

SemSR: Semantics aware robust Session-based Recommendations

Jyoti Narwariya*
jyoti.narwariya@tcs.com
TCS Research
New Delhi, India

Priyanka Gupta*
priyanka.g35@tcs.com
TCS Research
New Delhi, India

Muskan Gupta*
muskan.gupta4@tcs.com
TCS Research
New Delhi, India

Jyotsana Khatri
khatri.jyotsana@tcs.com
TCS Research
Pune, India

Lovekesh Vig
lovekesh.vig@tcs.com
TCS Research
New Delhi, India

ABSTRACT

Session-based recommendation (SR) models aim to recommend items to anonymous users based on their behavior during the current session. While various SR models in the literature utilize item sequences to predict the next item, they often fail to leverage semantic information from item titles or descriptions impeding session intent identification and interpretability. Recent research has explored Large Language Models (LLMs) as promising approaches to enhance session-based recommendations, with both prompt-based and fine-tuning based methods being widely investigated. However, prompt-based methods struggle to identify optimal prompts that elicit correct reasoning and lack task-specific feedback at test time, resulting in suboptimal recommendations. Fine-tuning methods incorporate domain-specific knowledge but incur significant computational costs for implementation and maintenance. In this paper, we present multiple approaches to utilize LLMs for session-based recommendation: (i) in-context LLMs as recommendation agents, (ii) LLM-generated representations for semantic initialization of deep learning SR models, and (iii) integration of LLMs with data-driven SR models. Through comprehensive experiments on two real-world publicly available datasets, we demonstrate that LLM-based methods excel at coarse-level retrieval (high recall values), while traditional data-driven techniques perform well at fine-grained ranking (high Mean Reciprocal Rank values). Furthermore, the integration of LLMs with data-driven SR models significantly outperforms both standalone LLM approaches and data-driven deep learning models, as well as baseline SR models, in terms of both Recall and MRR metrics.

CCS CONCEPTS

• Information systems → Recommender systems.

KEYWORDS

Session-based Recommendation, Large Language Models

ACM Reference Format:

Jyoti Narwariya*, Priyanka Gupta*, Muskan Gupta*, Jyotsana Khatri, and Lovekesh Vig. 2025. SemSR: Semantics aware robust Session-based Recommendations. In *Proceedings of EARL '25: Workshop on Evaluating and Applying Recommendation Systems with Large Language Models at RecSys '25*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Session-based Recommendation (SR) models are becoming increasingly popular due to their ability to recommend items based only on interactions in the current session for anonymous users. Several SR models employing deep learning architectures are proposed in the literature [11, 12, 15, 16, 21, 22, 26, 36, 38]. Recently, motivated by the notable achievements of LLMs, several works try to employ these large models in recommendations [18, 24, 25, 35]. SAID [18] aims to learn item embeddings that align with the textual descriptions of items. Authors proposed a two-stage training scheme. At first stage, SAID employs a projector module to transform an item ID into an embedding and feeds it into an LLM to explicitly to elicit the item's textual token sequence from the LLM. At second stage, embeddings are used for extracting the entire sequence's representation for recommendation. LLM-ESR [25] obtains semantic embeddings of items and users by encoding prompt texts from LLMs. Authors devise a dual-view modeling framework that combines semantic and collaborative information. Specifically, the embeddings derived from LLMs are frozen to avoid deficiency of semantics. Next, they propose a retrieval augmented self distillation method to enhance the sequence encoder of an SR model using similar users. [17] learns ID embeddings in the null space of language embeddings to combine semantic and collaborative knowledge in an optimal way.

In this work, our aim is also to utilize the capability of LLMs in SR models in an efficient and effective manner and propose **SemSR: Semantics aware robust Session-based Recommendations**; a framework where LLMs capability is incorporated in the form of embedding of items generated using pre-existing LLM. Our approach is different from the existing above approaches as we train **SemSR** end to end unlike SAID which has two stages of learning, and LLM-ESR that use self distillation method to enhance the sequence encoder of an SR model.

Through extensive experiments on two real-world publicly available datasets, we demonstrate that LLM-based methods excel at

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EARL Workshop at RecSys'25, September 22–26, 2025, Prague, Czech Republic

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

*These authors contributed equally to this work.

coarse-level retrieval (high recall values), while traditional data-driven techniques perform well at fine-grained ranking (high Mean Reciprocal Rank values). Furthermore, **SemSR** significantly outperforms both standalone LLM approaches and data-driven deep learning models, as well as baseline SR models, in terms of both Recall and MRR metrics.

Two distinct SR models are employed to show the efficacy of our approach i.e., MSGAT [29], and NISER [11]. However, the approach is model agnostic and can be employed for any SR model from the literature.

2 RELATED WORK

Session-based Recommendation: SR methods have evolved from traditional techniques such as Markov Chains [13, 19] and collaborative filtering [6, 33], to advanced deep learning-based approaches. Early methods struggled with capturing complex user behaviors and sequential dependencies, while deep learning models, such as GRU4Rec [15], SASRec[21], and SRGNN[36], have significantly improved predictive accuracy by leveraging sequential modelling, self-attention [32], and graph-based representations [9–11, 29, 34].

LLM as Recommender Systems (RS): LLMs have recently demonstrated unparalleled capabilities in natural language understanding, reasoning and beyond [27, 39]. The idea of directly utilizing the LLMs as RS have gained significant traction with models such as GPT [3], BERT [5], and LLaMA [31] which are trained on vast corpora of text, enabling a rich understanding of semantics. Due to their strong understanding of language and context, LLMs can generate more personalized, and context-aware recommendations compared to conventional models. Recent works [4, 7, 20, 30] have demonstrated that prompt-based LLMs can effectively recommend items by leveraging in-context learning. While prompt-based and fine-tuned LLM based techniques show strong potential, they are sensitive to the design of prompts and often lack up-to-date data knowledge which can lead to irrelevant recommendations in dynamic environments like e-commerce.

LLMs to augment Recommender Systems:

As more and more LLMs have been developed, the research has progressively explored how to utilize the knowledge in LLMs to improve the sequential recommendations. Recently, several works have highlighted the effectiveness of LLMs as components in recommendation tasks [8, 14, 18, 24, 25, 28, 35, 37]. KAR [37] proposes to use language embeddings as additional input in the learning of ID embeddings. LLM4SBR [28] utilizes a two step strategy. Firstly, session data is transformed into both textual and behavioral modalities, allowing LLMs to infer session intent from textual descriptions. Secondly, SR models use behavioral data to align and average session representations across two different modalities. LLM-ESR[25] retrieves semantic embeddings of items and users by encoding prompts from LLMs and uses a retrieval augmented self distillation method to enhance the sequence encoder of an SR model. SAID[18], on the other hand, evolves a two-stage training process: the first stage involves generating item embeddings by leveraging the projector module and LLM, and in the second stage, learned item embeddings are input into the sequential model to extract the entire sequence’s representation for recommendation. In contrast to above methods, our proposed method SemSR introduces an end-to-end framework that directly incorporates LLM-generated

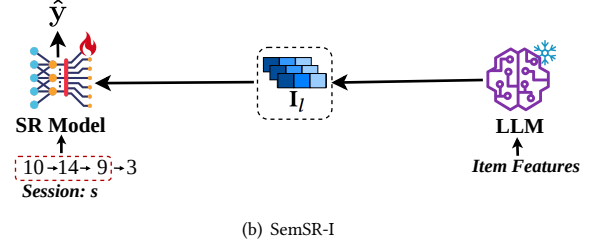
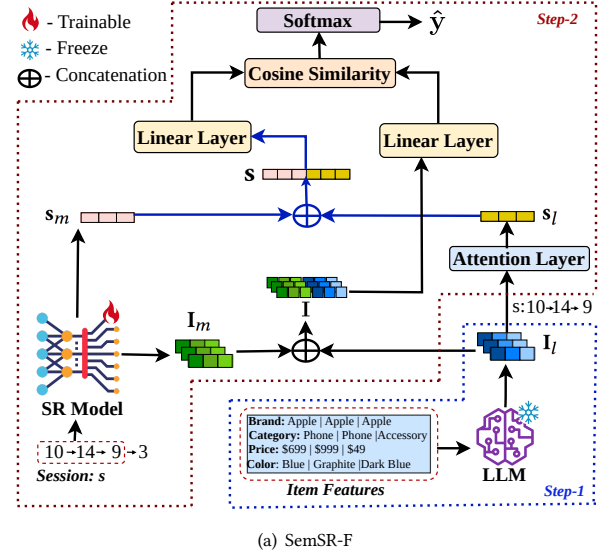


Figure 1: An overview of our proposed framework SemSR. (a) SemSR-F: For a given session s , SR model provides item ids based session representation s_m and item representations I_m . LLM model generates semantic item representations I_l for all items and s_l is the LLM session representation obtained using embeddings of items i_j where $\forall i_j \in s$ via attention layer. Further s and I_l session and item representations obtained in end-to-end framework. Step-1 represents semantically aligned embedding learning and Step-2 represents agnostic sequential recommender training. (b) SemSR-I: Leverage LLM generated item representations I_l in SR model.

item embeddings in SR model to obtain top- K recommendations. AlphaFuse [17] proposes an approach which injects collaborative signals into the null space of language embeddings which helps in preserving the semantic information. In AlphaFuse, trainable ID embeddings are learned in an orthogonal null space. Our approach is different, it tries to fuse semantic and collaborative signals (trainable) and the entire model is trained jointly.

3 PROPOSED FRAMEWORK

This section will introduce the problem definition and details of our approach:

3.1 Problem Definition

Suppose that \mathcal{S} denotes the set of all sessions in the logged data containing user-item interactions (e.g. click/view/order), and \mathcal{I} denotes the set of n items observed in \mathcal{S} . Any session $s \in \mathcal{S}$ is a

System Prompt: You are a helpful recommendation system for e-commerce platform.

User Prompt: Your task is to predict the title of the next product that a customer is going to engage with, based on their session data. The output should only contain the predicted title of the product and nothing else.

Here are a few examples:

Session: iphone max pro, iphone 13, iphone 13 cover
Predicted Next Product Title: Apple Air Pods
Session: {} **Predicted Next Product Title:** {}

Now, Predict Next Title for the following session
Session: 'Eau De Parfum', 'Perfuming Body Oil', 'Arbonne Bronzer'
Next Product Title:

Output by LLM: Arbonne Makeup Primer

Figure 2: Few-shot in-context prompt (FS-LLM) template

sequence of item-click events: $s = (i_{s,1}, i_{s,2}, \dots, i_{s,|s|})$, where $i_{s,j}$ ($j = 1 \dots |s|$) $\in \mathcal{I}$, denotes the j^{th} clicked item in session s . The goal of SR is to predict the next item $i_{s,|s|+1}$ as the target class in an n -way classification problem by estimating the n -dimensional item-probability vector $\hat{\mathbf{y}}_{s,|s|+1}$ corresponding to the relevance scores for the n items. The K items with the highest scores constitute the top- K recommendations.

3.2 LLM as Recommender System

We leverage an LLM model (Llama3-8B-Instruct^{*}) as an in-context text generation method to generate the title of the next recommended item for a given session which contains a sequence of clicked items titles. We use different in-context prompt design strategies such as few-shot in-context prompting, chain-of-thought (CoT) prompting to generate the next recommended title.

Prompt Design: We utilize different prompting strategies but use a common definition for designing prompts i.e., firstly, we define the role for system, then we define task for the model and get the response title from the LLM. Different prompting strategies are as follows:

- **Few-shot in-context prompt (FS-LLM):** In this prompt, in addition, we provide few shot input-output examples to the LLM, followed by the query session and allow the model to predict the next item title. The prompt template for FS-LLM is shown in figure 2.
- **Zero-shot Chain-of-Thought prompt (ZCoT-LLM):** In this prompt, we design a two step chain of thought prompt. First we generate rationale and then generate recommendations based on the generated rationale.
- **Few-shot Chain-of-Thought prompt (FSCoT-LLM):** Similar to ZCoT-LLM strategy, here we also provide some examples in the prompt to learn input and output relations.

LLM Inference: To retrieve top- k recommendations, we use a vector database (Chromadb^{*}) to store item embeddings obtained from the Llama3 model. We then retrieve top- k recommendations based on cosine similarity between the generated next item title and stored items in the database.

^{*}<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

^{*}<https://docs.trychroma.com/docs/overview/getting-started>

3.3 SemSR

The overall architecture of SemSR is shown in Figure 1. The following section specifies the implementation details of the architecture.

3.3.1 Semantic-aware Item and Session Representation. Each item is mapped to a d_1 -dimensional vector from the trainable embedding look-up table to obtain interaction-based item embeddings as $\mathbf{I}_m = [\mathbf{i}_{m_1}, \mathbf{i}_{m_2}, \dots, \mathbf{i}_{m_n}]^T \in \mathbb{R}^{n \times d_1}$ such that each row $\mathbf{i}_{m_j} \in \mathbb{R}^{d_1}$ is the d_1 -dimensional embedding vector corresponding to item $i_j \in \mathcal{I}$ ($j = 1, 2, \dots, n$). Further, each item is mapped to a d_2 -dimensional vector by inputting item features like brand, category, price, color, title, description, etc to an LLM model (in this work, we consider Angle [23] and LLM2Vec [1] embedding with BERT and Llama-3.1-8B as the backbone model respectively) and obtain LLM-based item embeddings as $\mathbf{I}_l = [\mathbf{i}_{l_1}, \mathbf{i}_{l_2}, \dots, \mathbf{i}_{l_n}]^T \in \mathbb{R}^{n \times d_2}$. These embeddings are frozen during model training.

We denote data-driven (interaction based) session embeddings by \mathbf{s}_m and LLM based session embedding by \mathbf{s}_l . Consider any function f (e.g., SR model's neural network from literature [11, 29]), parameterized by θ - maps sequence of items in session s to $\mathbf{s}_m = f(\mathbf{I}_{m,s}, \theta)$, where, $\mathbf{I}_{m,s} = [\mathbf{i}_{m,s,1}, \mathbf{i}_{m,s,2}, \dots, \mathbf{i}_{m,s,|s|}]^T \in \mathbb{R}^{|s| \times d_1}$. For obtaining \mathbf{s}_l , we compute the soft-attention weight of the j -th item in session s as $\alpha_j = \mathbf{q}^T \text{sigmoid}(\mathbf{W}_1 \mathbf{i}_{l,s,|s|} + \mathbf{W}_2 \mathbf{i}_{l,s,j} + c)$, where ($j = 1, 2, \dots, |s|-1$), $\mathbf{q}, \mathbf{c} \in \mathbb{R}^{d_2}$, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_2}$, and $\mathbf{i}_{l,s,|s|}$ is the most recent item in session s . The α_j 's are further normalized using a softmax operation yielding intermediate session embedding $\mathbf{s}' = \sum_{j=1}^{|s|-1} \alpha_j \mathbf{i}_{l,s,j}$. The LLM-based session embedding \mathbf{s}_l is a linear transformation over the concatenation of intermediate session embedding \mathbf{s}' and the embedding of the most recent item $\mathbf{i}_{l,s,|s|}$, s.t. $\mathbf{s}_l = \mathbf{W}_3 [\mathbf{s}'; \mathbf{i}_{l,s,|s|}]$, where $\mathbf{W}_3 \in \mathbb{R}^{d_2 \times 2d_2}$.

Finally, semantic aware session embeddings are obtained by linear transformation over the concatenation of \mathbf{s}_m and \mathbf{s}_l as $\mathbf{s} = \mathbf{W}_4 [\mathbf{s}_m; \mathbf{s}_l]$, where $\mathbf{W}_4 \in \mathbb{R}^{d \times (d_1 + d_2)}$. Further, semantic aware item embeddings are computed as linear transformation over the concatenation of \mathbf{I}_m and \mathbf{I}_l as $\mathbf{I} = \mathbf{W}_5 [\mathbf{I}_m; \mathbf{I}_l]$, where $\mathbf{W}_5 \in \mathbb{R}^{d \times (d_1 + d_2)}$. The semantic aware item and session embeddings are then used to obtain the relevance score for next clicked item i_k computed as,

$$\hat{\mathbf{y}}_k = \frac{\exp(\mathbf{i}_k^T \mathbf{s})}{\sum_{j=1}^n \exp(\mathbf{i}_j^T \mathbf{s})}. \quad (1)$$

3.3.2 Training and Inferencing of SemSR. The goal is to obtain \mathbf{s} that is close to the embedding $\mathbf{i}_{s,|s|+1}$ of the target item $i_k = i_{s,|s|+1}$, where k is estimated class for the target item, $k = \arg \max_j \mathbf{i}_j^T \mathbf{s}$ with $j = 1 \dots n$. For this n -way classification task, softmax (cross-entropy) loss is used during training for estimating θ by minimizing the sum of $\mathcal{L}(\hat{\mathbf{y}}) = -\sum_{j=1}^n y_j \log(\hat{y}_j)$ over all training samples, where $\mathbf{y} \in \{0, 1\}^n$ is a 1-hot vector with $y_k = 1$ corresponding to the correct (target) class k .

During inference, the final recommendation scores for the n items are computed by eq. 1. The top- K items are considered as the recommended items.

4 EXPERIMENTAL EVALUATION

In this section, we conduct extensive experiments to answer the following research questions:

Table 1: Statistics of the datasets used for experiments.

Datasets	#train	#test	#items	Avg. s
Amazon-M2 (UK)	1172181	115936	499611	4.12
Amazon-Beauty	290512	21580	54615	6.40

- **RQ1:** Are in-context LLMs’ predictions competitive with data driven SR methods?
- **RQ2:** Does semantic-aware representations with data-driven models improve performance?
- **RQ3:** Does re-ranking top-K recommendations improve performance in-terms of MRR?

We consider two SR models from the literature to demonstrate the efficacy of **SemSR**, i.e., MSGAT [29], and NISER [11]. We compared **SemSR** with exiting SR models from literature e.g., NARM [22] and SRGNN [2].

Dataset Details: We consider two datasets i.e., english dataset from Amazon KDDCup challenge 2023 (AmazonKDD-M2 (UK))* and Amazon Beauty 2014 review dataset (Beauty)* to evaluate our approach. After preprocessing, the statistics of the datasets are shown in table 1.

AmazonKDD-M2 (UK): We use a real-world multilingual recommendation dataset from KDDCup challenge 2023 to evaluate the effectiveness of SemSR. The dataset provides items and sessions in six different languages but for this paper non-english sessions are filtered out yielding 4, 99, 611 items with average sessions length of 4.12. We consider data from task 1 of Amazon KDDCup challenge 2023. The number of given training/testing sessions considered are 11, 72, 181/1, 15, 936 respectively. For LLM as RS experiments, we selected a subset of the data to minimize cost. We selected the most recent 10k sessions as the test set based on chronological splits to minimize the computational cost.

Amazon-Beauty: We consider the Amazon-Beauty sub-category from the Amazon 2014 review dataset which contains timestamped user-item interactions from May 1996 to July 2014, and metadata contains items’ title, descriptions, categories, brands, price, etc. We only consider users having more than 5 reviews, filtering out less popular items that have a frequency less than 5 and removed sessions of length less than 2 from the data. We split sessions based on user ids and consider 80%:10%:10% split of data into training, validation and test set, respectively. Average sessions length is 6.40. Further, we create incremental sessions to improve training of the SR model.

SemSR and its variants: We proposed the following variants of our approach:

- **Semantic Initialization via LLM embeddings (SemSR-I):** Item embeddings are initialized by LLM embeddings for MSGAT [29], and NISER [11] (denoted as SemMSGAT-I, SemNISER-I) respectively.
- **Semantically initialized model fused with SR (SemSR-F):** LLM based item embeddings are concatenated with data driven item embeddings followed by a linear transformation. Similarly, LLM

*<https://www.aicrowd.com/challenges/amazon-kdd-cup-23-multilingual-recommendation-challenge>

*<https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html>

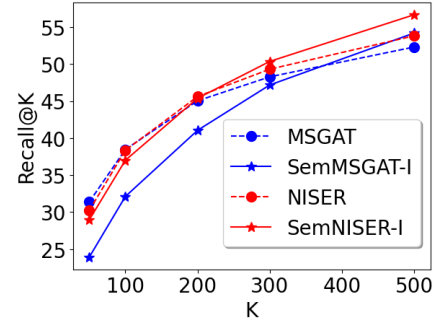


Figure 3: Recall@K with varying K indicating larger gains by using SemSR-I over SR with increasing K on Amazon-Beauty dataset.

based session embedding are fused with data driven session embeddings. SemMSGAT-F, SemNISER-F are variants for MSGAT, and NISER, respectively.

- **SemSR-I+, SemSR-F+:** Re-ranked recommendations list obtained from SemSR-I, and SemSR-F using NISER model.

Evaluation Metrics: We use the standard offline evaluation metrics Recall@K and Mean Reciprocal Rank (MRR@K). **Recall@K** represents the proportion of test instances which has the desired item in the top-K items. **MRR@K** (Mean Reciprocal Rank) is the average of reciprocal rank of desired item in recommendation list.

Hyperparameter Setup: We use validation data for hyperparameter selection using Recall@100 as the performance metric for all approaches except in-context LLMs based approach as this does not require model training. We use the adam optimizer with mini-batch size 100, momentum 0.9, $d_1 = 100$, $d_2 = 1024$, and $d = 100$. For NISER and its SemSR variants, we use the same parameters [11] i.e., scaling factor = 16.0 and learning rate 0.001. For MSGAT, we grid-search over learning rate in $\{(0.1, 0.01, 0.001, 0.0001)\}$. The best learning rate on the validation set is $lr = 0.001$.

4.1 Results and Discussion

From 2, we observe that MSGAT, and NISER performs significantly better than existing baselines NARM and SRGNN.

- **RQ1: LLM as RS vs data-driven SR model,** In table 3, we compare LLM as recommendation methods FS-LLM, ZCOT-LLM, and FSCoT-LLM with traditional data-driven SR models. We observe that while LLM as RS methods demonstrate some capability in recommendation task, their performance significantly lags behind data-driven models such as MSGAT and NISER. For example, FSCoT-LLM achieves R@20 and MRR@20 score of 31.70 and 10.00, respectively on the Amazon-M2 dataset, and only 7.04 and 1.86, respectively on Amazon-Beauty dataset. These results suggest that although LLMs may offer generalization benefits, they currently lack the collaborative knowledge which is necessary for session-based recommendation.
- **RQ2: Effectiveness of Semantic Integration,** From table 2, we observe that semantic initializations of embeddings significantly enhances performance at a coarse level of retrieval, i.e., SemMSGAT-I and SemNISER-I perform significantly better than respective vanilla SR models in terms of recall on the Amazon-M2 (UK) dataset and are comparable on the Amazon-Beauty

Table 2: Evaluation on Amazon-M2(UK) and Beauty datasets. We report overall performance in terms of Recall@20, MRR@20, Recall@100 and MRR@100. Bold numbers are for the best.

	Method	Amazon-M2 (UK)				Amazon-Beauty			
		R@20	MRR@20	R@100	MRR@100	R@20	MRR@20	R@100	MRR@100
	NARM	32.49	17.00	42.55	17.26	15.68	4.49	29.36	4.82
	SRGNN	41.77	23.54	51.67	23.54	19.67	6.14	34.59	6.05
	MSGAT	38.86	21.41	49.95	21.69	21.55	6.41	38.45	6.83
Angle embeddings with BERT	SemMSGAT-I	46.00	14.90	63.49	15.33	15.30	3.93	32.06	4.32
	SemMSGAT-I+	46.00	25.82	63.49	27.25	15.30	5.77	32.06	7.08
	SemMSGAT-F	51.55	27.36	65.81	27.73	22.80	7.44	41.22	7.88
	SemMSGAT-F+	51.55	27.37	65.81	27.48	22.80	7.66	41.22	7.94
Llama 3 embeddings	SemMSGAT-I	47.48	16.46	65.59	16.91	18.00	4.80	37.23	5.26
	SemMSGAT-I+	47.48	26.15	65.59	27.45	18.00	6.19	37.23	7.32
	SemMSGAT-F	52.66	26.12	68.82	26.53	24.07	6.93	43.64	7.40
	SemMSGAT-F+	52.66	27.42	68.82	27.52	24.07	7.72	43.64	7.85
Angle embeddings with BERT	NISER	49.84	26.45	61.77	26.76	20.84	6.95	38.35	7.36
	SemNISER-I	46.96	14.25	64.44	14.68	19.62	5.07	37.00	5.49
	SemNISER-I+	46.96	26.06	64.44	27.33	19.62	6.94	37.00	7.66
	SemNISER-F	54.98	26.41	69.90	26.79	22.67	6.10	42.41	6.57
	SemNISER-F+	54.98	27.71	69.90	27.57	22.67	7.33	42.41	7.63
Llama 3 embeddings	SemNISER-I	50.05	16.47	67.62	16.91	19.95	4.80	40.23	5.28
	SemNISER-I+	50.05	26.85	67.62	27.57	19.95	6.54	40.23	7.38
	SemNISER-F	54.26	25.19	69.87	25.59	20.95	5.42	40.36	5.89
	SemNISER-F+	54.26	27.66	69.87	27.60	20.95	7.24	40.36	7.56

Table 3: LLM as RS comparison with data-driven SR methods on subset of Amazon-M2(UK) and Beauty dataset.

Method	Amazon-M2 (UK)				Amazon-Beauty			
	R@20	MRR@20	R@100	MRR@100	R@20	MRR@20	R@100	MRR@100
FS-LLM	31.83	9.92	47.50	10.30	7.07	1.90	14.48	2.07
ZCoT-LLM	25.16	7.65	39.29	7.99	7.12	1.82	15.08	1.99
FSCoT-LLM	31.70	10.00	47.97	10.44	7.04	1.86	14.32	1.89
MSGAT	26.28	14.90	34.74	15.09	21.55	6.41	38.46	6.83
SemMSGAT-I	44.28	14.76	61.09	15.18	18.00	4.80	37.23	5.26
SemMSGAT-F	47.64	21.55	63.79	21.96	24.07	6.93	43.64	7.40
NISER	38.02	20.21	47.56	20.46	20.84	6.95	38.35	7.36
SemNISER-I	46.64	14.77	62.97	15.18	19.95	4.80	40.23	5.28
SemNISER-F	47.17	20.83	63.83	21.25	20.95	5.42	40.36	5.89

dataset. From table 2, we also observe that fusion-based models SemMSGAT-F and SemNISER-F consistently outperform their vanilla SR models as well as other existing methods in terms of recall as well as MRR which further emphasizes the advantages of incorporating semantic information into the data-driven SR models.

- **RQ3: Assessing the benefits of Re-ranking**, From table 2, we observe that re-ranking using vanilla SR methods helps to improve fine-grained ranking i.e., it further improves MRR for all semantic variants SemMSGAT-I+, SemMSGAT-F+, SemNiser-I+ and SemNiser-F+ over SemMSGAT-I, SemMSGAT-F, SemNiser-I and SemNiser-F, respectively.
- From figure 3, we observe that for lower K value (50 and 100), vanilla SR models MSGAT and NISER perform better than MSGAT-I and NISER-I, respectively. However, for higher K values ($K=200$,

..., 500) trend is reversed. Moreover, the performance gap between MSGAT vs MSGAT-I and NISER vs NISER-I widens as K increases. This suggests that semantic initialization via LLM embeddings excel at coarse-level retrieval.

5 CONCLUSION

In this work, we highlighted various methods of leveraging LLMs for SR, i.e., in-context prompting and retrieval via LLMs, and integrating LLM-based embeddings with deep learning based state-of-the-art SR models i.e., MSGAT [29] and NISER [11]. We showed the comparison of different approaches incorporating LLMs, and demonstrated that LLMs can help to improve the performance of SR models by understanding semantics of items and their features/meta-information.

REFERENCES

- [1] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961* (2024).
- [2] Juan Belieni and Diego Mesquita. 2025. SRGNN: simple recurrent graph neural network. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics* 11, 1 (2025), 1–2.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [4] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt’s capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1126–1132.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [6] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends® in Human-Computer Interaction* 4, 2 (2011), 81–173.
- [7] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*. 299–315.
- [8] Naicheng Guo, Hongwei Cheng, Qianqiao Liang, Linxun Chen, and Bing Han. 2024. Integrating Large Language Models with Graphical Session-Based Recommendation. *arXiv preprint arXiv:2402.16539* (2024).
- [9] Muskan Gupta, Priyanka Gupta, Jyoti Narwariya, Lovekesh Vig, and Gautam Shroff. 2024. SCM4SR: Structural Causal Model-based Data Augmentation for Robust Session-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2609–2613.
- [10] Muskan Gupta, Priyanka Gupta, and Lovekesh Vig. 2024. Guided Diffusion-based Counterfactual Augmentation for Robust Session-based Recommendation. *arXiv preprint arXiv:2410.21892* (2024).
- [11] Priyanka Gupta, Diksha Garg, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2019. NISER: Normalized Item and Session Representations with Graph Neural Networks. *arXiv preprint arXiv:1909.04276* (2019).
- [12] Priyanka Gupta, Ankit Sharma, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2021. Causer: Causal session-based recommendations for handling popularity bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3048–3052.
- [13] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 191–200.
- [14] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 720–730.
- [15] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [16] Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022. Core: simple and effective session-based recommendation within consistent representation space. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1796–1801.
- [17] Guoqing Hu, An Zhang, Shuo Liu, Zhibo Cai, Xun Yang, and Xiang Wang. 2025. Alphafuse: Learn id embeddings for sequential recommendation in null space of language embeddings. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1614–1623.
- [18] Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM Web Conference 2024*. 103–111.
- [19] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*. 135–142.
- [20] Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2024. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*. Springer, 494–502.
- [21] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [22] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.
- [23] Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871* (2023).
- [24] Qidong Liu, Xian Wu, Wanyu Wang, Yejing Wang, Yuanshao Zhu, Xiangyu Zhao, Feng Tian, and Yefeng Zheng. 2025. LLMEmb: Large Language Model Can Be a Good Embedding Generator for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12183–12191.
- [25] Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. 2024. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 26701–26727.
- [26] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1831–1839.
- [27] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.
- [28] Shutong Qiao, Chen Gao, Junhao Wen, Wei Zhou, Qun Luo, Peixuan Chen, and Yong Li. 2024. LLM4SBR: A Lightweight and Effective Framework for Integrating Large Language Models in Session-based Recommendation. *arXiv preprint arXiv:2402.13840* (2024).
- [29] Shutong Qiao, Wei Zhou, Junhao Wen, Hongyu Zhang, and Min Gao. 2023. Bi-channel multiple sparse graph attention networks for session-based recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 2075–2084.
- [30] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*. 890–896.
- [31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [33] Meirui Wang, Pengjie Ren, Lei Mei, Zhumin Chen, Jun Ma, and Maarten De Rijke. 2019. A collaborative session-based recommendation approach with parallel memory modules. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 345–354.
- [34] Shoujin Wang, Liang Hu, Yan Wang, Xiangnan He, Quan Z Sheng, Mehmet A Orgun, Longbing Cao, Francesco Ricci, and Philip S Yu. 2021. Graph learning based recommender systems: A review. *arXiv preprint arXiv:2105.06339* (2021).
- [35] Ziyan Wang, Yingpeng Du, Zhu Sun, Haoyan Chua, Kaidong Feng, Wenya Wang, and Jie Zhang. 2024. Re2LLM: Reflective Reinforcement Large Language Model for Session-based Recommendation. *arXiv preprint arXiv:2403.16427* (2024).
- [36] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based Recommendation with Graph Neural Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.
- [37] 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*. 12–22.
- [38] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.
- [39] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).