

TDGE

Human-inspired, Task-Dimension-Guided Exploration for Efficient Learning in High Dimensions

Learning where to explore

Fanyu Zhu, Jiahui An, Ni Ji

NeurIPS 2026

Obstacle in recommendation tasks

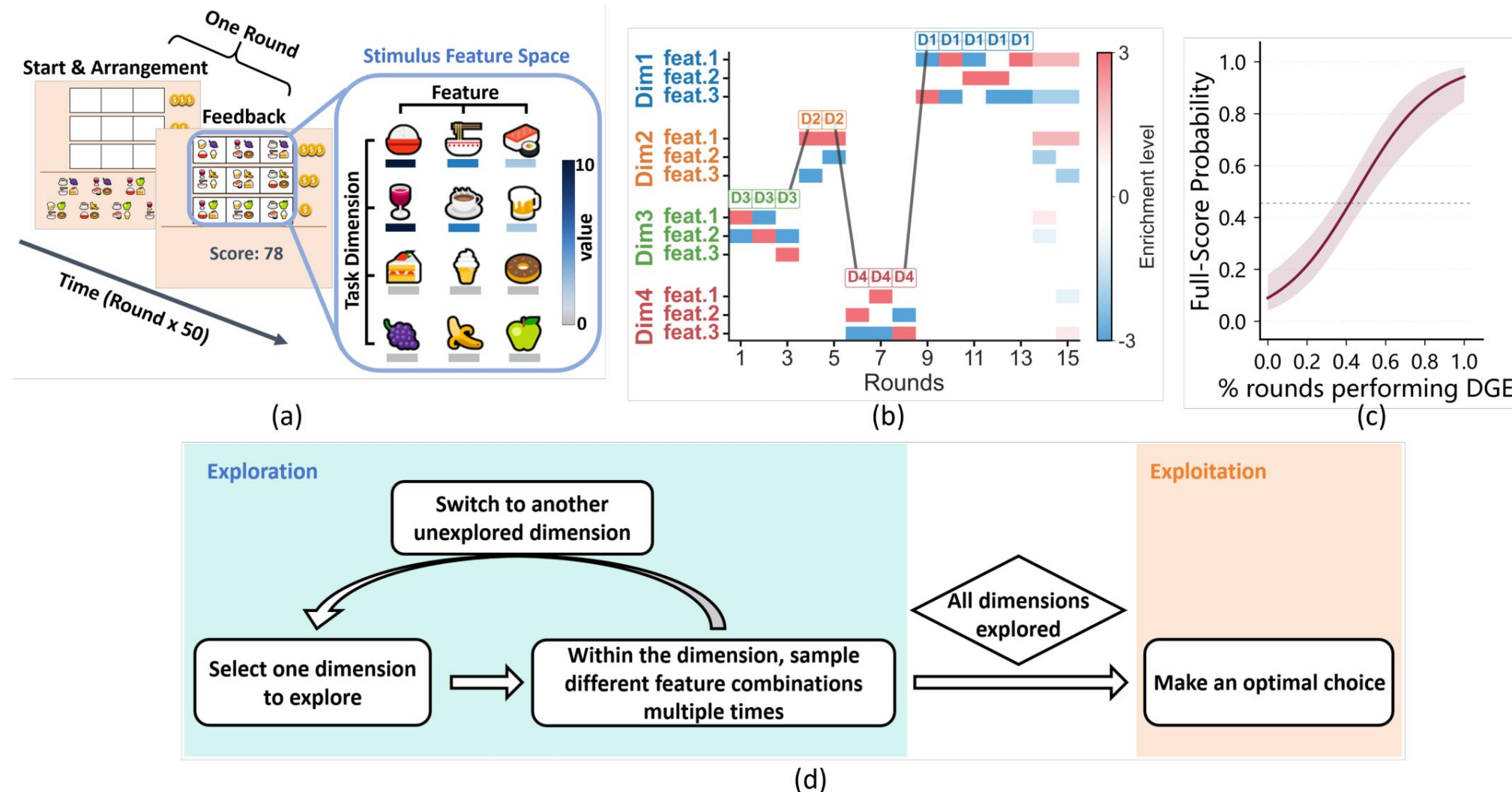
Hard to quickly identify new user's interest in large decision space.

Easy to trap old users in an information cocoon.

Lack interpretable exploration strategy.



Human exploration strategy in large decision space



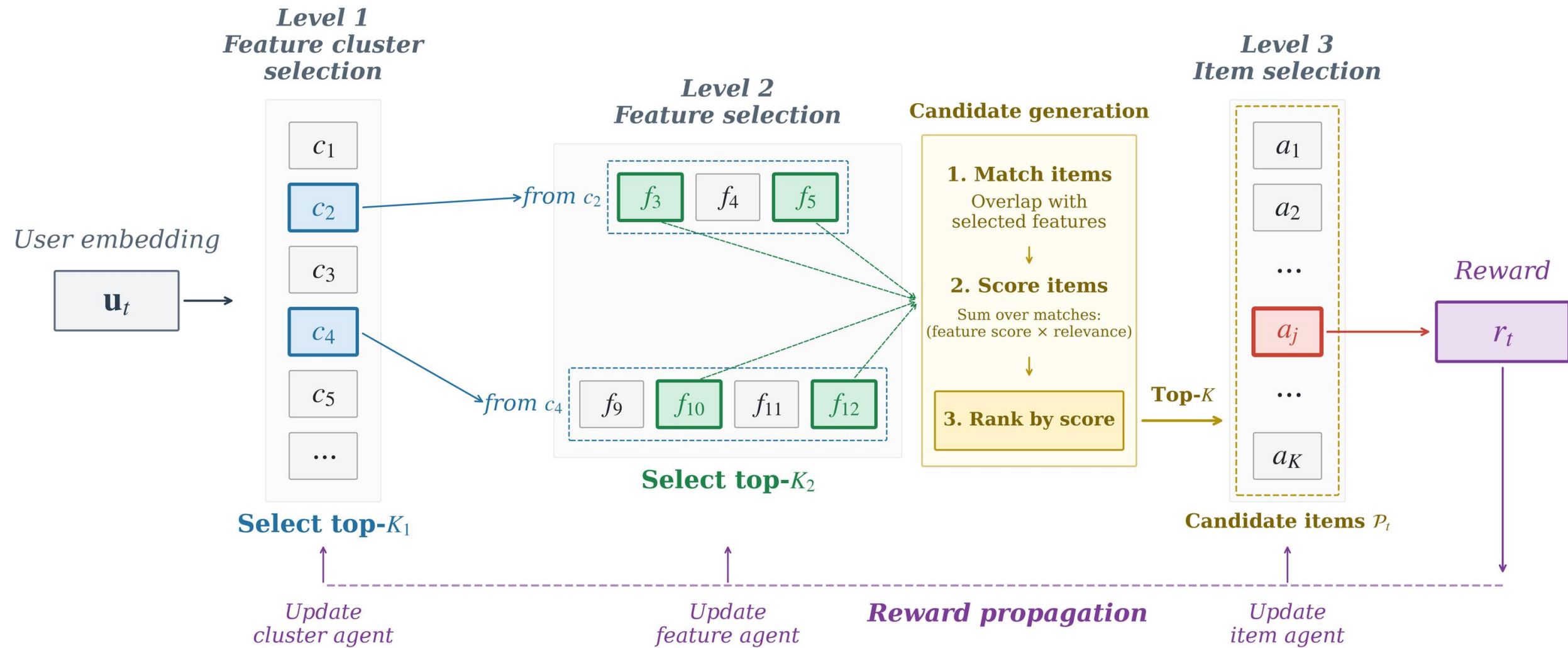
Probe one dimension.

Try several features within it.

Switch exploration dimension and repeat the procedure.

Inspiration: reduce the effective search space by learning which dimensions predict reward.

Human-inspired exploration framework for recommendation task



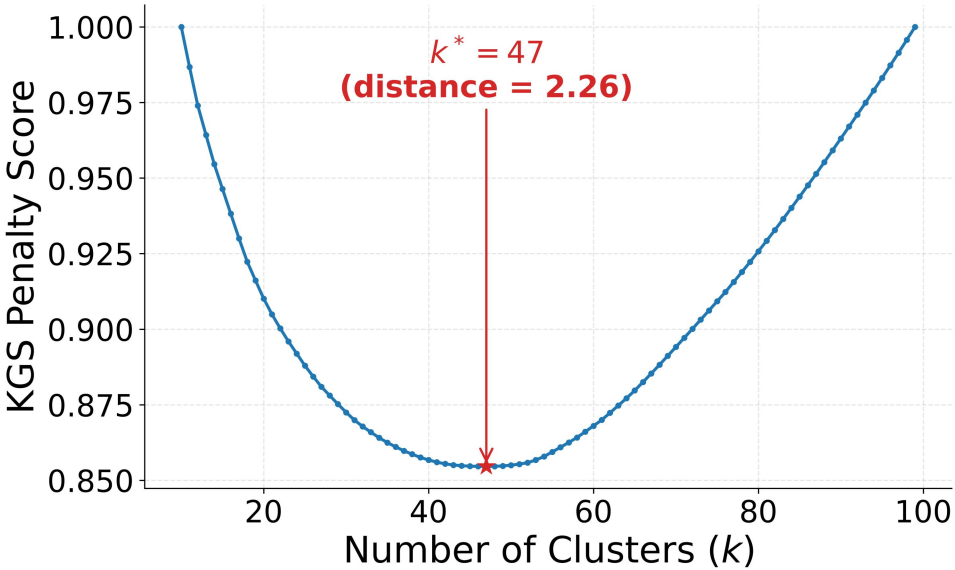
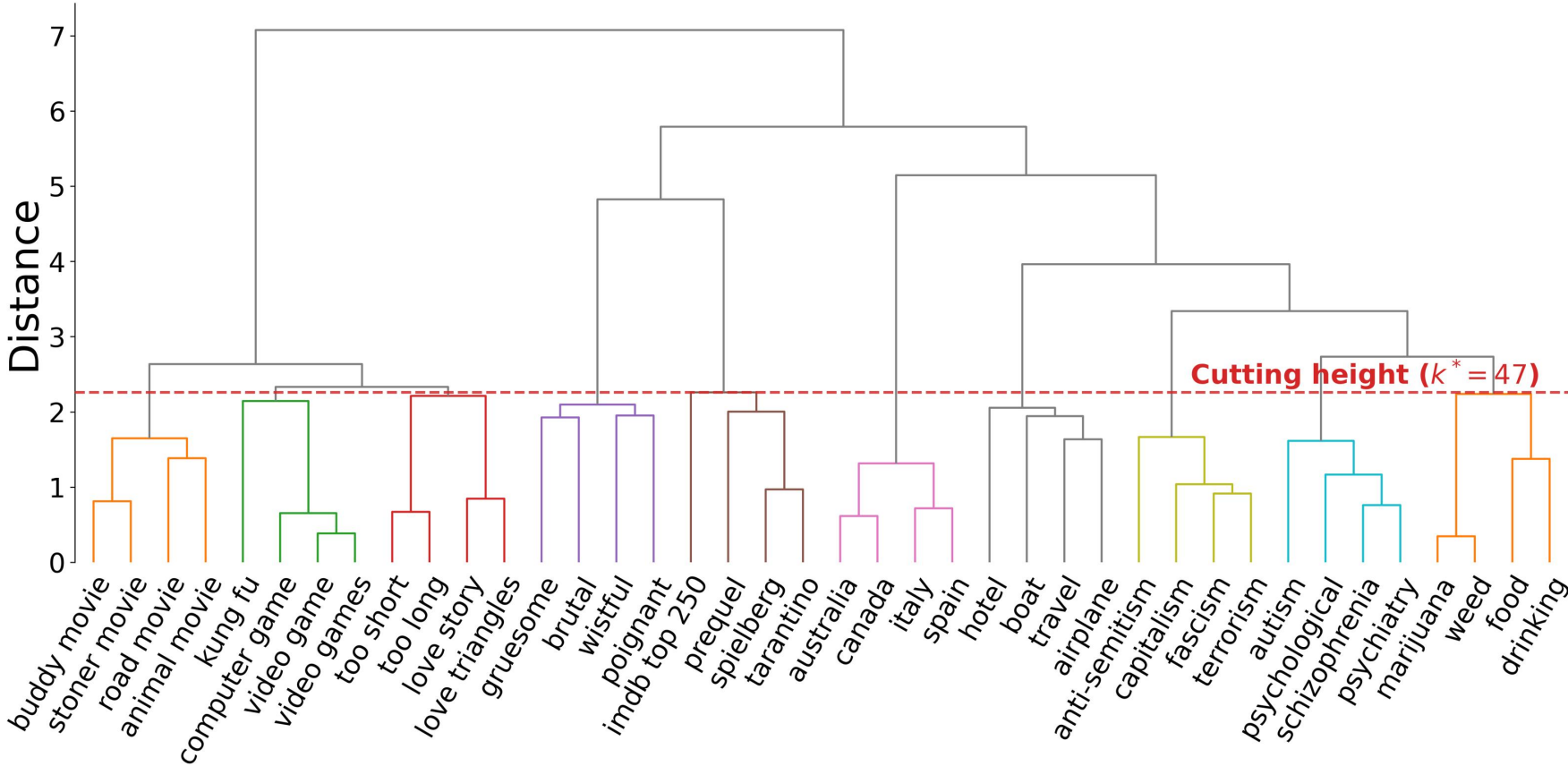
Dimension and feature decisions guide item selection

Building semantic dimensions from metadata

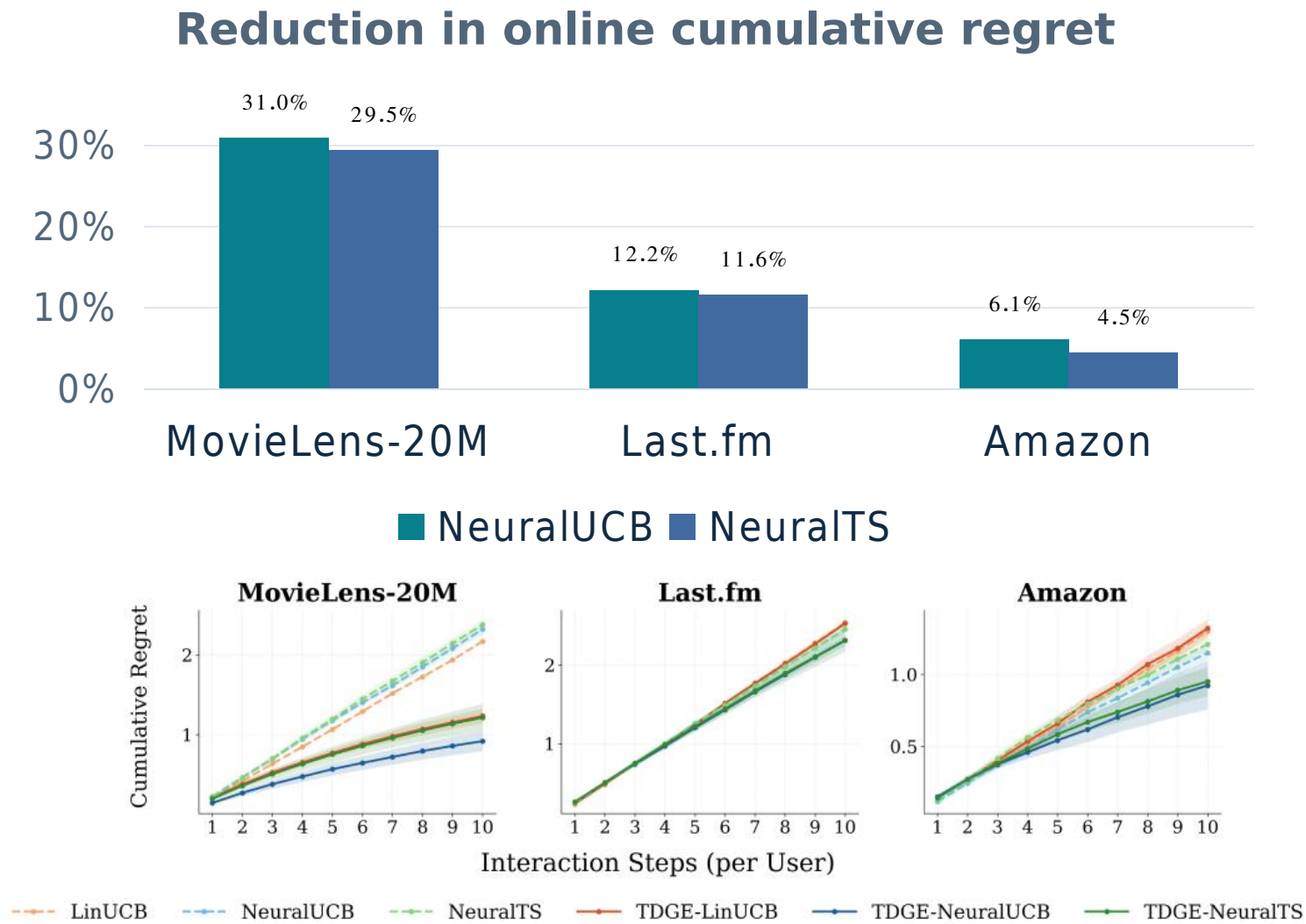
Example of hierarchical clustering of Movielens-20M tags.

Semantic embeddings using pretrained Sentence-Bert model.

Automatic clusters using KGS-inspired criterion.



Lower regret across online training

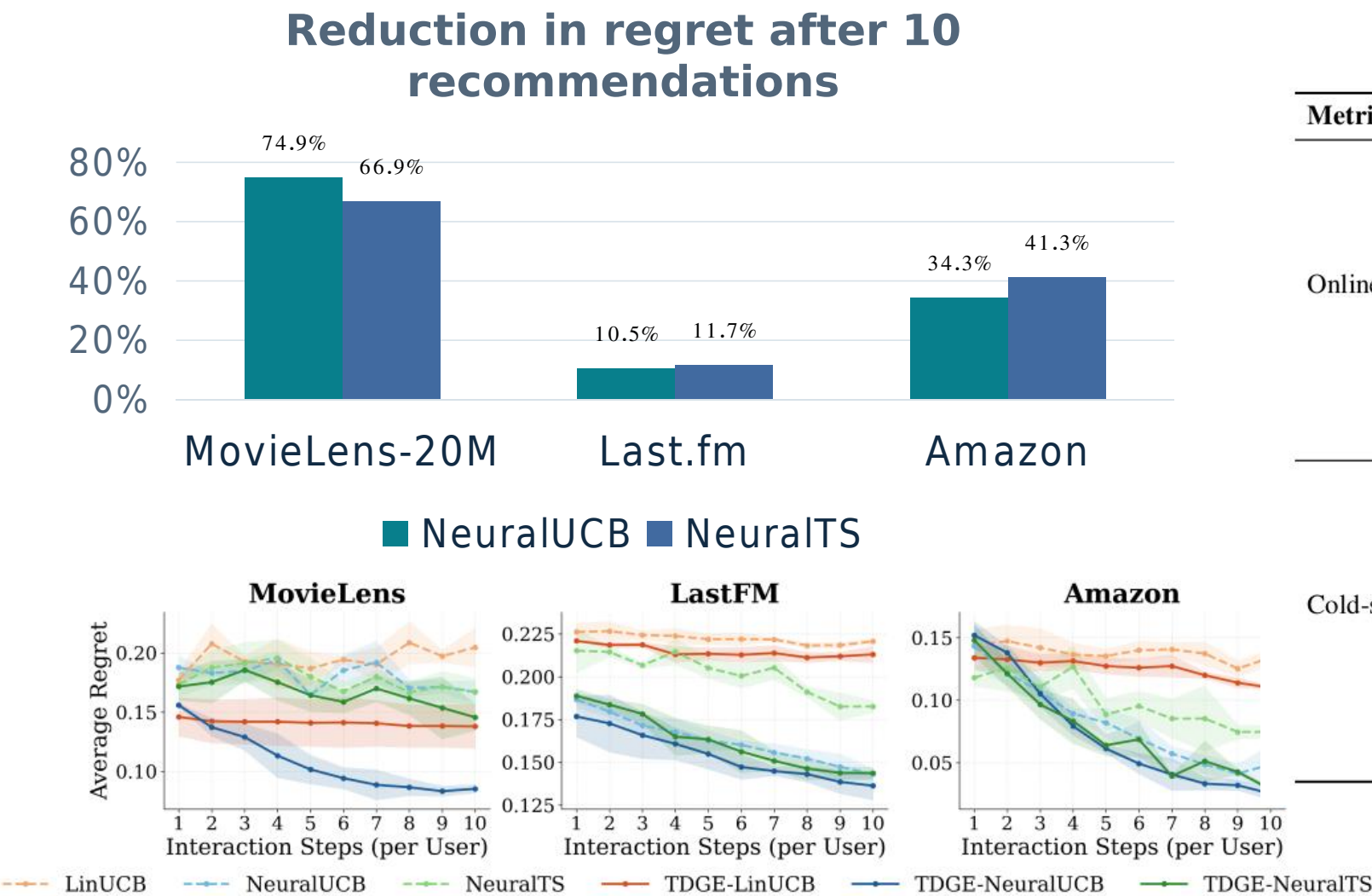


Metric	Dataset	Backbone	Baseline	TDGE
Online CReg ↓	MovieLens-20M	LinUCB	2299.697±20.439	1573.142±43.499***
		NeuralUCB	2417.798±15.161	1669.303±91.162***
		NeuralTS	2418.077±15.964	1705.542±52.769***
	Last.fm	LinUCB	2438.154±43.877	2453.269±27.544 n.s.
		NeuralUCB	2233.387±13.230	1960.563±57.620***
		NeuralTS	2234.352±16.528	1975.365±81.538**
	Amazon	LinUCB	1314.280±30.398	1303.841±29.050 n.s.
		NeuralUCB	1245.925±22.689	1169.567±32.945**
		NeuralTS	1233.443±30.148	1177.886±26.246*
Cold-start Regret ($t = 10$) ↓	MovieLens-20M	LinUCB	0.234±0.004	0.080±0.013***
		NeuralUCB	0.239±0.013	0.060±0.015***
		NeuralTS	0.236±0.014	0.078±0.026***
	Last.fm	LinUCB	0.260±0.006	0.260±0.005 n.s.
		NeuralUCB	0.239±0.007	0.214±0.007***
		NeuralTS	0.240±0.010	0.212±0.024 n.s.
	Amazon	LinUCB	0.137±0.023	0.140±0.028 n.s.
		NeuralUCB	0.099±0.012	0.065±0.022*
		NeuralTS	0.104±0.010	0.061±0.012***

4.5%–31.0% lower online cumulative regret for neural backbones.

10,000 rounds, K = 10, five seeds.

Faster adaptation to new users

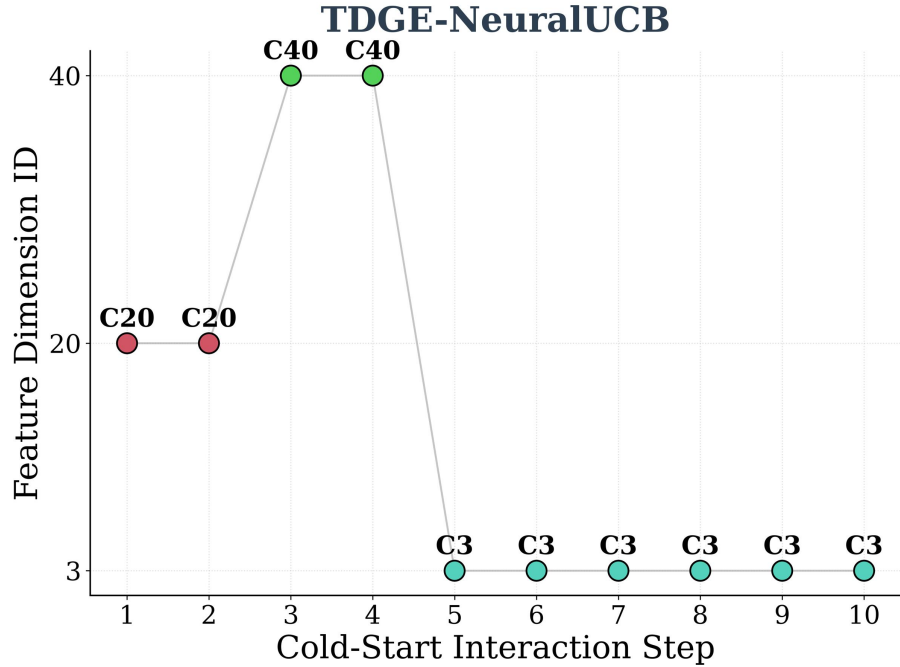


Metric	Dataset	Backbone	Baseline	TDGE
Online CReg ↓	MovieLens-20M	LinUCB	2299.697±20.439	1573.142±43.499***
		NeuralUCB	2417.798±15.161	1669.303±91.162***
		NeuralTS	2418.077±15.964	1705.542±52.769***
	Last.fm	LinUCB	2438.154±43.877	2453.269±27.544 n.s.
		NeuralUCB	2233.387±13.230	1960.563±57.620***
		NeuralTS	2234.352±16.528	1975.365±81.538**
	Amazon	LinUCB	1314.280±30.398	1303.841±29.050 n.s.
		NeuralUCB	1245.925±22.689	1169.567±32.945**
		NeuralTS	1233.443±30.148	1177.886±26.246*
Cold-start Regret ($t = 10$) ↓	MovieLens-20M	LinUCB	0.234±0.004	0.080±0.013***
		NeuralUCB	0.239±0.013	0.060±0.015***
		NeuralTS	0.236±0.014	0.078±0.026***
	Last.fm	LinUCB	0.260±0.006	0.260±0.005 n.s.
		NeuralUCB	0.239±0.007	0.214±0.007***
		NeuralTS	0.240±0.010	0.212±0.024 n.s.
	Amazon	LinUCB	0.137±0.023	0.140±0.028 n.s.
		NeuralUCB	0.099±0.012	0.065±0.022*
		NeuralTS	0.104±0.010	0.061±0.012***

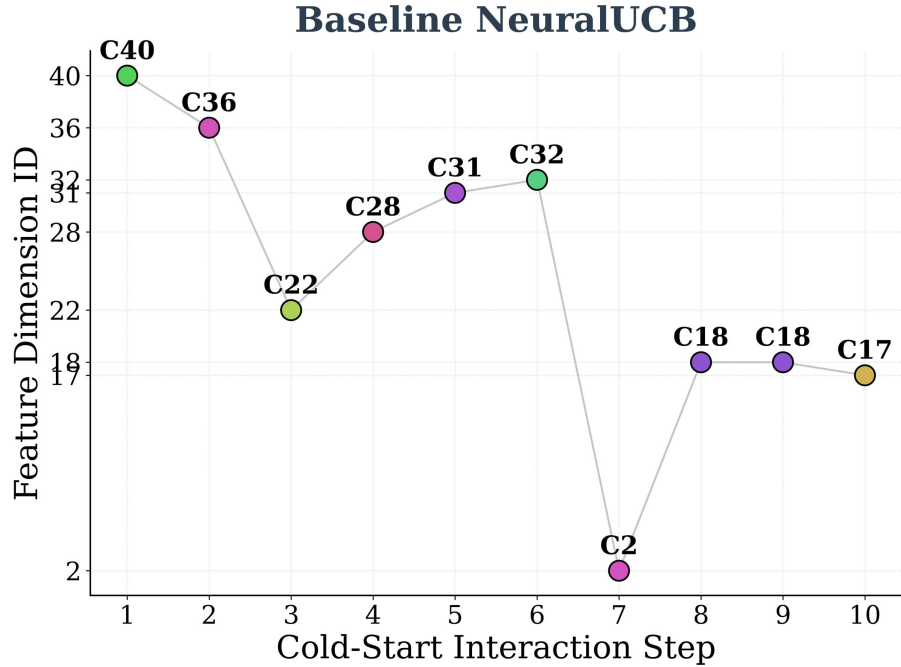
Up to 74.9% less final-round cold-start regret

100 held-out users per seed, five seeds.

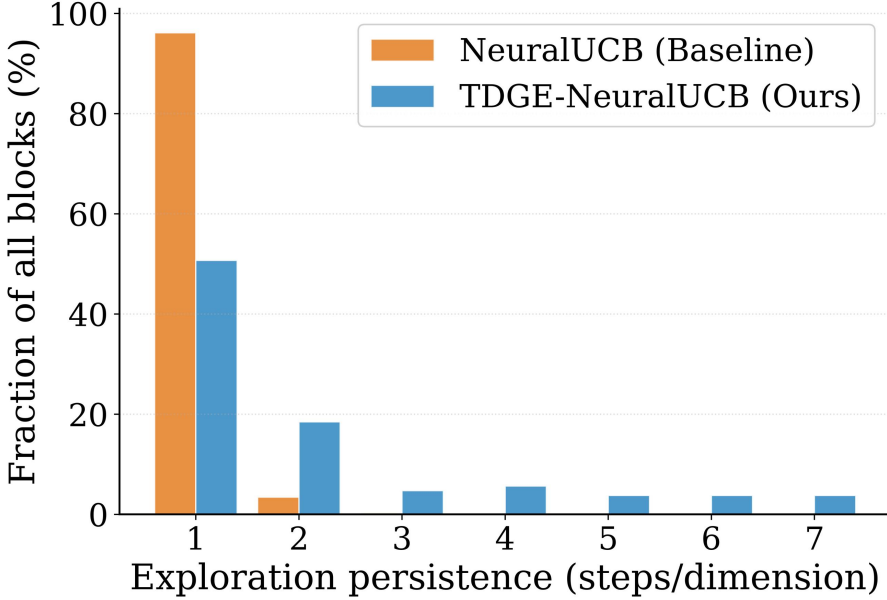
TDGE exhibits human-like dimension-persistent probing



TDGE-NeuralUCB



NeuralUCB

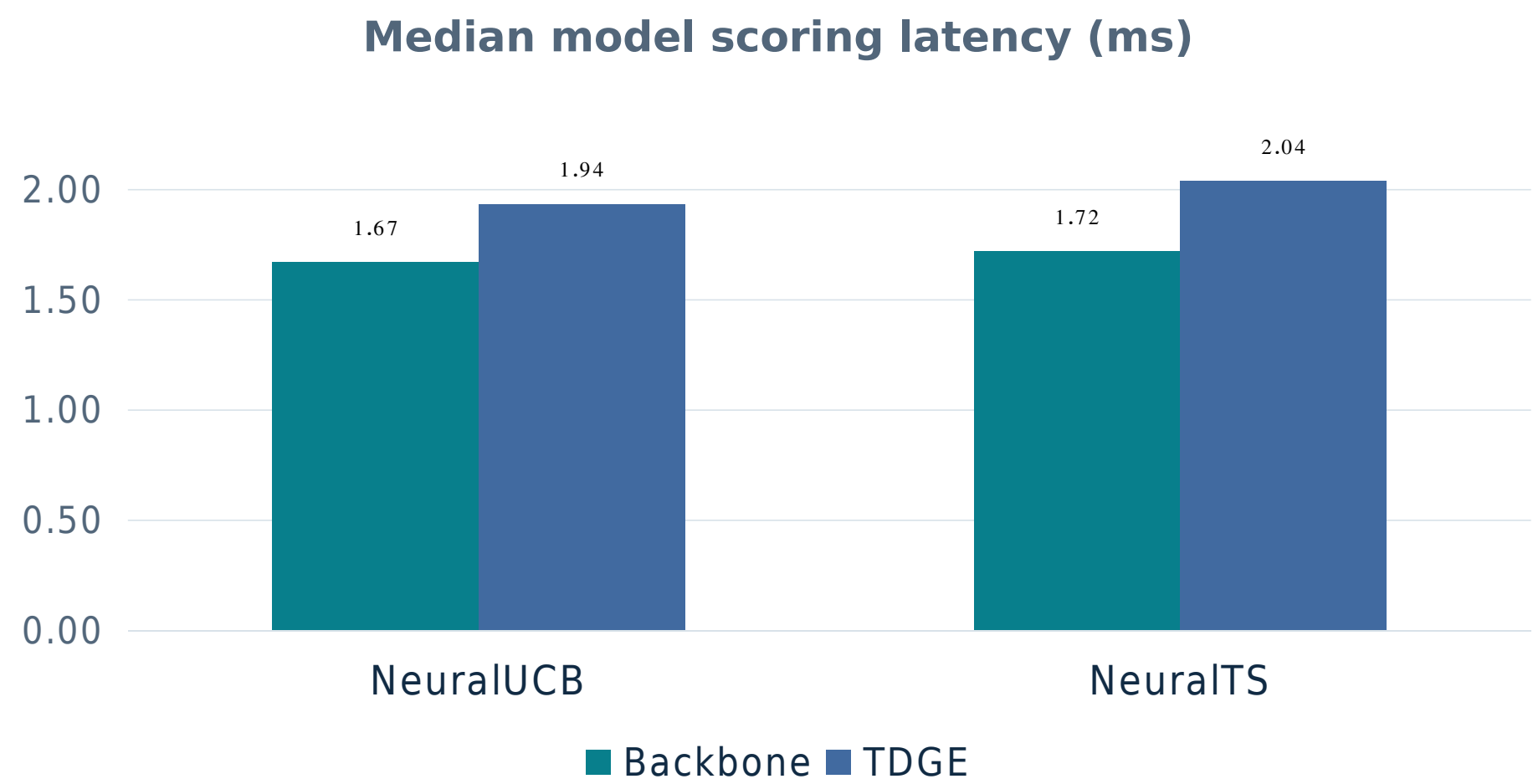


Persistence across users

More within-dimension probing before switching

Tested in new user adaptation period using MovieLens-20M dataset.

Small inference time overhead



Only +0.26 to +0.32 ms overhead for inference.

RTX 4090, 100,000 item arms plus 5,000 dimensions and 5,000 features. Model scoring only.

Contributions and limitations

Our contributions

Model-agnostic framework

TDGE is a model-agnostic framework that can combine with multiple contextual bandit backbones.

High sample efficiency and rapid adaptation

TDGE significantly reduces both online and cold-start adaptation regret on all three real-world datasets.

Robust to the choice of clustering method and hierarchy depth

Ablation studies demonstrate that TDGE is robust to the choice of clustering method and hierarchy depth within the evaluated settings.

Our limitations

Fixed routing budgets

K_1 and K_2 stay fixed during learning. Exploration breadth does not adapt online.

Limited effect of hierarchy depth

The limited effect of hierarchy depth in our ablations may be due to the structure of the evaluated data.

Contextual-bandit evaluation

Test on recommendation task only. Hope to extend to general reinforcement learning algorithms.

Code and implementation

Thank you!

