Randomized Antipodal Search Done Right for Data Pareto Improvement of LLM Unlearning

Ziwen Liu [1] Huawei Lin [2] Yide Ran [3] Denghui Zhang [3] Jianwen Xie [4] Chuan Li [4] Weijie Zhao [2] Zhaozhuo Xu [3*]

[1]Rice University [2]Rochester Institute of Technology [3]Stevens Institute of Technology [4]Lambda

Abstract

LLM unlearning is often triggered by an undesired generation at inference time, making retrieval the central challenge.

We introduce data Pareto improvement and propose RASLIK, a randomized retrieval algorithm based on permutation–projection hashing and antipodal search.

Across models, datasets, and unlearning algorithms, RASLIK consistently improves the forgetting–retention trade-off over deterministic baselines and oracle sampling.

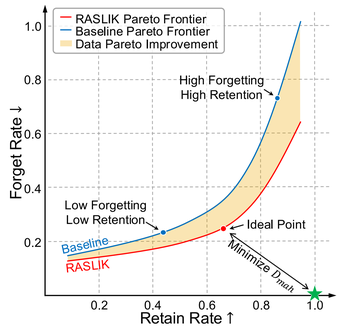

Figure 1. Pareto trade-off between forgetting and retention in LLM unlearning.

Motivation & Problem Setting

Most unlearning methods focus on optimization, assuming the forget and retain sets are already available. In practice, unlearning begins with an undesired generation and a massive training corpus, making retrieval the real bottleneck.

Contributions

- Identify retrieval as the key bottleneck in practical LLM unlearning.
- Introduce data Pareto improvement for forgetting–retention trade-offs.
- Propose RASLIK, a randomized retrieval method with reduced variance and sublinear complexity.
- Demonstrate consistent Pareto improvements across models, datasets, and unlearning algorithms.

Data Pareto Improvement

Unlearning induces a fundamental trade-off: stronger forgetting often harms retention, while prioritizing retention risks incomplete forgetting. We formalize this as a Pareto trade-off between forgetting and retention, and define data Pareto improvement as retrieval choices that shift the frontier outward.

A retrieval mechanism is *Pareto-improving* if it enables:

- stronger forgetting without disproportionate loss of retention;
- better retention without sacrificing forgetting performance.

Method: RASLIK

We propose Randomized Antipodal Search on Linearized Influence Kernel (RASLIK), a retrieval algorithm for influence-based unlearning.

RASLIK constructs randomized gradient sketches via permutation–projection hashing and performs antipodal search to identify samples to forget and retain.

Randomization reduces selection variance, while sketching achieves sublinear complexity.

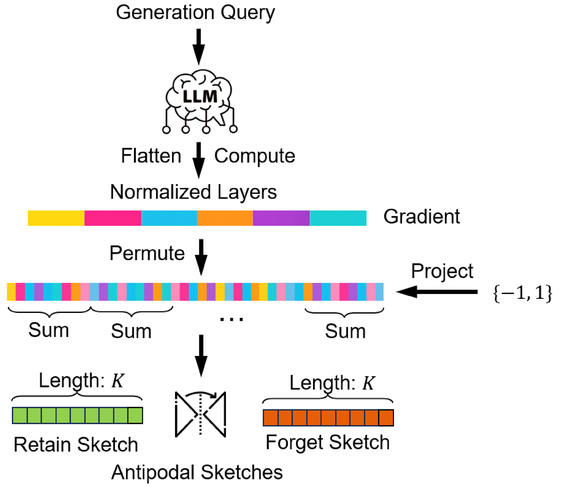

Figure 2. RASLIK retrieval pipeline.

Retrieval in Sketch Space

Retrieval is performed entirely in sketch space. Because $h(-q_y) = -h(q_y)$, the same sketch supports both forget and retain retrieval, making the procedure simple and efficient.

Linearized Influence Kernel

Definition: $\rho(y, x) = \cos(q_y, g_x)$
Measures similarity between query q_y and data gradient g_x.

Antipodal Search Intuition

Strategy: Directional alignment in gradient space.
• Forget F: $\max \cos(q_y, g_x)$ • Retain R: $\max \cos(-q_y, g_x)$

The Symmetry Trick

Linearity Benefit: $h(\text{perm}(\text{proj}(-q_y))) = -h(q_y)$.
One sketch supports both sets: $h(-q_y) = -h(q_y)$.

Experiments and Results

RASLIK is consistently Pareto-optimal or near-best across settings, and outperforms deterministic retrieval baselines overall.

Table 1. Main quantitative results on Howdy-Alpaca and Virtual-Alpaca.

Method	OLMo-2-1124-7B								Pythia-2.8B							
	GA_GDR				GA_KLR				GA_GDR				GA_KLR			
	F↓	R↑	D_{mah} ↓	Non-SF↑	F↓	R↑	D_{mah} ↓	Non-SF↑	F↓	R↑	D_{mah} ↓	Non-SF↑	F↓	R↑	D_{mah} ↓	Non-SF↑
Random Selection	0.569	0.844	10.856	0.040	0.249	0.487	39.468	0.987	0.162	0.274	38.868	0.222	0.135	0.202	253.495	0.683
Embedding Sim.	0.236	0.485	10.167	0.633	0.257	0.574	38.822	0.990	0.092	0.149	39.764	0.893	0.133	0.204	252.630	0.881
BM25	0.282	0.460	11.181	0.573	0.263	0.538	40.234	0.994	0.085	0.150	39.322	0.940	0.135	0.203	253.276	0.372
Oracle Sampling	0.239	0.418	11.083	0.874	0.248	0.525	38.629	0.985	0.103	0.207	38.081	0.982	0.132	0.196	254.341	0.674
RASLIK-F	0.290	0.511	10.660	0.466	0.265	0.561	39.990	0.974	0.086	0.165	38.783	0.992	0.137	0.201	254.199	0.647
RASLIK	0.272	0.555	9.813	0.911	0.246	0.572	37.573	0.994	0.084	0.166	38.622	0.992	0.117	0.186	253.884	0.886

Method	OLMo-2-1124-7B						Pythia-2.8B					
	GA_GDR			GA_KLR			GA_GDR			GA_KLR		
	F↓	R↑	D_{mah} ↓	F↓	R↑	D_{mah} ↓	F↓	R↑	D_{mah} ↓	F↓	R↑	D_{mah} ↓
Random Selection	0.174	0.264	87.590	0.149	0.250	92.907	0.440	0.506	54.346	0.131	0.221	28.514
Embedding Sim.	0.193	0.282	88.102	0.145	0.240	93.062	0.421	0.485	56.388	0.134	0.180	30.040
BM25	0.188	0.263	89.380	0.150	0.260	92.340	0.419	0.481	56.762	0.186	0.179	30.189
Oracle Sampling	0.201	0.299	87.546	0.149	0.257	92.417	0.080	0.468	56.113	0.138	0.229	28.243
RASLIK-F	0.199	0.299	87.333	0.150	0.277	90.937	0.153	0.470	56.314	0.141	0.204	29.150
RASLIK	0.176	0.272	87.166	0.139	0.251	90.915	0.098	0.476	55.458	0.160	0.247	27.670