SIGIR 2025
Padova
ITALY

You Are What You Bought: Generating Customer Personas for E-commerce Applications

Yimin Shi, Yang Fei, Shiqi Zhang,
Haixun Wang, Xiaokui Xiao

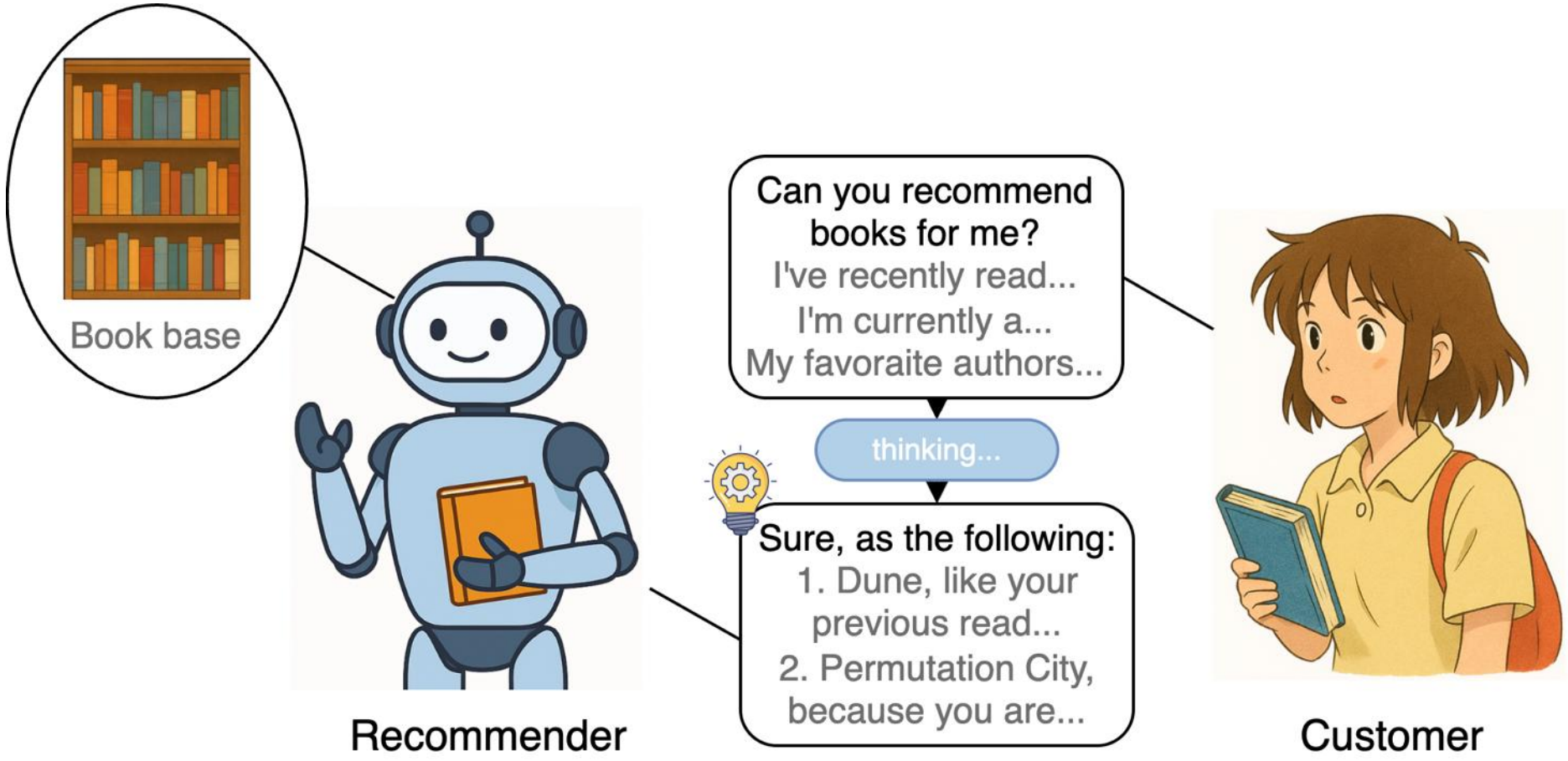
NUS National University of Singapore

EvenUp

PyroWis

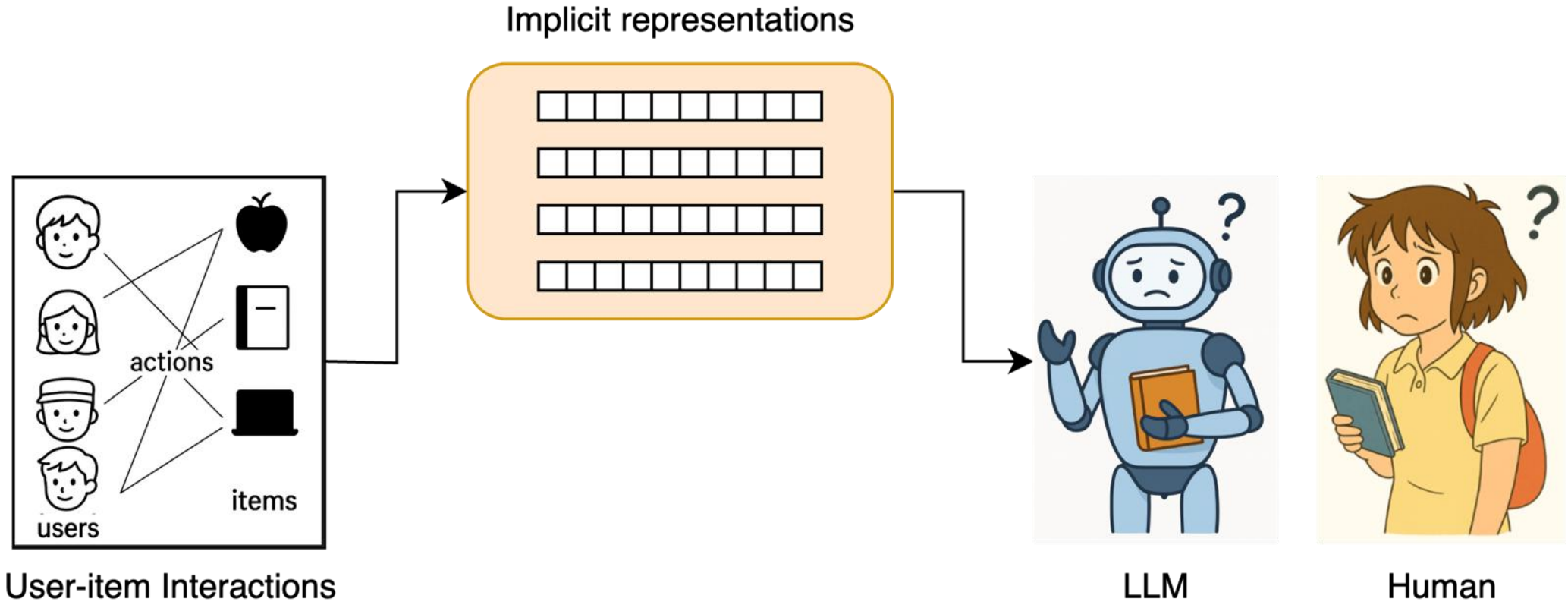
Background: E-commerce with LLM

Example: online book retailer
Agent: personalized, explainable, convincing recommendations



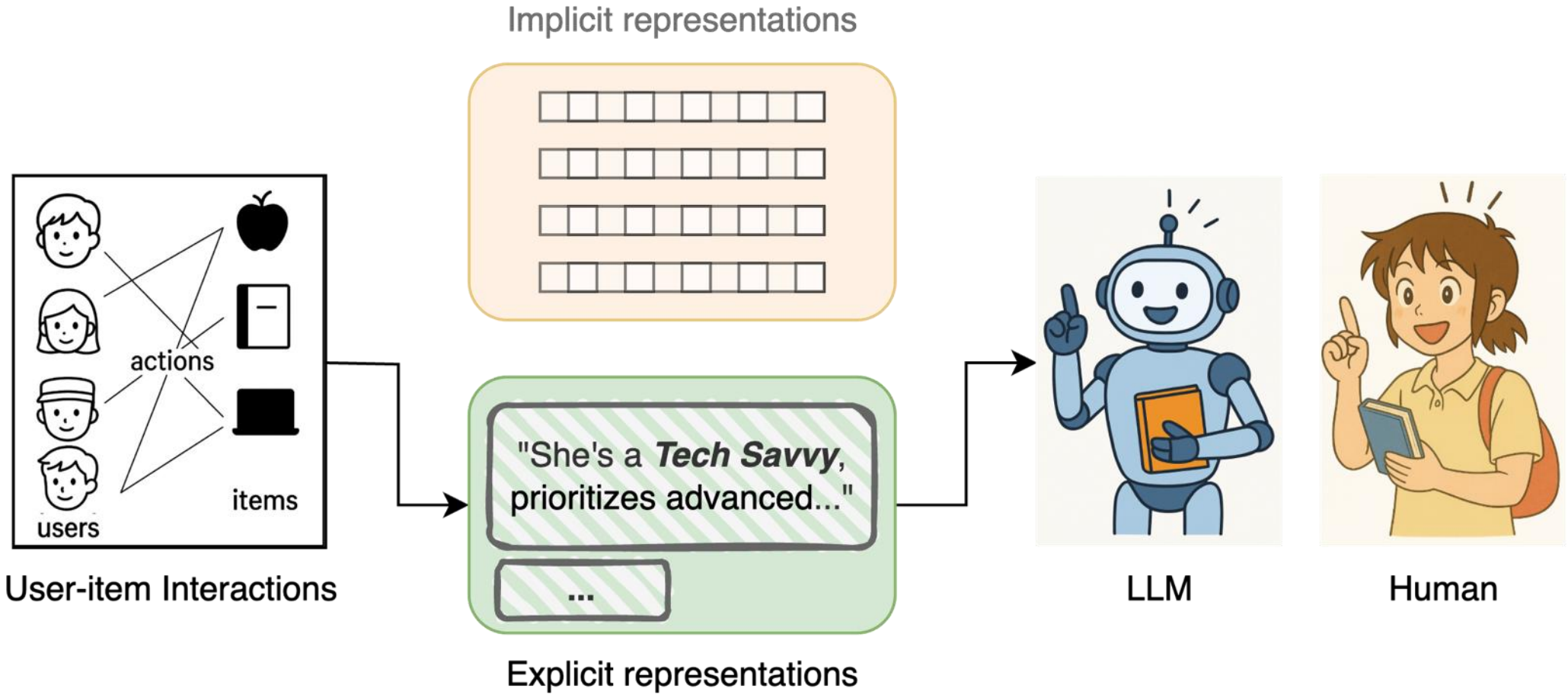
Question: How to represent customers?

Current: implicit representations, (e.g., latent embeddings)



Question: How to represent customers?

Current: implicit representations, (e.g., latent embeddings)
Agents need: explicit representations (e.g., a piece of text)



Motivation: Customer Personas

Persona examples:

Busy Parents

frequently purchases kid-friendly products... look for convenience, buying pre-made meals...

Health Enthusiast

prioritizes healthy, often organic or non-GMO food, supplements, and health-related products...

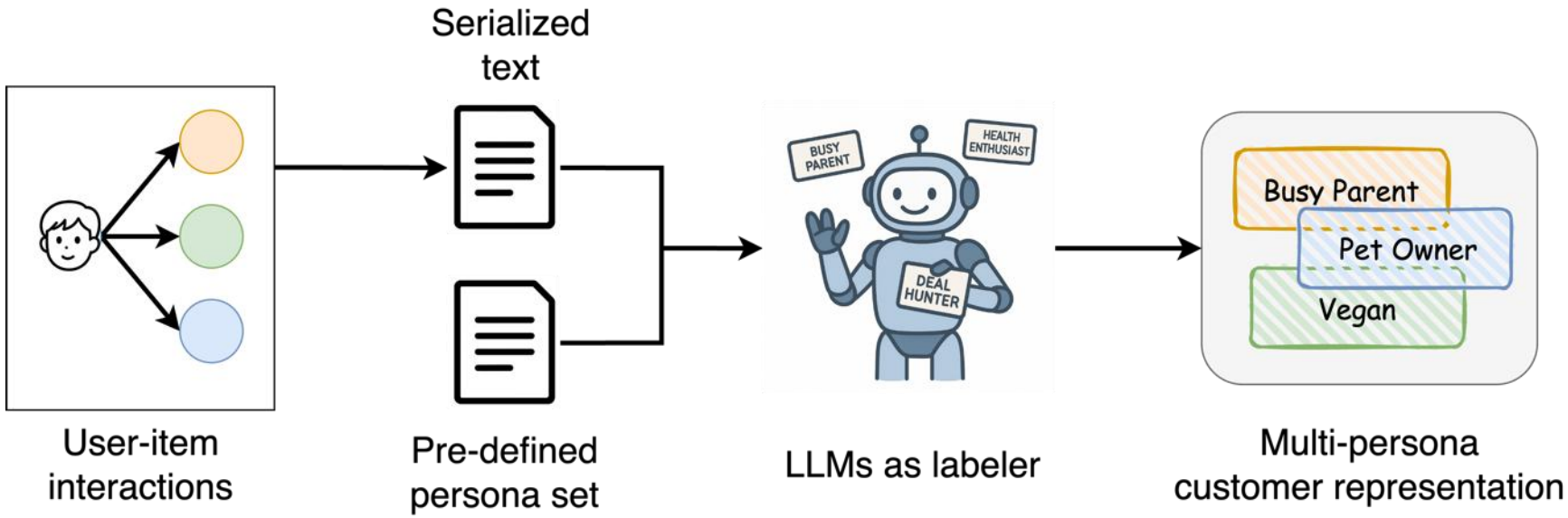
Bargain Hunter

always on the lookout for the best deals and discounts... purchase in bulk to save money...

Properties: Informativeness, Readability, Robustness

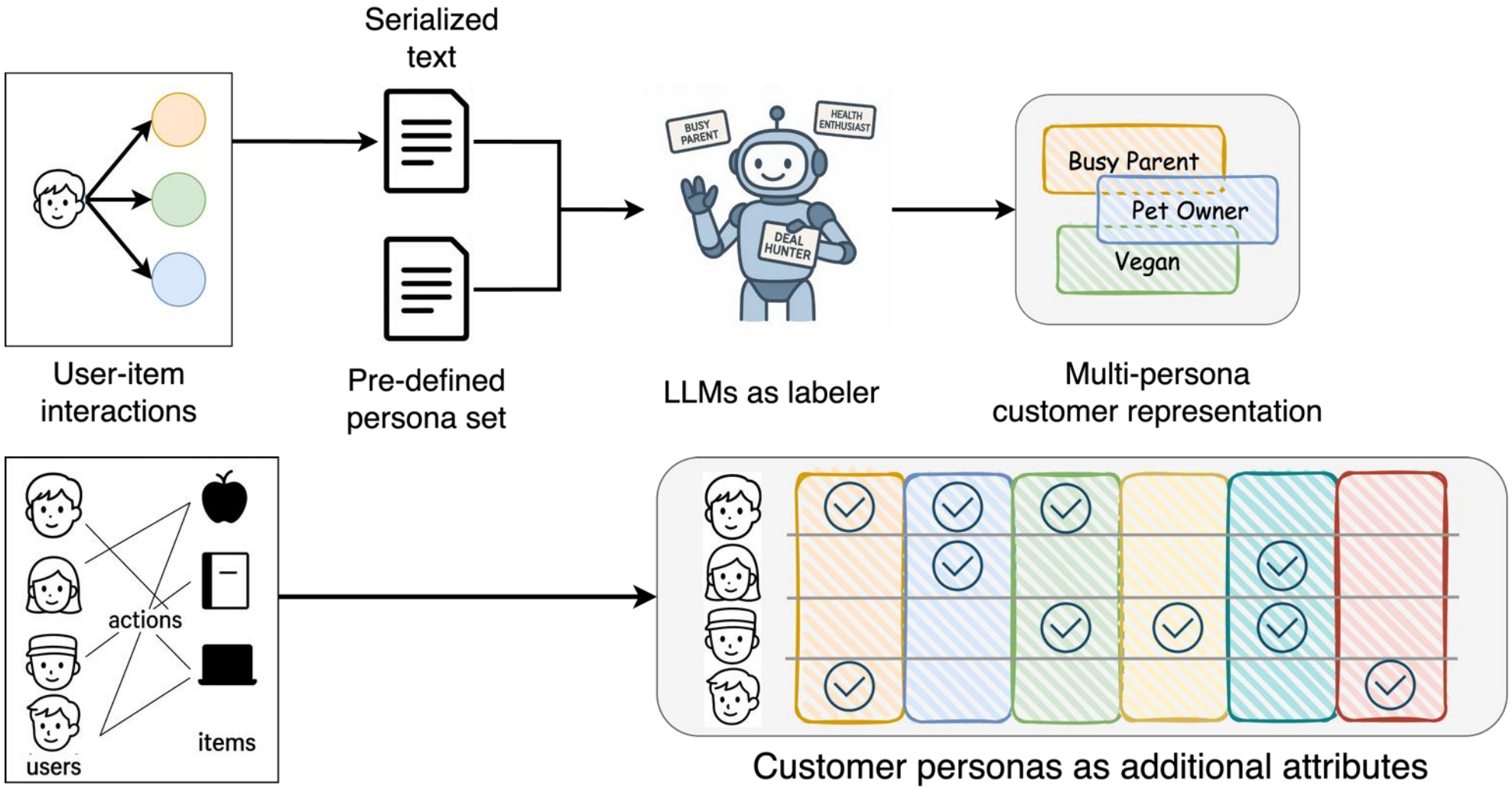
Motivation: Persona Representation

Persona generation pipeline:



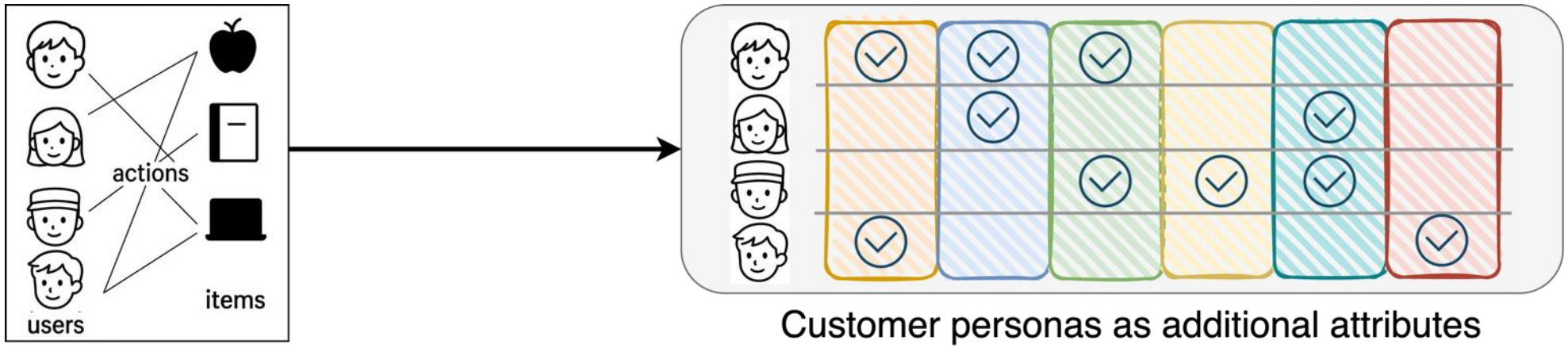
Motivation: Persona Representation

Persona generation pipeline:



Motivation: Persona Representation

Persona generation pipeline:



Downstream applications:

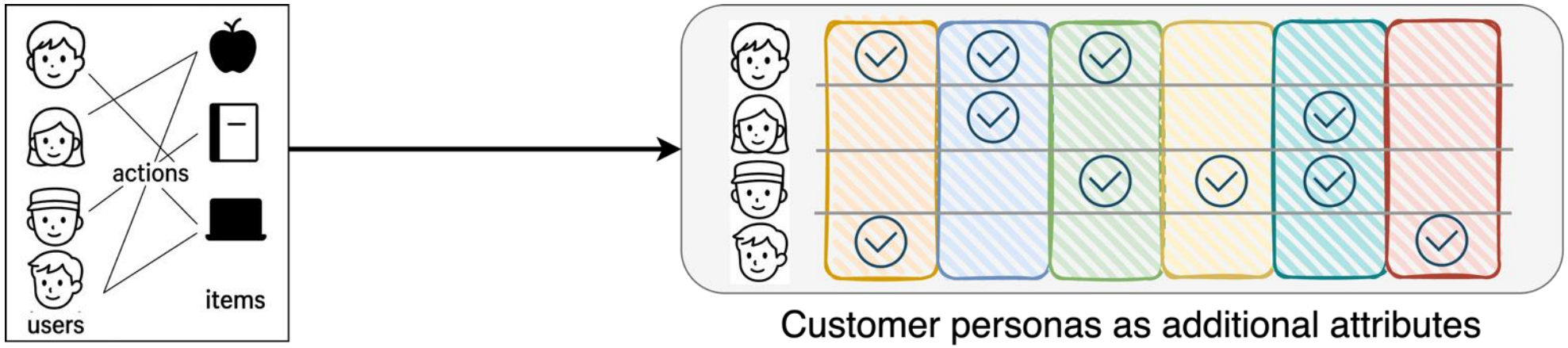
Product Recommendation

Customer Segmentation

Customer-centric Search Navigation

Motivation: Persona Representation

Persona generation pipeline:



Scalability issues:

- e-commerce platforms has millions of users
- user representations need to be dynamically updated
- LLMs are still expensive…

Our method: GPLR

which **G**enerates customers' **P**ersona representation through leveraging **L**arge language models and **R**andom walk-based affinities

Method overview:

Our method: GPLR

which **G**enerates customers' **P**ersona representation through leveraging **L**arge language models and **R**andom walk-based affinities

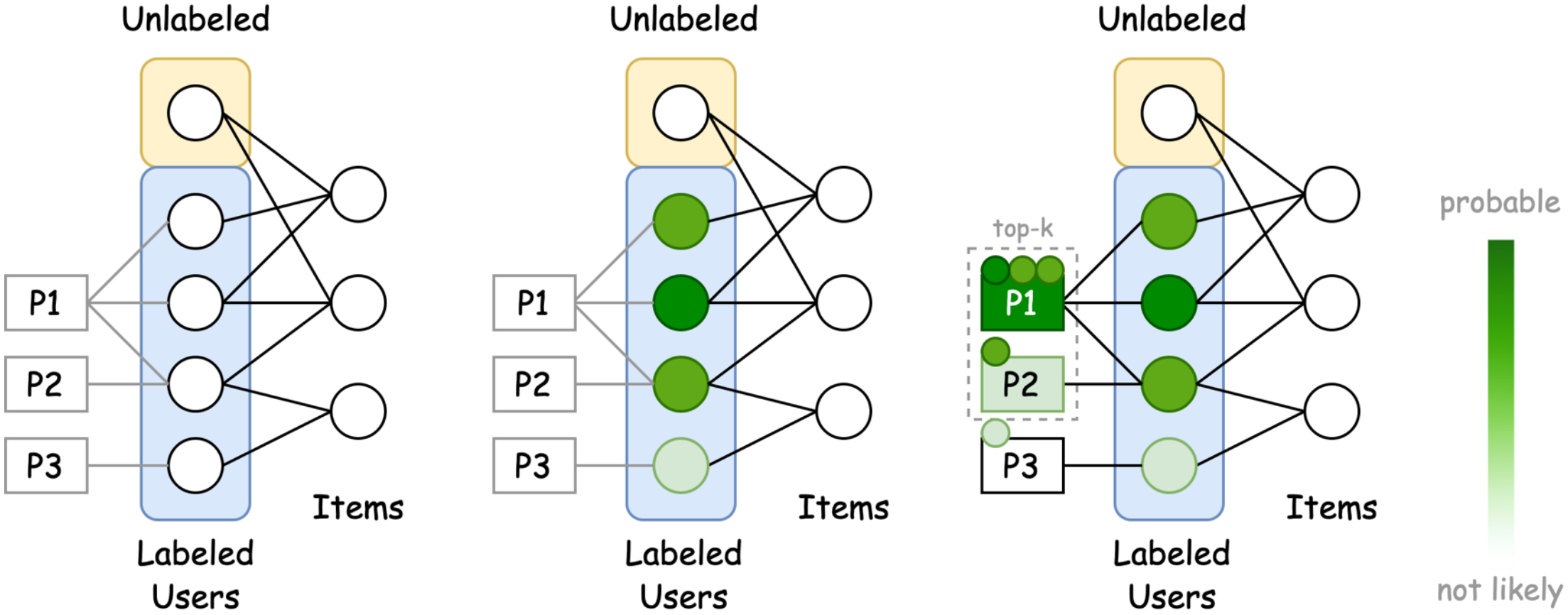
Method overview:

(1) DU-Sampling:

- iteratively sample 5~10% customers, label their personas via LLMs
- prefer customers with minor persona labels (**D**iversity)
- prefer those with greater **U**ncertainty

Our method: GPLR

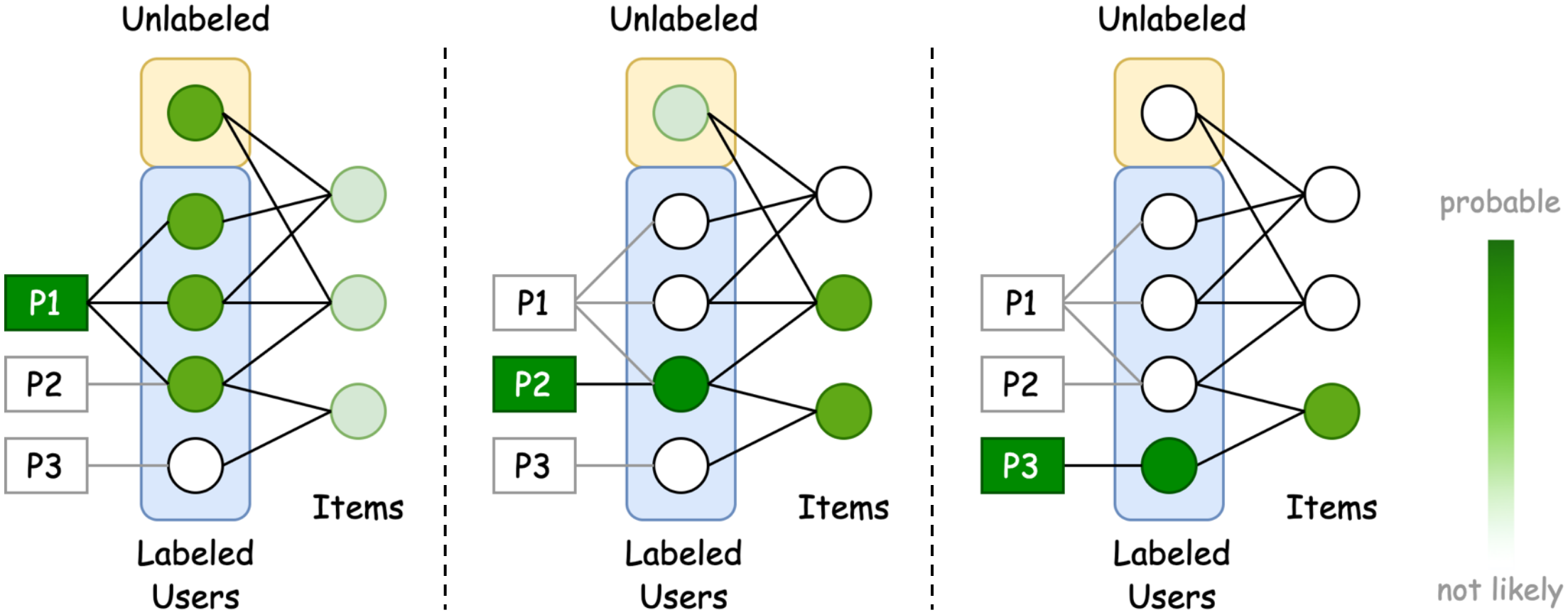
(2) Random walk-based affinities:



- Unlabeled user: find labeled neighbors via random walk
- then aggregate their personal labels

Our method: GPLR

(3) **Approximation:** RevAff



- Reverse random walk + Prune out minor walks
- Lower computational complexity

Application: Product Recommendation

- **Objective:**
  - Recommend relevant new products to customers
- **Solution**:
  - Integrate persona labels generated by GPLR into the original user–item bipartite graph to construct a tripartite graph
  - Apply GNN-based recommendation models on the resulting tripartite graph (e.g., LGCN3, AFDGCF)
- **Competitors**:
  - MF, GCMC, LCFN, LightGCN, LGCN, AFDGCF
- **Datasets:**

Dataset	User#	Item#	Interaction#	Sparsity
OnlineRetail	4,297	3,846	263,267	98.4070%
Instacart	20,620	41,521	1,333,805	99.8442%
Instacart Full	206,209	49,677	13,307,953	99.8701%

Application: Product Recommendation

- **Main results**
  - NDCG@K, F1-Score@K; K={2, 5, 10, 20, 50, 100}
  - LGCN3/LGCN: up to 11.7%; Ours/SOTA: up to 6.4%;

Table 3: Performance evaluation in NDCG@K (N@K) and F1-Score@K (F@K).

	OnelineRetail						Instacart					
	MF	Light	LGCN	AFD	LGCN3	A-LGCN3	MF	Light	LGCN	AFD	LGCN3	A-LGCN3
N@2	0.2391	0.2801	0.2686	0.2898	0.2933	**0.2940**	0.1166	0.1477	0.1405	0.1535	0.1570	**0.1634**
N@5	0.2143	0.2497	0.2356	0.2578	**0.2602**	0.2549	0.1006	0.1273	0.1208	0.1308	0.1319	**0.1357**
N@10	0.2104	0.2443	0.2274	0.2489	**0.2503**	0.2497	0.0916	0.1180	0.1106	0.1205	0.1197	**0.1230**
N@20	0.2221	0.2575	0.2383	0.2595	0.2577	**0.2617**	0.0934	0.1198	0.1117	0.1222	0.1206	**0.1242**
N@50	0.2612	0.2942	0.2772	0.2978	0.2940	**0.2996**	0.1150	0.1448	0.1351	0.1476	0.1451	**0.1491**
N@100	0.3011	0.3337	0.3167	0.3346	0.3325	**0.3392**	0.1389	0.1733	0.1618	0.1761	0.1732	**0.1766**
F@2	0.0859	0.1080	0.0955	0.1066	0.1052	**0.1092**	0.0326	0.0415	0.0375	0.0427	0.0411	**0.0433**
F@5	0.1123	0.1333	0.1207	0.1348	0.1332	**0.1349**	0.0478	0.0609	0.0564	**0.0629**	0.0611	0.0628
F@10	0.1225	0.1416	0.1352	**0.1436**	0.1431	0.1432	0.0560	0.0716	0.0672	0.0730	0.0722	**0.0738**
F@20	0.1239	0.1398	0.1341	0.1393	0.1406	**0.1415**	0.0602	0.0752	0.0710	0.0768	0.0758	**0.0780**
F@50	0.1081	0.1181	0.1159	0.1186	0.1201	**0.1210**	0.0565	0.0684	0.0644	0.0692	0.0681	**0.0701**
F@100	0.0886	0.0955	0.0934	0.0945	0.0967	**0.0976**	0.0470	0.0562	0.0528	0.0568	0.0557	**0.0570**

Application: Product Recommendation

- **Main results**
  - Sample rate ={5%, 10%, 20%, 100%}
  - Even a 5% sample achieves performance comparable to the full

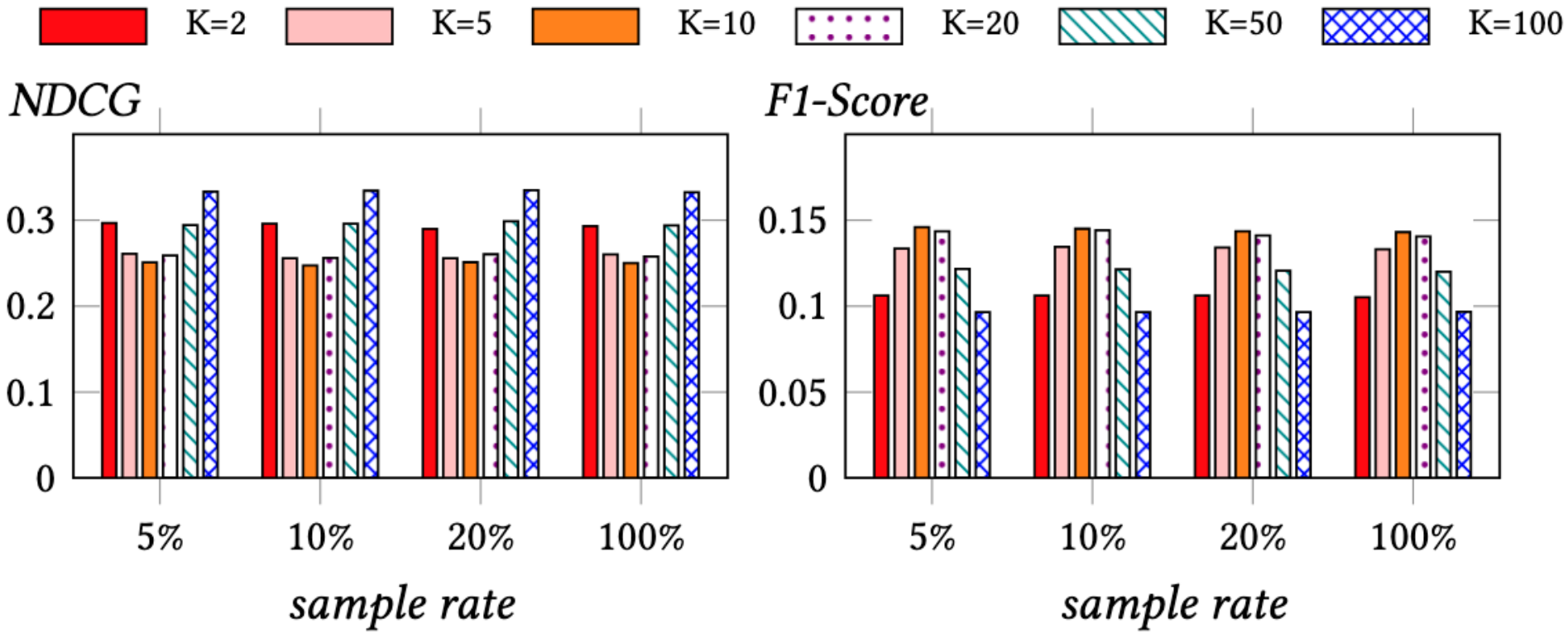


Figure 1: LGCN3 with different sample rates on OnlineRetail.

Application: Customer Segmentation

- **Objective**:
  - Cluster customers into groups with similar attributes
- **Solution**:
  - Encode each customer using a one-hot persona representation
  - Reduce the dimension via PCA, then apply K-Means
- **Competitor**:
  - RFM
- **Evaluation dimensions**:
  - Robustness, cluster quality
- **Datasets**:
  - OnlineRetail
  - sampled 300 customers with at least 10 transactions over all year

Application: Customer Segmentation

- **Main results:**
- **Robustness**
  - consistent: does not change in first and second half-year
  - 13.8×

Table 8: Robustness on OnlineRetail.

Method	# of Consistent Customers
RFM	4 (1.3%)
Persona	54 (18%)

- **Cluster quality**
  - Cluster# ={5, 15, 25, 35}
  - avg. 61.3% better

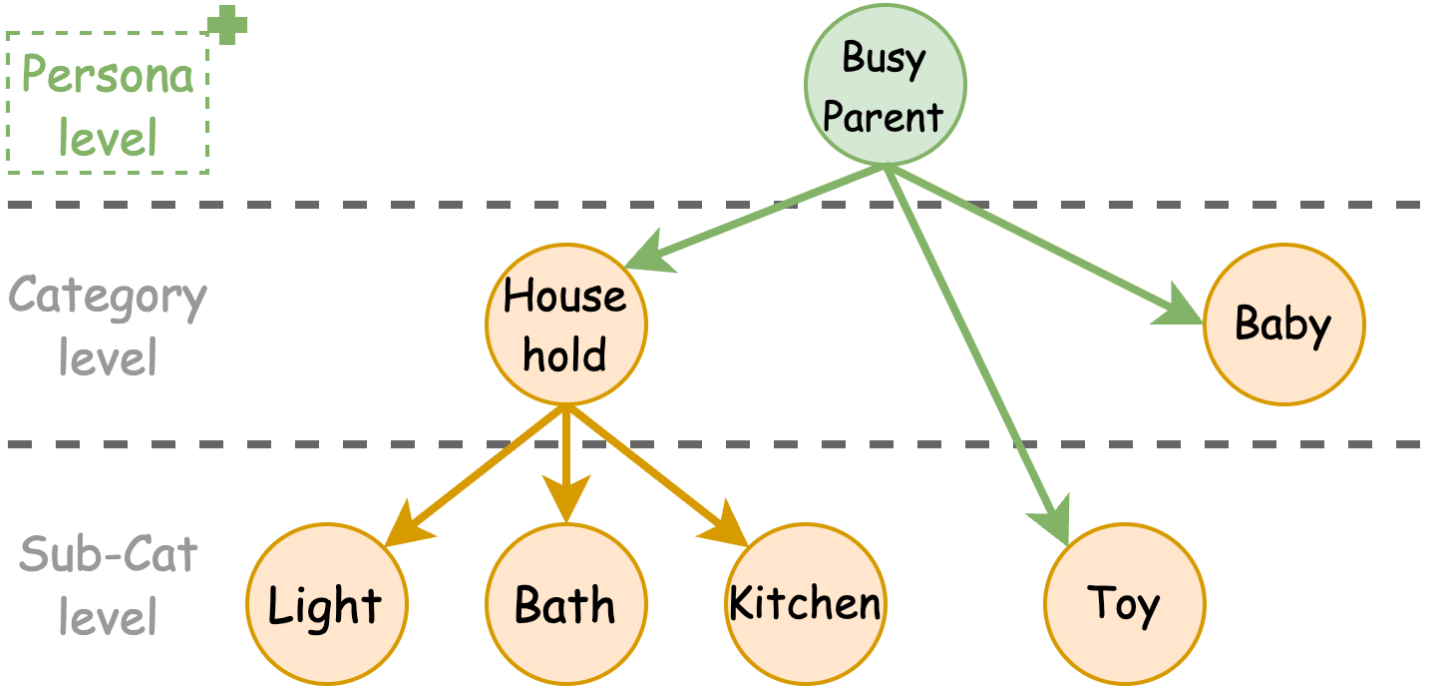
Table 9: Silhouette scores on OnlineRetail.

Clusters#	5	15	25	35
RFM	0.366	0.404	0.431	0.445
Persona	**0.451**	**0.671**	**0.771**	**0.788**

Future work

- **Real-word A/B tests**

- **New applications**
- Customer-centric Search Navigation
  - Add persona-level on the original category-based taxonomy



AI Usage Disclosure:
some images generated by ChatGPT