SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

Collaborative Cross-modal Fusion with Large Language Model for Recommendation

Zhongzhou Liu, Hao Zhang, Kuicai Dong, Yuan Fang

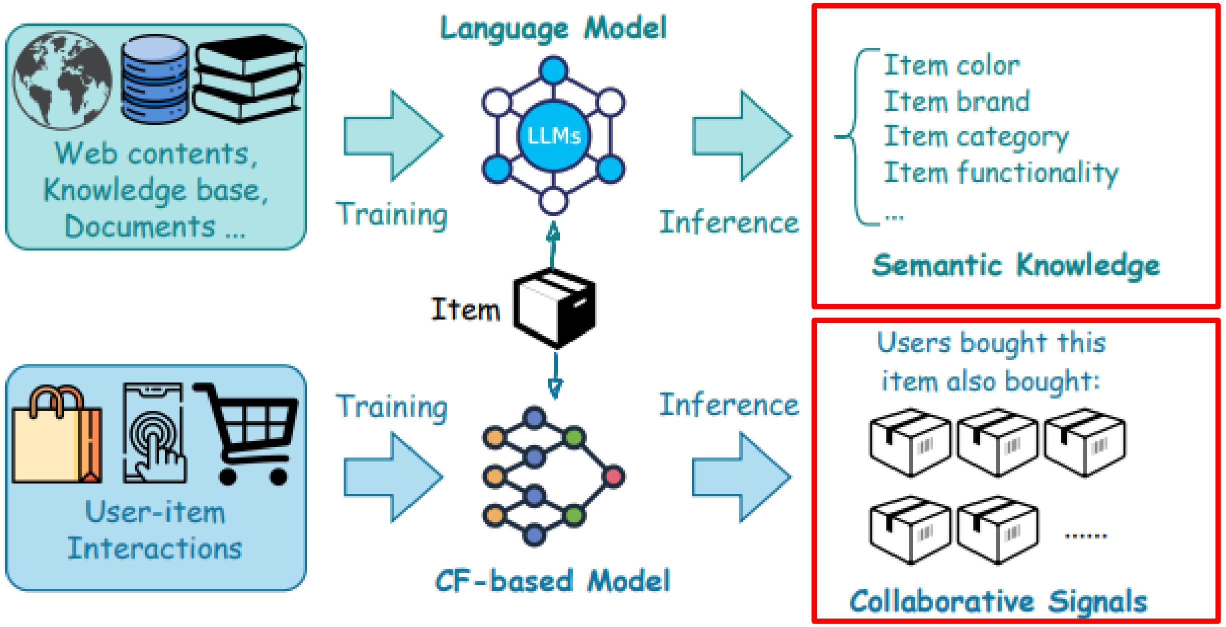
The FANG Lab @ SMU

Functional Algorithms for Networks and Graphs

Motivation

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

- Traditional CF-based model: capture collaborative signal but struggle to process rich semantic knowledge in user/item features.
- LLM: understand semantic knowledge but can not extract collaborative signals simply from textual descriptions.



How to integrate collaborative signals into LLM4Rec?

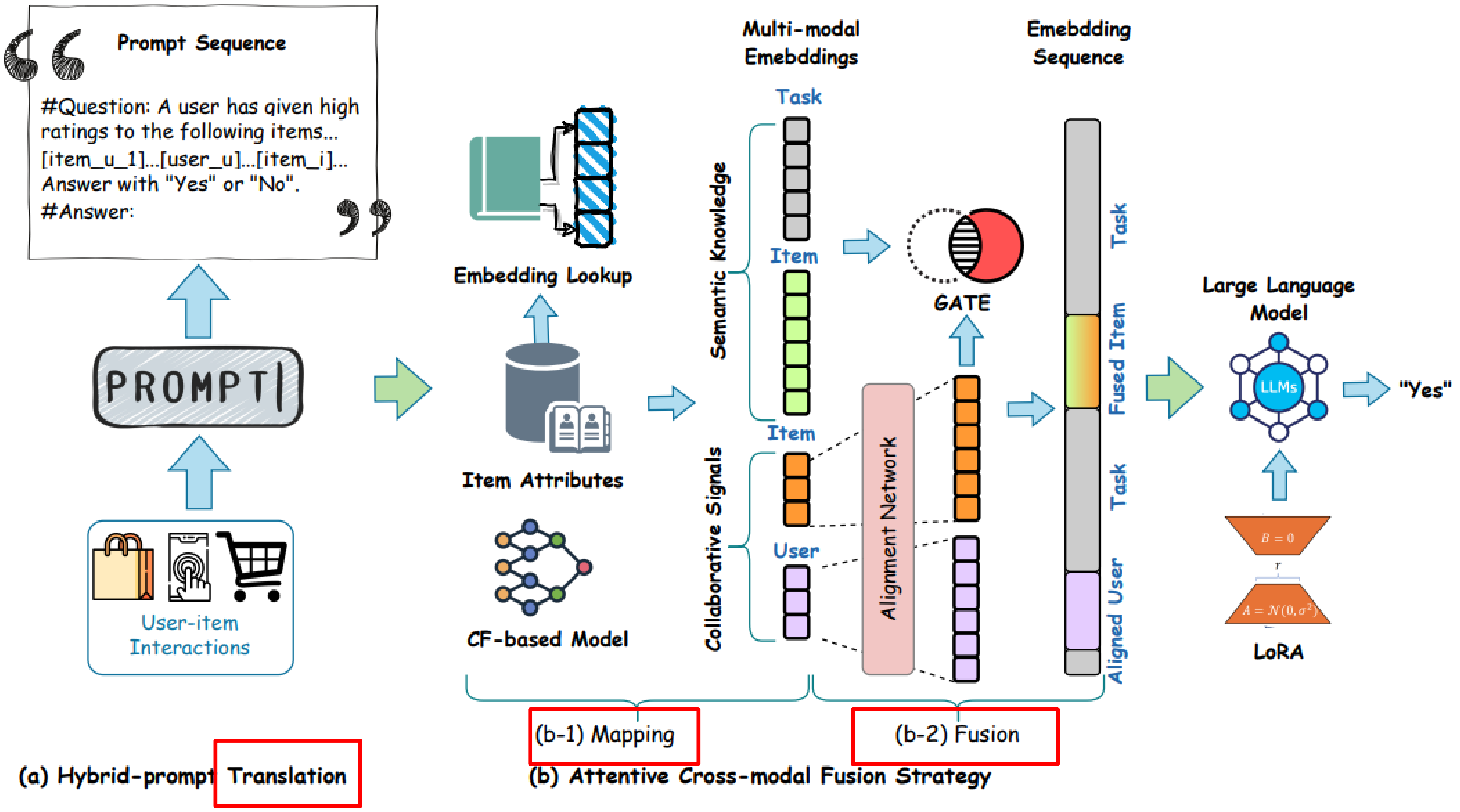
Related works

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

- Collaborative signals in natural language descriptions [Bao et al, 2023]
  - **Idea:** user-item interaction as plain text.
  - E.g., Input: Will Tom like to buy milk? Output: Yes.
  - **Limitation:** representability of plain text vs. high-dimensional non-linear dense vector
- Collaborative signals in embeddings [Zhang et al, 2023]
  - **Idea:** insert collaborative signals into prompts.
  - **E.g.,** Input: will Tom like to buy milk <extra_token_milk> ?
    <extra_token_milk> is a pre-defined special token. Its embedding is initialized from CF model.
  - **Limitation:** heterogenous collaborative signals

Objective: to assist LLMs to encode and fuse collaborative signals and semantic knowledge.

Framework


Prompt Sequence
#Question: A user has given high ratings to the following items... [item_u_1]...[user_u]...[item_i]... Answer with "Yes" or "No".
#Answer:
PROMPT
User-item Interactions
Embedding Lookup
Item Attributes
CF-based Model
Multi-modal Emebddings
Semantic Knowledge
Task
Item
Collaborative Signals
Item
User
GATE
Alignment Network
Emebdding Sequence
Task
Fused Item
Task
Aligned User
Large Language Model
LLMs
"Yes"
LoRA
(b-1) Mapping
(b-2) Fusion
(a) Hybrid-prompt Translation
(b) Attentive Cross-modal Fusion Strategy

Translation

SMU Singapore Management University | School of Computing and Information Systems

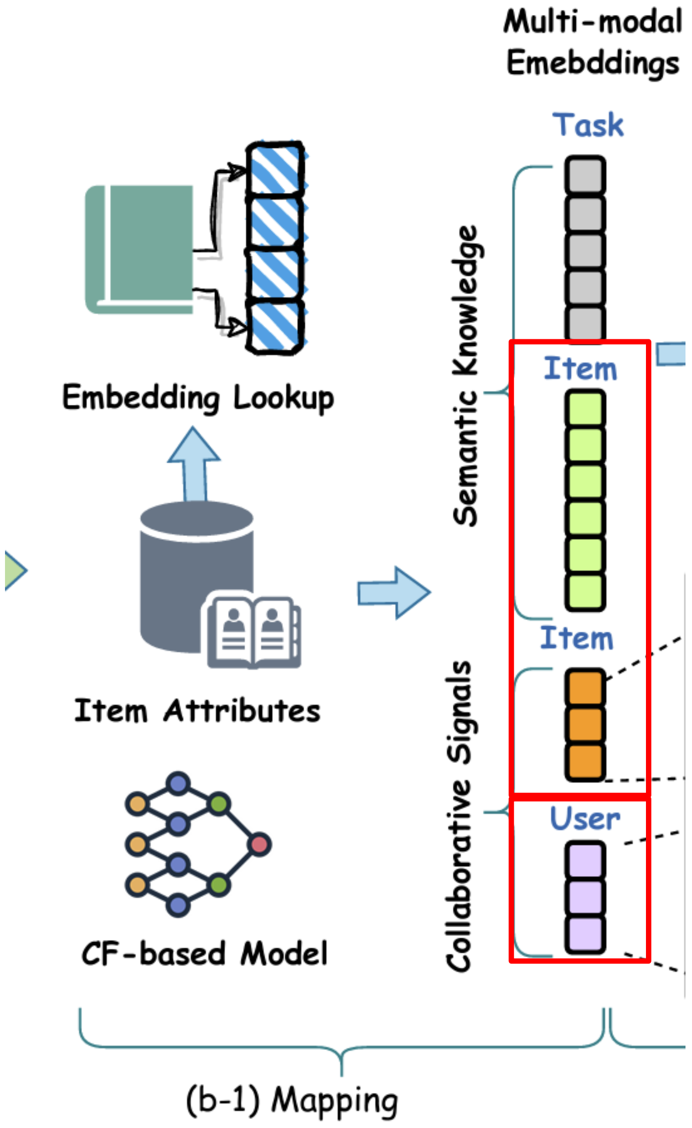
> #Question: A user has given high ratings to the following items: {[Item_u]}. Additionally, we have information about the user's preferences encoded in the feature [User_u]. Using all available information, make a prediction about whether the user would enjoy the item [Item_i]. Answer with "Yes" or "No". #Answer:

- [User_u]: special token as placeholder for user u's feature.
- [Item_i]: special token as placeholder for item i's feature.
- {[Item_u]} = [Item_1], [Item_2],…: the sequential set of special tokens for historical items for user u.

Mapping

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

- Item features
  - $x_i^{CF} \in R^l$: embedding for collaborative signal
  - $x_i^{SM} \in R^{T_i \times d}$: embeddings for semantic knowledge
  - T_i is the number of tokens in item i's textual description.
- User feature
  - Only x_i^{CF}
  - textual descriptions in different datasets vary a lot.



Fusion

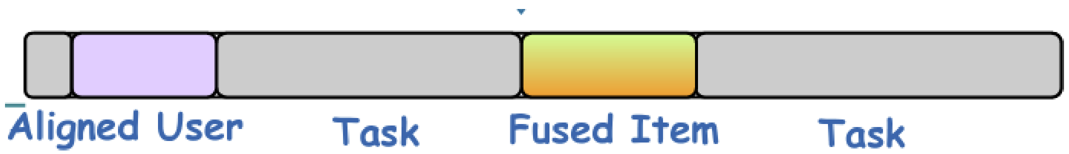
SMU Singapore Management University | School of Computing and Information Systems

- Alignment network ALG
  - Transform all embeddings into an identical dimension.
  - $\mathrm{ALG}: R^{l} \rightarrow R^{d}$
  - aligned user and item embeddings $\tilde{x}_u^{CF}$ and $\tilde{x}_i^{CF}$
- Gate network GATE
  - Fuse the two modality embeddings into one

$$\alpha = \mathrm{GATE}(\tilde{\mathrm{x}}_i^{CF}, \mathrm{x}_i^{SM}; \Theta_{\mathrm{G}}) = \mathrm{MLP}(\tilde{\mathrm{x}}_i^{CF}; \Theta_{\mathrm{G}_1}) + \mathrm{MLP}(\mathrm{x}_i^{SM}[t]; \Theta_{\mathrm{G}_2}), \tag{5}$$

$$\tilde{\mathrm{x}}_i[t] = \mathrm{x}_i^{SM}[t] + \alpha \cdot \tilde{\mathrm{x}}_i^{CF},$$

t^{th} token



Training

SMU Singapore Management University | School of Computing and Information Systems

- Learning objectives
  - Output: multinomial distribution over whole vocab
    - $\{p_{\text{yes}}, p_{\text{no}}\}$
  - Classification loss L1 and ranking loss L2.

$$\min_{\Theta} \mathcal{L} = \mathcal{L}_1(p_{yes}, y) + \mathcal{L}_1(p_{no}, 1-y) + k \times \mathcal{L}_2(p_{yes}, p_{no}, y), \quad (7)$$

- Two stage training
  - Stage 1: Fine-tuning only the LLM with LoRA.
  - Stage 2: Fine-tuning only ALG and GATE modules.

Experiments

SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

Method	MovieLens-1M		Amazon-Book	
	AUC	RelaImpr	AUC	RelaImpr
MF	0.6482†	-	0.7134†	-
CoLLM (MF)	0.7295†	54.86%	0.8109†	45.69%
CCF-LLM (MF)	**0.7315**	56.21%	**0.8150**	47.61%
LightGCN	0.5959†	-	0.7103†	-
CoLLM (LightGCN)	0.7100†	118.98%	0.7978†	41.61%
CCF-LLM (LightGCN)	**0.7427**	153.08%	**0.8049**	44.98%
SASRec	0.7078†	-	0.6887†	-
CoLLM (SASRec)	0.7235†	7.56%	0.7746†	45.52%
CCF-LLM (SASRec)	**0.7526**	21.56%	**0.7792**	47.96%
Softprompt	0.7071†	-	0.7224†	-
TallRec	0.7097†	1.25%	0.7375†	6.79%
CoLLM (Best)	0.7295†	10.82%	0.8109†	39.79%
CCF-LLM (Best)	**0.7526**	21.97%	**0.8150**	41.64%

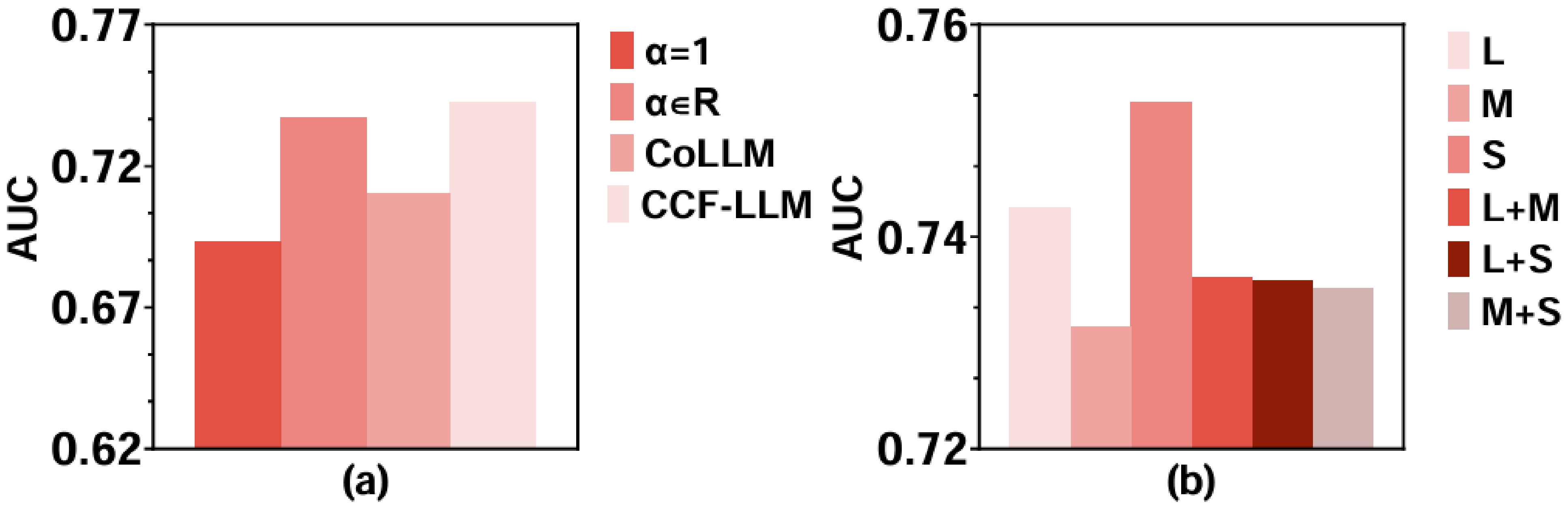
Results are reported as the average of 5 runs.
†Results are obtained from Zhang et al. [48].

No / Inadequate fusion

Other LLM4Rec

1. Semantic knowledge is effective.
2. Collaborative signals is useful.
3. Fusion strategy contributes to a more comprehensive integration.
4. Improper tuning of the embeddings can lead to a negative impact.

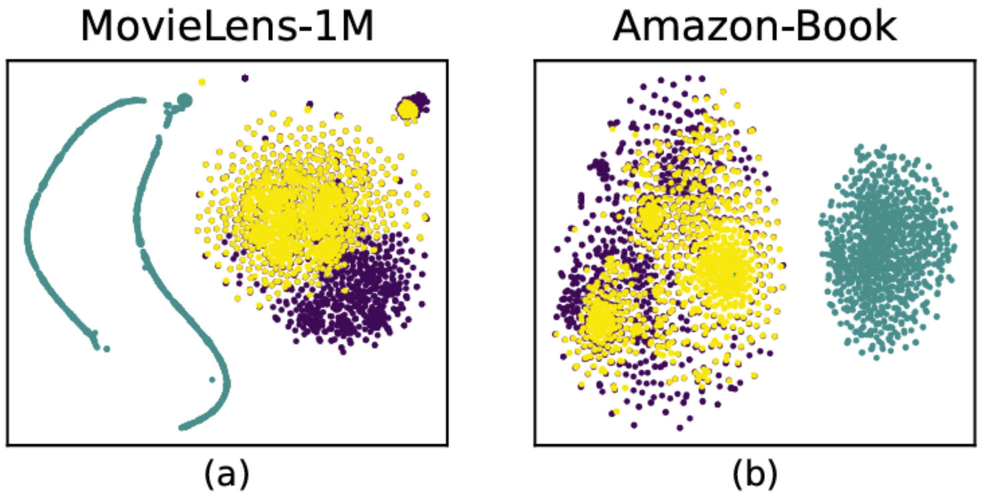
Experiments



Left: the ablation of different fusion strategy.
Right: the ablation of different backbone CF-based model.

1. Our finer dimensional-level fusion led to the optimal performance.
2. Backbone CF-based model can influence the results. Using multiple backbone models do not improve as introducing redundant collaborative signals may not offer additional insights.

Experiments



Green: aligned collaborative signals
Yellow: semantic knowledge
Violet: fused embedding.

Two types of modalities are better fused with the proposed attentive cross-modal fusion strategy.

Conclusion

- A novel framework for collaborative cross-modal fusion with large language models for recommendation.
  - Hybrid prompt translation, mapping, fusion

- Pros:
  - Integrate collaborative signals and semantic knowledge for recommendation.
  - Proposed a fusion strategy to let language model better understand the collaborative signals

- Limitations & future works:
  - The semantic knowledge for user-side is not incorporated.
  - More modalities (such as image) could be considered.
  - Analysis of different LLMs.

Thank you!
Q&A

https://arxiv.org/abs/2408.08564