Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures
Final Report for JHPCN Joint Research of FY 2024

jh241004

Large Language Models for Recommender Systems

Toyotaro Suzumura（The University of Tokyo）

Abstract
In this research, we explored the application of Large Language Models (LLMs) to advance recommender systems by integrating textual information, user behavioral patterns, and sequential interactions. Specifically targeting various applications in e-commerce, news platforms, and travel recommendations, we proposed recommendation methods and models utilizing LLMs. First, we proposed the Prompting-based Representation Learning (P4R) method, leveraging LLMs to generate comprehensive user-item representations enhanced by graph neural network (GNN) alignment. Second, we developed an LLM-driven Avoidance-Aware Recommender System (AWRS) that addresses user avoidance behaviors and popularity biases in news recommendation through sophisticated textual counterfactual modeling. Third, we enhanced numerical encoding methods within sequential recommendation systems (SRS) by employing LLM-generated embeddings to predict user behaviors more accurately. Finally, we introduced ReviewGNN, a model that combines LLM-based text embeddings of user-generated reviews with GNNs, significantly improving recommendation accuracy. Our methods exhibited performance improvements on various real-world datasets, emphasizing the capability of LLMs in enriching recommender systems. Overall, this research contributes foundational insights into effectively utilizing LLMs for personalized recommendations and a deeper understanding of user preferences.

1. Basic Information

(1) Collaborating JHPCN Centers

mdx

(2) Theme Area

Data science/data usage area

(3) Project Members and Their Roles

	Name	Role
Representative	Toyotaro Suzumura (1)(2)	Model Design, Paper Writing
Deputy Representative	Hiroki Kanezashi (2)	Design, Implementation Experiments, Paper Writing
Researcher	Refaldi Intri Dwi Putra Junyi Chen Igor L.R. Azevedo David Pohl Limin Wang (1)	

(1)Graduate School of Information Science and Technology, The University of Tokyo
(2)Information Technology Center, The University of Tokyo

2. Purpose and Significance of the Research

In recent years, recommender systems have become an integral part of online services, particularly in delivering personalized recommendations that reflect individual user behaviors and preferences. These systems have demonstrated a significant impact across various domains, from e-commerce to content platforms, making their optimization crucial for real-world applications. The emergence of Large Language Models (LLMs) has opened new possibilities in recommendation systems, particularly in leveraging rich textual information such as item descriptions, user reviews, and contextual data that traditional systems often struggled to incorporate effectively.

While existing approaches and models in recommender systems using Graph Neural Networks (GNNs) and Transformers have shown promising outcomes, the full potential of LLMs in recommendation systems remains largely unexplored. This research addresses the gap by investigating how LLMs can enhance recommendation systems beyond current capabilities. Our research focus is in that it explores novel ways to leverage LLMs' advanced language understanding and reasoning capabilities in the context of personalized recommendations.

The academic significance of our research lies in three key innovations: First, we develop methods to predict user interest persistence and item relevance using LLMs, incorporating temporal dynamics and contextual understanding that traditional models often miss. Second, we tackle the challenge of optimizing numerical encoding for LLM-based recommender systems, bridging the gap between language understanding and

computational efficiency. Third, we explore the development of an LLM-based multi-behavior recommendation model that can capture and interpret complex user interactions beyond simple engagement metrics.

3. Significance as JHPCN Joint Research Project

Our work significantly benefits from and demonstrates the value of high-performance computing resources, particularly in developing and evaluating LLM-based recommender systems. The computational demands of our research, which involves training and fine-tuning large-scale models like LLaMA-2-7B and processing extensive user-item interaction data, require substantial GPU computing power. The GPU-equipped mdx environment has been important to conduct comprehensive experiments and develop our proposed frameworks and models efficiently.

The significance of our project is as follows. First, the high-performance computing infrastructure allows us to process and analyze large-scale datasets from real-world applications, particularly in our work with the large-scale open datasets for recommendation applications. Second, the GPU resources enable us to experiment with various model architectures and hyperparameters, leading to the development of our novel Profile Representation Learning and GNN-based Alignment components. Third, the computing environment supports the intensive computational requirements of our prompt-based approach, which involves generating and processing multiple item profiles through LLMs.

Furthermore, our research directly contributes to advancing real-world applications in recommendation systems. By leveraging computing resources, we have demonstrated significant improvements in recommendation accuracies compared to existing methods. These results have important implications for various online services, from e-commerce platforms to content delivery systems, where efficient and accurate personalized recommendations are crucial for user experience and business success.

4. Outline of Research Achievements until FY2023

This project is not a continuous project.

5. Details of FY2024 Research Achievements

5.1. Prompting-Based Representation Learning Method for Recommendation (P4R)

In recent years, graph-based collaborative filtering (CF) methods have demonstrated significant success in recommender systems. However, most of these approaches have overlooked the rich textual information associated with users and items, which can offer valuable semantic cues. With the advent of Large Language Models (LLMs) in the field of natural language processing, it has become increasingly feasible to extract high-quality representations from contextual data. Despite this progress, traditional fine-tuning of LLMs for recommendation tasks remains computationally prohibitive, prompting the need for more efficient and flexible integration strategies.

To tackle this issue, we proposed **P4R**, a Prompting-Based Representation Learning method for recommendation tasks that effectively leverages LLMs' inference capabilities. The P4R framework has two main components: Profile Representation Learning and GNN-based Alignment. Figure 1 illustrates the overall architecture of P4R.

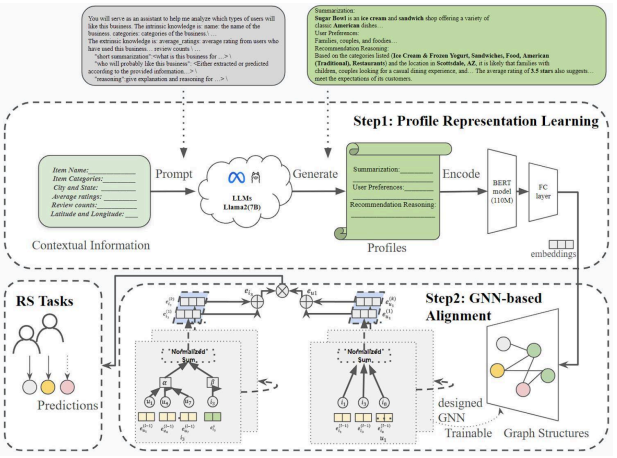

Fig. 1 Architecture of P4R.

Profile Representation Learning generates informative item profiles by prompting a pre-trained LLaMA-2-7B model with item-related textual information such as names, categories, locations, and user reviews. The prompts are designed with a reasoning format inspired by Chain-of-Thought prompting, allowing the LLM to generate profiles that include item summaries, user preference predictions, and recommendation-oriented rationales. These natural language outputs are then encoded using a BERT model to produce semantically rich vector representations.

GNN-based Alignment integrates the embeddings into a GCN-based collaborative filtering framework, where user-item interactions are modeled through message propagation on a bipartite graph. This dual embedding structure enables the system to capture structural user-item relations and textual semantics jointly. We adopt the Bayesian Personalized Ranking (BPR) loss to train the model, optimizing for personalized ranking quality.

We evaluated our method on two widely used public datasets: Yelp2018 and Amazon-VideoGames. Experimental results show that P4R consistently outperforms strong baselines such as NGCF, LightGCN, and SGL across multiple metrics, including Recall, NDCG, MRR, and Hit Rate. Notably, P4R achieves up to 51.5% improvement in NDCG and 68.6% in MRR on the Amazon dataset compared to lower-embedding variants. Ablation studies further confirmed the effectiveness of prompt-based profile generation; removing this component leads to a significant drop in performance.

P4R presents a promising direction for building lightweight yet expressive recommendation systems powered by LLMs. By bridging textual reasoning and collaborative filtering within a unified framework, this approach lays the foundation for more interpretable and context-aware recommender models in real-world applications. This work has been accepted and presented in an international workshop, RecSys-WS@24.

5.2. Understanding and enhancing user interaction with news recommendations

News recommender systems have become essential tools for delivering personalized information to users amid an overwhelming volume of digital content. However, despite advances in click-based modeling and collaborative filtering techniques, complex real-world representations such as news avoidance, the systemic bias toward popular articles, and the influence of major political events on user engagement have been unexplored. Traditional recommendation methods often overlook the nuanced semantics and contextual dynamics that underlie user decisions, whether to engage with or avoid certain content. Recent research has moved toward leveraging richer behavioral signals and external context through models beyond click-based learning to address these challenges. In this research, we worked on three complementary approaches—AWRS, POPK, and EDSMF—tackling a specific limitation in conventional news recommendation. These methods represented a shift toward avoidance-aware, debiased, and context-sensitive recommendation strategies.

(1) AWRS: Traditional recommender systems in the news domain relied heavily on click-through data to model user preferences. However, they often ignored the behavioral signal when users deliberately avoided or disengaged from specific news content. Motivated by rising trends in news avoidance driven by psychological fatigue and distrust in media, this study introduced **AWRS (Avoidance-Aware Recommender System)**, a novel framework that incorporated news avoidance as a fundamental component of the recommendation process. AWRS modeled news content using three principal elements: Exposure, Avoidance, and Relevance. Exposure was defined by the number of impressions a news article received; avoidance was calculated as the proportion of exposures that did not result in a click; and relevance was a learned score that integrated these signals alongside time and content embeddings. A key technique in AWRS is the User Engagement Embedding, which encodes the spatial relationship between exposure and avoidance for each clicked item, capturing cases where a user clicked on a widely avoided article—an indicator of strong personal interest. Figure 2 illustrates the overall architecture of the AWRS.

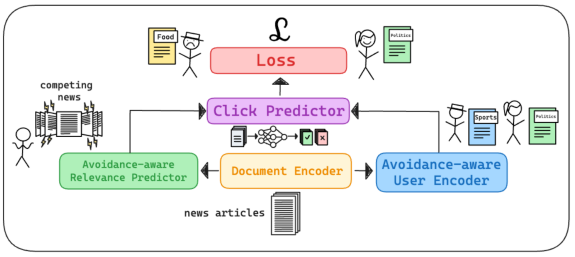

Fig. 2 Figure 2: Overview of the AWRS architecture.

The AWRS architecture combines language model-based news embeddings (e.g., RoBERTa, NB-BERT, DeBERTa), Time2Vec temporal embeddings, and graph-based self-attention to build an avoidance-aware user representation. A final relevance predictor then adjusts the candidate scores using a combination of time-sensitive relevance and user interest matching. We evaluated AWRS on three multilingual datasets — MIND-small (English), Adressa (Norwegian), and Nikkei (Japanese) — demonstrating consistent improvements over state-of-the-art baselines such as NRMS, NAML, and LSTUR. Notably, AWRS achieved AUC of 65.75% and NDCG@5 of 39.07% on the Nikkei dataset. These results confirmed that avoidance behavior, when properly modeled, provided valuable insight into user interests and

significantly enhanced recommendation quality. This work has been accepted and presented at the international conference SDM 2025.

(2) POPK: Popularity bias is a significant challenge in news recommendation, where widely read articles are disproportionately recommended, potentially suppressing diverse or niche content. In this study, we proposed **POPK (Popularity-based Counterfactual Knowledge)**, a temporal-counterfactual method that explicitly incorporates popular articles into the negative sampling process, thereby mitigating their implicit influence.

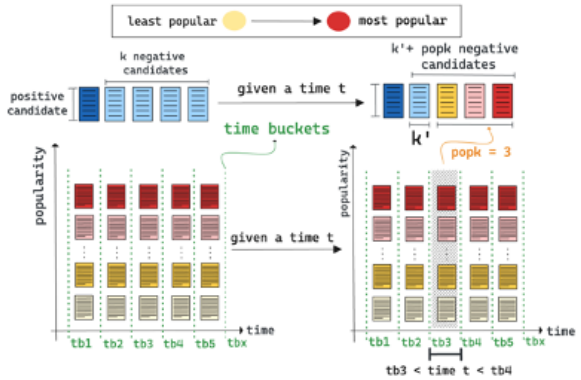

Fig. 3 Process of selecting the *popk* most popular news articles.

The core insight of POPK is that popular articles compete for user attention, even if they are not directly shown in the impression list. To reflect this, POPK modifies the training process by selecting a subset of the most popular news articles at a time, defined via click count, click ratio, or click variance, and inserting them into the set of negative samples. We employed two sampling strategies: acc (accumulated) and ptb (per-time-bucket), which allow flexible temporal evaluation of article popularity (Figure 3). This counterfactual injection tunes the model to make the user's decision to ignore a popular article meaningful. As a result, the recommendation algorithm became less prone to recommending items purely based on popularity and instead focused on actual user preferences.

We evaluated POPK on three benchmark datasets as AWRS—MIND-small, Adressa one-week, and Nikkei one-week—integrated into baseline models such as NRMS, NAML, and LSTUR. We trained the POPK model using a single NVIDIA A100 GPU in mdx GPU instances. The method achieved substantial improvements in both accuracy and diversity metrics. For example, on the Adressa dataset, POPK improved MRR by up to 54.92% and NDCG@5 by 49.95%, enhancing category diversity (Dctg@k). These results demonstrated that modeling the implicit influence of popular items via counterfactual reasoning could significantly improve both personalization and diversity in news recommendations.

(3) EDSMF: Election periods introduce unique volatility into the stock market due to policy uncertainties and shifting public sentiment. We applied recommender methodologies to a political forecasting context by proposing **EDSMF (Election Day Stock Market Forecasting)**—a model that integrated political signals into high-frequency stock price prediction.

EDSMF enhances the StockMixer architecture by incorporating candidate-aware political signals extracted using a multi-agent LLM pipeline. These signals are derived from over 90,000 U.S. news articles and official campaign documents using a structured framework of agents (News Analyst, Policy Analyst, Market Analyst, and Synthesis Analyst). Each candidate was associated with a Candidate Impact Vector that quantified expected economic sector performance under their leadership.

The model also includes a Candidate Context variable to simulate real-world electoral outcomes. During training and validation, candidate contexts were randomly assigned or ensembled; during testing, they were aligned with the actual election result. The final prediction integrated the learned interest score and the context-aware political relevance.

We evaluated EDSMF using minute-level trading data for all S&P 500 stocks from October 30 to November 6, 2024, encompassing the U.S. presidential election. Compared to the baseline StockMixer, the EDSMF model achieved improved performance across key financial metrics, including RIC (0.2306) and Sharpe Ratio (1.8163). Ablation studies confirmed that political signals contributed to predictive performance, validating the value of integrating news-based contextual knowledge into market forecasting.

5.3. Numerical Encoding in SRS

Sequential recommendation systems (SRS) aim to predict the recommended item based on the user's past sequential behavior. The assumption is that the user-item interactions happen under sequential behavior. For example, if user A has a past sequence of purchases such as an iPhone, iPad, and iWatch, then the next recommended item would most likely be a MacBook rather than

a Samsung Galaxy. This is because, according to the buying pattern, we can tell that the user has an affinity for the Apple brand. Furthermore, sequential behavior can also model the dynamics of user-item interaction over time. This is more realistic as the user preferences and item popularity should change over time. Hence, in recent years, many researchers have been developing SRS and have shown that they could outperform other recommendation methods, for example [Neupane,et al., in NeurIPS, 2024].

To perform the sequential recommendation, the attention-based methods have gained traction following their success on language modeling [Vaswani,et al., in NeurIPS, 2017]. First, it was pioneered by [Kang and McAuley, in ICDM, 2018] that uses a unidirectional attention-based model on the SRS task named SASRec. Then, subsequent models were developed, such as BERT4Rec by [Sun,et al., in CIKM, 2019], based on the bi-directional model. Most recently, [Shin,et al., in AAAI, 2024] developed BSARec that incorporates a frequency module into the attention block. Currently, BSARec is the top-ranked model in the public benchmark and is regarded as the current state-of-the-art (SOTA) model.

This study aims to provide an analysis based on the numerical encoding of the item that is used by the attention-based models, including the SOTA model. As a background, to process the past sequential interaction, a way to do it is to convert the items into natural numbers like 1,2,3,.... This encoding will be used as the indices in the item embeddings, where each item is assigned its vector representation. Reflecting on the study in language models for item (token) embeddings, some structure can emerge after a model reaches generalization, for example, the study from [Zhong,et al.,. in NeurIPS, 2023]. To our knowledge, such a study has not been conducted yet for SRS, despite the similarity of the models and training objectives. Therefore, in this study, *for the first time*, we conduct such an analysis.

We first conducted a preliminary study where we analyzed the item embeddings on SASRec under training. We observed an interesting pattern where some adjacent items have their indices also in adjacency, for example, 6-7. We plotted the similarity evolution for some pairs and observed some correlation between their similarity and generalization (Figure 4).

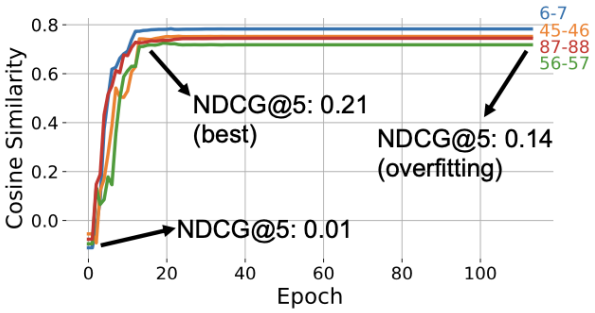

Fig. 4 Correlation of similarities between item pairs.

Furthermore, we also observed that some ordered patterns emerge as well with longer sequences of adjacent items. For example, 9-10-11-12 was found in our t-SNE visualization. These observations lead us to two questions: 1) *Does such an ordered structure occur naturally in attention-based models?*, and 2) *Can we use such a structure in embeddings to improve the models?*. Based on these questions, we study these observed patterns akin to numerical encoding, which we named **Ordered Item Embeddings** (OIE).

As a result, our research on OIE gives contributions to: 1) Characterization of OIE with novel metrics named AAE and SA, 2) Measurement of OIE on SOTA models (SASRec, BERT4Rec, and BSARec) which showed that OIE formation occurs to some degree, 3) Novel training method to induce OIE and can be used as a pre-trained model on transfer learning, and 4) Improved models by using OIE initialization, including BSARec, the current SOTA model, by as much as 81% of NDCG@5, in the LastFM dataset as well as LLM-based model.

Overall, these results suggest that some attention-based models have some inductive bias toward OIE formation, and OIE initialization can help models to reach generalization as a part of the inductive bias source in the SRS task. Inductive bias refers to a certain prior assumption that the model (hypothesis class) uses for machine learning generalization. Future studies will look into a more controllable method to drive OIE formation from the objective function, as the current method is hard to control.

5.4. ReviewGNN with User Review Embedding

In recommender systems, many models rely on “implicit” feedback, such as user clicks or purchase histories, to predict future preferences. However, “explicit” feedback—particularly in written user reviews and five-star ratings—can provide more direct and nuanced insights into

user opinions. While incorporating such textual data holds potential to represent user preferences, it poses two key challenges: sparsity, since only a subset of users write reviews, and alignment, because review texts and behavioral signals are fundamentally different data types that are difficult to represent in a unified latent space without introducing noise.

To address these challenges, we proposed **ReviewGNN**, a hybrid recommendation model that integrates explicit textual feedback with an LLM and implicit user behaviors within a unified GNN framework. ReviewGNN has two main components: Review Embedding and Purchase-Review-GNN, as shown in Figure 5.

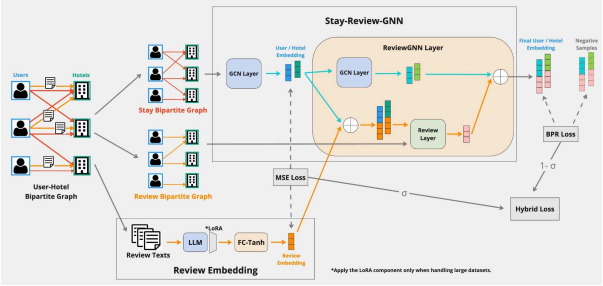

Fig 5. Overall architecture of ReviewGNN.

Review Embedding transforms natural-language reviews into high-quality semantic vectors using a BERT-based LLM. These embeddings are then passed through an FC-Tanh layer, which performs dimensionality reduction and normalization to align them with user and item embeddings in the model's latent space. This step ensures that review representations are compact, comparable, and semantically rich.

Purchase-Review-GNN learns the user and item representations from a bipartite graph where purchase and review edges connect users and items. Standard GCN propagation is applied in the first layer based solely on user-item purchase data. In the second layer, the model integrates Review Embeddings via a ReviewGNN Layer, effectively enriching the user/item representations with explicit textual signals derived from reviews. This two-step design enables the model to represent both behavioral and semantic preferences in a unified manner.

Finally, to coordinate the learning of these heterogeneous signals, we introduced a Hybrid Loss function that combines BPR Loss and Mean Squared Error (MSE) Loss. The BPR Loss emphasizes user behavior-based learning, while the MSE Loss aligns review embeddings with user/item representations. The balance between these two components is controlled by an adaptive weighting parameter, which dynamically adjusts based on the current performance of collaborative filtering. When behavioral signals are strong, the model prioritizes BPR Loss; when weak, it leverages textual review data more heavily. This dynamic loss balancing ensures robust learning across various data conditions.

We evaluated ReviewGNN on a semi-public travel site dataset, as well as public datasets from Amazon Reviews and Yelp. In parallel, we utilized NVIDIA A100 GPUs in mdx GPU instances for experiments with public datasets. Experimental results demonstrated substantial performance gains, especially in sparse settings where explicit feedback is limited. For example, in the "Beauty and Personal Care" category of the Amazon dataset, ReviewGNN outperformed the baseline GCN model by 32.7% in Hit Ratio@10 and 20.6% in NDCG@10. Conversely, in denser datasets, such as Office Products and Yelp, the improvement was less significant, indicating that traditional collaborative filtering remains sufficient when abundant user interaction data is available.

ReviewGNN provides a flexible and effective solution for integrating review information into recommendation systems. Its design allows for robust performance in sparse environments, making it a promising approach for real-world applications where explicit user feedback is valuable but limited. Future work will involve scaling to larger datasets, addressing long-tail user behavior, and expanding to new domains such as e-commerce platforms and travel recommendations.

6. Self-review of Current Progress and Future Prospects

In this research, we aimed to build and enhance recommender systems based on LLMs, addressing the primary objectives outlined in our initial goal. The application detailed a plan to pursue several research themes in parallel, culminating in the design and construction of a unified LLM-based model specialized for recommender systems.

For the first research theme on predicting item interest persistence using LLMs, we developed two key models. The P4R model generates personalized item profiles using LLMs aligned with GNN interaction patterns, showing significant improvements over baselines at RecSys-WS@24. We also created AWRS and POPK for news recommendations - AWRS uses news avoidance signals to achieve 33% better AUC on the Nikkei News Dataset (accepted at SDM@25). In comparison, POPK improves MRR by 7.11%

through temporal-aware sampling, helping model long-term article interest beyond temporary popularity.

For the second theme, "Optimizing numerical encoding for LLM-based recommender systems," which focused on understanding and improving how LLMs encode numerical information for recommendation tasks, we conducted the first analysis and demonstration of the importance of Ordered Item Embeddings (OIE) in sequential recommendation systems. We proposed a novel training method to induce OIE, leading to substantial performance improvements. This method achieved up to an 81% increase in NDCG@5 on the LastFM dataset for existing state-of-the-art models (like BSARec) and LLM-based models, directly investigating how LLMs can better handle numerical information.

Towards the third theme, "Building an LLM-based Multi-Behavior Recommendation model," which aimed to leverage auxiliary user behaviors using LLMs, GNNs, and Transformers to address the long-tail problem, we proposed the ReviewGNN framework. While initially focusing on user review text, this framework integratively handles textual information (a form of user behavior/feedback) with LLMs and interaction data with GNNs. It yielded up to a 36% improvement in HR@10 and a 33% improvement in NDCG@10, particularly in scenarios with sparse review data, and has been submitted to ACM RecSys 25. This work incorporates richer user feedback, a significant step towards full multi-behavior modeling.

These research achievements contributed to the goal, "Design and construction of a unified LLM-based model specialized for recommender systems by integrating GNNs and Transformers." Each developed model—P4R, AWRS, POPK, OIE, and ReviewGNN—explores different facets of combining LLM capabilities with structured data representations (graphs, sequences) and user behaviors, laying the groundwork for a comprehensive, unified model. We developed specific models for each planned objective and demonstrated quantitative performance improvements, clearly indicating the potential of integrating LLMs, GNNs, and other neural network models for recommender systems.

As future work, for the advancement of recommender systems using LLMs and other models, we need to conduct detailed analyses to strengthen our methodologies and apply them to real-world applications in each of our research areas. Regarding OIE, a key challenge remains in controlling OIE formation with the current training methods. Developing techniques to control OIE formation more directly and precisely at the objective function level is crucial. For ReviewGNN, we must investigate how review embeddings representing user review evaluations contribute to accurate item prediction. In particular, analyzing how positive and negative review comments influence user embeddings based on purchase behavior will enable more sophisticated embedding alignment. We also aim to further pursue the complementary strengths of LLMs in deep contextual understanding and GNNs/Transformers in collaborative and sequential pattern learning. By integrating these insights, we will strive to build a unified, high-performance LLM-based foundational model for recommendations adaptable to a broader range of recommendation tasks and diverse datasets.