Findings in Language Model for Recommendation

Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation (WSDM 2025)

Language Representations Can be What Recommenders Need: Findings and Potentials (ICLR 2025)

Xinyu He, Xinrui He

Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation

Zhankui He, Zhouhang Xie, Harald Steck, Dawen Liang, Rahul Jha, Nathan Kallus, and Julian McAuley. 2025. Reindex-Then-Adapt: Improving Large Language Models for Conversational Recommendation. Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, 866–875. https://doi.org/10.1145/3701551.3703573

Xinyu

Conversational Recommender System (CRS)

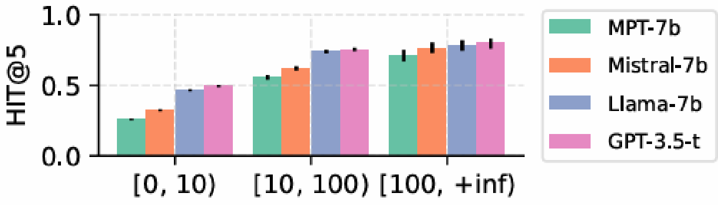
- A conversation is represented by $C = (u_t, s_t, \mathcal{I}_t)_{t=1}^{T}$ involving user $u_t \in \mathcal{U}$, user utterance s_t and ranked item list $\mathcal{I}_t \in \mathcal{I}$ with T conversational turns.
- Goal: generate ranked list of items $\hat{\mathcal{I}}_k$ that align with $\mathcal{I}_k$, based on the preceding context $(u_t, s_t, \mathcal{I}_t)_{t=1}^{k-1}$.

Motivation: Item-Indexing Ability Analysis

- Learn to Index task(L2I) : LLMs index items by using the item titles as the item identifiers

- Finding: LLMs Show Sufficient Item Content Knowledge.

"A 2014 American science fiction action film starring Tom Cruise and Emily Blunt with ..." → Edge of Tomorrow

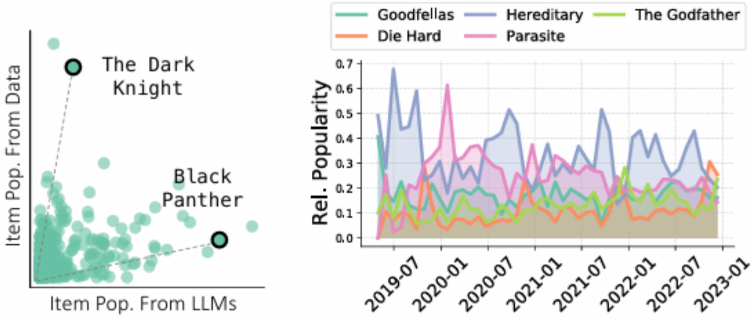


ReDIAL [38]	Cold – [0, 10)	Warm – [10, 100)	Pop. – [100, +inf)
#Items	**4,960 (78.97%)**	1,193 (18.99%)	128 (2.04%)
#Occurences	12,523 (18.03%)	**33,304 (47.94%)**	**23,647 (34.04%)**

Item popularities

Motivation: Recommendation Ability Analysis

- Learn to Retrieve task (L2R): LLMs use conversational context as queries to generate item indices.
- Finding: LLMs Show Severe Distribution Misalignment.



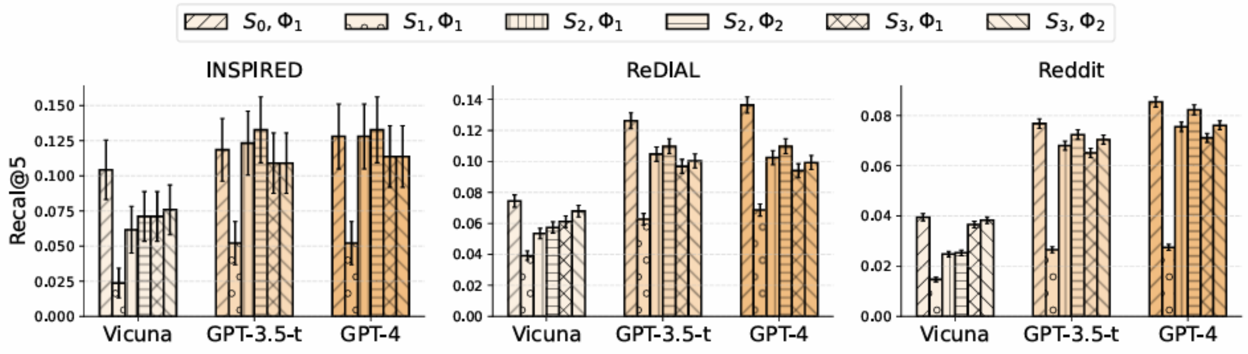
(a) Static Item Popularities (b) Item Popularities Over Time

Static and dynamic perspective

"I'm feeling bored today and looking for a sci-fi action movie, preferably starring Tom Cruise." → Edge of Tomorrow

Motivation: Recommendation Ability Analysis

- Finding: LLMs Struggle to Use Collaborative Information.
- Masking or randomly replacing mentioned movies in conversation contexts has little negative impact on the final recommendation accuracy



S0(Original)
S1(ItemOnly)
S2(ItemRemoved)
S3(ItemRandom)
Φ1(in-dataset titles)
Φ2(excludes seen titles)

Motivation

- Room for improvement:
  - adapting towards the item distributions in target platforms
  - enables the synergy between LLMs and traditional methods
- LLMs represent (i.e., index) each item with typically multiple tokens, which makes obtaining recommendation probabilities for a large-scale set of items extremely inefficient, thus hindering the potential of LLMs

Re-index

- To better control LLM item recommendation, Reindex step condense multi-token item embeddings into single-token item embeddings.

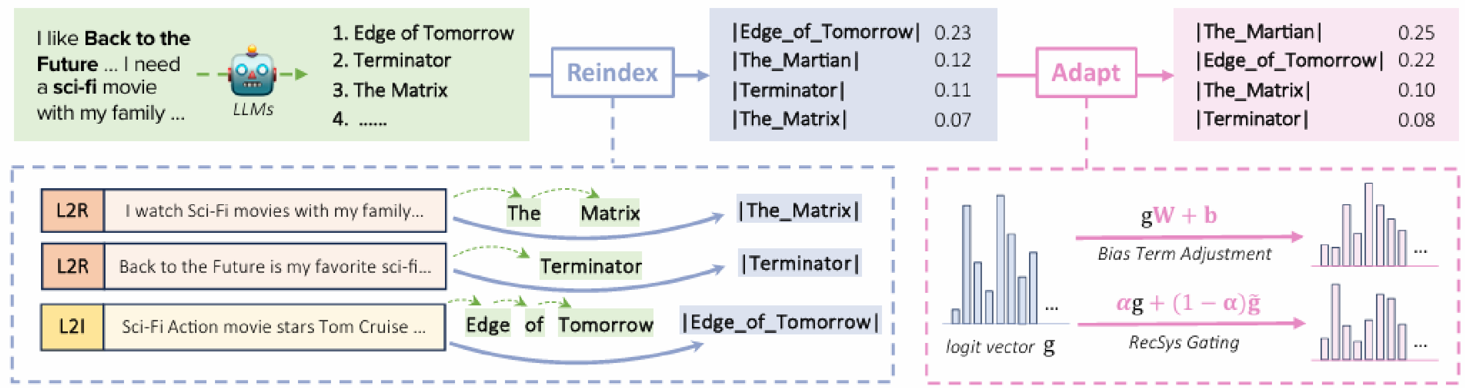
$$(\mathbf{v}_i)_{i=j}^{j+n} = \mathsf{Embed}\left((v_i)_{i=j}^{j+n}\right), \quad \tilde{\mathbf{v}} = \mathsf{Aggregator}\left((\mathbf{v}_i)_{i=j}^{j+n}\right),$$

- $(v_i)_{i=j}^{j+n}$: tokens representing the item, e.g., tokens of 'Edge of Tomorrow'.
- Aggregator: RNN/transformer/weighted pooling

$$\mathcal{L}_{reindex} = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{q},\tilde{\mathbf{v}} \in \mathcal{D}} \log \left[\frac{\exp\left(\mathbf{q}^\top \tilde{\mathbf{v}}\right)}{\exp\left(\mathbf{q}^\top \tilde{\mathbf{v}}\right) + \sum_{\mathbf{n} \in \mathcal{N}} \exp\left(\mathbf{q}^\top \mathbf{n}\right)} \right],$$

Adapt

- Bias Term Adjustment: $\hat{\mathbf{p}} = \mathtt{softmax}\left(\mathbf{g}\mathbf{W} + \mathbf{b}\right),$
- Recsys gating: $\hat{\mathbf{p}} = \mathtt{softmax}\left(\alpha\mathbf{g} + (1-\alpha)\tilde{\mathbf{g}}\right),$
- $\mathcal{L}_{adapt} = -\frac{1}{|\mathcal{D}^*|}\sum_{i=1}^{|\mathcal{D}^*|} \log \hat{\mathbf{p}}_{i,*},$

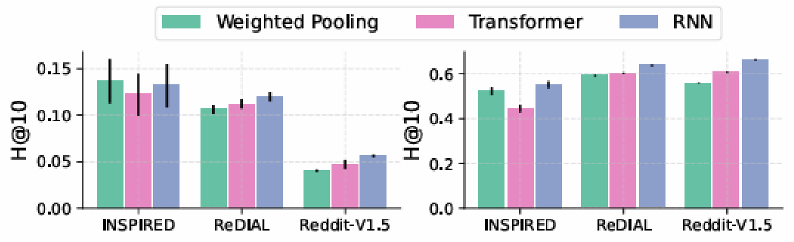


Experiment

Table 3: The main results for our models on conversational recommendation accuracy performance, compared against (1) traditional recommendation models; (2) zero-shot large language models (LLMs); (3) traditional conversational recommendation models; and (4) zero-shot dense retrievers. The size of the reported LLMs used here is 7B. We denote the model metrics with the best performance in bold. Llama2-R denotes the Llama2-7b model after our `reindex` step. We also show the results after the `adapt` step with bias terms (+Bias) or RecSys (here we use FISM [31]) model combination with Gating mechanism (+RecSys).

	INSPIRED				ReDIAL				Reddit-V1.5			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
Popuplarity	.089 .020	.065 .015	.103 .021	.070 .015	.035 .003	.025 .002	.052 .003	.030 .002	.008 .001	.004 .000	.014 .001	.006 .000
FISM [31]	.090 .017	.065 .012	.104 .021	.072 .011	.065 .004	.040 .003	.112 .005	.054 .003	.022 .001	.012 .001	.043 .001	.019 .001
SASRec [32]	.092 .014	.068 .011	.103 .020	.073 .013	.068 .004	.041 .002	.116 .005	.056 .003	.022 .001	.013 .001	.039 .001	.018 .001
MPT [52]	.075 .018	.045 .011	.099 .020	.052 .012	.072 .004	.045 .003	.116 .005	.059 .003	.026 .001	.017 .001	.040 .001	.021 .001
Mistral [30]	.061 .016	.040 .011	.066 .017	.041 .012	.082 .004	.056 .003	.111 .005	.065 .003	.029 .001	.020 .001	.038 .001	.023 .001
Llama2 [53]	.080 .019	.050 .012	.122 .022	.064 .013	.094 .004	.059 .003	.145 .005	.075 .003	.042 .001	.027 .001	.064 .001	.034 .001
ReDIAL [38]	.060 .016	.041 .012	.106 .021	.056 .012	.067 .004	.044 .003	.106 .005	.057 .003	.029 .001	.019 .001	.044 .001	.024 .001
UniCRS [57]	.091 .019	.055 .011	.132 .019	.073 .014	.085 .003	.058 .003	.112 .004	.071 .003	.028 .001	.017 .001	.040 .001	.021 .001
SBERT [47]	.038 .013	.026 .010	.066 .017	.036 .010	.016 .002	.010 .001	.026 .002	.013 .001	.003 .000	.002 .000	.005 .000	.002 .000
Instructor [50]	.052 .015	.034 .011	.085 .019	.045 .011	.025 .002	.013 .001	.043 .003	.019 .001	.009 .001	.006 .000	.017 .001	.008 .000
Llama2-R	.075 .018	.045 .012	.131 .023	.063 .012	.072 .004	.044 .003	.120 .005	.059 .003	.034 .001	.020 .001	.056 .001	.028 .001
w/ Bias	.117 .022	.077 .016	.174 .026	.095 .016	.096 .004	.061 .003	**.165** .006	.083 .003	.052 .001	.033 .001	.088 .002	.044 .001
w/ RecSys	**.131** .023	**.081** .015	**.197** .027	**.101** .016	**.103** .005	**.066** .003	**.165** .006	**.085** .003	**.057** .001	**.035** .001	**.093** .002	**.047** .001

Ablations



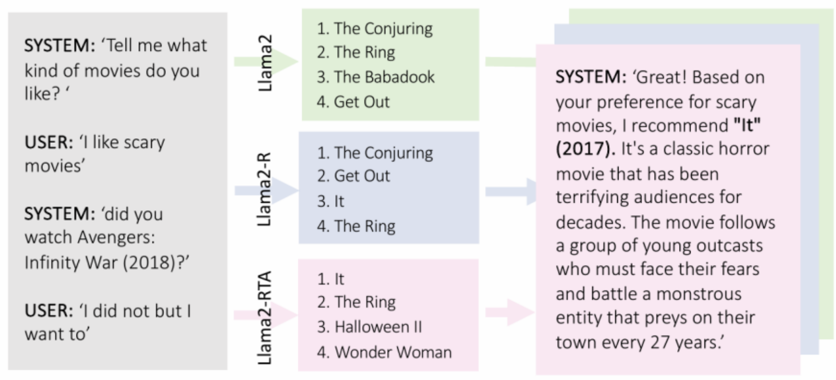
(a) Labels from Data (b) Labels from LLMs

	INSPIRED		ReDIAL		Reddit-V1.5	
Model	H@10	N@10	H@10	N@10	H@10	N@10
Llama2	.122 .022	.064 .013	.145 .005	.075 .003	.064 .001	.034 .001
→ RTA w/ Bias	.174 .026	.095 .016	.165 .006	.083 .003	.088 .002	.044 .001
→ RTA w/ RecSys	.197 .027	.101 .016	.165 .006	.085 .003	.093 .002	.047 .001
GPT-3.5-t [49]	.150 .024	.089 .016	.163 .006	.089 .003	.104 .002	.055 .001

	Agg. Frozen?	INSPIRED	ReDIAL	RedditV1.5
Llama2-R	–	.131 .023	.120 .005	.056 .001
Bias Term Adjustment (w/ Bias)				
w/ gW	×	.150 .025	.165 .006	.082 .002
	✓	.146 .024	.141 .006	.058 .001
w/ b	×	.174 .026	.165 .006	.087 .002
	✓	.136 .235	.117 .005	.058 .001
w/ gW+b	×	.155 .025	.167 .006	.088 .002
	✓	.160 .025	.144 .005	.058 .001
RecSys Model Gating (w/ RecSys)				
w/ FISM	×	.197 .027	.146 .005	.093 .002
	✓	.207 .028	.165 .006	.074 .002
w/ SASRec	×	.178 .026	.148 .005	.093 .002
	✓	.188 .027	.157 .006	.075 .002

Conclusion and Comments

- Reindex technique provides potential for re-ranking and post-processing
- LLM is sufficient as item indexers for popular items
- Adapting popularity distribution may increase popularity bias



(a) Context (b) Recommendation (c) Generation

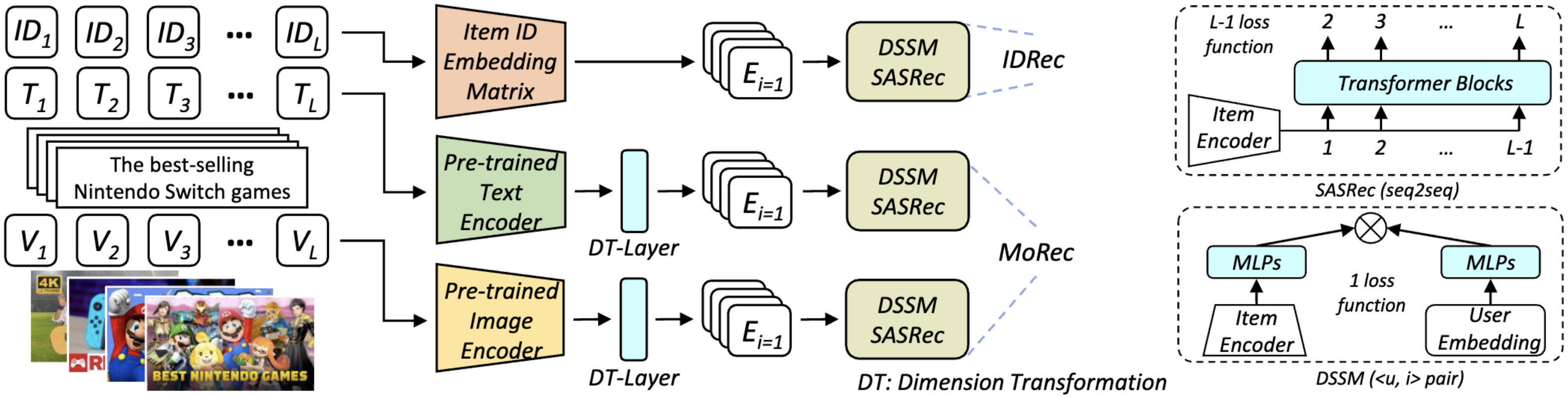
Language Representations Can be What Recommenders Need: Findings and Potentials

Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, Tat-Seng Chua. ICLR 2025

ID- vs. Modality-based Recommnder Models

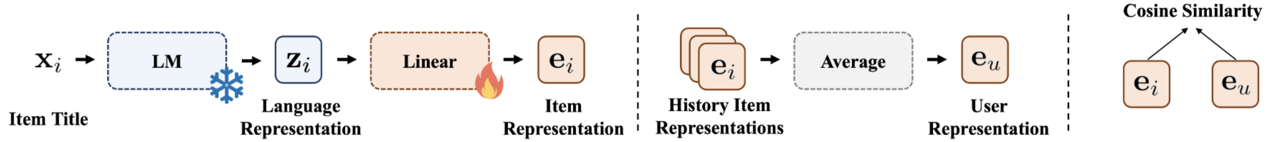
- **ID-based Recommender**: a learnable ID embedding matrix $X^I \in \mathbb{R}^{|I| \times d}$ is initialized . Each vector X^I represents the latent space of an item i.

- **LLm-based Recommender**: For an Item i, a pre-trained encoder is used to generate the representation to replace the ID embedding in IDRec.



RQ1: Do LMs inherently encode collaborative signals

- Linear Mapping:

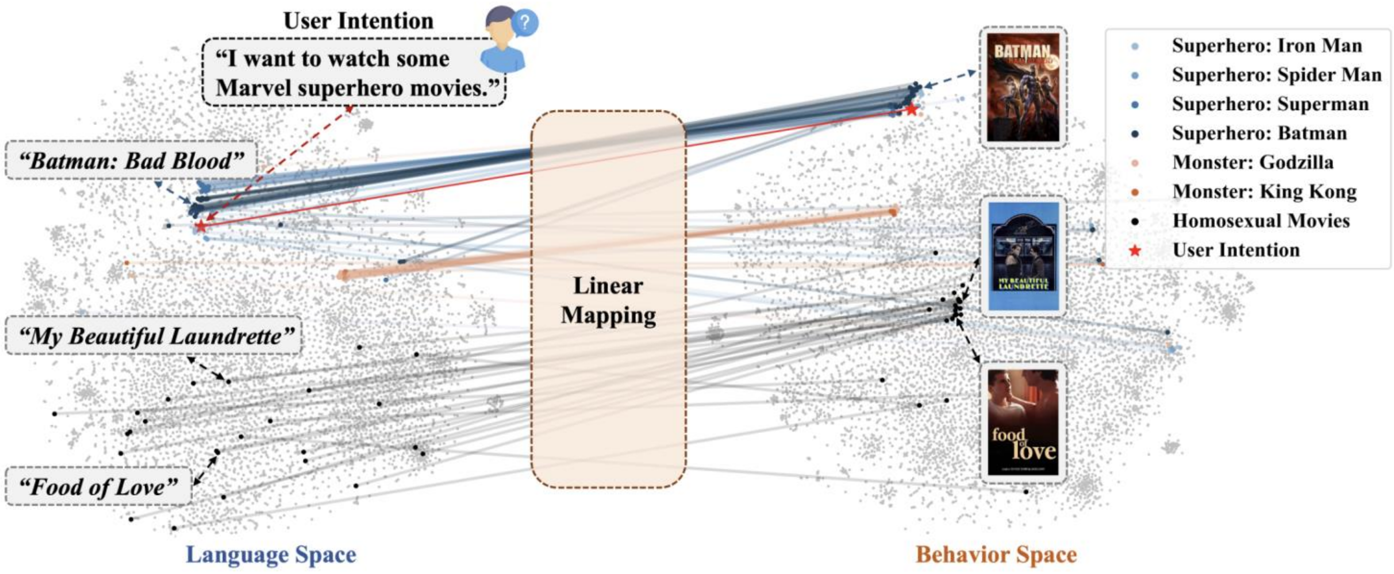


- Finding 1: Post-mapping representations of advanced LMs achieve superior recommendation performance in most cases, suggesting the possible homomorphism between language spaces and behavior spaces.

		Movies & TV			Video Games			Books		
		Recall	NDCG	HR	Recall	NDCG	HR	Recall	NDCG	HR
CF	MF (Rendle et al., 2012)	0.0568	0.0519	0.3377	0.0323	0.0195	0.0864	0.0437	0.0391	0.2476
	MultVAE (Liang et al., 2018)	0.0853	0.0776	0.4434	0.0908	0.0531	0.2211	0.0722	0.0597	0.3418
	LightGCN (He et al., 2021)	0.0849	0.0747	0.4397	0.1007	0.0590	0.2281	0.0723	0.0608	**0.3489**
Linear Mapping	BERT	0.0415	0.0399	0.2362	0.0524	0.0309	0.1245	0.0226	0.0194	0.1240
	RoBERTa	0.0406	0.0387	0.2277	0.0578	0.0338	0.1339	0.0247	0.0209	0.1262
	Llama2-7B	0.1027	0.0955	0.4952	0.1249	0.0729	0.2746	0.0662	0.0559	0.3176
	Mistral-7B	0.1039	0.0963	0.4994	0.1270	0.0687	0.2428	0.0650	0.0544	0.3124
	text-embedding-ada-v2	0.0926	0.0874	0.4563	0.1176	0.0683	0.2579	0.0515	0.0436	0.2570
	text-embeddings-3-large	0.1109	0.1023	0.5200	0.1367	**0.0793**	**0.2928**	0.0735	0.0608	0.3355
	SFR-Embedding-Mistral	**0.1152**	**0.1065**	**0.5327**	**0.1370**	0.0787	0.2927	**0.0738**	**0.0610**	0.3371

RQ1: Do LMs inherently encode collaborative signals

❖ Finding 2: Language representations encode user preference similarities beyond semantic textual similarities.



RQ2: Scaling and Robustness

- Finding 3: The encoding of user preference similarities becomes more refined as model size increases, leading to better linear mapping performance.
- Finding 4: Language representations are relatively robust to prompt disturbances.

Table 2: The robustness of language representations for recommendation.

	Movies & TV			Video Games			Books		
	Recall	NDCG	HR	Recall	NDCG	HR	Recall	NDCG	HR
Title + Random Noise	0.0952	0.0887	0.4731	0.1213	0.0706	0.2722	0.0632	0.0525	0.3099
Title Only	0.1027	0.0955	0.4952	0.1249	0.0729	0.2746	0.0662	0.0559	0.3176

RQ3: How powerful are such language representations for building advanced CF models that can outperform prevailing ID-based CF methods

- AlPHAREC
  - Nonlinear projection

$$\mathbf{e}_i^{(0)} = \boldsymbol{W}_2 \, \mathrm{LeakyReLU}\left(\boldsymbol{W}_1 \mathbf{z}_i + \boldsymbol{b}_1\right) + \boldsymbol{b}_2, \quad \mathbf{e}_u^{(0)} = \boldsymbol{W}_2 \, \mathrm{LeakyReLU}\left(\boldsymbol{W}_1 \mathbf{z}_u + \boldsymbol{b}_1\right) + \boldsymbol{b}_2.$$

  - Graph convolution

$$\mathbf{e}_u^{(k+1)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \mathbf{e}_i^{(k)}, \quad \mathbf{e}_i^{(k+1)} = \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} \mathbf{e}_u^{(k)}.$$

  - Contrastive learning objective

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{(u,i) \in \mathcal{O}^+} \log \frac{\exp\left(s(\mathbf{e}_u, \mathbf{e}_i)/\tau\right)}{\exp\left(s(\mathbf{e}_u, \mathbf{e}_i)/\tau\right) + \sum_{j \in \mathcal{S}_u} \exp\left(s(\mathbf{e}_u, \mathbf{e}_j)/\tau\right)}.$$

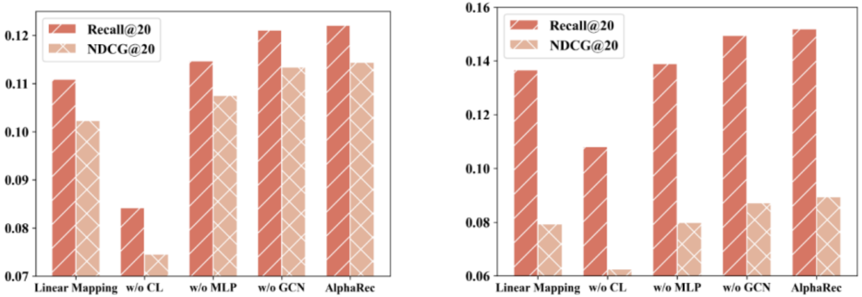
RQ3: Recommendation capabilities

❖ Finding 5: Advanced language representations shows strong potentials for recommendation, which can be unleashed by appropriate model design.

	Movies & TV			Video Games			Books		
	Recall	NDCG	HR	Recall	NDCG	HR	Recall	NDCG	HR
MF (Rendle et al., 2012)	0.0568	0.0519	0.3377	0.0323	0.0195	0.0864	0.0437	0.0391	0.2476
MultVAE (Liang et al., 2018)	0.0853	0.0776	0.4434	0.0908	0.0531	0.2211	0.0722	0.0597	0.3418
LightGCN (He et al., 2021)	0.0849	0.0747	0.4397	0.1007	0.0590	0.2281	0.0723	0.0608	0.3489
SGL (Wu et al., 2021)	0.0916	0.0838	0.4680	0.1089	0.0634	0.2449	0.0789	0.0657	0.3734
BC Loss (Zhang et al., 2022)	0.1039	0.0943	0.5037	0.1145	0.0668	0.2561	0.0915	0.0779	0.4045
XSimGCL (Yu et al., 2024)	0.1057	0.0984	0.5128	0.1138	0.0662	0.2550	0.0879	0.0745	0.3918
KAR (Xi et al., 2023)	0.1084	0.1001	0.5134	0.1181	0.0693	0.2571	0.0852	0.0734	0.3834
RLMRec (Ren et al., 2024b)	0.1119	0.1013	0.5301	0.1384	0.0809	0.2997	0.0928	0.0774	0.4092
AlphaRec	**0.1221***	**0.1144***	**0.5587***	**0.1519***	**0.0894***	**0.3207***	**0.0991***	**0.0828***	**0.4185***
Imp.% over the best baseline	6.79%	5.34%	2.27%	9.12%	10.75%	5.40%	9.75%	10.51%	7.01%

KAR and RLMRec: the combination of ID-based embeddings and language representations in these methods does not yield higher results than purel language-representation-based AlphaRec

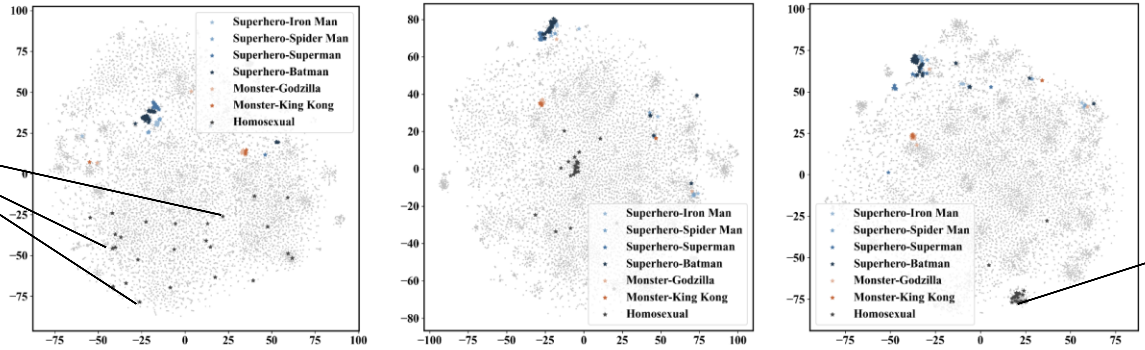
Ablation Study



(a) Ablation study on Movies & TV (b) Ablation study on Video Games

Figure 6: Ablation study

Homosexual



Homosexual

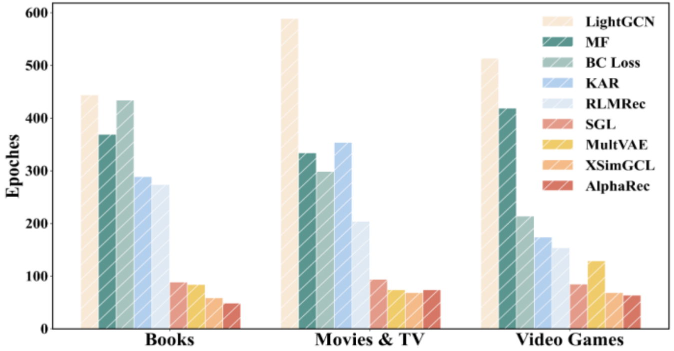
(a) LM representations (b) AlphaRec (w/o MLP) (c) AlphaRec

Potentials of Language Representations for Recommendation

Table 4: The zero-shot recommendation performance comparison on entirely new datasets. The improvement achieved by AlphaRec is significant (p-value $<< 0.05$).

		Industrial & Scientific			MovieLens-1M			Book Crossing		
		Recall	NDCG	HR	Recall	NDCG	HR	Recall	NDCG	HR
full	MF (Rendle et al., 2012)	0.0344	0.0225	0.0521	0.1855	0.3765	0.9634	0.0316	0.0317	0.2382
	MultVAE (Liang et al., 2018)	0.0751	0.0459	0.1125	0.2039	0.3741	0.9740	0.0736	0.0634	0.3716
	LightGCN (He et al., 2021)	0.0785	0.0533	0.1078	0.2019	0.4017	0.9715	0.0630	0.0588	0.3475
zero-shot	Random	0.0148	0.0061	0.0248	0.0068	0.0185	0.2611	0.0039	0.0036	0.0443
	Pop	0.0216	0.0087	0.0396	0.0253	0.0679	0.5439	0.0119	0.0101	0.1157
	ZESRec (Ding et al., 2021)	0.0326	0.0272	0.0628	0.0274	0.0787	0.5786	0.0155	0.0143	0.1347
	UniSRec (Hou et al., 2022)	0.0453	0.0350	0.0863	0.0578	0.1412	0.7135	0.0396	0.0332	0.2454
	AlphaRec	**0.0913***	**0.0573**	**0.1277***	**0.1486***	**0.3215***	**0.9296***	**0.0660***	**0.0545***	**0.3381***
	Imp.% over the best zero-shot baseline	157.09%	127.69%	30.29%	66.67%	64.16%	37.78%	101.55%	63.71%	47.97%

❖ **Potential 1**: Good initialization for item representations



(b) Training efficiency comparison

❖ **Potential 1**: Zero-shot ability

➢ provide opportunities for learning transferable item representations

Potentials of Language Representations for Recommendation

❖ **Potential 3**: The language understanding ability in advanced language representations enables recommenders to perceive user intentions and refine recommendations.

➢ User intention representation

$$e_u^{Intention}$$

➢ New user representation:

$$\tilde{e}_u^{(0)} = (1-\alpha)e_u^{(0)} + \alpha e_u^{Intention}$$

Table 5: The performance comparison in user intention capture.

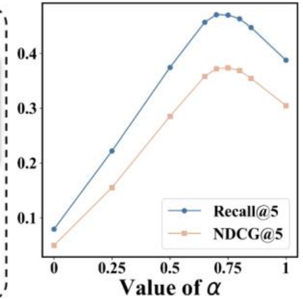
	MovieLens-1M		Video Games	
	HR@5	NDCG@5	HR@5	NDCG@5
TEM (Bi et al., 2020)	0.2738	0.1973	0.2212	0.1425
AlphaRec (w/o Intention)	0.0793	0.0498	0.0663	0.0438
AlphaRec (w Intention)	**0.4704***	**0.3738***	**0.2569***	**0.1862***

User Intention: I'm looking for a classic movie that delves into the world of organized crime, family loyalty, and power struggles, with iconic performances and unforgettable quotes.
Target: Godfather, The (1972)

Recommendation List (w/o Intention)
- Braveheart (1995)
- Schindler's List (1993)
- Star Wars: Episode V (1980)
- Pulp Fiction (1994)
- L.A. Confidential (1997)

Recommendation List (w Intention)
- Godfather, The (1972)
- L.A. Confidential (1997)
- Schindler's List (1993)
- Pulp Fiction (1994)
- Braveheart (1995)

(a) Case study of user intention capture



(b) Effect of α

Thank you!
Q & A