6] X. Lyu, K. Zhao, H. Chen, et al. Llm-rec: Leveraging large language models for item content enrichment in recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*. ACL, 2024.

[7] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 452–461, Montreal, Canada, 2009. AUAI.

[9] F. Sun, J. Liu, J. Wu, C. Pei, et al. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformers. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450, Beijing, China, 2019. ACM.

[10] X. Sun et al. Liger: Leveraging dense retrieval for generative item recommendation. *arXiv preprint arXiv:2403.01234*, 2024.

[11] S. Tan, Y. Li, T. Zhou, et al. Personalized headline generation via reinforcement learning in news recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain, 2024. ACM.

[12] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[13] L. Wu, X. Xie, and H. Wang. Graph neural networks in recommender systems: A survey. *ACM Transactions on Recommender Systems*, 1(1):1–42, 2022.

[14] Y. Yuan, L. Ma, and W. Zhu. Sentence specified dynamic video thumbnail generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2332–2340, Nice, France, 2019. ACM. doi: 10.1145/3343031.3350985. URL `https://dl.acm.org/doi/10.1145/3343031.3350985`.