



## Introduction

- Product tagging is essential in E-commerce for organizing content, **enhancing recommendations**, and **improving search**.
- As platforms grow, **scalable** and **adaptive tagging** becomes increasingly important.
- Traditional methods—**manual**, **rule-based**—struggle with **cost**, **consistency**, and **semantic depth**.
- LLMs improve semantic understanding but often ignore **user behavior** and **item relationships**, leading to **generic**, **non-personalized tags**.

## Main Contributions

- Leverage large language models (LLMs) to generate tags in a more intelligent and automated way.
- Incorporate **user behavior data** during finetuning to go **beyond surface-level semantics** and **model real user interests**.
- Produce **interpretable BETags** that reveal **implicit preferences**.
- Ensure **offline tag generation**, allowing **seamless integration** without adding overhead to existing systems.

## Base Tag Generation

- Prompt an off-the-shelf LLM with structured item descriptions (e.g., title, plot summary).
- Generate base tags as **basic semantic representations** of items.

## Behavior-enhanced Finetuning

- Construct finetuning examples from **user behavior sequences**.
- Represent each item with its **base tags**, formatted as numbered tag lists.

### Finetuning Prompt:

You are a helpful recommendation assistant tasked with predicting the next likely item based on user interactions. Given a sequence of previously interacted items, where each item is represented by multiple tags (comma-separated), predict the tags for the item the user is most likely to interact with next.

### Finetuning Answer:

- Serial Killer Thriller, FBI Investigation, Small Town Mystery
- Psychological Thriller, Mind Games, Wealth and Isolation
- Historical Drama, Romance, Tragedy, Epic Adventure
- Historical Drama, WWII Holocaust, Humanitarianism
- Classic Mystery, Psychological Thriller, Murder Mystery

## BETag Generation

- Prompt the finetuned model with **base tags** to generate **BETags**.
- Use **diverse multi-beam generation** to encourage outputs that cover different aspects of the item
- Maintain consistent prompting format between finetuning and inference.

## Downstream Applications

- We evaluate the generated BETags **across diverse retrieval and recommendation tasks** to demonstrate their versatility and impact on real-world systems.
- Retrievers** rely solely on textual signals (tags), while **recommenders** combine item IDs with textual content in a deep model.
- User-based** settings use interaction histories to model preferences, while **item-based** settings rely only on item content, without access to user behavior data.

## Methodology

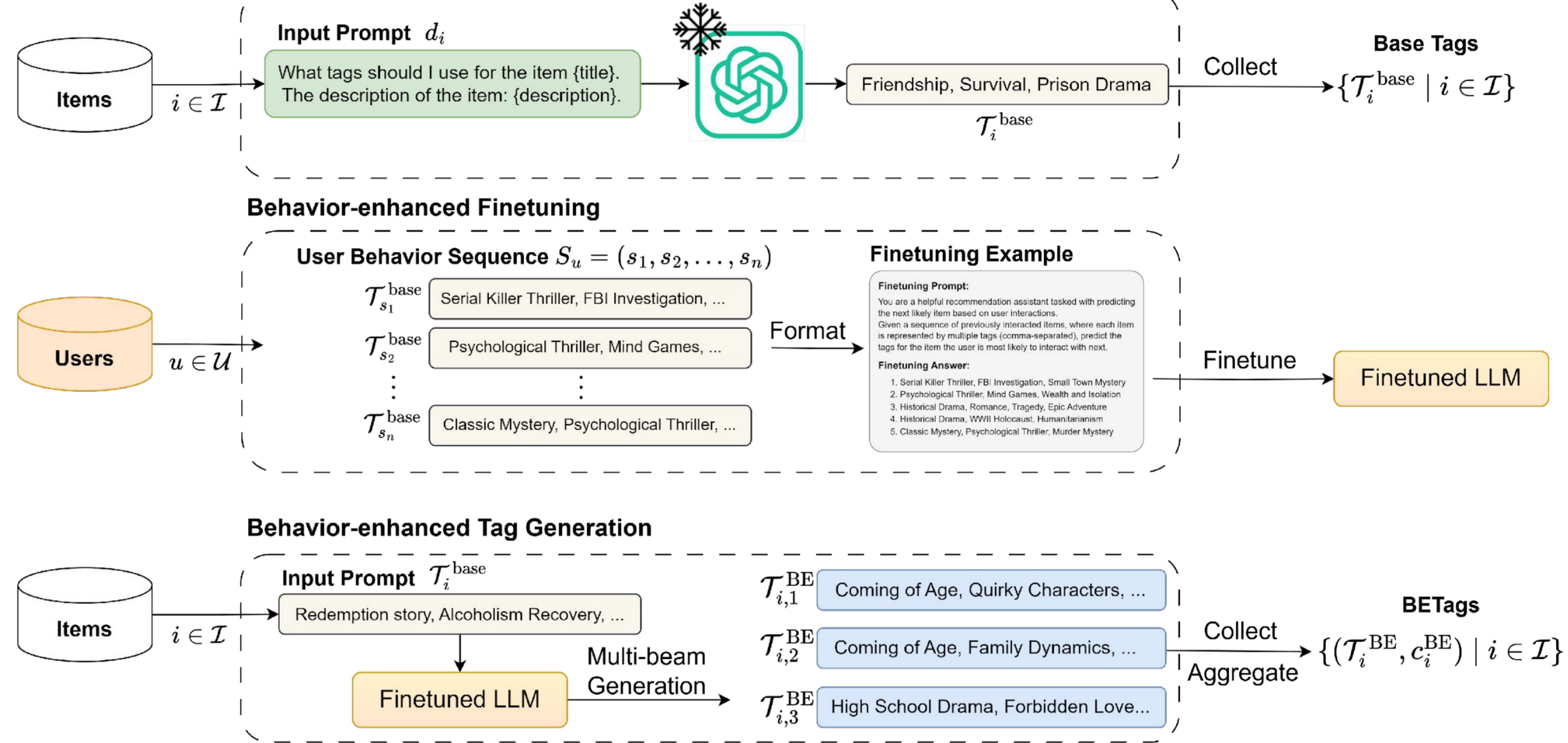


Fig. 1. The overall framework of BETag

## Experimental Results

- BETag consistently **outperforms TagGPT and Native Tags** on both retrieval and recommendation tasks, highlighting its versatility. (Table 1-2)

Table 1. Retrieval tasks

User-Based	Scientific		Movielens-1M		FreshFood	
	H@10	N@10	H@10	N@10	H@10	N@10
<b>Popular</b>	0.2871	0.1672	0.3866	0.2115	0.4158	0.2292
<b>BM25</b>						
- Native Tags	-	-	0.3488	0.1910	0.5145	0.3216
- TagGPT	0.3430	0.2241	0.2929	0.1601	0.4666	0.3165
- <b>BETag (Ours)</b>	<b>0.4883</b>	<b>0.3098</b>	<b>0.4723</b>	<b>0.2791</b>	<b>0.5152</b>	<b>0.3350</b>
<b>BiRank</b>						
- Native Tags	-	-	0.3939	0.2200	0.5287	0.3493
- TagGPT	0.3741	0.2409	0.3095	0.1663	0.4832	0.3284
- <b>BETag (Ours)</b>	<b>0.4681</b>	<b>0.2991</b>	<b>0.4667</b>	<b>0.2732</b>	<b>0.5526</b>	<b>0.3569</b>

Item-Based	Scientific		Movielens-1M		FreshFood	
	H@10	N@10	H@10	N@10	H@10	N@10
<b>Popular</b>	0.2871	0.1672	0.3467	0.1784	0.4158	0.2292
<b>BM25</b>						
- Native Tags	-	-	0.3179	0.1718	0.3990	0.2361
- TagGPT	0.2862	0.1930	0.2104	0.1116	0.2928	0.1896
- <b>BETag (Ours)</b>	<b>0.4249</b>	<b>0.2747</b>	<b>0.3779</b>	<b>0.2160</b>	<b>0.4603</b>	<b>0.2712</b>
<b>BiRank</b>						
- Native Tags	-	-	0.2927	0.1563	0.4062	0.2432
- TagGPT	0.3040	0.2024	0.2244	0.1170	0.3260	0.2026
- <b>BETag (Ours)</b>	<b>0.4239</b>	<b>0.2735</b>	<b>0.3764</b>	<b>0.2176</b>	<b>0.4791</b>	<b>0.2779</b>

Table 2. Recommendation tasks

User-Based	Scientific		Movielens-1M		FreshFood	
	H@10	N@10	H@10	N@10	H@10	N@10
<b>SASRec</b>	0.5057	0.3342	0.7335	0.4904	<b>0.5698</b>	<b>0.3399</b>
<b>UniSRec</b>						
- Native Tags	-	-	0.7462	0.5039	0.5689	0.3271
- TagGPT	0.5308	0.3400	0.7474	0.5066	0.5618	0.3254
- <b>BETag (Ours)</b>	<b>0.5801</b>	<b>0.3742</b>	<b>0.7523</b>	<b>0.5106</b>	<b>0.5741</b>	<b>0.3417</b>
Item-Based	Scientific		Movielens-1M		FreshFood	
	H@10	N@10	H@10	N@10	H@10	N@10
<b>SASRec</b>	0.3930	0.2409	<b>0.6636</b>	<b>0.4304</b>	<b>0.5396</b>	<b>0.3157</b>
<b>UniSRec</b>						
- Native Tags	-	-	0.6634	0.4244	0.4680	0.2639
- TagGPT	0.4415	0.2754	0.6547	0.4254	0.5168	0.3041
- <b>BETag (Ours)</b>	<b>0.4740</b>	<b>0.2972</b>	<b>0.6703</b>	<b>0.4344</b>	<b>0.5590</b>	<b>0.3245</b>

## Conclusion

- BETag successfully bridges the gap between **semantic content** and **user behavior** in product tagging, creating tags that are both **descriptive** and **aligned with real user preferences**.
- Consistently **outperforms both human-annotated tags and existing automated methods** across multiple datasets and downstream tasks.
- Provides an **efficient, scalable solution** for E-commerce platforms that enhances recommendation quality and search relevance **without imposing real-time computational burdens**.

## Impact of Finetuning and Multi-Beam

- Removing either fine-tuning or multi-beam generation leads to reduced performance, with the largest drop when both are absent.

Table 3. Impact on item-based retrieval

	Scientific		Movielens-1M		FreshFood	
	H@10	N@10	H@10	N@10	H@10	N@10
<b>BM25</b>						
- <b>BETag (Ours)</b>	<b>0.4249</b>	<b>0.2747</b>	<b>0.3779</b>	<b>0.2160</b>	<b>0.4603</b>	<b>0.2712</b>
- w/o Finetuning	0.3582	0.2562	0.2382	0.1288	0.2918	0.1898
- w/o Multi-beam	0.2234	0.1508	0.2093	0.1138	0.3493	0.2153
- w/o Both	0.1570	0.1064	0.1740	0.0848	0.2270	0.1437
<b>BiRank</b>						
- <b>BETag (Ours)</b>	<b>0.4239</b>	<b>0.2735</b>	<b>0.3764</b>	<b>0.2176</b>	<b>0.4791</b>	<b>0.2779</b>
- w/o Finetuning	0.3907	0.2723	0.2430	0.1309	0.3210	0.2001
- w/o Multi-beam	0.2863	0.1821	0.2277	0.1203	0.3693	0.2213
- w/o Both	0.2059	0.1340	0.1426	0.0744	0.2432	0.1564

## Effect of Beam Count

- Increasing beam count improves performance, but the effect saturates (Fig. 2a).
- More beams **increase tag diversity**, and the distribution gradually stabilizes (Fig. 2b).

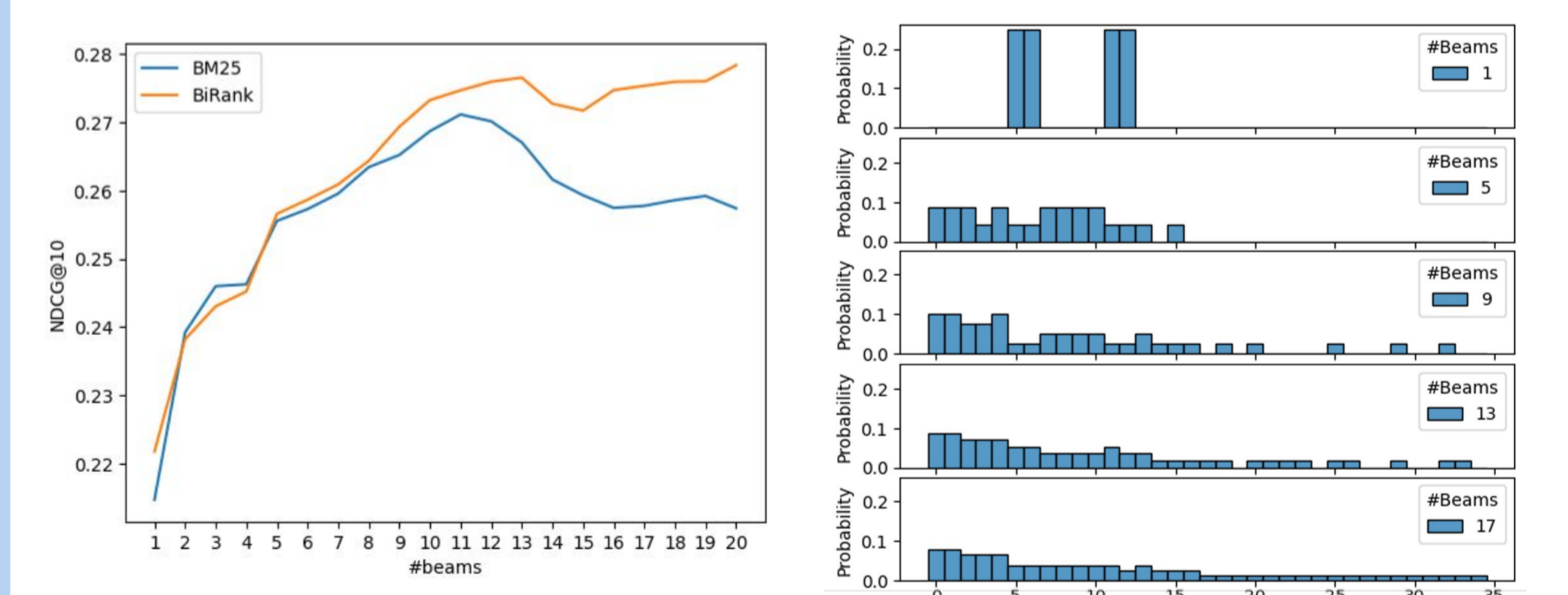


Fig. 2. Effect of beam count (FreshFood)

## Reducing Tag Sparsity

- Raw LLM tags are sparse, but post-processing makes the distribution more balanced (Fig. 3a).
- BETag **reduces sparsity without post-processing**, aided by finetuning and multi-beam generation (Fig. 3b).

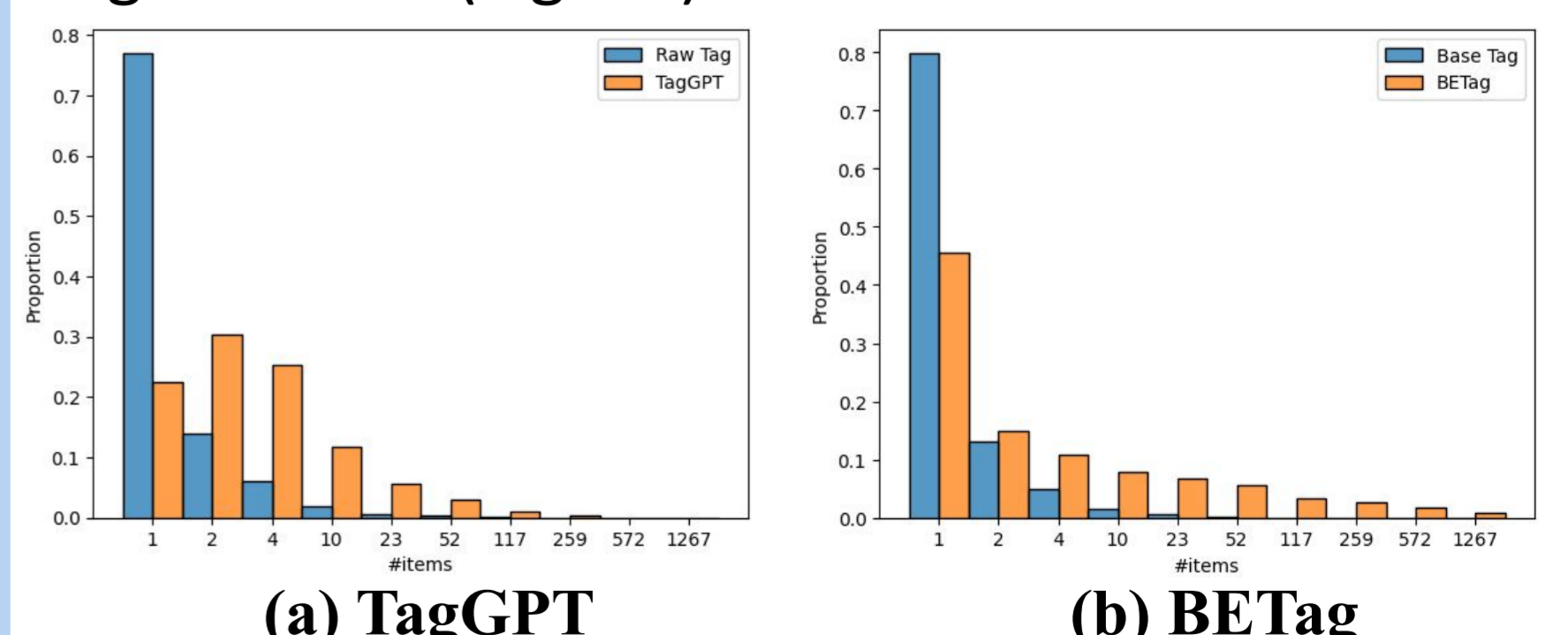


Fig. 3. Popularity (Movielens-1M)

## Acknowledgement

This work was supported in part by the National Science and Technology Council (NSTC) of Taiwan under grant number 113-2622-E-002-015. We also thank AviviD.ai for providing the dataset for FreshFood.

