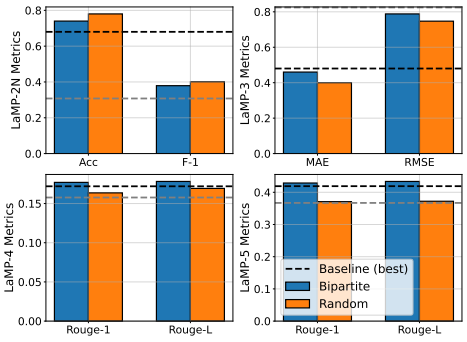

Figure R1: Experiments on LaMP benchmark of **significant disparities in user behavior history**, training over a mixture of 50% most active users and 50% most inactive users (blue) compared to the random selection (orange), respectively. The black and gray dashed lines represent the best-performing baselines on first and second metrics.

Dataset (→)	**MovieLens-1M**		**Recipe**	
Method (↓)	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
`gpt-3.5-turbo`	0.780	1.208	0.880	1.149
ICL-Random (k=1)	0.880	1.2649	0.960	1.233
ICL-Random (k=2)	0.800	1.095	0.980	1.192
ICL-Random (k=4)	0.820	1.225	0.760	1.114
RAG (k=1)	0.800	1.166	0.980	1.192
RAG (k=2)	0.700	1.030	0.880	1.114
RAG (k=4)	0.640	1.000	0.820	1.049
PAG (k=0)	0.820	1.147	0.860	1.128
PAG (k=1)	0.760	1.010	0.800	1.063
`HYDRA`	**0.600**	**0.908**	**0.720**	**0.938**

Table R1: Experiments on **two additional personalization datasets, MovieLens-1M and Recipe,** focusing on predicting users' personal ratings for movies or recipes based on their historical rating patterns.

Dataset (→)	**LaMP-2N**		**LaMP-2M**		**LaMP-3**		**LaMP-4**			**LaMP-5**		
Method (↓)	Acc. ↑	F-1 ↑	Acc. ↑	F-1 ↑	MAE ↓	RMSE ↓	R-1 ↑	R-L ↑	BLEU ↑	R-1 ↑	R-L ↑	BLEU ↑
`gpt-3.5-turbo`	0.638	0.499	0.412	0.347	0.540	0.851	0.133	0.119	1.043	0.439	0.371	6.018
ICL-Random (k=1)	0.598	0.476	0.392	0.335	0.676	0.959	0.147	0.132	1.330	0.457	0.396	8.118
ICL-Random (k=2)	0.632	0.499	0.376	0.311	0.562	0.871	0.151	0.137	2.388	0.451	0.393	8.550
ICL-Random (k=4)	0.630	0.518	0.392	0.352	0.440	0.740	0.161	0.146	2.418	0.457	0.396	8.404
RAG (k=1)	0.610	0.486	0.408	0.345	0.602	0.871	0.154	0.138	1.649	0.468	0.405	7.820
RAG (k=2)	0.624	0.479	0.380	0.315	0.559	0.836	0.161	0.149	2.958	0.480	0.419	9.021
RAG (k=4)	0.656	0.524	0.392	0.339	0.391	0.716	0.167	0.155	3.615	0.479	0.418	9.108
PAG (k=0)	0.618	0.489	0.404	0.340	0.583	0.872	0.161	0.141	1.950	0.460	0.405	7.372
PAG (k=1)	0.630	0.500	0.418	0.357	0.414	0.787	0.163	0.153	2.934	0.474	0.414	8.372
`HYDRA`	**0.748**	**0.551**	**0.446**	**0.373**	**0.328**	**0.656**	**0.175**	**0.167**	**4.772**	**0.508**	**0.442**	**9.519**

Table R2: **Scale-up experiment results** on the LaMP benchmark, including 1000 and 500 users during training and testing, respectively.

Method	**Mode**	**Time Complexity**	**LaMP-2N**	**LaMP-2M**	**LaMP-3**	**LaMP-4**	**LaMP-5**
`HYDRA`-Reranker	Training	$O(N_{\text{train}}(M^2+1)TL^2d)$	31m10s	41m51s	50m37s	1h1m31s	1h8m16s
`HYDRA`-Reranker	Fit New User	$O(N_{\text{test}}(M^2+1)TL^2d)$	18m8s	21m17s	25m36s	33m52s	31m25s
`HYDRA`-Reranker	Inference	$O(N_{\text{test}}L^2d)$	3m4s	3m1s	3m7s	4m38s	5m14s
`HYDRA`-Adapter	Training	$O(N_{\text{train}}k\bar{H}TL^2d)$	1h10m17s	2h2m16s	2h1m59s	3h56m47s	3h19m42s
`HYDRA`-Adapter	Fit New User	$O(N_{\text{test}}k\bar{H}TL^2d)$	28m15s	1h7m27s	1h19s	2h23m10s	1h59m2s
`HYDRA`-Adapter	Inference	$O(N_{\text{test}}kL^2d)$	4m16s	4m17s	4m17s	5m53s	5m59s

Table R3: **Time complexity analysis with running time summary** on the LaMP benchmark.

Dataset (→)	**LaMP-2N**		**LaMP-2M**		**LaMP-3**		**LaMP-4**			**LaMP-5**		
Method (↓)	Acc. ↑	F-1 ↑	Acc. ↑	F-1 ↑	MAE ↓	RMSE ↓	R-1 ↑	R-L ↑	BLEU ↑	R-1 ↑	R-L ↑	BLEU ↑
`gpt-3.5-turbo` (Random)	0.680	0.287	0.460	0.316	0.600	0.872	0.140	0.128	1.303	0.380	0.340	5.516
ICL (Random)	0.660	0.278	0.480	0.353	0.580	0.921	0.153	0.141	1.494	0.375	0.320	6.139
RAG (Random)	0.640	0.284	0.480	0.349	0.640	0.956	0.171	0.156	1.710	0.420	0.363	8.001
`HYDRA` (Random)	0.700	0.316	0.460	0.381	0.500	0.838	0.172	0.158	1.836	0.394	0.342	5.257
`HYDRA` (SC)	0.740	0.360	0.500	0.394	0.420	0.762	-	-	-	-	-	-
`HYDRA`	**0.780**	**0.401**	**0.540**	**0.458**	**0.400**	**0.747**	**0.178**	**0.169**	**2.396**	**0.434**	**0.372**	**7.531**

Table R4: Effect of inherent randomness in LLM generation. We also add two additional baselines by substituting the `HYDRA`-Adapter with random selection (random) and self-consistency (SC).