End-to-End Personalization: Unifying Recommender Systems with Large Language Models

Danial Ebrat
Ebrat@uwindsor.ca
University of Windsor
Windsor, Ontario, Canada

Tina Aminian
aminiant@uwindsor.ca
University of Windsor
Windsor, Ontario, Canada

Sepideh Ahmadian
ahmadia3@uwindsor.ca
University of Windsor
Windsor, Ontario, Canada

Luis Rueda
lrueda@uwindsor.ca
University of Windsor
Windsor, Ontario, Canada

ABSTRACT

Recommender systems are essential for guiding users through the vast and diverse landscape of digital content by delivering personalized and relevant suggestions. However, improving both personalization and interpretability remains a challenge, particularly in scenarios involving limited user feedback or heterogeneous item attributes. In this article, we propose a novel hybrid recommendation framework that combines Graph Attention Networks (GATs) with Large Language Models (LLMs) to address these limitations. LLMs are first used to enrich user and item representations by generating semantically meaningful profiles based on metadata such as titles, genres, and overviews. These enriched embeddings serve as initial node features in a user–movie bipartite graph, which is processed using a GAT-based collaborative filtering model. To enhance ranking accuracy, we introduce a hybrid loss function that combines Bayesian Personalized Ranking (BPR), cosine similarity, and robust negative sampling. Post-processing involves reranking the GAT-generated recommendations using the LLM, which also generates natural-language justifications to improve transparency. We evaluate our model on benchmark datasets, including MovieLens 100k and 1M, where it consistently outperforms strong baselines. Ablation studies confirm that LLM-based embeddings and the cosine similarity term significantly contribute to performance gains. This work demonstrates the potential of integrating LLMs to improve both the accuracy and interpretability of recommender systems.

KEYWORDS

Recommender Systems, Large Language Models, Graph Neural Networks, Vector Embeddings

ACM Reference Format:
Danial Ebrat, Tina Aminian, Sepideh Ahmadian, and Luis Rueda. 2025. End-to-End Personalization: Unifying Recommender Systems with Large Language Models . In *Second Workshop on Generative AI for Recommender Systems and Personalization at the ACM Conference on Knowledge Discovery and Data Mining (GenAIRecP@KDD 2025), August 4, 2025, Toronto, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The growing demand for personalized, context-aware, and interpretable recommendation systems has led to a surge of interest in integrating LLMs into the recommendation pipeline. Conventional recommender systems struggle with limitations such as data sparsity, shallow contextual understanding, and the need for extensive manual feature engineering. LLMs represent a paradigm shift, offering richer feature representations, adaptive reasoning, and the ability to enhance transparency throughout the recommendation process.

Recent research has explored various avenues through which LLMs contribute to recommendation systems, including data augmentation, feature generation, re-ranking, and explanation generation. For instance, a comprehensive taxonomy of LLM-augmented recommenders, highlighting their role across different stages of the pipeline, provided by [36]. In addition, the authors of [19, 29], emphasize the integration of LLMs in machine learning workflows, underscoring their utility in preprocessing, knowledge alignment, and ranking optimization.

Building upon this body of work, we propose a novel recommendation framework that strategically incorporates LLMs into both preprocessing and post-processing stages. In the preprocessing phase, semantically enriched user and item profiles are generated through multi-turn interactions and schema-aligned transformations. Post-processing involves an LLM-driven re-ranking module that re-evaluates top-N recommendations for semantic alignment and diversity. At the core of the pipeline, a GAT is employed to capture user-item interactions using LLM-derived embeddings, thereby enhancing cold-start robustness and contextual sensitivity.

This section reviews relevant literature across three key dimensions: (1) the use of LLMs for feature engineering and representation learning, (2) LLMs for ranking refinement and interpretability, and (3) the use of Graph Attention Networks in recommender systems.

1.1 LLMs for Feature Engineering and Representation Learning

Feature engineering is central to the performance of recommender systems, traditionally requiring substantial domain knowledge and manual effort. The advent of LLMs has shifted this paradigm by enabling automated, context-aware transformation of raw textual data into semantically enriched representations. These capabilities are particularly valuable in scenarios with sparse or noisy inputs, where traditional techniques struggle to capture nuanced relationships.

Recent research has explored a variety of LLM-based approaches for representation enhancement. KAR proposes auxiliary feature

generation for user-item modeling, though its reliance on static feature construction reduces adaptability [30]. SAGCN employs chain-based prompting to reveal semantic relationships, although its performance is highly sensitive to prompt design and consistency [20]. CUP addresses input length limitations through compact summarization, yet often sacrifices fine-grained user preferences [24]. In domain-specific applications, LLaMA-E and EcomGPT apply LLMs for attribute extraction, achieving promising results within narrow verticals, yet their generalizability remains limited [16, 23]. Additionally, LLMs have been applied in preprocessing pipelines for knowledge graph completion [2, 4, 27, 28], text refinement [5, 22, 38], and synthetic data generation [17]. While these methods mitigate data sparsity and cold-start issues, they risk introducing semantic noise or bias if outputs are not rigorously validated.

In this paper, we introduce End-to-End Personalization, a unified pipeline that integrates GATs and LLMs to enhance recommendation quality. Our methodology employs LLMs in a principled, schema-aligned manner to ensure semantic coherence, consistency, and interpretability throughout the recommendation process. This structured preprocessing pipeline enhances embedding initialization for graph-based models, offering improvements in both personalization depth and generalizability. Building on our prior framework [6], which demonstrated the effectiveness of LLMs in generating dynamic, explainable feedback for evolving user states, we further refine their role in pretraining representations for robust recommendation.

1.2 LLMs for Ranking and Interpretability

Ranking plays a critical role in shaping user experience, influencing not only which items are recommended but also the order in which they are prioritized. Conventional models, including matrix factorization [12], sequence-based predictors [3, 21], and GNNs [26], offer strong performance while often lacking transparency and interpretability—qualities increasingly demanded in sensitive domains such as health, finance, and education.

Existing LLM-based ranking systems fall into two primary categories. Scoring models, such as E4SRec and ClickPrompt use modified architectures to output relevance scores, offering efficiency while limiting the expressive capabilities of the underlying model [15, 18]. Two-tower frameworks like CoWPiRec are scalable, though model shallow interactions, relying on fixed similarity metrics [32]. Classification approaches like TALLRec reformulate ranking as a prediction task, yet often struggle with score calibration in multi-item settings [1]. Generative models, such as LANCER and LlamaRec, offer more flexibility, yet are prone to hallucination and are constrained by retrieval quality [13, 34].

To address these limitations, our approach combines graph-based representation learning with a lightweight LLM reranker. By decoupling ranking from generation and leveraging semantically structured profiles, we achieve higher interpretability, lower computational cost, and stronger alignment with user intent.

1.3 Graph Attention Networks in Recommendation Systems

Graph-based collaborative filtering methods have become foundational in modern recommender systems due to their ability to model high-order interactions in user–item bipartite graphs. Traditional techniques like matrix factorization (e.g., SVD, ALS) struggle to incorporate contextual signals and often underperform in cold-start scenarios [8, 14]. Early propagation-based models, including ItemRank and BiRank, improved upon this by diffusing preferences across the graph structure while lacked trainable parameters, reducing their expressiveness [9, 11].

The introduction of Graph Neural Networks (GNNs) transformed this landscape by enabling message-passing architectures to capture both local and global interaction patterns. GC-MC, PinSage, and SpectralCF demonstrated how incorporating node features and graph topology can enhance recommendation accuracy [31, 33, 37]. NGCF introduced explicit multi-hop connectivity through stacked convolutions, though its complexity raised concerns about overfitting [12]. LightGCN addressed this by simplifying the architecture, removing activation functions, and emphasizing pure neighborhood aggregation [10]. GATs further extend this paradigm by assigning adaptive weights to neighbors during message passing, allowing for more selective and context-sensitive representation learning.

Attention-based recommendation models like IGAT and TKGAT improved flexibility at the cost of scalability and interoperability [7, 35]. In this work, we adopt a lightweight GAT architecture that integrates LLM-derived semantic profiles as node features, combining the relational strength of GNNs with the contextual richness of LLMs. This fusion improves performance in sparse regimes and supports personalization through adaptive, meaningful attention. performance in sparse regimes and supports personalization through adaptive, meaningful attention.

1.4 Contributions

This paper introduces a unified architecture that integrates LLMs into both the preprocessing, which also directly affects model training, and post-processing stages of the recommendation pipeline.

- **Semantic Profiling via LLMs**: We propose a structured preprocessing pipeline that transforms raw movie and user data into semantically enriched profiles using multi-turn interactions and schema alignment.
- **Iterative User Preference Modeling**: A novel multi-turn LLM dialogue system is introduced to incrementally capture complex user preferences, overcoming token constraints while maintaining continuity and coherence.
- **Structured Embedding Initialization for Graph Models**: The enriched profiles are embedded and used to initialize a GAT, including a combined loss function that integrates BPR with a cosine similarity term and robust negative sampling to optimize ranking performance and the alignment of semantically similar embeddings, enhancing representation quality in sparse data conditions and cold-start scenarios.
- **Post-Hoc Reranking and Explainability**: An LLM-based reranker evaluates top-N recommendations, offering fine-grained diversity and semantic alignment alongside human-interpretable explanations.

Collectively, these components yield a transparent, generalizable, and high-performing recommendation framework that addresses longstanding challenges in feature sparsity, cold-start handling,

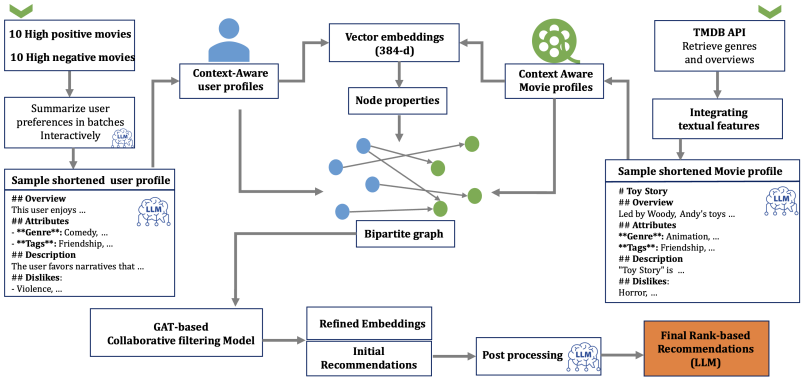

Figure 1: Schematic view of the pipeline utilized by the proposed method.

and interpretability. By bridging LLM-driven reasoning with graph-based relational learning, this work sets a new benchmark for scalable and explainable recommender systems.

2 METHODOLOGY

This section outlines our proposed methodology, systematically structured into three main distinct yet interconnected phases: Preprocessing and LLM-Based Profile Generation, Collaborative Filtering Model and training procedures, and LLM-Based Post-processing, and Explainability. Each phase emphasizes clear integration and justification of LLMs within our recommender system pipeline. Figure 1, depicts a schematic view of the methodology pipeline and how we integrate these steps.

2.1 Preprocessing and Profile Generation

We utilized the widely recognized MovieLens 100K and 1M datasets, enhancing them through additional metadata acquired from The Movie Database (TMDB) API. Specifically, movie titles, genres, and textual overviews were integrated to form coherent, semantically enriched descriptions. This initial preprocessing stage provides the foundational textual resources necessary for subsequent advanced semantic analysis. Furthermore, standardized normalization and cleaning techniques were applied to ensure data consistency and quality.

2.1.1 Item Profile Generation. Movie profiles were enriched through an advanced LLM-driven process. The textual metadata (titles, genres, overviews) served as inputs to an LLM agent tasked with extracting structured, nuanced descriptors. These descriptors included specific narrative elements (e.g., "time travel," "heist scenarios") and character-driven attributes (e.g., "anti-hero," "strong female protagonist"), significantly enhancing profile specificity and semantic depth.

2.1.2 User Profile Generation. To capture complex and multifaceted user preferences accurately, an iterative, conversational refinement process was developed leveraging multi-turn LLM interactions. We begin by selecting the user's 10 highest-rated (4-5) and 10 lowest-rated (1-2) movies, processed incrementally in batches of 5. Each batch triggered sequential LLM interactions, with prompt engineering explicitly incorporating previously refined profile elements to ensure continuity and semantic consistency across conversational turns. This iterative methodology effectively addressed context-length constraints, resulting in highly nuanced user profiles reflective of sophisticated preference patterns.

2.1.3 Structured Schema Alignment. Both movie and user profiles followed an aligned, structured schema emphasizing semantic coherence. This schema included consistent and clearly defined, enabling accurate comparison, vector embedding representation, and robust matching accuracy between user interests and movie characteristics.

The United profile structure using Markdown language contains:

- **Overview**: Summary capturing the core narrative or thematic essence.
- **Attributes**: Concise genres and descriptive tags relevant to content.
- **Description**: Expanded narrative and character-focused insights.
- **Dislikes**: Explicitly identified non-relevant attributes.

2.2 Model Embedding Initialization

Semantic vectors for the textual profiles were first generated with the pretrained SentenceTransformer allMiniLML6v2 [25], producing 384dimensional embedding representations that seeded a bipartite useritem graph. We then refine these embeddings with a threelayer Graph Attention Network tailored for collaborative filtering. Each layer featured 64 hidden units, four attention heads, layer normalization, LeakyReLU activations, residual skips, and dropout, enabling bidirectional message passing so that user and item nodes updated one another simultaneously. Edges captured explicit feedback—ratings ≥ 4 as positive and ≤ 2 as negative—while neutral scores were omitted to limit noise. Training optimized a blend of BPR loss and a cosinebased alignment regularizer that pulls embeddings of positively rated useritem pairs closer together. Using AdamW with adaptive learning rates, weight decay, and early stopping secured stable convergence, and final relevance scores were computed via the dot product of the refined user and item vectors.

2.3 Post-processing and Explainability

Following the GAT-based collaborative filtering stage, we introduce a series of LLM-driven post-processing methods designed to refine recommendation rankings and generate transparent, user-facing rationales. Each method begins with the initial candidate pool produced by graph-based signals, subsequently refined through semantic reasoning by an LLM. The refined rankings are then fused with the original GAT scores using an 80:20 weighted hybrid scheme—assigning 80% weight to the LLM output—ensuring enhanced semantic alignment without disregarding interaction patterns.

We implemented and evaluated several post-processing variants to rerank the top 20 recommendations for each user:

- **Prompt-level Re-ranking**: A context-rich prompt containing the user's genre preferences, thematic leanings, recent interactions, and detailed metadata of candidate films is provided to the LLM. The model performs a single-pass semantic alignment and re-ranking of all 20 candidates simultaneously.
- **Pairwise BST-based Re-ranking**: Leveraging a balanced Binary Search Tree (BST), the LLM conducts pairwise comparisons, evaluating two items at a time against user preferences to determine relative suitability. This method significantly reduces redundant comparisons by preserving logical order.
- **Batch-of-5 Re-ranking with Overlaps** items from bigger pool are partitioned into overlapping batches of five to maintain prompt conciseness and minimize hallucinations. Each batch is independently re-ranked by the LLM, and global rankings are subsequently merged based on overlapping results to ensure consistency and accuracy.
- **Relevancy Scoring Across Batches**: Items are organized into batches of five from bigger pool, each batch evaluated separately by the LLM, which assigns relevancy scores (0-100) to individual movies. To mitigate scoring biases, each film appears in three different batches. Final scores are averaged across the batches.

For each post-processing variant, the same prompt context is reused immediately after re-ranking to generate succinct, natural-language explanations articulating the rationale behind each recommendation. Explanations explicitly reference shared narrative themes, stylistic elements, and cast or crew relevance, enhancing transparency without additional inference overhead.

We utilized both the OpenAI 4o-mini and the o4-mini models and the open-source Gemma-3 4B model for all methodologies. This comparative analysis facilitates understanding trade-offs between proprietary and open-source models, assessing both performance and deployment practicality.

3 EXPERIMENTAL RESULTS

We conducted experimental evaluations using the MovieLens 100K and 1M datasets, employing a 5-fold cross-validation framework with an 80:20 train-test split. Due to dataset sparsity ($\tilde{9}7\%$), evaluations focused strictly on explicit user-item interactions to accurately measure model effectiveness. All code and resources are publicly accessible via our GitHub repository.[1]

Our proposed methodology consistently outperformed established baselines (NGCF, LightGCN, ALS, and a GAT model without LLM enhancements) across Precision, Recall, NDCG, and MAP metrics, as shown in Figure 2. Key performance improvements resulted primarily from two factors: LLM-generated profiles providing semantically rich embeddings and a cosine similarity term that enhanced latent space alignment. Moreover, structured unified textual representations consistently outperformed integrated textual representations, confirming structured schemas' efficacy in improving semantic coherence and embedding initialization.

Due to operational constraints, comprehensive post-processing evaluations using LLM architectures (Gemma-3 4B, OpenAI 4.1-mini, and o4-mini) were conducted only on the MovieLens 100K dataset. Surprisingly, OpenAI's 4.1-mini consistently achieved the best performance, surpassing the reasoning-focused o4-mini, which performed comparably to the Gemma-3 model despite its larger size. This result underscores Gemma-3's efficiency given its smaller parameter count.

Among the LLM-based post-processing strategies, the relevancy scoring method consistently yielded the highest Precision, NDCG, and MAP in the cold start scenario (users with less than ten interactions). Its strong performance can be attributed to averaging scores across multiple evaluations, reducing biases and enhancing stability. The pairwise Binary Search Tree (BST) approach ranked second, excelling notably in Recall and NDCG, demonstrating the reliability of simplified binary judgments from LLMs, especially beneficial in cold-start scenarios.

However, when excluding cold-start scenarios, our semantic re-ranking methods did not improve our baseline GAT method except for precision. This indicates that highly optimized embeddings from our core approach leave limited scope for further improvement through re-ranking when ample interaction data is present. Conversely, as shown in Table 1, semantic re-ranking substantially improved relevance and rankings in cold-start conditions, validating the critical role of semantic insights when user-item interaction data are sparse.

Additionally, we generated natural-language explanations for each recommendation using structured user and item profiles. Manual evaluation of selected recommendations confirmed these explanations as accurate, contextually relevant, and aligned with user preferences and historical interactions. However, the absence of standardized automated metrics highlights a need for further research into evaluation methodologies for explainable recommendations.

Key insights from our experiments include:

- Structured profiles substantially improve semantic embedding quality.
- Relevancy scoring effectively balances precision and consistency.
- Pairwise BST comparisons offer strong recall and semantic reliability.
- Semantic re-ranking is especially beneficial in cold-start conditions, although less impactful with ample historical data.

[1] https://github.com/anonymous-public/End_to_End_Personalization

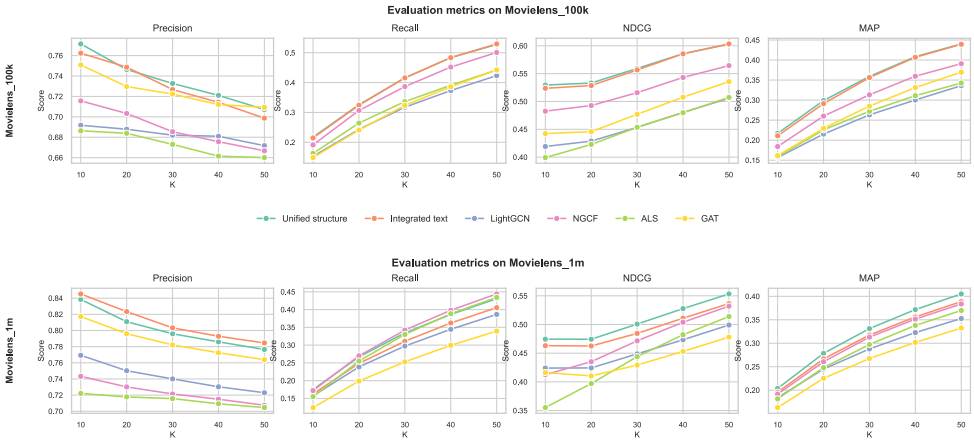

Figure 2: Evaluation Metrics on Movielens 100k and 1M datasets.

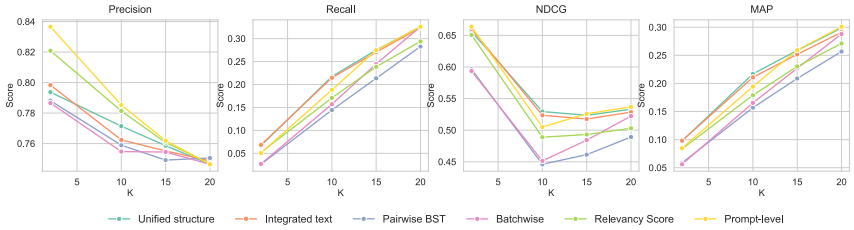

Figure 3: Comparison between different post-processing methods on Movielens 100k.

Table 1: Performance of cold-start users (fewer than 10 interactions) for Movielens100k and $k = 10$.

Method	Precision	Recall	NDCG	MAP
ALS	0.541	0.082	0.197	0.127
NGCF	0.633	0.126	0.253	0.162
LightGCN	0.700	0.121	0.347	0.196
GAT	0.562	0.105	0.271	0.202
Integrated text	0.555	0.241	0.314	**0.319**
Unified structure	0.566	0.183	0.335	0.271
Prompt-level	0.600	0.166	0.319	0.250
Pairwise BST	0.693	**0.244**	0.380	0.234
Batchwise	0.650	0.166	0.289	0.214
Relevancy score	**0.750**	0.237	**0.387**	0.255

- Standardized metrics for evaluating explanation quality remain a critical area for future research.

Future efforts should focus on balancing semantic depth, diversity, computational efficiency, fairness-aware constraints, and novel evaluation methods for explainability to enhance recommendation quality and user satisfaction.

4 CONCLUSION AND FUTURE WORK

Our study demonstrates the efficacy of systematically integrating LLMs into hybrid recommender systems, achieving state-of-the-art performance on MovieLens datasets. By synergizing LLM-generated semantic profiles with graph-based collaborative filtering, we address critical limitations in traditional systems, particularly in capturing nuanced user preferences and item characteristics. The iterative LLM-driven profile refinement, coupled with a GAT architecture optimized via BPR loss and cosine alignment, enables robust representation learning even under extreme sparsity. Post-processing with LLM-based re-ranking further enhances recommendation quality, outperforming baselines across precision, recall, and ranking metrics while maintaining explainability. Notably, our method excels in cold-start scenarios, proving its adaptability to sparse interaction data. The structured schema alignment and hybrid reranking strategies (e.g., BST-based comparisons) ensure semantic coherence while mitigating LLM hallucinations, validating the practicality of combining neural graph models with LLM reasoning.

The proposed method can be extended in various ways. One of these is mitigating biases introduced by metadata richness and enhancing diversity without sacrificing accuracy on more datasets. Techniques such as fairness-aware regularization, dynamic user

preference adaptation, and lightweight LLM fine-tuning for domain-specific tasks could further optimize efficiency and scalability. Additionally, developing end-to-end training pipelines that jointly optimize graph embeddings and LLM reranking—rather than treating them as separate stages—could reduce computational overhead and improve alignment. Finally, user studies are needed to validate the perceived quality of explanations and ensure ethical transparency in LLM-driven recommendations.

REFERENCES

[1] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 1007–1014. doi:10.1145/3604915.3608857

[2] Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason H. D. Cho, Kaushiki Nag, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs. *CoRR* abs/2305.09858 (2023). doi:10.48550/ARXIV.2305.09858 arXiv:2305.09858

[3] Mingyue Cheng, Qi Liu, Wenyu Zhang, Zhiding Liu, Hongke Zhao, and Enhong Chen. 2024. A general tail item representation enhancement framework for sequential recommendation. *Frontiers Comput. Sci.* 18, 6 (2024), 186333. doi:10.1007/S11704-023-3112-Y

[4] Zhixuan Chu, Yan Wang, Qing Cui, Longfei Li, Wenqing Chen, Sheng Li, Zhan Qin, and Kui Ren. 2024. LLM-Guided Multi-View Hypergraph Learning for Human-Centric Explainable Recommendation. *CoRR* abs/2401.08217 (2024). doi:10.48550/ARXIV.2401.08217 arXiv:2401.08217

[5] Yingpeng Du, Di Luo, Rui Yan, Xiaopei Wang, Hongzhi Liu, Hengshu Zhu, Yang Song, and Jie Zhang. 2024. Enhancing Job Recommendation through LLM-Based Generative Adversarial Networks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 8363–8371. doi:10.1609/AAAI.V38I8.28678

[6] Danial Ebrat and Luis Rueda. 2024. Lusifer: LLM-based User SImulated Feedback Environment for online Recommender systems. *CoRR* abs/2405.13362 (2024). doi:10.48550/ARXIV.2405.13362 arXiv:2405.13362

[7] Ehsan Elahi, Sajid Anwar, Mousa Al-Kfairy, Joel J. P. C. Rodrigues, Alladoumbaye Ngueilbaye, Zahid Halim, and Muhammad Waqas. 2025. Graph attention-based neural collaborative filtering for item-specific recommendation system using knowledge graph. *Expert Syst. Appl.* 266 (2025), 126133. doi:10.1016/J.ESWA.2024.126133

[8] Gene Howard Golub. 1968. Least Squares, Singular Values and Matrix Approximations. *Aplikace matematiky* 13, 1 (1968), 44–51. doi:10.21136/AM.1968.103138 https://doi.org/10.21136/AM.1968.103138.

[9] Marco Gori and Augusto Pucci. 2007. ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, Manuela M. Veloso (Ed.). 2766–2771. http://ijcai.org/Proceedings/07/Papers/444.pdf

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 639–648. doi:10.1145/3397271.3401063

[11] Xiangnan He, Ming Gao, Min-Yen Kan, and Dingxian Wang. 2017. BiRank: Towards Ranking on Bipartite Graphs. *IEEE Trans. Knowl. Data Eng.* 29, 1 (2017), 57–71. doi:10.1109/TKDE.2016.2611584

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 173–182. doi:10.1145/3038912.3052569

[13] Junzhe Jiang, Shang Qu, Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, Kai Zhang, Rui Li, Jiatong Li, and Min Gao. 2024. Reformulating Sequential Recommendation: Learning Dynamic User Interest with Content-enriched Language Modeling. In *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 14852)*, Makoto Onizuka, Jae-Gil Lee, Yongxin Tong, Chuan Xiao, Yoshiharu Ishikawa, Sihem Amer-Yahia, H. V. Jagadish, and Kejing Lu (Eds.). Springer, 353–362. doi:10.1007/978-981-97-5555-4_25

[14] Kiryung Lee and Dominik Stöger. 2023. Randomly Initialized Alternating Least Squares: Fast Convergence for Matrix Sensing. *SIAM J. Math. Data Sci.* 5, 3 (2023), 774–799. doi:10.1137/22M1506456

[15] Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. E4SRec: An Elegant Effective Efficient Extensible Solution of Large Language Models for Sequential Recommendation. *CoRR* abs/2312.02443 (2023). doi:10.48550/ARXIV.2312.02443 arXiv:2312.02443

[16] Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Haitao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. EcomGPT: Instruction-Tuning Large Language Models with Chain-of-Task Tasks for E-commerce. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 18582–18590. doi:10.1609/AAAI.V38I17.29820

[17] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 10443–10461. doi:10.18653/V1/2023.EMNLP-MAIN.647

[18] Jianghao Lin, Bo Chen, Hangyu Wang, Yunjia Xi, Yanru Qu, Xinyi Dai, Kangning Zhang, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. ClickPrompt: CTR Models are Strong Prompt Generators for Adapting Language Models to CTR Prediction. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (Eds.). ACM, 3319–3330. doi:10.1145/3589334.3645396

[19] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2023. How Can Recommender Systems Benefit from Large Language Models: A Survey. *CoRR* abs/2306.05817 (2023). doi:10.48550/ARXIV.2306.05817 arXiv:2306.05817

[20] Fan Liu, Yaqi Liu, Zhiyong Cheng, Liqiang Nie, and Mohan S. Kankanhalli. 2023. Understanding Before Recommendation: Semantic Aspect-Aware Review Exploitation via Large Language Models. *CoRR* abs/2312.16275 (2023). doi:10.48550/ARXIV.2312.16275 arXiv:2312.16275

[21] Weiwen Liu, Wei Guo, Yong Liu, Ruiming Tang, and Hao Wang. 2023. User Behavior Modeling with Deep Learning for Recommendation: Recent Advances. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 1286–1287. doi:10.1145/3604915.3609496

[22] Zhenghao Liu, Zulong Chen, Moufeng Zhang, Shaoyang Duan, Hong Wen, Liangyue Li, Nan Li, Yu Gu, and Ge Yu. 2024. Modeling User Viewing Flow using Large Language Models for Article Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 83–92. doi:10.1145/3589335.3648305

[23] Kaize Shi, Xueyao Sun, Dingxian Wang, Yinlin Fu, Guandong Xu, and Qing Li. 2023. LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following. *CoRR* abs/2308.04913 (2023). doi:10.48550/ARXIV.2308.04913 arXiv:2308.04913

[24] Ghazaleh Haratinezhad Torbati, Anna Tigunova, Andrew Yates, and Gerhard Weikum. 2023. Recommendations by Concise User Profiles from Review Text. *CoRR* abs/2311.01314 (2023). doi:10.48550/ARXIV.2311.01314 arXiv:2311.01314

[25] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[26] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 165–174. doi:10.1145/3331184.3331267

[27] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y. Zhang, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2023. Enhancing Recommender Systems with Large Language Model Reasoning Graphs. *CoRR* abs/2308.10835 (2023). doi:10.48550/ARXIV.2308.10835 arXiv:2308.10835

[28] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, Luz Angelica Caudillo-Mata, Silvio Lattanzi, Andrés Muñoz Medina, Leman Akoglu, Aristides Gionis, and Sergei Vassilvitskii (Eds.). ACM, 806–815. doi:10.1145/3616855.3635853

[29] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *World Wide Web (WWW)* 27, 5 (2024), 60. doi:10.1007/S11280-024-01291-2

[30] Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London (Eds.). ACM, 12–22. doi:10.1145/3640457.3688104

[31] Qiang Yang, Shuxin Zhang, Suyu Dong, Long Xu, Weihe Dong, Xiaokun Li, Pengzhong Sun, Feng Jiang, Xianyu Zhang, and Gongning Luo. 2023. Graph Convolutional Network with Neural Inductive Matrix Completion for Predicting Disease-Related LncRNA Genes. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Istanbul, Turkiye, December 5-8, 2023*, Xingpeng Jiang, Haiying Wang, Reda Alhajj, Xiaohua Hu, Felix Engel, Mufti Mahmud, Nadia Pisanti, Xuefeng Cui, and Hong Song (Eds.). IEEE, 3595–3601. doi:10.1109/BIBM58861.2023.10386047

[32] Shenghao Yang, Chenyang Wang, Yankai Liu, Kangping Xu, Weizhi Ma, Yiqun Liu, Min Zhang, Haitao Zeng, Junlan Feng, and Chao Deng. 2023. Collaborative Word-based Pre-trained Item Representation for Transferable Recommendation. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, Guihai Chen, Latifur Khan, Xiaofeng Gao, Meikang Qiu, Witold Pedrycz, and Xindong Wu (Eds.). IEEE, 728–737. doi:10.1109/ICDM58522.2023.00082

[33] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 974–983. doi:10.1145/3219819.3219890

[34] Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. 2023. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. *CoRR* abs/2311.02089 (2023). doi:10.48550/ARXIV.2311.02089 arXiv:2311.02089

[35] Shaowei Zhang, Zhao Li, Xin Wang, Zirui Chen, and Wenbin Guo. 2023. TKGAT: Temporal Knowledge Graph Representation Learning Using Attention Network. In *Advanced Data Mining and Applications - 19th International Conference, ADMA 2023, Shenyang, China, August 21-23, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14177)*, Xiaochun Yang, Heru Suhartanto, Guoren Wang, Bin Wang, Jing Jiang, Bing Li, Huaijie Zhu, and Ningning Cui (Eds.). Springer, 46–61. doi:10.1007/978-3-031-46664-9_4

[36] 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 Trans. Knowl. Data Eng.* 36, 11 (2024), 6889–6907. doi:10.1109/TKDE.2024.3392335

[37] Lei Zheng, Chun-Ta Lu, Fei Jiang, Jiawei Zhang, and Philip S. Yu. 2018. Spectral collaborative filtering. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 311–319. doi:10.1145/3240323.3240343

[38] Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. 2023. Generative Job Recommendations with Large Language Model. *CoRR* abs/2307.02157 (2023). doi:10.48550/ARXIV.2307.02157 arXiv:2307.02157