Table 1: Comparison of DPR, BM-25, LLM Reranking and GPR-LLM (RBF kernel, $\epsilon = 0.1$) with all-MiniLM-L6-v2 encoder across four datasets (TravelDest, POINTREC, Yelp Restaurant, TripAdvisor Hotel) at varying budget of LLM labels. Underline indicates statistically significant improvements over the best-performing baseline (paired t-test, $p < 0.05$).

Budget	Method	TravelDest				POINTREC				Yelp Restaurant				TripAdvisor Hotel			
		P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30
N/A	DPR	0.360	0.366	0.314	0.332	0.164	0.179	0.104	0.182	0.346	0.362	0.282	0.331	0.231	0.297	0.166	0.365
	BM25	0.234	0.238	0.237	0.239	0.025	0.032	0.025	0.038	0.309	0.327	0.236	0.283	0.205	0.257	0.153	0.325
25	LLM Reranking	0.294	0.309	0.238	0.219	**0.175**	**0.206**	0.106	0.158	0.324	0.374	0.237	0.260	0.251	0.349	0.178	0.332
	GPR-LLM	**0.376**	**0.401**	**0.340**	**0.364**	0.157	0.200	**0.108**	**0.177**	**0.360**	**0.402**	**0.273**	**0.340**	**0.282**	**0.382**	**0.182**	**0.426**
50	LLM Reranking	0.356	0.371	0.248	0.281	0.196	0.234	0.105	0.208	0.386	0.426	0.254	0.332	0.289	0.397	0.169	0.417
	GPR-LLM	**0.432**	**0.445**	**0.362**	**0.389**	**0.236**	**0.262**	**0.113**	**0.222**	**0.408**	**0.451**	**0.304**	**0.377**	**0.324**	**0.439**	**0.197**	**0.482**
100	LLM Reranking	0.398	0.406	0.296	0.328	0.225	0.255	0.124	0.231	0.418	0.459	0.298	0.379	0.322	0.438	0.189	0.472
	GPR-LLM	**0.448**	**0.472**	**0.380**	**0.412**	**0.246**	**0.269**	**0.130**	**0.239**	**0.444**	**0.487**	**0.334**	**0.413**	**0.358**	**0.481**	**0.218**	**0.529**

Table 2: Comparison of DPR, BM-25, LLM Reranking and GPR-LLM (RBF kernel, $\epsilon = 0.1$) with msmarco-distilbert-base-tas-b encoder across four datasets (TravelDest, POINTREC, Yelp Restaurant, TripAdvisor Hotel) at varying budget of LLM labels. Underline indicates statistically significant improvements over the best-performing baseline (paired t-test, $p < 0.05$).

Budget	Method	TravelDest				POINTREC				Yelp Restaurant				TripAdvisor Hotel			
		P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30	P@10	NDCG@10	P@30	NDCG@30
N/A	DPR	0.358	0.365	0.318	0.333	0.143	0.161	0.094	0.165	0.363	0.385	0.294	0.351	0.224	0.275	0.165	0.349
	BM25	0.234	0.238	0.237	0.239	0.025	0.032	0.025	0.038	0.309	0.327	0.236	0.283	0.205	0.257	0.153	0.325
25	LLM Reranking	0.344	0.379	0.313	0.339	0.154	0.188	0.094	0.178	0.340	0.382	**0.289**	**0.355**	0.227	0.302	0.165	0.369
	GPR-LLM	**0.370**	**0.402**	**0.325**	**0.356**	**0.159**	**0.192**	**0.098**	**0.183**	**0.359**	**0.397**	0.272	0.343	**0.286**	**0.378**	**0.177**	**0.419**
50	LLM Reranking	0.380	0.419	0.298	0.339	0.171	0.215	0.098	0.193	0.368	0.410	0.282	0.355	0.251	0.336	0.166	0.389
	GPR-LLM	**0.416**	**0.453**	**0.348**	**0.386**	**0.176**	**0.217**	**0.103**	**0.197**	**0.376**	**0.414**	**0.286**	**0.361**	**0.325**	**0.423**	**0.199**	**0.470**
100	LLM Reranking	0.428	0.467	0.323	0.371	0.214	0.245	0.102	0.200	0.391	0.432	0.285	0.366	0.273	0.366	0.170	0.408
	GPR-LLM	**0.444**	**0.479**	**0.344**	**0.389**	**0.215**	**0.246**	**0.121**	**0.227**	**0.430**	**0.472**	**0.311**	**0.398**	**0.360**	**0.476**	**0.225**	**0.535**

Table 3: Per-query time complexity and latency (in seconds) for different retrieval methods under the following setup: $R = 50$, $N = 100{,}000$, using `MiniLM` embeddings. (N: number of passages, D: the embedding dimension, R: the LLM budget, C_{LLM}: cost of a single LLM querying.) For GPR, the time complexity includes: kernel matrix computation $\mathcal{O}(R^2D)$, matrix inversion $\mathcal{O}(R^3)$, and inference over N passages $\mathcal{O}(NRD)$. Latency values include 95% confidence intervals in $[\cdot]$.

Method	Per-query Complexity	Latency (sec)
DPR	$O(ND)$	0.165 [0.161, 0.168]
LLM Rerank	$O(ND + R \cdot C_{\text{LLM}})$	0.678 [0.671, 0.685]
GPR-LLM	$O(ND + R \cdot C_{\text{LLM}} + R^2D + R^3 + NRD)$	Dot Product: 0.782 [0.769, 0.795] Cosine Similarity: 0.774 [0.762, 0.787] RBF Kernel: 0.754 [0.730, 0.780]

System specifications: CPU—Intel(R) Core(TM) i7-14700HX; GPU—NVIDIA GeForce RTX 4070 Laptop GPU; average CPU utilization during measurement: ∼5%.

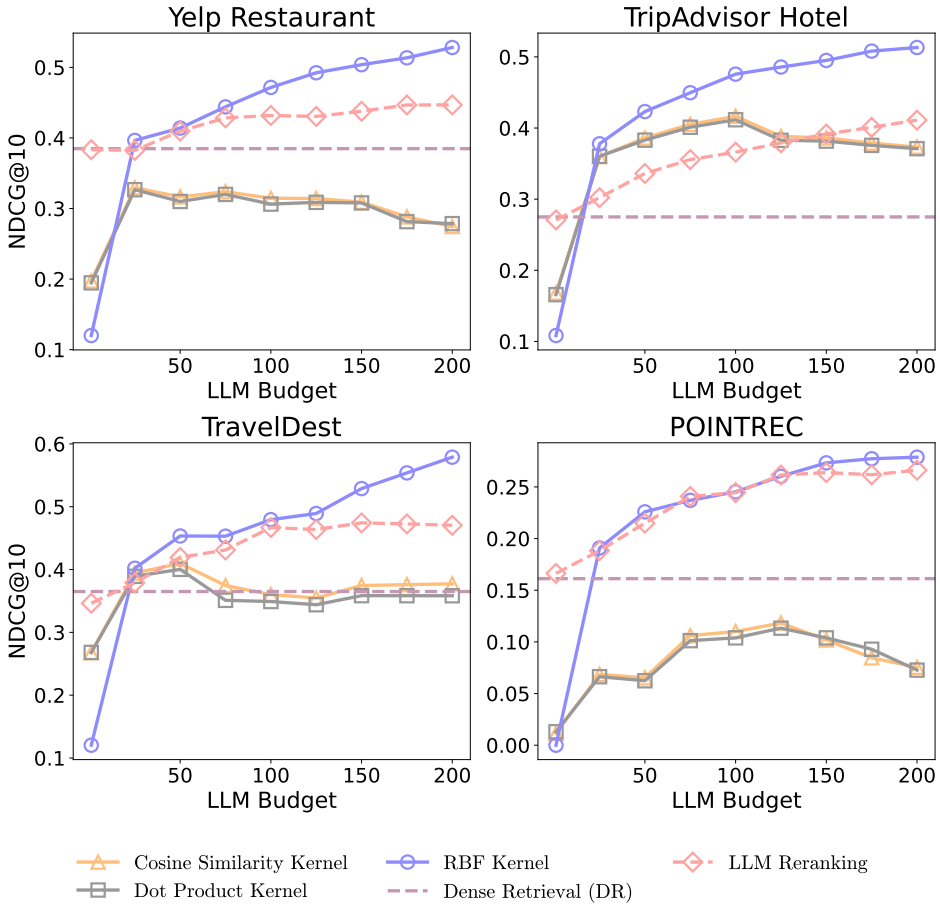

Figure 1: Performance comparison of different kernel functions with Greedy Sampling ($\epsilon = 0$) under varying number of passages with LLM judgments (budget of LLM labels) with `TAS-B` embedding.