Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

AIRI SBER AI Lab

Danil Gusak, Anna Volodkevich, Anton Klenitskiy, Alexey Vasilev, Evgeny Frolov

Github

Paper

1 Motivation

Modern sequential recommender systems, ranging from lightweight transformer variants to LLMs, **dominate at next-item prediction** and are widely adopted in academia and industry. However, common **evaluation protocols remain underdeveloped**, often misaligned with real-world scenarios.

Popular **LOO split** aligns with NIP but allows overlap between training and test periods, causing **temporal leakage and unrealistic test horizons**. In contrast, **GTS** better reflects real-world deployment by evaluating on future time periods. Yet its **application to SeqRec is loosely defined**, especially regarding target interaction selection and consistent validation construction.

We show that evaluation strategies significantly **impact model performance rankings and deployment decisions**. To improve reproducibility, we compare splitting strategies across datasets and baselines, revealing that prevalent splits, such as leave-one-out, **may be insufficiently aligned with more realistic evaluation strategies.**

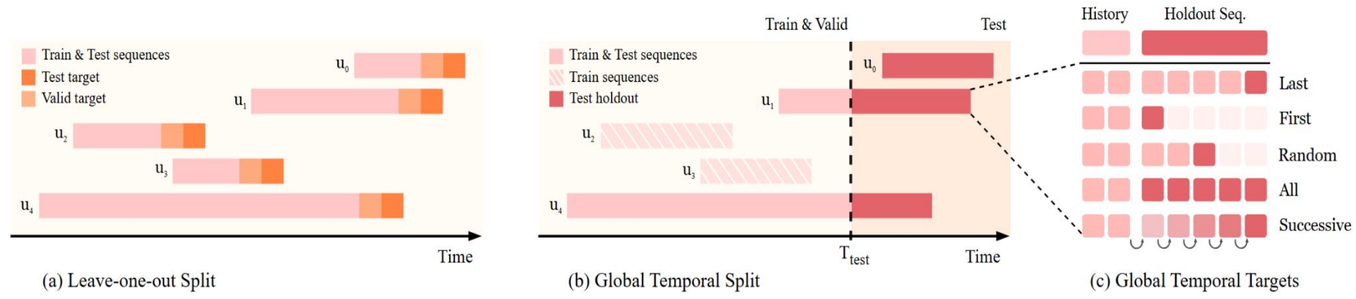


Figure 1: Data splitting and target selection strategies for sequential recommendations. (a) **Leave-one-out split**. (b) **Global temporal split:** all interactions after timepoint T_{test} are placed in the holdout set, targets for these holdout sequences are chosen according to *(c)*. (c) **Target items selection options** for each holdout sequence (applicable for both test and validation sequences).

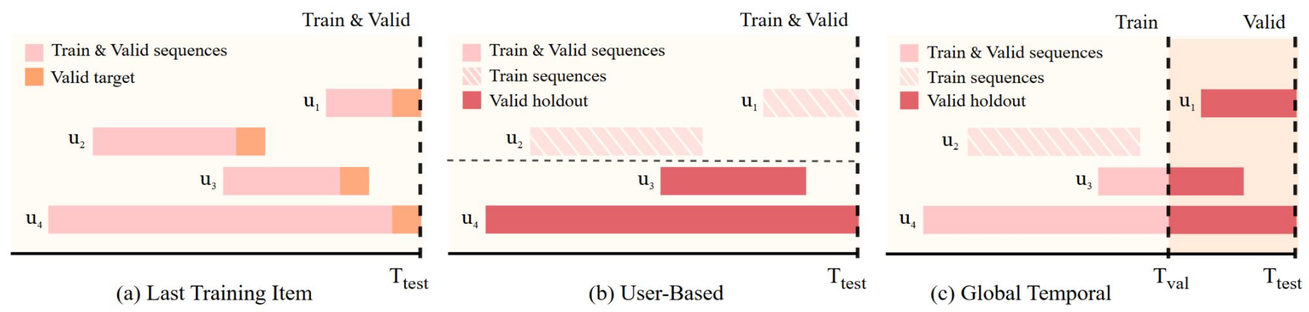
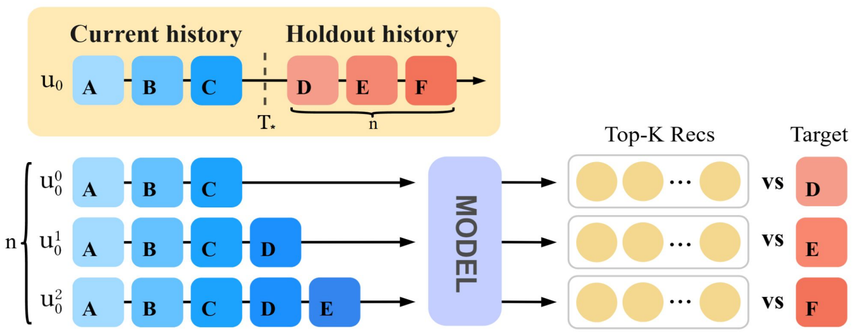


Figure 3: Validation split options for **GTS (Fig. 1b)**: (a) each user' **Last training item** is a target, (b) **User-based**: interactions of n random users are reserved for holdout, (c) **Global temporal**: interactions after T_{val} are reserved for holdout. Targets for holdout sequences are chosen according to **Figure 1c.**

Figure 2: Successive evaluation scheme applied to one user with n = 3 holdout interactions.



2 Research Questions

RQ1 What are the important properties of subsets obtained with different splitting strategies?

RQ2 What is a distribution of time delta between consecutive user interactions, and how does it affect target item selection for GTS?

RQ3 How consistent are recommendation metrics for different splitting strategies in terms of correlation?

RQ4 How do different data splitting strategies influence the final model rankings?

RQ5 Which validation strategies are more appropriate for GTS?

RQ6 How does retraining model on the combined training and val data influence final test performance?

3 RQ1

Table 1: Test subset statistics for GTS for different quantiles →

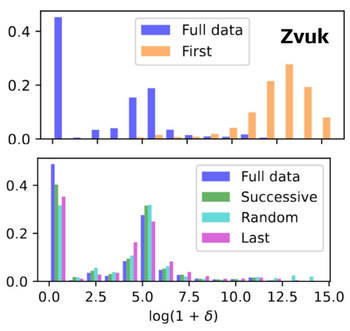
Dataset	Len.	Holdout Len.				#Users (K)					#Days				
	Full	$q_{0.8}$	$q_{0.9}$	$q_{0.95}$	$q_{0.975}$	*Full*	$q_{0.8}$	$q_{0.9}$	$q_{0.95}$	$q_{0.975}$	*Full*	$q_{0.8}$	$q_{0.9}$	$q_{0.95}$	$q_{0.975}$
Beauty	8.88	3.88	3.25	2.76	2.45	22.4	10.2	6.11	3.52	1.91	4,424	138	71	35	19
BeerAdv	101	42.5	28.8	18.7	12.0	14.6	6.94	5.12	3.94	3.07	5,620	354	183	94	48
Diginetica	7.93	7.68	7.66	7.38	6.55	61.3	12.7	6.35	3.29	1.86	152	20	9	4	2
ML-1M	166	112	82.7	61.5	45.6	6.04	1.78	1.21	0.81	0.55	1,038	818	790	617	400
ML-20M	144	126	108	92.8	86.9	139	31.7	18.6	10.8	5.75	7,385	1,994	1,100	569	201
Sports	8.32	3.52	2.89	2.61	2.60	35.6	16.7	10.2	5.63	2.79	4,521	163	88	43	22
YooChoose	8.33	8.49	8.55	8.57	8.79	335	65.8	32.7	16.3	7.94	181	34	17	10	5
Zvuk	420	150	95.9	61.6	42.8	19.3	10.8	8.43	6.57	4.73	91	16	8	4	2

Table 2: Holdout statistics for different splits ($q_{0.9}$ for GTS) →

Set	Split	Stats.↓	Beauty	BeerAdv	Diginetica	ML-1M	ML-20M	Sports	YooChoose	Zvuk
Test	LOO	#Days (%)	84.0	66.9	100	100	94.5	68.1	100	100
		#Users (%)	100	100	100	100	100	100	100	100
	GTS	#Days (%)	1.60	3.26	5.92	76.1	14.9	1.95	9.39	8.79
		#Users (%)	27.3	35.0	10.4	20.0	13.4	28.7	9.74	43.8
		Holdout Len.	3.25	28.8	7.66	82.7	108	2.89	8.55	95.9

4 RQ2

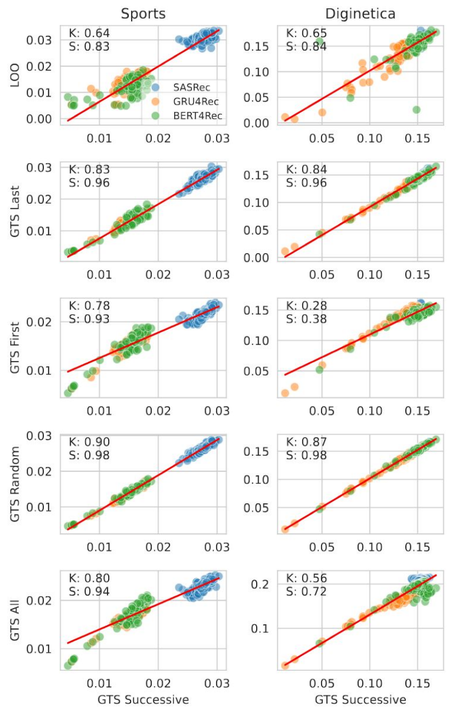
Distributions of time gaps δ between interactions



Set	Setup↓	Beauty	BeerAdv	Diginetica	ML-1M	ML-20M	Sports
Full Data	–	345,600	73,182	58	0	11	172,800
(a) Valid	LOO	172,800	360,900	63	18	17	86,400
	GT Last	1,036,800	446,371	71	27	41	1,209,600
	UB	604,800	691,188	70	15	19	518,400
	LTI	604,800	690,794	70	15	21	518,400
(b) Test	LOO	604,800	737,140	70	17	20	518,400
	Last	1,382,400	508,452	70	67	29	1,296,000
	First	8,640,000	4,921,729	186	7,153,214	21,145,894	11,577,600
	Rand.	3,628,800	439,805	65	35	15	4,752,000
	Succ.	172,800	75,916	58	22	14	86,400

5 RQ3

Figure 3: Scatterplots for NDCG@10 between GTS Sucv. target and other options. K and S denote Kendall and Spearman →



Test Split	Kendall			Spearman		
	HR@10	MRR@10	NDCG@10	HR@10	MRR@10	NDCG@10
LOO	0.71	0.70	0.71	0.87	0.86	0.87
GTS Last	0.83	0.82	0.83	0.93	0.94	0.94
GTS First	0.70	0.60	0.62	0.82	0.70	0.72
GTS Random	**0.91**	**0.90**	**0.91**	**0.98**	**0.98**	**0.98**
GTS All	0.57	0.37	0.43	0.68	0.46	0.53

Table 3: Mean (across datasets) correlations between test GTS Successive target and other options for different metrics.

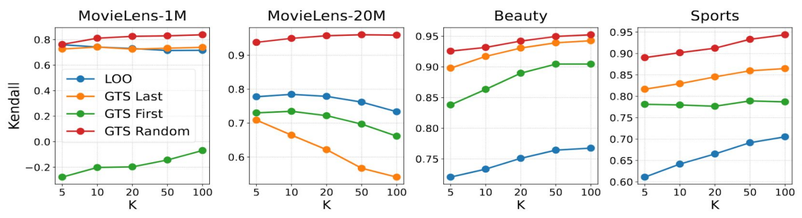
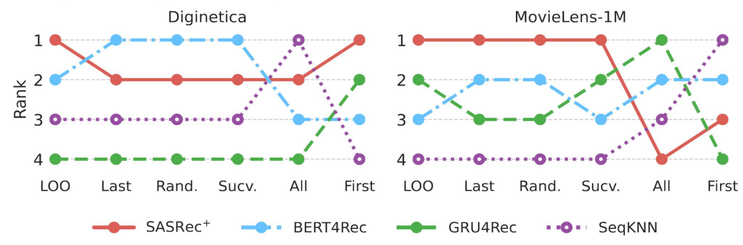


Figure 4: Kendall correlation between test NDCG@K for GTS with Successive target and other options.

6 RQ4

Figure 5: Model rankings based on test NDCG@10 for LOO split, and GTS split with global temporal validation.



7 RQ5

Table 4: Mean (across datasets) correlations between test and validation metrics for GTS with (a) Last and (b) Successive test targets and different validation types. Best values are in **bold**, second best are underlined.

Correlation Target	Valid. Type↓	Kendall HR@10	Kendall MRR@10	Kendall NDCG@10	Spearman HR@10	Spearman MRR@10	Spearman NDCG@10
(a) Test Last	UB	0.72	0.72	0.74	0.87	0.88	0.89
	LTI	0.73	0.75	0.75	0.88	0.90	0.90
	GT Last	**0.78**	**0.79**	**0.79**	**0.93**	**0.93**	**0.93**
	GT First	0.61	0.54	0.57	0.77	0.69	0.73
	GT Rand.	0.75	0.75	0.76	0.90	0.91	0.92
	GT Sucv.	0.76	0.77	0.77	0.91	0.92	0.92
	GT All	0.46	0.37	0.43	0.59	0.50	0.56
(b) Test Sucv.	UB	0.78	0.78	0.80	0.93	0.92	0.94
	LTI	0.80	**0.83**	0.82	0.94	**0.95**	**0.95**
	GT Last	0.81	0.81	0.82	0.94	0.94	**0.95**
	GT First	0.64	0.56	0.59	0.80	0.72	0.75
	GT Rand.	0.80	0.81	0.81	0.94	0.94	0.94
	GT Sucv.	**0.83**	**0.83**	**0.83**	**0.95**	**0.95**	**0.95**
	GT All	0.48	0.37	0.44	0.60	0.49	0.56

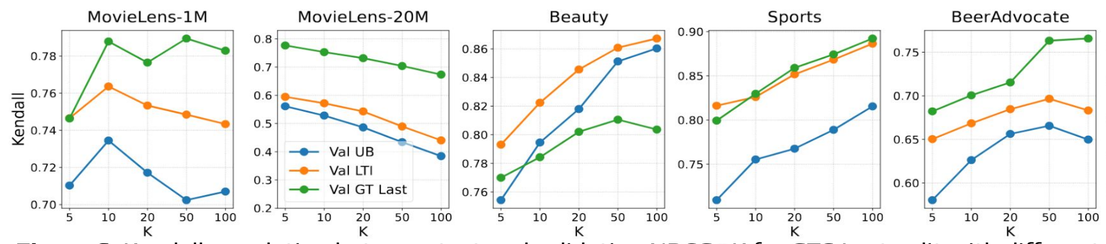


Figure 6: Kendall correlation between test and validation NDCG@K for GTS Last split with different validation strategies.

8 RQ6

Table 5: Validation and test NDCG@10 of SASRec+ at optimal val. config for different splits. Test R. denotes setup with retraining on combined training and validation data. LTI and UB in this study use only Last validation target.

Dataset		Diginetica				Amazon Beauty			
Split	Target↓	Valid	Test	Test R.	Δ Test	Valid	Test	Test R.	Δ Test
GT	Last	0.154	0.154	0.161	4.55%	0.046	0.024	0.037	54.2%
	Sucv.	0.154	0.149	0.160	7.38%	0.044	0.022	0.040	81.8%
UB	Last	0.180	0.152	0.155	1.97%	0.074	0.036	0.037	2.78%
	Sucv.	–	0.159	0.158	-0.63%	–	0.040	0.040	0.00%
LTI	Last	0.187	0.135	0.126	-6.67%	0.067	0.031	0.036	16.1%
	Sucv.	–	0.147	0.129	-12.2%	–	0.036	0.039	8.33%
LOO	Last	0.179	0.181	0.157	-13.3%	0.073	0.059	0.065	10.2%

Key Takeaways

- **LOO** split often misaligns with real-world scenarios and **can distort model rankings**
- **GTS All** target option suffers from a full **task mismatch** with standard next-item prediction
- **GTS First** exhibits **weak correlation** with more realistic evaluation strategies due to significant shifts in time-gap distributions between interactions
- **GTS with Last or Random target yields strong agreement with the more complex but close-to-reality Successive evaluation scheme**

RecSys 2025 conference