Leveraging Emerging LLM Capability for Constructing an Advancing Sequential Recommendation System

Group 2
Project Type: Research

Presentation Requirement

- **Requirement (Duration of Presentation)**
  - For a 6-member group, each member must give at least 3 minutes and at most 3.33 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 18 minutes and at most 20 minutes.
  - For a 5-member group, each member must give at least 3 minutes and at most 4 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 15 minutes and at most 20 minutes.
  - For a 4-member group, each member must give at least 3 minutes and at most 5 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 12 minutes and at most 20 minutes.
  - For a 3-member group, each member must give at least 3 minutes and at most 6.66 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 9 minutes and at most 20 minutes.
  - For a 2-member group, each member must give at least 3 minutes and at most 10 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 6 minutes and at most 20 minutes.
  - For a 1-member group, each member must give at least 3 minutes and at most 20 minutes for the entire presentation. Thus, the duration of the entire presentation is at least 3 minutes and at most 20 minutes.

  - An over-time presentation may lead to **mark deductions**. Thus, please time your presentation.

PPT File Requirement

- On each slide of the PPT file, please include the name of the presenter at the bottom left corner of this slide.

Submission Details

- Each group must need to submit the filled presentation order form in Canvas **before** your presentation time slot where the form could be found in this link.
- Each group must need to submit a zipped file containing the following to Canvas on the day of the presentation (23:59).
  1. a PPT file
  2. some "optional" materials to be described next
- If your group has some source code files, please also include them in the zipped file. In your source files, please write a readme file which includes the following.
  1. how to compile
  2. how to execute
  3. the description of each source file
  4. an example to show how to run the program
  5. the operating system you tested your program (e.g., linux and Windows)
  6. anything you want to include

  Alternatively, you could include the URL of an online code repository (e.g., Github) in a readme file.

General Outline

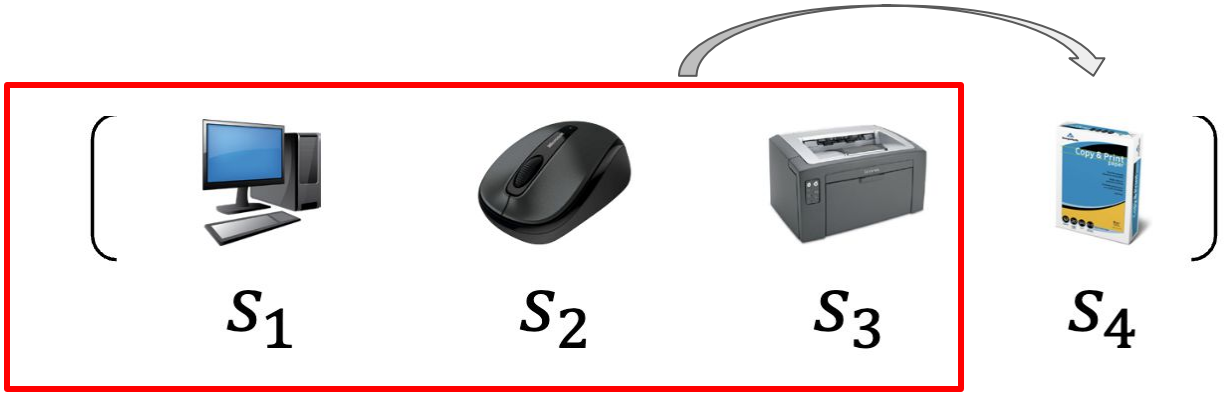
1. Introduction (3 mins) wenbin
   a. Background info
   b. Motivation
   c. Problem statement and research gap (2 parts)
   d. Related work
2. Knowledge distillation framework to create a lightweight and efficient student model (3 mins **Huihao Jing**)
   a. Methodology
3. A Noise-Robustness Module (3 mins **Zaifei YANG**)
   a. Methodology
4. Experiment and Evaluation (3 mins + 3 mins) (**Shuhao CHEN (2), Hoi Ying LAU (1)**)
   a. Setup
   b. Result
   c. Ablation Study
5. Result and discussion (5+6 ~ 3 mins)（JIN Mengxi）
   a. Limitation & future direction
6. Conclusion

Introduction

From Sequential Recommendation to LLM4Rec.

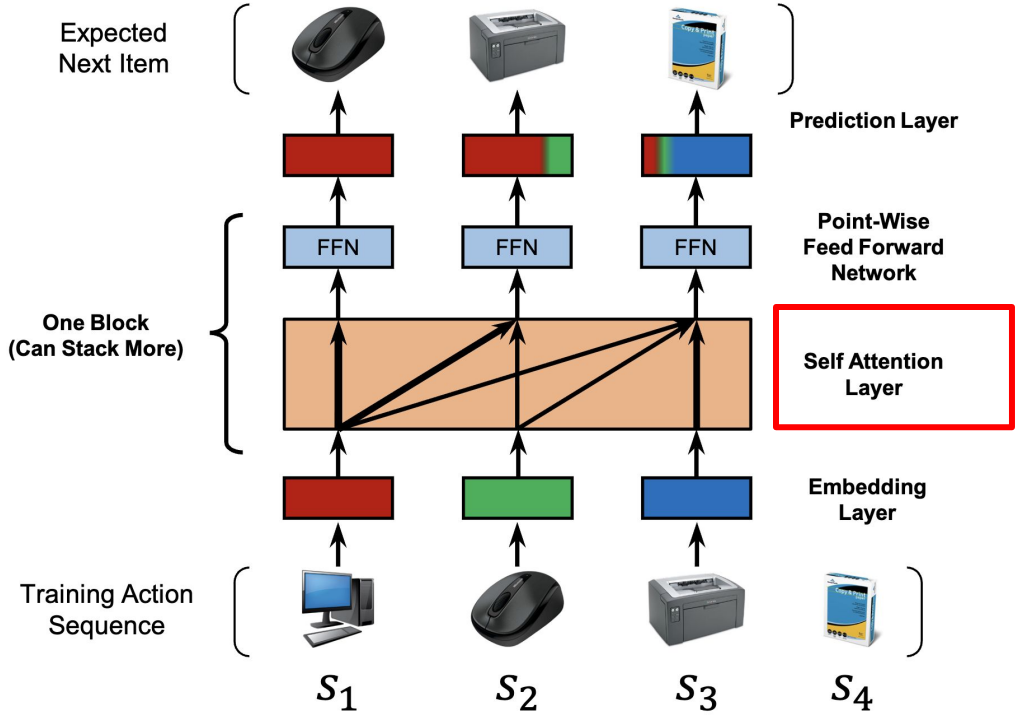
Wenbin Hu

Sequential Recommandation



- Goal: Predict next item with a past sequence.
- Items contain some textual information.

Sequential Recommandation (continue)



These layers can be attention or RNN.

SASrec: use attention layers as backbone for sequential modeling.

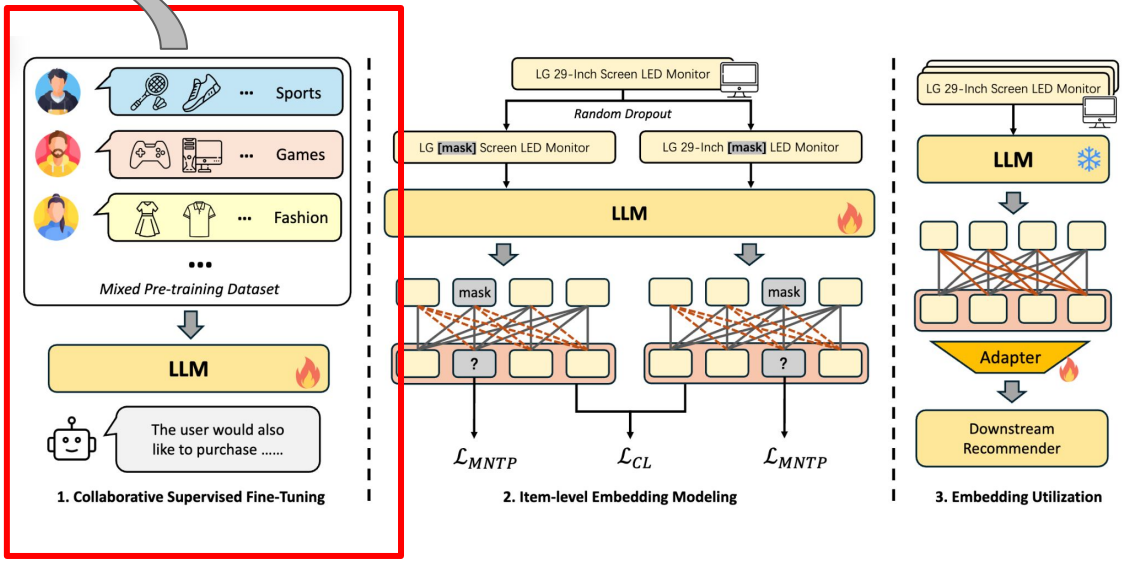
Wenbin Hu

Wang-Cheng Kang, Julian McAuley. Self-Attentive Sequential Recommendation. ICDM 2018.

Sequential Recommandation with LLMs

Input:
Logitech G13 Gameboard,
Tamron 70-200mm Camera Lens (Nikon) ,
Pelican SD Card Case,
YONGNUO Flash Trigger (Canon) ,
TAKSTAR SGC-598 Microphone,
STK EN-EL3e Charger for Nikon Camera ,
VGA to HDMI Cable,
Allstate 2-Year Protection Plan,
BLACKRAPID Lock Star Cover.

Output: Neewer Wireless Flash Trigger for Camera



LLM2Rec: LLM as embedder with SFT on sequential items.

- Leverage language understanding capability of LLMs, to capture textual information.
- Better generalization.

Wenbin Hu

Yingzhi He, Xiaohao Liu, An Zhang, Yunshan Ma, Tat-Seng Chua. LLM2Rec: Large Language Models Are Powerful Embedding Models for Sequential Recommendation. KDD 2025.

Research Directions in Our Work

Cons of existing LLM4Rec:

1. Model weights are heavy (≥1.5B). Usually, recommendation system in industry should handle billions of request.
   - RD1: We leverage a distillation method on a 135M LLM, which can maintain decent performance.
2. We find that our introduced item-level noise can hurt LLMs' performance.
   - RD2: We introduce our rerank strategy against the noise.

Efficiency Enhancement Module

Accelerating LLM2Rec via Knowledge Distillation

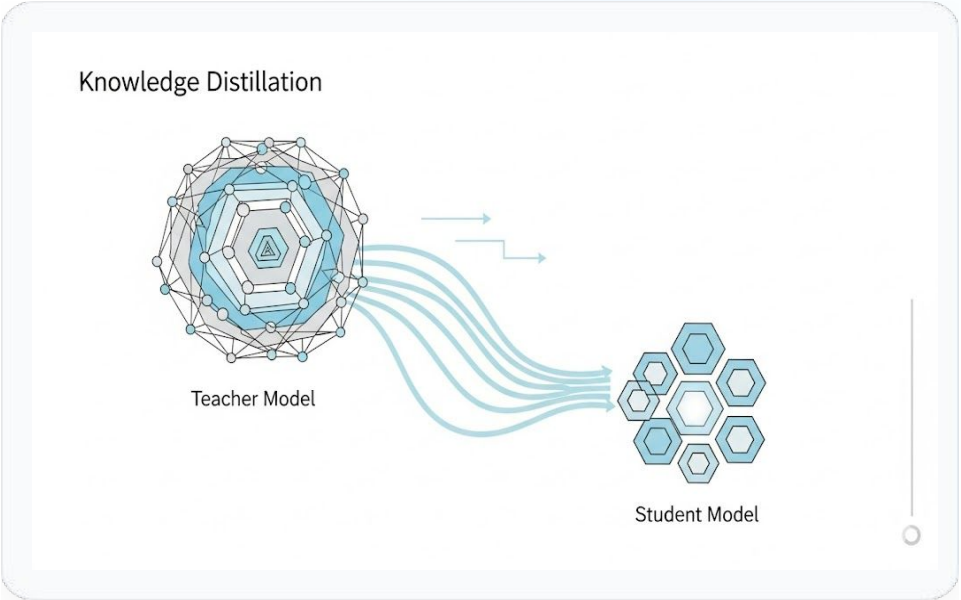
Jing, Huihao

The Distillation Framework

The Challenge: Real-world deployment is hindered by high parameter counts, leading to slow inference and excessive storage costs.

The Solution: A distillation approach to transfer performance from a large Teacher model to a lightweight Student model.

- **Teacher Model :** A large, fixed embedding model that guides training.
- **Student Model :** A lightweight model learning to generate effective vector representations.



Distillation Objective & Loss

Objective

The goal is to align the relational structure of Student embeddings with the Teacher embeddings.

Mechanism

We minimize the **KL Divergence** between the pairwise similarity distributions of both models, controlled by a temperature parameter τ.

KL Divergence Loss (L_{KL})

$$L_{\mathrm{KL}} = \mathrm{KL}\left(\mathrm{softmax}\left(\frac{T_X T_X^{\mathrm{T}}}{\tau}\right) \middle\| \mathrm{softmax}\left(\frac{S_X S_X^{\mathrm{T}}}{\tau}\right)\right)$$

Final Loss Function

$$L = \lambda_1 L_{\mathrm{IC}} + \lambda_2 L_{\mathrm{KL}}$$

* L_{IC} is the stabilization term from LLM2Rec design.

Jing, Huihao

Noise-Robustness Module: From Generative Imputation to Post-hoc Re-ranking

YANG, Zaifei

Problem Formulation: Defining "Deletion Noise"

- **Motivation:** In real-world recommendation scenarios, user interaction sequences are **frequently incomplete**. This incompleteness is rarely due to a lack of user interest; instead, it stems from external factors such as tracking failures, cross-device usage gaps, or interactions occurring during unlogged sessions.
- **Challenge:** We define this phenomenon as **"Deletion Noise"**. Unlike insertion noise which adds irrelevant items, deletion noise removes critical intermediate steps, creating "logical discontinuities" in the user's history.
- **Simulation:** we simulate deletion noise by defining a **dropout mechanism**. For a sequence S_u of length T, we randomly sample a deletion index k from a uniform distribution and remove item i_k, This results in a corrupted sequence with length T-1, which serves as the input for our robustness experiments.

$$S_u = [i_1, i_2, \dots, i_T] \longrightarrow k \sim \text{Uniform}(\{1, 2, \dots, T\}) \longrightarrow \tilde{S}_u = [i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_T]$$

sample a deletion index

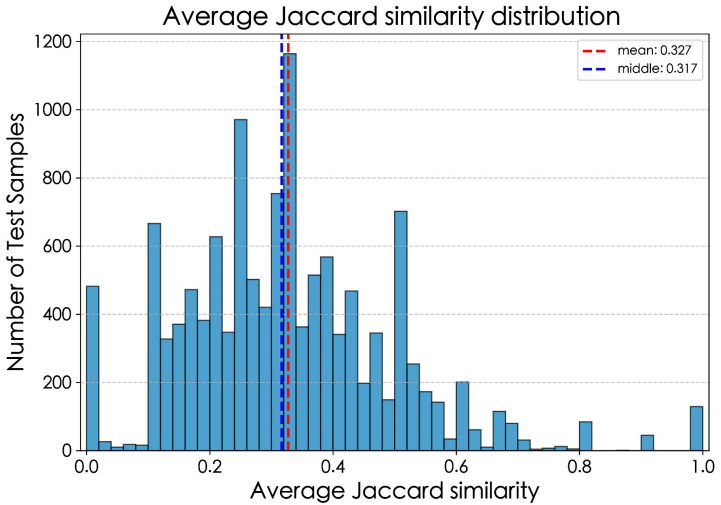
Initial Exploration: Pre -processing Denoising with a LLM

- **Concept:** Our first hypothesis was to repair the data upstream using RAG. We proposed a "Generative Causal Imputation" mechanism. The goal was to leverage the reasoning capabilities of LLMs to detect these logical gaps and fill in the missing items based on semantic necessity rather than simple co-occurrence.
- **Mechanism:** For a noisy target sequence, we utilized a retriever to fetch a reference set of similar, complete sequences from the training corpus. The LLM was then prompted to analyze the target sequence using this reference set to identify and insert the missing interaction that best restores causal continuity.

Failure Analysis: The Retrieval Bottleneck

Quantitative Insight: The retrieval quality was the bottleneck. A quantitative analysis of the jaccard similarity between target sequences and their retrieved references revealed that **the retrieved reference sequences had very low overlap with the target sequences**.

retrieval relevance gap: : Relying on divergent references caused "contextual hallucination", where the LLM modifies sequences toward irrelevant user intents, compounding the noise rather than fixing it.



Adopted Approach: LLM-based Post-hoc Re-ranking

- **Paradigm Shift:** Instead of repairing the input upstream, the focus shifted to correcting the output downstream.
- **Two-Stage Strategy:**
  - **Base Recommender:** Trained exclusively on clean, complete data to learn optimal sequential patterns. It generates a high-recall candidate set from the noisy input.
  - **LLM Re-ranker:** The LLM acts as a discriminator. It takes the noisy history and re-ranks without needing external retrieved context, thus avoiding the retrieval bottleneck.
- **Prompt Engineering Guidelines:**

  To perform this re-ranking, the LLM is guided by a prompt enforcing three specific criteria:
  - **Sequential Coherence:** Prioritizing items that logically follow the most recent interactions (bridging logical gaps).
  - **Interest Consistency:** Aligning items with the user's dominant long-term interests.
  - **Specificity Matching:** Favoring items that match granular attributes (e.g., brand, material) seen in history

Experiment
Effectiveness of our advancement

How effective is our distillation module over the sequential recommendation system?

What extent can the proposed re-ranking mechanism improve the performance, even with noisy data or tasks?

LAU Hoi Ying

Experimental Setup

Dataset

Pre-training the embedding model:

1. Video Games (Games)
2. Arts, Crafts and Sewing (Arts)
3. Movies and TV (Movies)
4. Home and Kitchen (Home)
5. Electronics (Electronics)
6. Tools and Home Improvement (Tools)

*From Amazon reviews 2023 datasets and Goodreads

- Applying 5-core filtering
- Capping sequence length at 10

Evaluation

1. In-domain: Games, Arts, Movies
2. Out-of-domain: Sports, Baby from Amazon, Goodreads from a separate book review platform
    a. Novel item types and different data distribution

- Allow assessment of both performance and cross-domain generalization

Experimental Setup

Evaluation Metrics

Full ranking evaluation

- Recall@10
- Recall@20
- NDCG@10
- NGCG@20

Recall → Whether the ground-truth next item appears in the top-K predictions

NDCG → Ranking quality with position-aware weighting

All results are averaged over three random seeds to ensure statistical reliability and reduce variance due to model initialization or sampling

Experimental Setup

Baselines

1. General-purpose encoders

a. BERT (Bidirectional Encoder Representations from Transformers)

b. BGE (BAAI General Embedding)

c. GTE (General Text Embeddings)

d. LLM2Vec (Large Language Model to Vector)

2. Recommendation-specific encoders

a. BLaiR (Bridging Language and Items for Retrieval and Recommendation)

b. EasyRec

c. LLMEmb (Large Language Model Embedding)

*For fair comparison: All LLM-based methods employ a backbone no smaller than Qwen2-0.5B

Main Results

		In-Domain Datasets											
		Games				Arts				Movies			
	Models	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20
SASRec	BERT	0.0585	0.0311	0.0863	0.0381	0.0650	0.0405	0.0869	0.0460	0.0447	0.0240	0.0646	0.0290
	GTE	0.0641	0.0349	0.0911	0.0418	0.0644	0.0394	0.0880	0.0454	0.0570	0.0300	0.0817	0.0363
	BGE	0.0733	0.0410	0.1022	0.0483	0.0748	0.0475	0.1006	0.0540	0.0626	0.0350	0.0847	0.0406
	LLM2Vec	0.0740	0.0407	0.1029	0.0480	0.0770	0.0506	0.1007	0.0566	0.0662	0.0384	**0.0874**	0.0438
	BLAIR	0.0654	0.0361	0.0954	0.0437	0.0648	0.0379	0.0906	0.0444	0.0581	0.0315	0.0801	0.0370
	EasyRec	0.0647	0.0357	0.0926	0.0428	0.0658	0.0395	0.0929	0.0463	0.0528	0.0278	0.0739	0.0331
	LLMEmb	0.0813	0.0487	**0.1085**	0.0555	0.0865	0.0601	**0.1086**	0.0657	0.0659	0.0390	0.0837	0.0435
	LLM2Rec*	0.0779	0.0477	0.1062	0.0548	0.0882	0.0593	0.1102	0.0648	0.0651	0.0394	0.0814	0.0435
	Ours(*w/ Qwen2-0.5B*)	0.0844	0.0570	0.1048	0.0621	**0.0896**	0.0654	**0.1086**	0.0702	0.0677	0.0469	0.0807	0.0502
	Ours(*w/ Smollm2-135M*)	**0.0860**	**0.0597**	0.1059	**0.0647**	0.0878	**0.0658**	0.1056	**0.0704**	**0.0684**	**0.0490**	0.0799	**0.0519**

		Out-Of-Domain Datasets											
		Sports				Baby				Goodreads			
	Models	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20
SASRec	BERT	0.0860	0.0649	0.1017	0.0689	0.0114	0.0050	0.0232	0.0080	0.1479	0.0858	0.1929	0.0972
	GTE	0.0823	0.0584	0.1001	0.0629	0.0242	0.0142	0.0387	0.0173	0.1488	0.0851	0.1944	0.0957
	BGE	0.0974	0.0738	0.1141	0.0778	0.0428	0.0250	0.0569	0.0286	0.1445	0.0813	0.1972	0.0945
	LLM2Vec	0.1079	0.0854	0.1234	0.0893	0.0561	0.0359	0.0732	0.0414	0.1430	0.0790	0.1906	0.0911
	BLAIR	0.0934	0.0614	0.1091	0.0664	0.0332	0.0194	0.0484	0.0218	**0.1518**	**0.0860**	**0.2000**	**0.0984**
	EasyRec	0.0887	0.0627	0.1061	0.0671	0.0271	0.0154	0.0381	0.0182	0.1445	0.0825	0.1908	0.0941
	LLMEmb	0.1131	0.0936	**0.1257**	0.0969	0.0659	0.0439	0.0807	0.0476	0.1374	0.0778	0.1838	0.0895
	LLM2Rec*	0.1118	0.0904	0.1256	0.0939	**0.0698**	0.0478	**0.0832**	0.0512	0.1468	0.0859	0.1946	0.0980
	Ours(*w/ Qwen2-0.5B*)	**0.1202**	0.0959	0.1220	0.0989	0.0638	0.0512	0.0799	0.0552	0.1267	0.0844	0.1496	0.0902
	Ours(*w/ Smollm2-135M*)	0.1189	**0.0965**	0.1191	**0.0990**	0.0650	**0.0520**	0.0793	**0.0556**	0.1141	0.0763	0.1303	0.0803

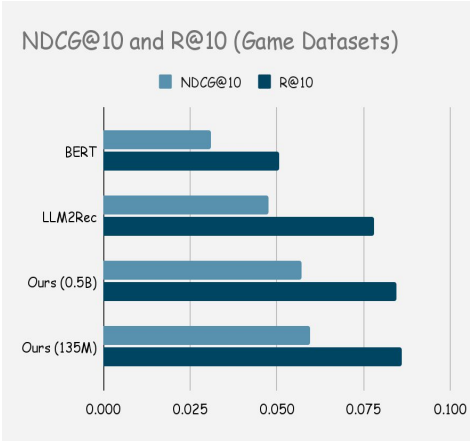


Table 1. Performance comparison of different embedding methods under in-domain and out-of-domain datasets. R is shorts for Recall, N is short for NDCG. Results marked with ∗ indicate values reproduced by our implementation.

Ablation Study: Effectiveness of the Distillation Module

- Distillation consistently improves performance across all datasets.
- Large improvements on in-domain datasets.
- Strong generalization gains on out-of-domain datasets.

Dataset	Model	R@10	N@10	R@20	N@20
In-domain Datasets					
Games	Base model	0.0600	0.0350	0.0822	0.0406
	Distillated model	0.0856	0.0524	0.1122	0.0592
Arts	Base model	0.0738	0.0490	0.0948	0.0543
	Distillated model	0.0868	0.0622	0.1081	0.0676
Movies	Base model	0.0447	0.0281	0.0567	0.0311
	Distillated model	0.0679	0.0442	0.0824	0.0478
Out-of-domain Datasets					
Sports	Base model	0.1020	0.0813	0.1158	0.0846
	Distillated model	0.1118	0.0904	0.1256	0.0939
Baby	Base model	0.0608	0.0431	0.0713	0.0458
	Distillated model	0.0698	0.0478	0.0832	0.0512
Goodreads	Base model	0.1443	0.0869	0.1885	0.0966
	Distillated model	0.1468	0.890	0.1946	0.0980

Table 2. Performance comparison between base models (SmolLM2-135M-Instruct) and distillated models on both in-domain and out-of-domain datasets. Metrics reported include Recall@10/20 (R@10, R@20) and NDCG@10/20 (N@10, N@20). This table highlights the effect of distillation on model performance.

Ablation Study: Improvement in Robustness from the Re-ranking Mechanism

- Re-ranking significantly improves robustness under noisy user histories.
- Works for both large (Qwen2-0.5B) and distilled (Smollm2-135M) models

Dataset	noise	re-rank	R@10	N@10	R@20	N@20
In-domain Datasets						
Games	✗	✗	0.0779	0.0477	0.1062	0.0548
	✗	✓	0.0844	0.0570	0.1048	0.0621
	✓	✗	0.0687	0.0412	0.0939	0.0475
	✓	✓	0.0731	0.0490	0.0925	0.0539
Arts	✗	✗	0.0882	0.0593	0.1102	0.0648
	✗	✓	0.0896	0.0654	0.1086	0.0702
	✓	✗	0.0761	0.0505	0.0982	0.0560
	✓	✓	0.0763	0.0560	0.0944	0.0606
Movies	✗	✗	0.0651	0.0394	0.0814	0.0435
	✗	✓	0.0677	0.0469	0.0807	0.0502
	✓	✗	0.0591	0.0350	0.0753	0.0391
	✓	✓	0.0605	0.0417	0.0731	0.0449
Out-of-domain Datasets						
Sports	✗	✗	0.1118	0.0904	0.1256	0.0939
	✗	✓	0.1202	0.0959	0.1220	0.0989
	✓	✗	0.0953	0.0761	0.1085	0.0794
	✓	✓	0.0940	0.0803	0.1047	0.0830
Baby	✗	✗	0.0698	0.0478	0.0832	0.0512
	✗	✓	0.0638	0.0512	0.0799	0.0552
	✓	✗	0.0611	0.0417	0.0759	0.0454
	✓	✓	0.0571	0.0444	0.0731	0.0484
Goodreads	✗	✗	0.1468	0.0859	0.1946	0.0980
	✗	✓	0.1267	0.0844	0.1496	0.0902
	✓	✗	0.1424	0.0807	0.1882	0.0923
	✓	✓	0.1211	0.0808	0.1444	0.0867

Qwen2-0.5B

Dataset	noise	re-rank	R@10	N@10	R@20	N@20
In-domain Datasets						
Games	✗	✗	0.0856	0.0524	0.1122	0.0592
	✗	✓	0.0860	0.0597	0.1059	0.0647
	✓	✗	0.0750	0.0449	0.0991	0.0510
	✓	✓	0.0754	0.0508	0.0953	0.0559
Arts	✗	✗	0.0868	0.0622	0.1081	0.0676
	✗	✓	0.0878	0.0658	0.1056	0.0704
	✓	✗	0.0747	0.0529	0.0934	0.0576
	✓	✓	0.0751	0.0564	0.0907	0.0603
Movies	✗	✗	0.0679	0.0442	0.0824	0.0478
	✗	✓	0.0684	0.0490	0.0799	0.0519
	✓	✗	0.0603	0.0392	0.0727	0.0424
	✓	✓	0.0614	0.0428	0.0715	0.0454
Out-of-domain Datasets						
Sports	✗	✗	0.1115	0.0948	0.1235	0.0978
	✗	✓	0.1189	0.0965	0.1191	0.099
	✓	✗	0.0945	0.0789	0.1067	0.0819
	✓	✓	0.0918	0.0804	0.1018	0.0829
Baby	✗	✗	0.0699	0.0500	0.0821	0.0531
	✗	✓	0.0650	0.0520	0.0793	0.0556
	✓	✗	0.0627	0.0439	0.0743	0.0468
	✓	✓	0.0570	0.0447	0.0722	0.0485
Goodreads	✗	✗	0.1424	0.0836	0.1868	0.0948
	✗	✓	0.1141	0.0763	0.1303	0.0803
	✓	✗	0.1422	0.0810	0.1814	0.0909
	✓	✓	0.1129	0.0756	0.1271	0.0792

Smollm2-135M

Conclusion & Discussion

JIN, Mengxi

Summary & Core Contributions

Addressing Key Challenges of LLM2Rec

Efficiency:

- Problem: High Computation Cost
- Solution: Knowledge Distillation Module
- Key Result:
  - Lightweight model (Smollm2-135M) achieves performance of the larger teacher model (Qwen2-0.5B).
  - Enables efficient deployment without sacrificing quality.

Robustness:

- Problem: Sensitivity to Noisy Data
- Solution: LLM-based Post-hoc Re-ranking
- Key Results:
  - Universal resilience in noisy environments.
  - Proves LLM is a practical semantic denoising judge at the final stage.

JIN, Mengxi

Limitation & Future Work

Room for Improvement and Future direction

Future Direction 1: Multimodal Information Fusion

- Incorporate visual inputs (e.g., product images) alongside text.
- Goal: Capture richer item semantics for a more comprehensive understanding.

Future Direction 2: Advanced Noise Scenarios

- Move beyond single-item deletion and Explore more complex noise types.
- Goal: Further enhance model robustness and generalization for real-world chaos.

Thank you

Q & A