KDD 2024

Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System

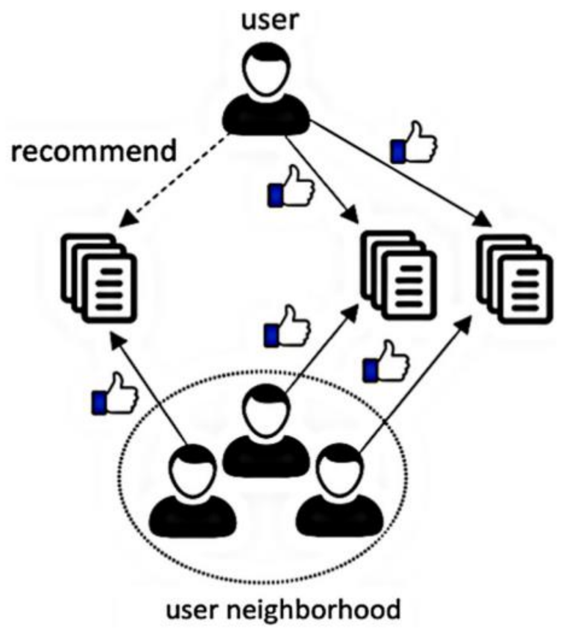
Sein Kim*, Hongseok Kang*, Seungyoon Choi
Donghyun Kim, Minchul Yang, Chanyoung Park†
Korean Advanced Institute of Science and Technology (KAIST)
NAVER Corporation

*: these authors contributed equally to this work
†: corresponding author

CF-RecSys

Collaborative Filtering based Recommender Systems (CF-RecSys)

- CF-RecSys recommends personalized items by **using user history and data from similar users**
- Since collaborative knowledge is based on user history (i.e., user-item interaction history), **CF-RecSys struggles to generate collaborative knowledge and proper recommendations** when there are **limited interactions** (e.g., **Cold-start problem, Cross-domain problem**)



Collaborative filtering [1]

[1] Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022). https://doi.org/10.1186/s40537-022-00592-5

Solutions for Limited Interactions

Multi-modality in recommendation tasks

- Various **textual information** such as item titles and descriptions **exist in recommendation data**
- Using this **text modality**, **address various problems** that arise from the **sparsity of user-item interactions**
- In particular, the text modality is primarily used to solve the **cold item** problem and the **cross-domain** problem

Solutions for Limited Interactions

Multi-modality in recommendation tasks

- Various **textual information** such as item titles and descriptions **exist in recommendation data**
- Using this **text modality**, **address various problems** that arise from the **sparsity of user-item interactions**
- In particular, the text modality is primarily used to solve the **cold item** problem and the **cross-domain** problem

Recommendations by LLM

- Provide the order of **item titles**, as an **input prompt**

- **Completion:** $item_{+}^{t,1}$, $item_{+}^{t,2}$, $item_{+}^{t,3}$, $item_{+}^{t,4}$, $item_{+}^{t,5}$, $\langle item_{*}^{t} \rangle$
- **Zero-shot:** I like the following movies: $item_{+}^{t,1}$, $item_{+}^{t,2}$, $item_{+}^{t,3}$, $item_{+}^{t,4}$, $item_{+}^{t,5}$. Then I would also like $\langle item_{*}^{t} \rangle$
- **Few-shot (k):** Repeat $r \in \{1, \ldots, k\}$ $\begin{cases} \text{User Movie Preferences: } item_{+}^{r,1}, item_{+}^{r,2}, item_{+}^{r,3}, item_{+}^{r,4} \\ \text{Additional User Movie Preference: } item_{+}^{r,5} \end{cases}$

  User Movie Preferences: $item_{+}^{t,1}$, $item_{+}^{t,2}$, $item_{+}^{t,3}$, $item_{+}^{t,4}$, $item_{+}^{t,5}$

  Additional User Movie Preference: $\langle item_{*}^{t} \rangle$

E.g., LLMs' input prompt [2]

Instruction Input	
Task Instruction:	Given the user's historical interactions, please determine whether the user will enjoy the target new movie by answering "Yes" or "No".
Task Input:	User's liked items: GodFather. User's disliked items: Star Wars. Target new movie: Iron Man
Instruction Output	
Task Output:	No.

TALLRec [3]

[2] Sanner, Scott, et al. "Large language models are competitive near cold-start recommenders for language-and item-based preferences." Proceedings of the 17th ACM conference on recommender systems. 2023.
[3] Bao, Keqin, et al. "Tallrec: An effective and efficient tuning framework to align large language model with recommendation." Proceedings of the 17th ACM Conference on Recommender Systems. 2023.

Motivations Knowledge gap: Collaborative/textual knowledge

Lack of collaborative knowledge

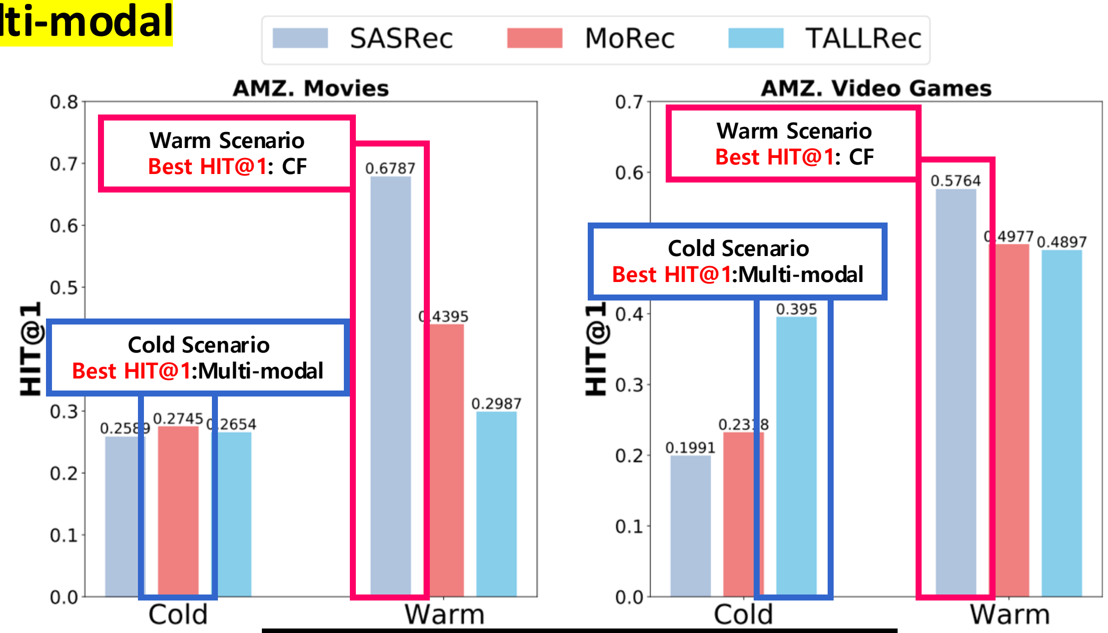
- Using only item titles and descriptions **overlooks user-item interactions** (collaborative signals/knowledge)
- Rich textual knowledge can hinder capturing collaborative knowledge [4]

Comparison between collaborative knowledge and textual knowledge (multi-modal data)

1) If **abundant** user-item interactions are **available**; **CF > Multi-modal** **(CF-Model: Collaborative filtering Model)**

2) If user-item interactions are **lacking**; **CF < Multi-modal**

→ **Knowledge gap** between collaborative/textual knowledge



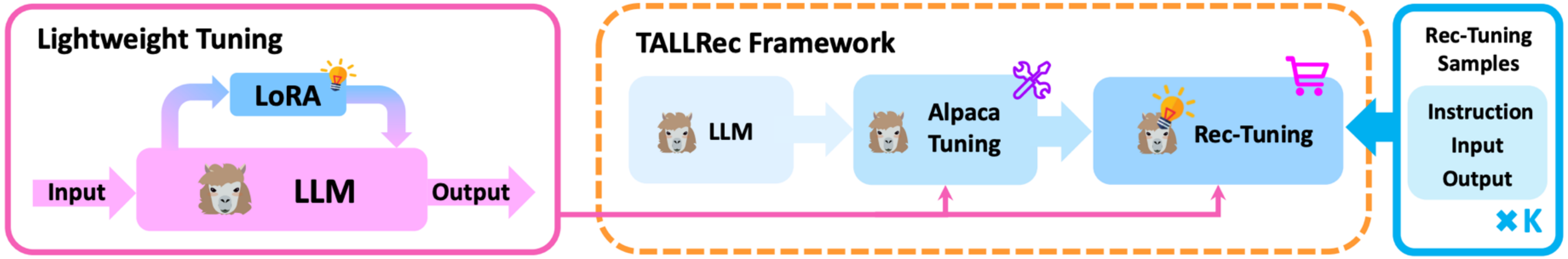
CF: SASRec
Multi-modal: MoRec/TALLRec

[4] Yuan, Zheng, et al. "Where to go next for recommender systems? id-vs. modality-based recommender models revisited." *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023

Motivations Inefficiency

Inefficiency of previous LLM based Recommenders

- In the case of TALLRec, **LoRA is used to tune the LLM** for recommendation tasks
- Many LLM recommendation papers, currently under review or published on arXiv, utilize LoRA for tuning LLMs



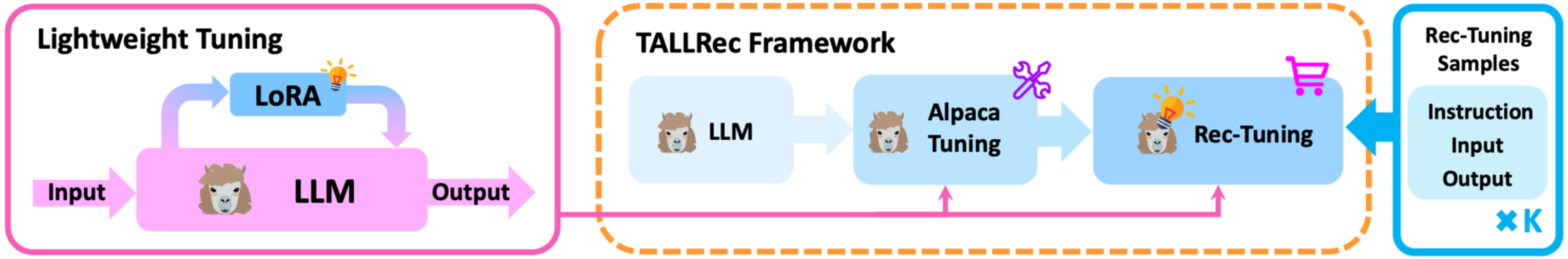
TALLRec [3]

[3] Bao, Keqin, et al. "Tallrec: An effective and efficient tuning framework to align large language model with recommendation." Proceedings of the 17th ACM Conference on Recommender Systems. 2023.

Motivations Inefficiency

Inefficiency of previous LLM based Recommenders

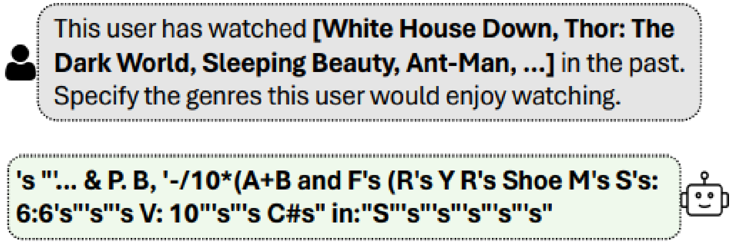
- In the case of TALLRec, **LoRA is used to tune the LLM** for recommendation tasks
- Many LLM recommendation papers, currently under review or published on arXiv, utilize LoRA for tuning LLMs



TALLRec [3]

Tuning with LoRA tends to be time-consuming and often results in over-fitting to the instruction prompt

- **Our model (A-LLMRec) is <u>2.53 times faster</u> in training and <u>1.71 times faster</u> in inference** than TALLRec



(c) TALLRec

Example of overfitting to the instruction prompt (TALLRec)

[3] Bao, Keqin, et al. "Tallrec: An effective and efficient tuning framework to align large language model with recommendation." Proceedings of the 17th ACM Conference on Recommender Systems. 2023.

Motivations

Proposing efficient/effective learning by training recommendation tasks without fine-tuning the LLM

- **Inject collaborative knowledge (pretrained CF-RecSys)** into LLMs to **address LLMs lower performance compared to CF-models** in scenarios with abundant user-item interactions
- **Faster training and inference time** w/o LoRA

Solving various tasks through the alignment of collaborative and textual knowledge:

- Cold/Warm item scenarios
- Cold user scenarios
- Few-shot scenarios
- Cross-domain scenarios
- (Model agnostic)
- (Text generation)

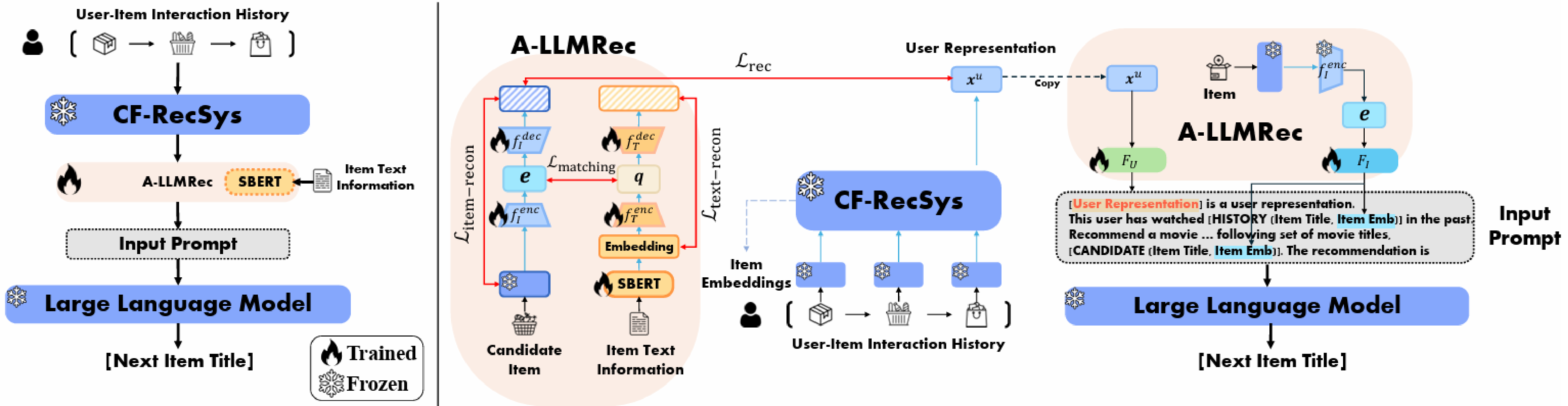
Proposed efficient and effective **A**ll-round **LLM**-based **Rec**ommender System (**A-LLMRec**)

Methods: Framework Overview (A-LLMRec)

Training Stage1: Align – Collaborative & Textual Knowledge

Training Stage2: Recommendation Tasks

LLM: OPT-6.7b



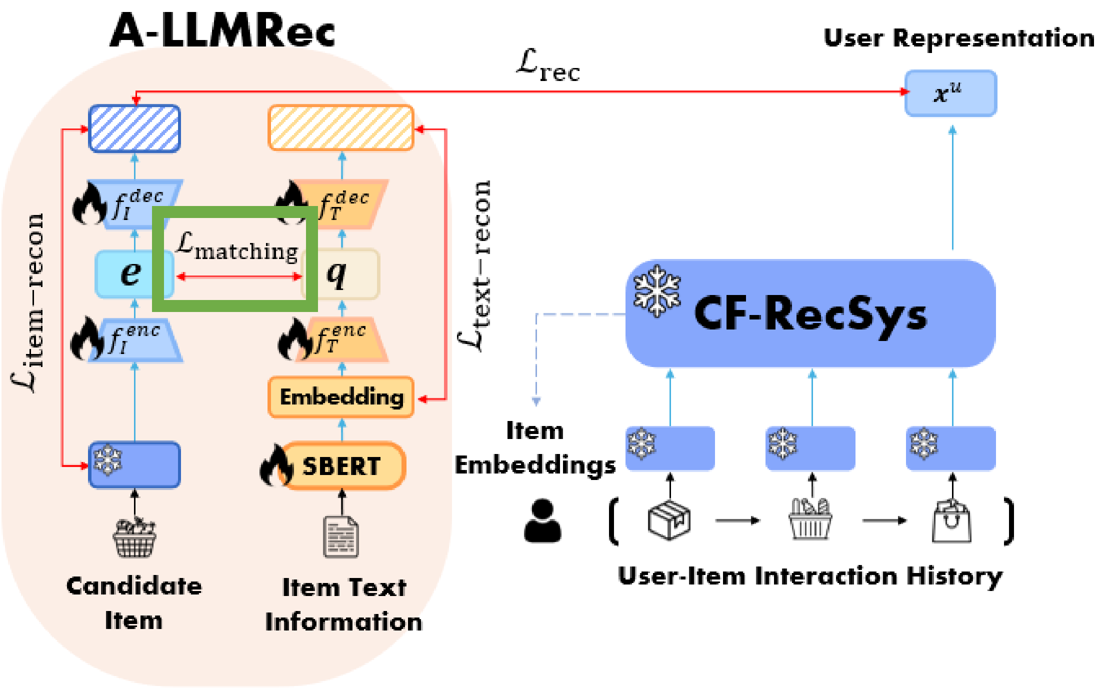
(a) Framework Overview (b) Stage 1 (c) Stage 2

Methods: Training stage 1 (A-LLMRec)

Matching Loss (auto-encoder based)

- **Matching item embeddings** derived from a pretrained collaborative model with **text embeddings** of the same items in **latent space**

$$\mathcal{L}_{\text{matching}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{e}_i, \mathbf{q}_i) \right] \right]$$
$$= \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(f_I^{enc}(\mathbf{E}_i), f_T^{enc}(\mathbf{Q}_i)) \right] \right]$$



(b) Stage 1

Methods: Training stage 1 (A-LLMRec)

Matching Loss (auto-encoder based)

- **Matching item embeddings** derived from a pretrained collaborative model with **text embeddings** of the same items in **latent space**

$$\mathcal{L}_{\text{matching}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} [MSE(\mathbf{e}_i, \mathbf{q}_i)] \right]$$

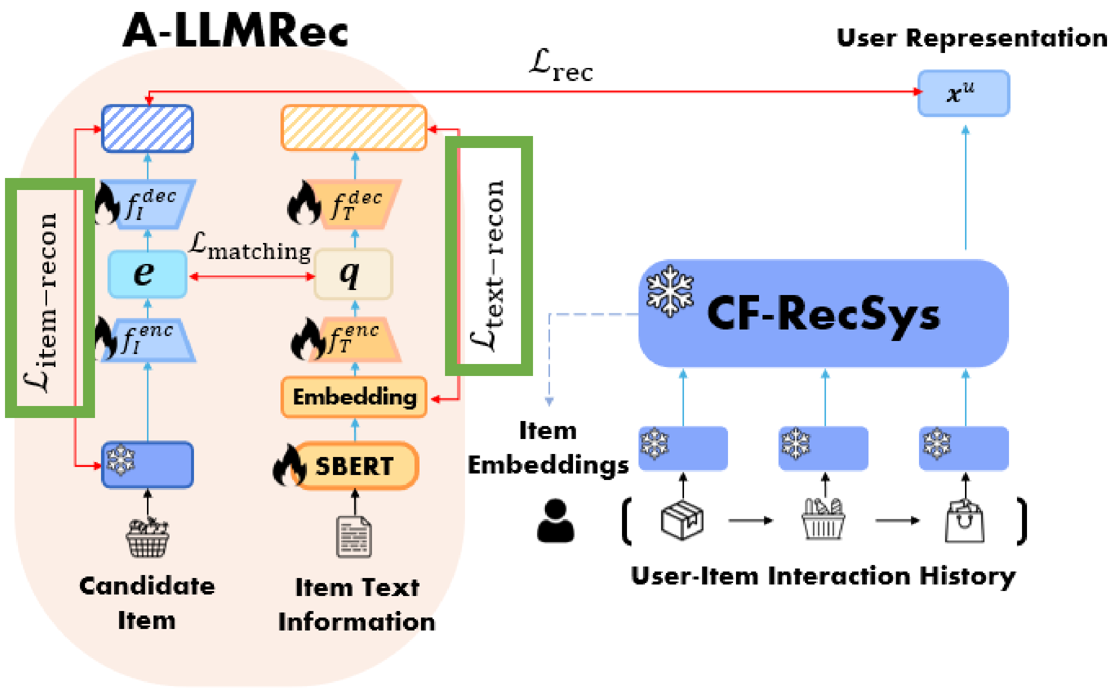
$$= \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} [MSE(f_I^{enc}(\mathbf{E}_i), f_T^{enc}(\mathbf{Q}_i))] \right]$$

Reconstruction Loss

- **Prevent the collapse** of item embeddings and text embeddings

$$\mathcal{L}_{\text{item-recon}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{E}_i, f_I^{dec}(f_I^{enc}((\mathbf{E}_i)))\right] \right]$$

$$\mathcal{L}_{\text{text-recon}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{Q}_i, f_T^{dec}(f_T^{enc}((\mathbf{Q}_i)))\right] \right]$$



(b) Stage 1

Methods: Training stage 1 (A-LLMRec)

Matching Loss (auto-encoder based)

- **Matching item embeddings** derived from a pretrained collaborative model with **text embeddings** of the same items in **latent space**

$$\mathcal{L}_{\text{matching}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{e}_i, \mathbf{q}_i) \right] \right] = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(f_I^{enc}(\mathbf{E}_i), f_T^{enc}(\mathbf{Q}_i)) \right] \right]$$

Reconstruction Loss

- **Prevent the collapse** of item embeddings and text embeddings

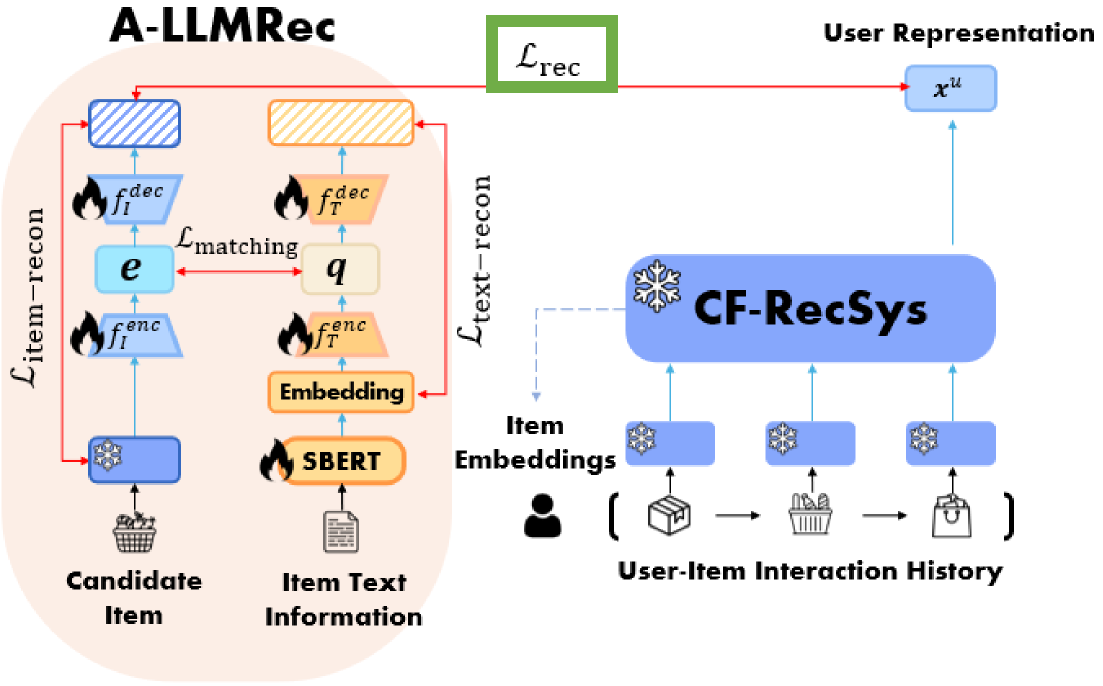
$$\mathcal{L}_{\text{item-recon}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{E}_i, f_I^{dec}(f_I^{enc}((\mathbf{E}_i)))) \right] \right]$$

$$\mathcal{L}_{\text{text-recon}} = \underset{\mathcal{S}^u \in \mathcal{S}}{\mathbb{E}} \left[\underset{i \in \mathcal{S}^u}{\mathbb{E}} \left[MSE(\mathbf{Q}_i, f_T^{dec}(f_T^{enc}((\mathbf{Q}_i)))) \right] \right]$$

Recommendation Loss

- **Inject explicit collaborative knowledge** in aligning procedure

$$\mathcal{L}_{\text{rec}} = -\sum_{\mathcal{S}^u \in \mathcal{S}} \Big[log(\sigma(s(\mathbf{x}^u_{|\mathcal{S}^u|-1}, f_I^{dec}(f_I^{enc}(\mathbf{E}_{i^u_{|\mathcal{S}^u|}}))))) + log(1 - \sigma(s(\mathbf{x}^u_{|\mathcal{S}^u|-1}, f_I^{dec}(f_I^{enc}(\mathbf{E}_{i^{u,-}_{|\mathcal{S}^u|}}))))) \Big]$$



(b) Stage 1

Methods: Training stage 2 (A-LLMRec)

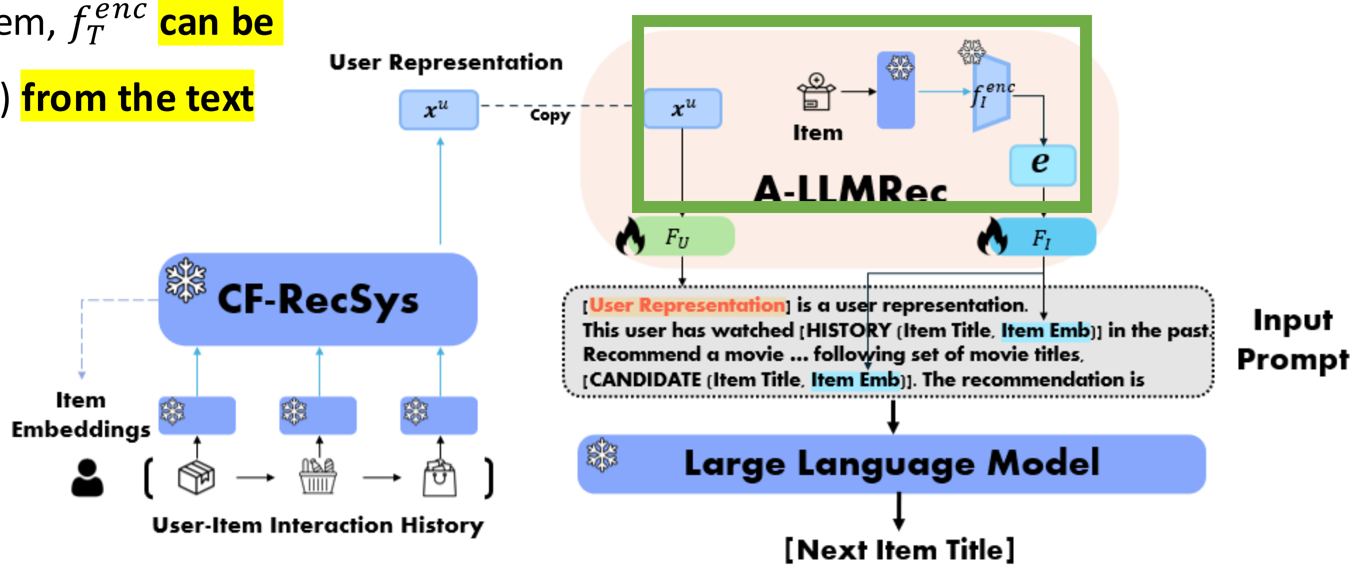
Using the prompt tuning approach, the LLM is trained for the recommendation task

- All parameters are fixed, only trains the projection layer

➢ The user embeddings derived from the Recommender (CF-RecSys) and item embeddings aligned with text (**joint collaborative-text embedding, denoted (*e*)**) are **used as soft prompts**

➢ If there is **insufficient collaborative information** about an item, f_T^{enc} **can be used to generate the joint collaborative-text embedding (*e*) from the text**

LLM Input:	[User Representation] is a user representation. This user has watched **[HISTORY (Item Titles, Item Emb)]** in the past. Recommend a movie for this user to watch next from the following set of movie titles, **[CANDIDATE (Item Titles, Item Emb)].** The recommendation is
LLM Output:	[Next Item Title]



(c) Stage 2

Experiments

Datasets

Datasets	#Users	#Items	#Interactions.	Avg. Len
Movies and TV	297,498	59,944	3,409,147	11.46
Video Games	64,073	33,614	598,509	8.88
Beauty	9,930	6,141	63,953	6.44
Toys	30,831	61,081	282,213	9.15

Baselines

- **Collaborative filtering - based baselines**
  - NCF
  - NextItNet
  - GRU4Rec
  - SASRec
- **Modality aware - based baselines**
  - MoRec (initialized the item embedding using BERT)
  - CTRL (initialized the backbone model using contrastive learning on textual information, then fine-tune to recommendation task)
  - RECFORMER (Transform the sequential recommendation into a task of predicting the next item as if predicting the next sentence)
- **LLM- based baselines**
  - LLM-Only (Use vanilla OPT-6.7b)
  - TALLRec
  - MLP-LLM (Ablation of A-LLMRec; Remove item-text matching)

Experiments: Main Results (Hit@1)

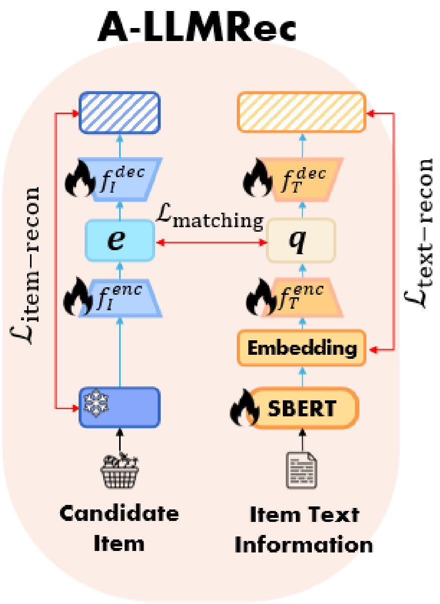
	Collaborative filtering				Modality-aware			LLM-based			
	NCF	NextItNet	GRU4Rec	SASRec	MoRec	CTRL	RECFORMER	LLM-Only	TALLRec	MLP-LLM	A-LLMRec
Movies and TV	0.4273	0.5855	0.5215	0.6154	0.4130	0.3467	0.4865	0.0121	0.2345	0.5838	**0.6237**
Video Games	0.3159	0.4305	0.4026	**0.5402**	0.4894	0.2354	0.4925	0.0168	0.4403	0.4788	0.5282
Beauty	0.2957	0.4231	0.4131	0.5298	0.4997	0.3963	0.4878	0.0120	0.5542	0.5548	**0.5809**
Toys	0.1849	0.1415	0.1673	0.2359	0.1728	0.1344	0.2871	0.0141	0.0710	0.3225	**0.3336**

<u>Observations</u>

- **Modality-aware models** typically **show lower performance** than collaborative models in general settings
- Proposed **A-LLMRec** can **outperform collaborative models** and modality-aware models in general setting

Experiments: Cold/Warm item Scenarios (Hit@1)

	Movies and TV		Video Games		Beauty	
	Cold	Warm	Cold	Warm	Cold	Warm
SASRec	0.2589	0.6787	0.1991	0.5764	0.1190	0.6312
MoRec	0.2745	0.4395	0.2318	0.4977	0.2145	0.5425
CTRL	0.1517	0.3840	0.2074	0.2513	0.1855	0.4711
RECFORMER	0.3796	0.5449	0.3039	0.5377	0.3387	0.5133
TALLRec	0.2654	0.2987	0.3950	0.4897	0.5462	0.6124
A-LLMRec	0.5714	**0.6880**	0.4263	**0.5970**	**0.5605**	**0.6414**
A-LLMRec (SBERT)	**0.5772**	0.6802	**0.4359**	0.5792	0.5591	0.6405



Cold (item): items with interaction counts in the bottom 35%
Warm (item): items with interaction counts in the top 35%

<u>**Observations**</u>

- **Modality and LLM-based** are quite satisfactory in cold scenarios, **they degrade in warm scenarios**
- The model proposed in this study **(A-LLMRec) outperforms** in both cold and warm scenarios
- Additionally, **A-LLMRec (SBERT)** (which generates embeddings from f_T^{enc}) shows **meaningful results for cold items**

Experiments: Cold user & few-shot Scenarios (Hit@1)

	Movies and TV	Video Games	Beauty
SASRec	0.2589	0.4048	0.4459
MoRec	0.3918	0.3572	0.4815
CTRL	0.2273	0.1737	0.3902
RECFORMER	0.4481	0.3989	0.4644
TALLRec	0.2143	0.3895	0.5202
MLP-LLM	0.4909	0.3960	0.5276
A-LLMRec	**0.5272**	**0.4160**	**0.5337**

Cold user scenario

	K	SASRec	MoRec	TALLRec	A-LLMRec	A-LLMRec (SBERT)
Movies and TV	256	0.2111	0.2208	0.1846	0.2880	**0.2963**
	128	0.1537	0.1677	0.1654	0.2518	**0.2722**
Video Games	256	0.1396	0.1420	0.2321	0.2495	**0.2607**
	128	0.1089	0.1157	0.1154	0.1608	**0.1839**
Beauty	256	0.2243	0.2937	0.3127	0.3467	**0.3605**
	128	0.1813	0.2554	0.2762	0.3099	**0.3486**

Few-shot scenario

Cold user scenario: evaluations under the cold user (interaction sequence = 3. Cold users are not used in training)
Few-shot scenario: Train the models using only K users

Observations

- **[Cold user] A-LLMRec outperforms** other models in the **cold user scenario**, while SASRec struggles to perform well, especially on a large dataset, due to the lack of collaborative knowledge
- **[Few-shot] A-LLMRec outperforms** all baselines **in few-shot scenarios** by combining CF-RecSys's collaborative knowledge with items' textual knowledge, achieving superior performance even with extremely small amount of user data
- **[Few-shot] A-LLMRec (SBERT)** (which generates embeddings from f_T^{enc}) shows **meaningful results for few-shot**

Experiments: Cross-domain & Model-agnostic (Hit@1)

	SASRec	MoRec	RECFORMER	TALLRec	A-LLMRec	A-LLMRec (SBERT)
Movies and TV → Video Games	0.0506	0.0624	0.0847	0.0785	0.0901	**0.1203**

Cross-domain scenario

Train on Movies and TV → Inference on Video Games (w/o fine-tune)

Model	Beauty	Toys
SASRec	0.5298	0.2359
A-LLMRec (SASRec)	**0.5809**	**0.3336**
NextItNet	0.4231	0.1415
A-LLMRec (NextItNet)	0.5642	0.3203
GRU4Rec	0.4131	0.1673
A-LLMRec (GRU4Rec)	0.5542	0.3089
NCF	0.2957	0.1849
A-LLMRec (NCF)	0.5431	0.3263

Model-agnostic

Observations

- [Cross domain] **A-LLMRec outperforms** all the baselines in the **cross-domain scenario**, and **A-LLMRec (SBERT) particularly performs well** → text encoder that becomes useful when collaborative information is lacking
- [Cross domain] **SASRec underperforms** all the baselines, indicating that **using textual knowledge is crucial for the cross-domain scenario** due to the lack of collaborative information
- [Model-agnostic] Transferring **high-quality collaborative knowledge** can **enhance the performance of A-LLMRec**
- [Model-agnostic] Adopting A-LLMRec to any backbone improves the performance of the vanilla model

Experiments: Ablation study (Hit@1)

Ablation	Movies and TV	Beauty	Toys
A-LLMRec	**0.6237**	**0.5809**	**0.3336**
w/o $\mathcal{L}_{\text{matching}}$	0.5838	0.5548	0.3225
w/o $\mathcal{L}_{\text{item-recon}}$&$\mathcal{L}_{\text{text-recon}}$	0.5482	0.5327	0.3204
w/o $\mathcal{L}_{\text{rec}}$	0.6130	0.5523	0.1541
Freeze SBERT	0.6173	0.5565	0.1720

Ablation studies on Stage-1

Row	Ablation	Movies and TV	Video Games	Beauty	Toys
(1)	A-LLMRec	**0.6237**	**0.5282**	**0.5809**	**0.3336**
(2)	A-LLMRec w/o user representation	0.5925	0.5121	0.5547	0.3217
(3)	A-LLMRec w/o joint embedding	0.1224	0.4773	0.5213	0.2831
(4)	A-LLMRec with random joint embedding	0.1200	0.4729	0.5427	0.0776

Ablation studies on Stage-2

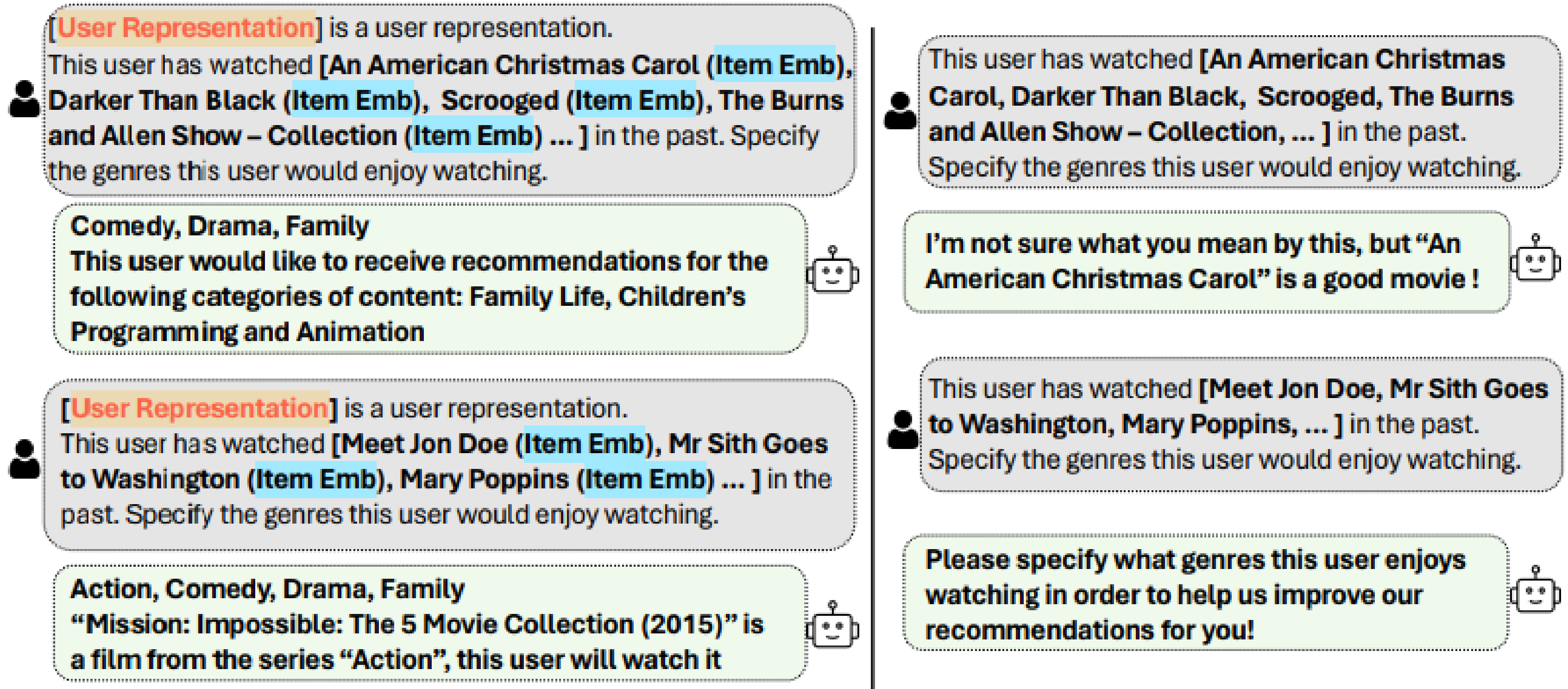
Observations

- [Stage-1] Removing $\mathcal{L}_{matching}$ results in a significant performance decline
  - → Alignment between the item and the text information is effective
  - → **LLM can comprehend item collaborative/textual information** in joint collaborative-text embedding
- [Stage-1] **Removing reconstruction loss** leads risk of **over-smoothed representation** → performance drop
- [Stage-2] Exclusion of the joint embedding results in a substantial decrease
  - → Underscoring **joint embeddings'** significant role in **transferring collaborative knowledge**

Experiments: Language Generation Task

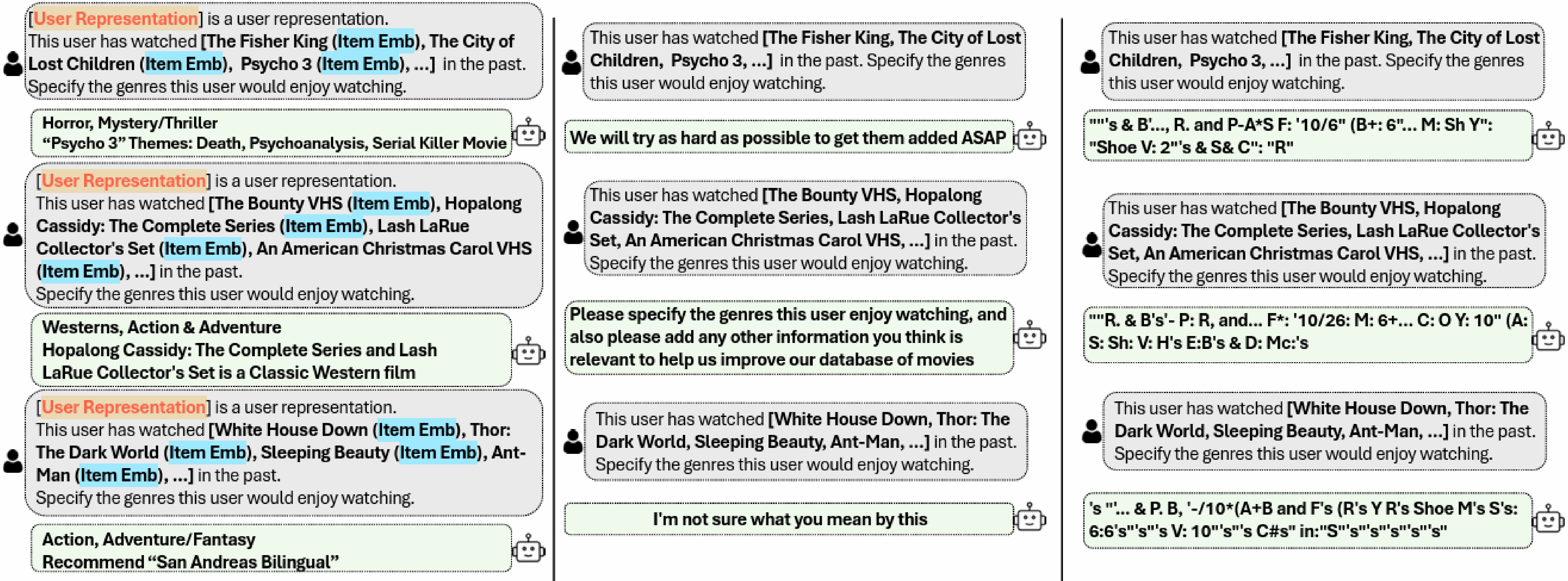
Beyond Recommendation: Language Generation Task (Favorite genre prediction)

- A-LLMRec indeed generates proper answers, while LLM-Only fails to do so
- Item embeddings of the CF-RecSys are well aligned with the token space of the LLM, which enables the LLM to understand and utilize collaborative knowledge



(a) A-LLMRec (b) LLM-Only

Experiments: Language Generation Task



(a) A-LLMRec (b) LLM-Only (c) TALLRec

Thank you!

[KDD’ 24] Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System
[Full Paper] https://arxiv.org/abs/2404.11343
[Source Code] https://github.com/ghdtjr/A-LLMRec

Paper

Code

[Lab Homepage] http://dsail.kaist.ac.kr

[Email] rlatpdlsgns@kaist.ac.kr

DSAIL Data Science & Artificial Intelligence