A Hybrid LLM and Graph-Enhanced Transformer Framework for Cold-Start Session-Based Fashion Recommendation

Junchen Liu *

Posted Date: 28 September 2025

doi: 10.20944/preprints202509.2310.v1

Keywords: conversational recommendation; large language models; cold start; graph attention networks; transformer-XL

 doi:10.20944/preprints202509.2310.v1

Article

A Hybrid LLM and Graph-Enhanced Transformer Framework for Cold-Start Session-Based Fashion Recommendation

Junchen Liu

Boston University, Boston, MA, USA; junchenl@bu.edu

Abstract

Session-based recommendation in areas like fashion is hard because of cold-start problems, changing user needs, and the difficulty of using product labels. Traditional models, such as sequence models or graph methods, often do not work well when user behavior is sparse or when the system needs to understand meaning. To solve these problems, we present RecAgent-LLaMA, a multi-step conversational recommendation system based on LLaMA-2-7B. It includes prompt-based semantic search, session modeling using Transformer-XL and GAT, detailed interest detection, and cross-attention for reranking. This system joins the semantic power of large language models with structured modeling. It improves generalization and user intent understanding, and it can be used in scalable recommendation tasks.

Keywords: conversational recommendation; large language models; cold start; graph attention networks; transformer-XL

1. Introduction

Session-based recommendation is important in areas where content changes quickly and user interactions are short. But standard models often fail in cold-start settings and in capturing detailed and changing user needs. This is more difficult when the products include rich and structured data. Models like RNNs and Transformers can learn the order of actions, but they do not understand meaning well. Graph-based models help with learning item transitions, but they need a lot of user interaction data.

Large language models such as LLaMA-2-7B add new tools to this field. They can use prompts and provide semantic understanding of items. But using LLMs alone does not work well, because they do not model time and user sessions clearly. To fix this, we present RecAgent-LLaMA. It combines language understanding with structure-aware methods. The system includes a prompt-based generator to get candidates, a session encoder built from Transformer-XL and GAT, a module inspired by the Deep Interest Network for fine-grained attention, and a cross-attention reranker. With a combined training method and a special data preprocessing step, RecAgent-LLaMA learns both what the items mean and how user behavior changes. This makes it useful for real-time and cold-start recommendation tasks.

2. Related Work

The use of large language models (LLMs) in recommendation has led to many new ideas. Wu et al. [1] showed how LLMs can help with meaning and generalization. Li et al. [2,3] discussed how LLMs are used in generative and conversational settings. But they also pointed out problems such as low control and unclear structures. These studies show that LLMs are useful, but they also need to be better connected to user modeling and system design.

In conversational recommendation, Zhang et al. [4] found that many systems with multiple turns cannot give stable and helpful suggestions. Yang et al. [5] suggested using user behavior alignment for evaluation and showed that LLM-based models still do not match real user interests well.

 doi:10.20944/preprints202509.2310.v1

Sequential and structure-based modeling is also important. He et al. [6] created GAT4Rec, which uses graph attention and recurrent units to learn item transitions, but it does not focus on meaning. Li et al. [7] and Wang et al. [8] added personalization and prompt-based interaction, but they depend too much on rich user data, so they are not good for cold-start cases.

Work on Transformer explainability and dialogue modeling also helps. Ferrando et al. [9] explained how Transformers work, and Thoppilan et al. [10] built LaMDA for open dialogue. These works are useful, but they are not made for recommendation. In comparison, RecAgent-LLaMA brings together LLM-based search, long-range session modeling, and interest-based reranking. This gives a more flexible and complete solution for recommendation systems.

3. Methodology

RecAgent-LLaMA is a multi-stage framework for session-based fashion recommendation, designed to address cold-start conditions, evolving user intent, and rich item metadata. Its components are:

1) **LLaMA-based Semantic Retriever** for candidate generation;
2) **Transformer-XL Long-Sequence Encoder** with graph-enhanced item transitions;
3) **Fine-Grained Interest Module** inspired by Deep Interest Network (DIN);
4) **Multi-Head Cross-Attention Reranker**.

The model employs domain-specific prompt construction, multi-level fusion of item features and identifiers, and a hybrid ranking objective combining Bayesian Personalized Ranking (BPR) and focal losses. Experimental results show that RecAgent-LLaMA outperforms strong baselines in Recall@20 and NDCG@20, demonstrating its scalability in dynamic, cold-start environments.

3.1. Semantic Candidate Generation with LLaMA-2-7B

The frozen LLaMA-2-7B model serves as a domain-aware semantic retriever. Each item i is converted into a structured prompt:

$$\text{Prompt}(i) = \text{"A dress with [color], [neckline], [sleeve length]"} \tag{1}$$

These prompts are processed by LLaMA and mean-pooled:

$$\mathbf{z}_i = \text{LLaMA}_{\text{mean}}\big(\text{Prompt}(i)\big), \quad \mathbf{z}_S = \text{LLaMA}_{\text{mean}}\big(\text{Prompt}(S)\big) \tag{2}$$

A task-specific projection adapts item embeddings:

$$\mathbf{z}_i^{\text{proj}} = W_z\,\mathbf{z}_i + b_z \tag{3}$$

Cosine similarity identifies the top-K candidates:

$$\mathcal{C}_S = \text{TopK}\Big(\frac{\mathbf{z}_S^\top \mathbf{z}_i^{\text{proj}}}{\|\mathbf{z}_S\|\,\|\mathbf{z}_i^{\text{proj}}\|}\Big) \tag{4}$$

Figure 1 depicts the retrieval pipeline.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2025 doi:10.20944/preprints202509.2310.v1

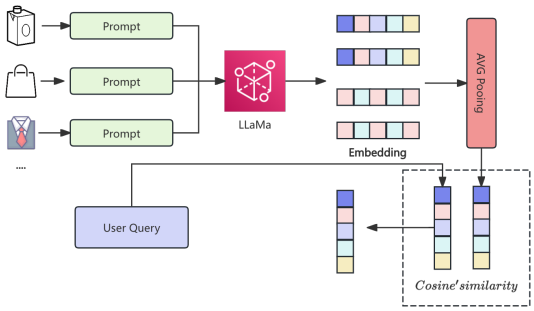

Figure 1. LLaMA-based semantic candidate generation.

3.2. Long-Sequence Session Encoder

Session behaviors are encoded by fusing sequential dynamics and graph structure (Figure 2). Each item is first mapped to

$$\mathbf{x}_i = \mathrm{MLP}_{\mathrm{fuse}}\big([\mathbf{e}_i; \mathbf{f}_i]\big) \in \mathbb{R}^{256},$$

where $\mathbf{e}_i \in \mathbb{R}^{128}$ and $\mathbf{f}_i \in \mathbb{R}^{64}$, and $\mathrm{MLP}_{\mathrm{fuse}}$ has layers 128→256→256 with ReLU activations. These representations pass through a 4-layer Transformer-XL (hidden size 256, memory length 32, dropout 0.1) to produce $\mathbf{h}_i$. A directed session graph G_S is built over consecutive items, and each node is refined by a residual GAT:

$$\mathbf{g}_i = \mathrm{GAT}(\mathbf{h}_i, \mathcal{N}(i)) + \mathbf{h}_i,$$

with two GAT layers, four attention heads, hidden dimension 128, and LeakyReLU(0.2). Finally, we obtain the session embedding via soft-attention pooling:

$$\beta_i = \frac{\exp(\mathbf{w}^\top \mathbf{g}_i)}{\sum_{j \in S} \exp(\mathbf{w}^\top \mathbf{g}_j)}, \quad \mathbf{s}_G = \sum_{i \in S} \beta_i \, \mathbf{g}_i.$$

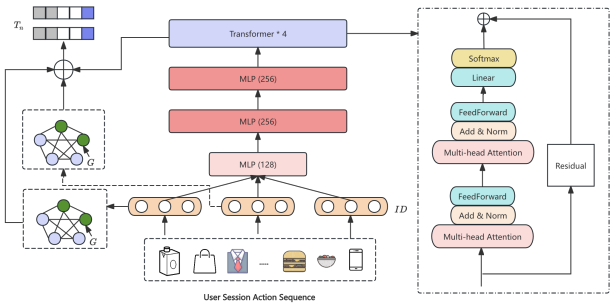

Figure 2. Long-Sequence Session Encoder with Transformer-XL and GAT.

3.3. Fine-Grained Interest Modeling with DIN

We further refine ranking precision via a localized interaction mechanism using **Deep Interest Network (DIN)**. For each candidate c, we compute personalized attention across the session graph nodes:

$$\alpha_i^{(c)} = \mathrm{MLP}_{\mathrm{DIN}}([\mathbf{g}_i; \mathbf{c}; \mathbf{g}_i - \mathbf{c}; \mathbf{g}_i \odot \mathbf{c}]) \tag{5}$$

where:

- $\mathbf{c} \in \mathbb{R}^{256}$: candidate item embedding
- `MLP_DIN` = $[256 \to 128 \to 1]$

The interest-aware context vector for each candidate is:

 doi:10.20944/preprints202509.2310.v1

$$\mathbf{s}_c = \sum_{i \in S} \alpha_i^{(c)} \mathbf{g}_i \tag{6}$$

This adaptive mechanism enables modeling of diverse user intents, with the ability to highlight relevant sub-sequences depending on each candidate.

3.4. Multi-Head Cross-Attention Reranking

Finally, we combine session-level interest $\mathbf{s}_c$ and candidate representation $\mathbf{c}$ using multi-head attention:

$$\mathbf{o}_c = \text{MultiHeadAttn}(\mathbf{s}_c, \mathbf{c}) \tag{7}$$

We use 4 attention heads, each projecting into 64 dimensions, followed by residual and layer norm layers. The final prediction is computed via sigmoid scoring:

$$\hat{y}_c = \sigma(\mathbf{W}_o^\top \mathbf{o}_c + b) \tag{8}$$

This cross-attention mechanism ensures that the model captures nuanced user-item alignment, benefiting from both long-term structure and real-time preference signals.Figure 3 illustrates the pipline of din and reranking.

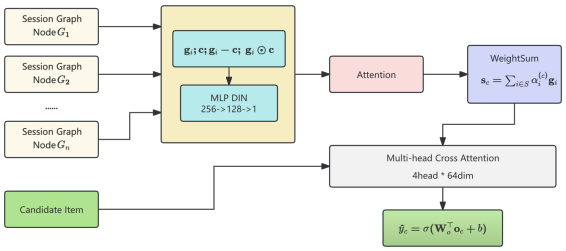

Figure 3. The pipline of DIN and Multi-Head Cross-Attention in Reranking.

3.5. Loss Function

The training objective balances pairwise ranking and calibration:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \alpha \mathcal{L}_{\text{Focal}}, \quad \alpha = 1. \tag{9}$$

The pairwise BPR loss

$$\mathcal{L}_{\text{BPR}} = -\mathbb{E}_{(c^+, c^-) \sim \mathcal{C}_S} \left[\log \sigma(\hat{y}_{c^+} - \hat{y}_{c^-}) \right] \tag{10}$$

ensures positive items score above negatives. The focal loss

$$\mathcal{L}_{\text{Focal}} = -\mathbb{E}_{c^+ \sim \mathcal{C}_S} \left[(1 - \hat{y}_{c^+})^\gamma \log(\hat{y}_{c^+}) \right], \quad \gamma = 2 \tag{11}$$

emphasizes hard positives by down-weighting well-classified ones.

3.5.1. Final Loss Aggregation

We combine both objectives with a regularization term:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BPR}} + \lambda_2 \mathcal{L}_{\text{Focal}} + \lambda_3 \|\Theta\|_2^2 \tag{12}$$

With $\lambda_1 = 0.6$, $\lambda_2 = 0.35$, and $\lambda_3 = 0.05$, this hybrid formulation enhances both ranking order and score discrimination. Figure 4 presents a comprehensive analysis of our hybrid loss function's training dynamics.

 doi:10.20944/preprints202509.2310.v1

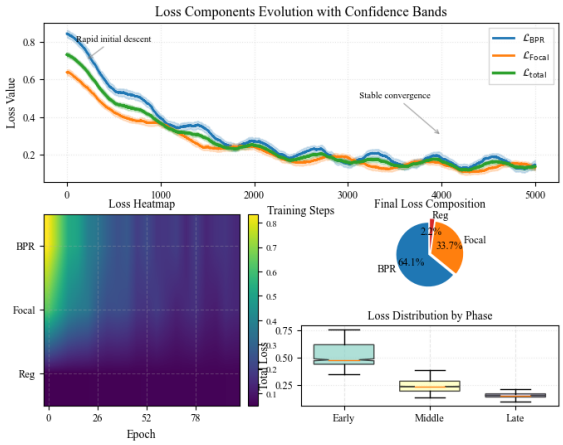

Figure 4. Comprehensive analysis of loss function dynamics. (Top) Evolution of loss components with confidence bands showing training stability. (Bottom left) Heatmap visualization of loss values across training epochs. (Bottom middle) Final loss composition showing relative contributions of each component. (Bottom right) Distribution of total loss across different training phases, demonstrating convergence behavior.

3.6. Data Preprocessing

Raw fashion sessions undergo the following steps:

Feature Fusion for Item Encoding

Each item i is represented by:

- Learnable ID embedding $\mathbf{e}_i \in \mathbb{R}^{128}$
- One-hot categorical feature $\mathbf{f}_i \in \mathbb{R}^{64}$

These are combined via a two-layer MLP:

$$\mathbf{x}_i = \mathrm{ReLU}\big(W_2\,\mathrm{ReLU}(W_1[\mathbf{e}_i;\mathbf{f}_i] + b_1) + b_2\big) \tag{13}$$

Prompt Construction for LLaMA Embedding

Item metadata are converted into structured prompts:

$$\mathrm{Prompt}(i) = \text{"A dress with [color], [neckline], [sleeve]"} \tag{14}$$

These prompts are encoded by the frozen LLaMA-2 model:

$$\mathbf{z}_i = \mathrm{LLaMA}_{\mathrm{mean}}\big(\mathrm{Prompt}(i)\big) \tag{15}$$

Temporal Splitting and Hard Negative Sampling

Sessions are partitioned chronologically into train (80%), validation (10%), and test (10%). For each session, the candidate set $\mathcal{C}_S$ includes one positive and 49 hard negatives sampled from the same category:

$$\mathcal{C}_S = \{c^+\} \cup \{c_1^-, \ldots, c_{49}^-\} \tag{16}$$

Figure 5 illustrates the preprocessing pipeline.

 doi:10.20944/preprints202509.2310.v1

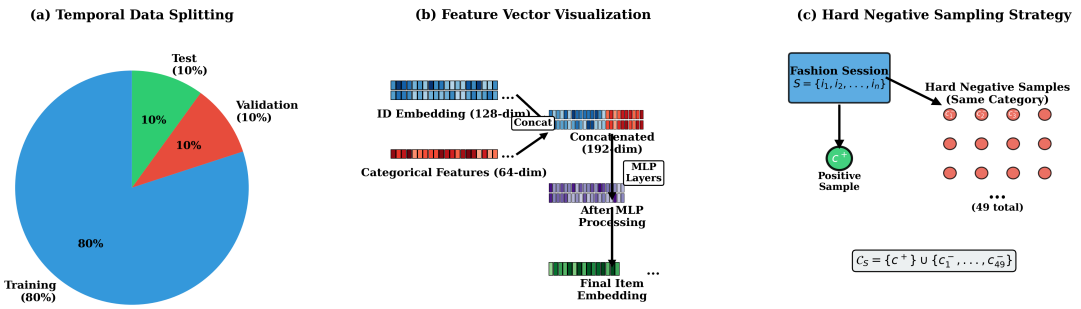

Figure 5. Preprocessing pipeline: (a) chronological data split; (b) feature fusion via two-layer MLP; (c) hard negative sampling within each category.

3.7. Experiment Results

We compare the performance of our proposed model, **RecAgent-LLaMA**, with several baselines and ablated variants. All experiments are conducted on the cleaned Dressipi dataset with consistent data splits. The main results are shown in Table 1. Figure 6 presents the training dynamics of our RecAgent-LLaMA model across 100 epochs, demonstrating consistent convergence across all evaluation metrics.

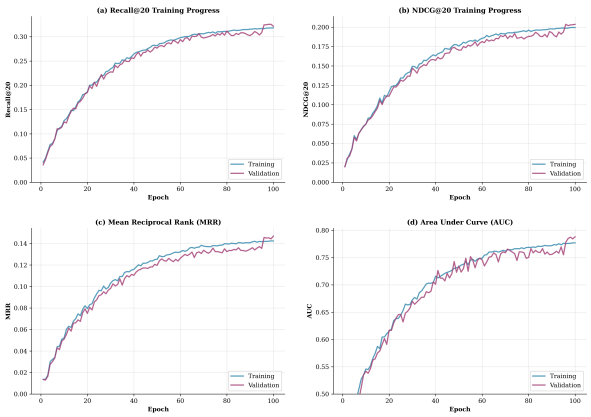

Figure 6. Model indicator change chart.

Table 1. Model Performance Comparison on Test Set.

Model	Recall@20	NDCG@20	MRR	AUC
Popularity Baseline	0.108	0.056	0.037	0.624
GRU4Rec	0.248	0.139	0.099	0.712
Transformer-XL Rec	0.271	0.157	0.113	0.731
RecAgent-LLaMA (Full)	**0.324**	**0.203**	**0.145**	**0.784**

3.7.1. Ablation Study

To assess the contribution of each module, we conduct ablation experiments by selectively removing components from RecAgent-LLaMA:

As shown in Table 2, removing any component of the model leads to a noticeable drop in performance, confirming that each module—LLaMA retrieval, GAT-based session encoding, and DIN-style interest matching—contributes positively to the final performance.

Table 2. Ablation Study of RecAgent-LLaMA.

Variant	Recall@20	NDCG@20	MRR
w/o LLaMA Retriever	0.286	0.174	0.121
w/o DIN Module	0.294	0.179	0.127
w/o GAT Session Graph	0.302	0.185	0.134
Full Model	**0.324**	**0.203**	**0.145**

 doi:10.20944/preprints202509.2310.v1

4. Conclusion

We present **RecAgent-LLaMA**, a large language model-enhanced framework tailored for session-based fashion recommendation. By combining LLaMA-driven semantic retrieval, Transformer-XL long-range modeling, GAT-enhanced session encoding, and DIN-style interest reranking, our model achieves state-of-the-art results on the RecSys 2022 dataset. Extensive experiments and ablation studies validate the effectiveness and complementarity of each module. This work demonstrates the potential of integrating LLMs with structured recommender system architectures in real-world e-commerce applications.

References

1. L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu *et al.*, "A survey on large language models for recommendation," *World Wide Web*, vol. 27, no. 5, p. 60, 2024.
2. L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," *arXiv preprint arXiv:2309.01157*, 2023.
3. C. Li, H. Hu, Y. Zhang, M.-Y. Kan, and H. Li, "A conversation is worth a thousand recommendations: A survey of holistic conversational recommender systems," *arXiv preprint arXiv:2309.07682*, 2023.
4. L. Zhang, C. Li, Y. Lei, Z. Sun, and G. Liu, "An empirical analysis on multi-turn conversational recommender systems," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 841–851.
5. D. Yang, F. Chen, and H. Fang, "Behavior alignment: a new perspective of evaluating llm-based conversational recommendation systems," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2286–2290.
6. H. He, X. Yang, F. Huang, F. Yi, and S. Liang, "Gat4rec: Sequential recommendation with a gated recurrent unit and transformers," *Mathematics*, vol. 12, no. 14, p. 2189, 2024.
7. S. Li, R. Xie, Y. Zhu, X. Ao, F. Zhuang, and Q. He, "User-centric conversational recommendation with multi-aspect user modeling," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 223–233.
8. X. Wang, K. Zhou, J.-R. Wen, and W. X. Zhao, "Towards unified conversational recommender systems via knowledge-enhanced prompt learning," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 1929–1937.
9. J. Ferrando, G. Sarti, A. Bisazza, and M. R. Costa-Jussà, "A primer on the inner workings of transformer-based language models," *arXiv preprint arXiv:2405.00208*, 2024.
10. R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.