ICML
International Conference
On Machine Learning

Less is More: Geometric Unlearning for LLMs with Minimal Data Disclosure

Chenchen Tan, Xinghao Li, Shujie Cui, Youyang Qu, Cunjian Chen, Longxiang Gao

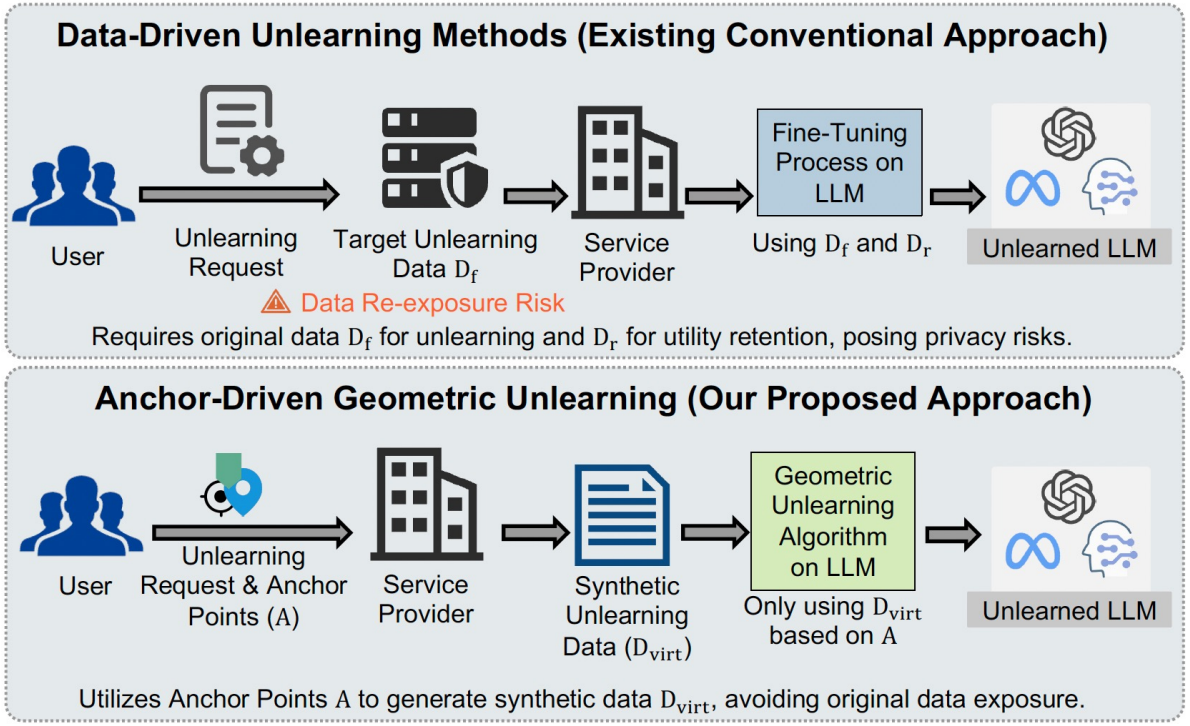
Presenter: Chenchen Tan

MONASH University

Motivation: Unlearning without original training corpus

Why LLM unlearning:

LLMs may memorize sensitive entities (*e.g.,* PII, biographies). We aim to remove such knowledge while preserving general capabilities.



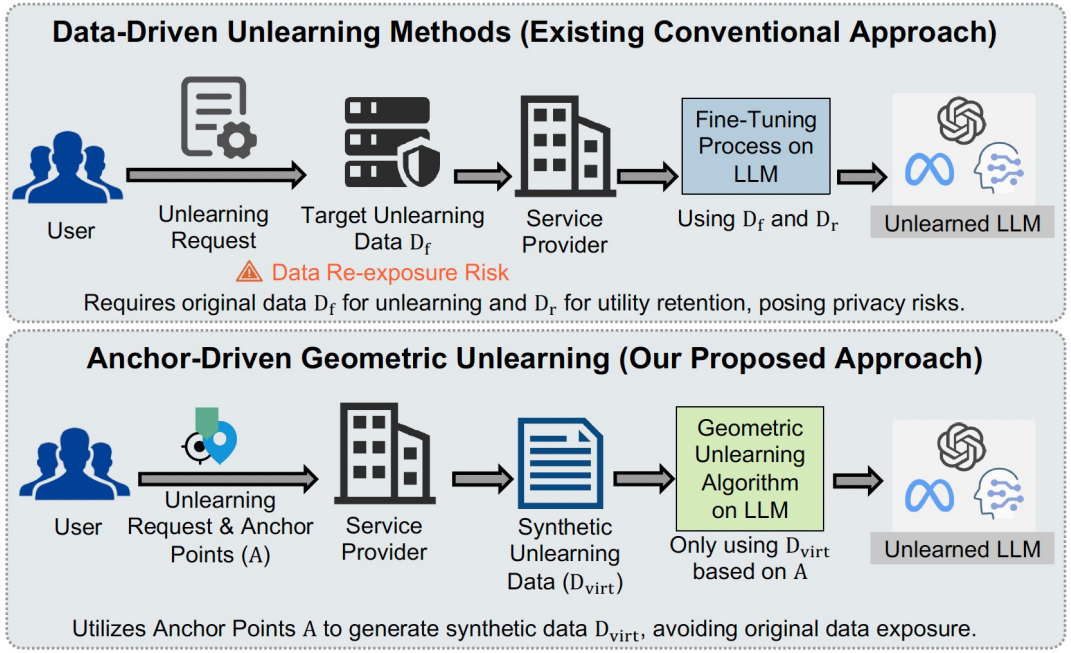
Our Goal

Effective unlearning with minimal synthetic data and no training corpus disclosure.

Key insight: hidden response plans

LLMs form an internal response plan after reading a prompt and before generating an answer [1].

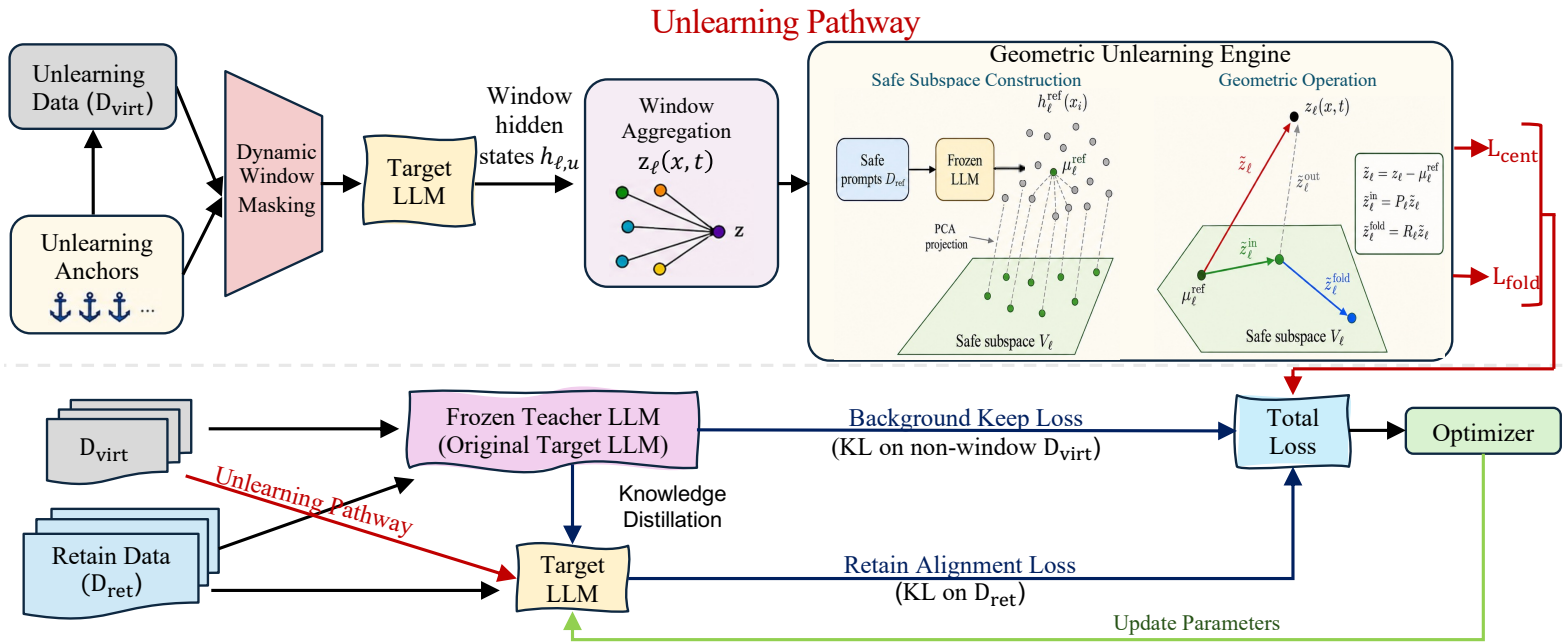
- Target-related prompts can activate a content-generating internal state.
- GU intervenes earlier by steering target-triggered hidden states toward uncertain/refusal-style behavior.
- This enables unlearning without revisiting the original training corpus.



[1] Dong, Z., Zhou, Z., Liu, Z., Yang, C., & Lu, C. (2025, October). Emergent Response Planning in LLMs. In International Conference on Machine Learning (pp. 14301-14320). PMLR.

Our Solution: Geometric Unlearning (GU)

Geometric Unlearning: safe region and target steering

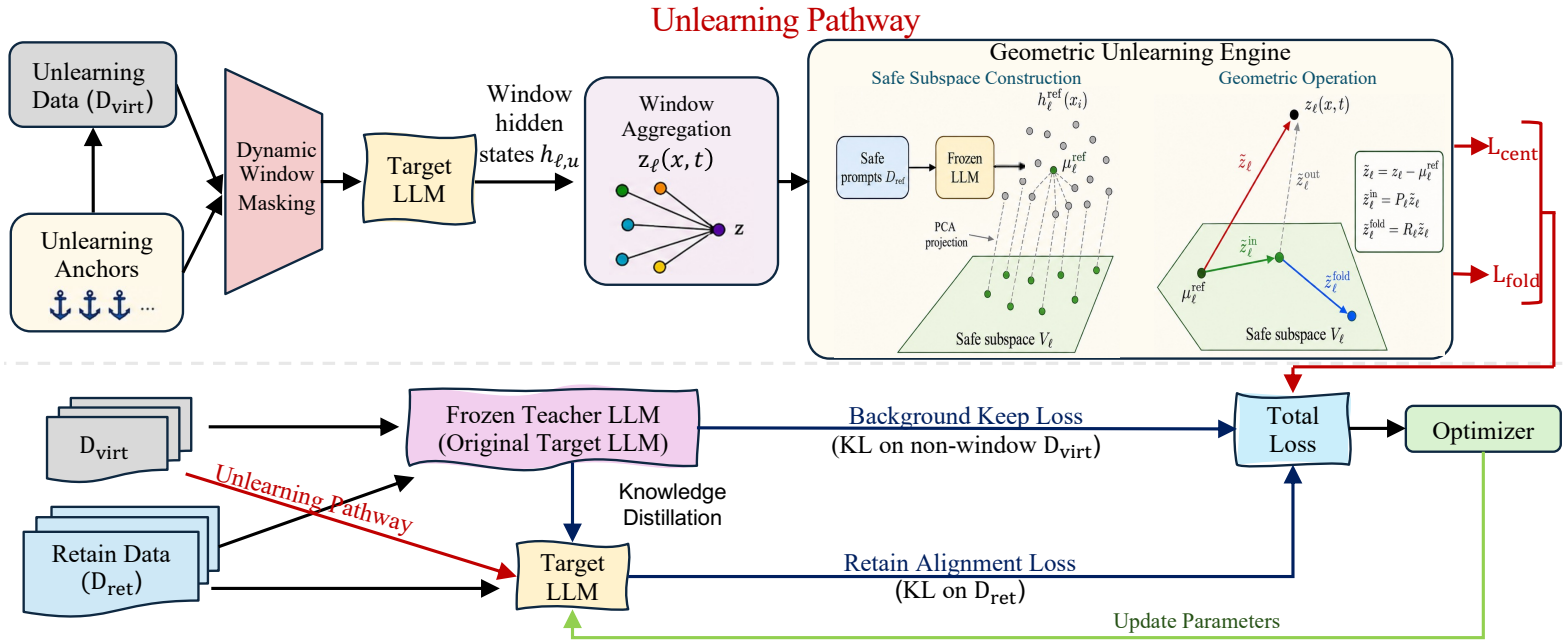


Align target hidden states to safe geometry with centroid pull loss:

$$L_{\text{cent}} = \mathbb{E}_{(x,t)\in\mathcal{D}_{\text{acc}}}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}||\tilde{z}_\ell(x,t)||_2^2\right]$$

Our Solution: Geometric Unlearning (GU)

Geometric Unlearning: safe region and target steering



However, under finite-step optimization, mean alignment alone does not explicitly control the relative out-of-subspace component of the residual. We further introduce fold-back confinement:

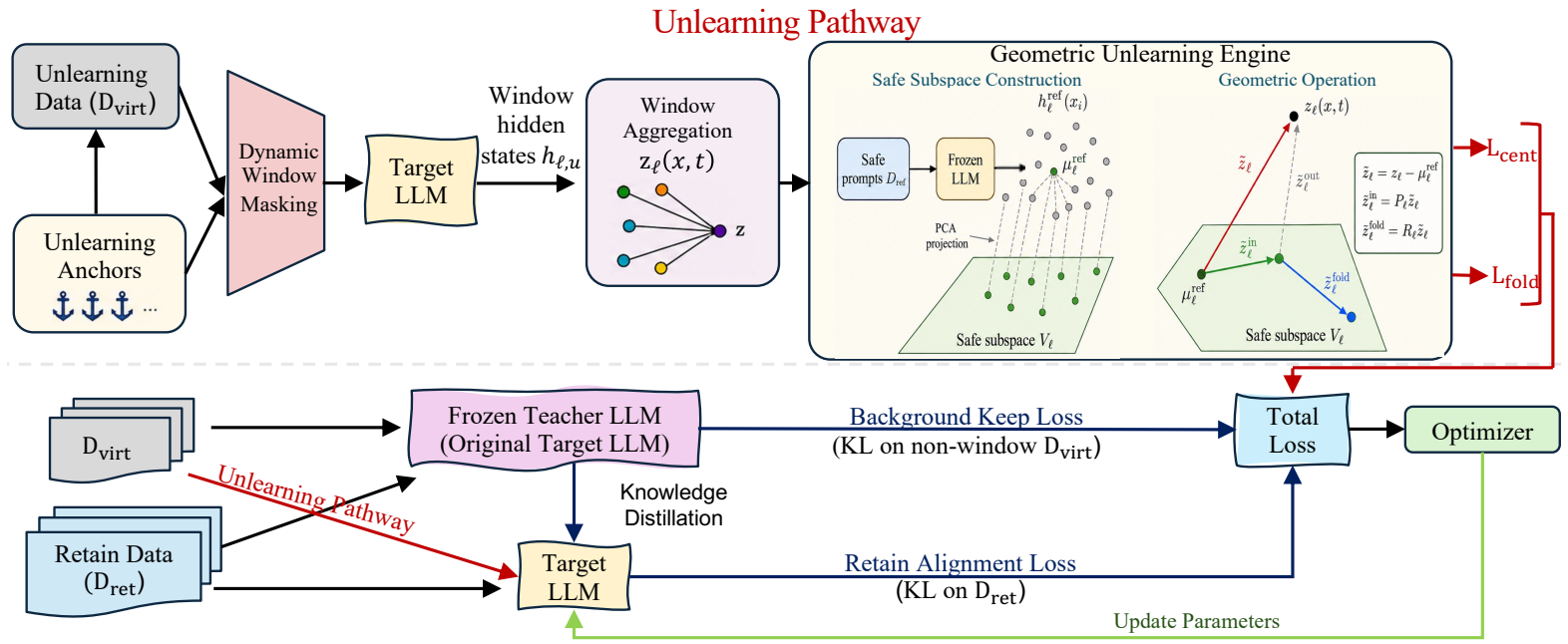
$$\tilde{z}_\ell = \tilde{z}_\ell^{\text{in}} + \tilde{z}_\ell^{\text{out}}, \qquad \tilde{z}_\ell^{\text{in}} := P_\ell \tilde{z}_\ell \text{ (in−subspace)},$$
$$\tilde{z}_\ell^{\text{out}} := P_\ell^{\perp} \tilde{z}_\ell \text{ (orthogonal residual)},$$

$$\tilde{z}_\ell^{\text{fold}} = \tilde{z}_\ell^{\text{in}} - \tilde{z}_\ell^{\text{out}},$$

$$L_{\text{fold}} = \mathbb{E}_{(x,t)\in\mathcal{D}_{\text{acc}}}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\left(1-\cos_\epsilon\left(\tilde{z}_\ell, \tilde{z}_\ell^{\text{fold}}\right)\right)\right]$$

Our Solution: Geometric Unlearning (GU)

Geometric Unlearning: model utility retaining



Avoiding unlearning induce collateral drift on non-target behavior we introduce stabilization components that (i) preserves the teacher behavior on background tokens of target prompts, and (ii) preserves behavior on a separate synthetic retain pool.

$$L_{\text{bg}} = \mathbb{E}_{(x,t)\in\mathcal{D}_{\text{acc}}}\left[\frac{1}{\sum_{u=1}^{T-1} m_u}\sum_{u=1}^{T-1} m_u \,\text{KL}\left(\sigma\left(\ell_u^{(T)}(x)\right)\,||\,\sigma\left(\ell_u(x)\right)\right)\right]$$

$$L_{\text{ret}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{ret}}}\left[\frac{1}{T_x}\sum_{u=1}^{T_x}\text{KL}\left(\sigma\left(\ell_u^{(T)}(x)\right)\,||\,\sigma\left(\ell_u(x)\right)\right)\right]$$

Experimental Results

- **Best overall unlearning-retaining trade-off** under the OpenUnlearning evaluation framework.

Table 1. The evaluation results comparison between the proposed unlearning method and baseline methods on the ToFU benchmark for unlearning 10% dataset.

Methods	LLaMA3.2-1B				LLaMA2-7B			
	Unlearning Effectiveness			M.U.	Unlearning Effectiveness			M.U.
	E.S ↓	F.R. ↓	Privleak (≈ 0)	Avg. ↑	E.S ↓	F.R. ↓	Privleak (≈0)	Avg. ↑
GA (Jang et al., 2023)	**0.032**	**0.000**	**-21.38**	0.000 (-0.598)	**0.027**	0.023	-25.84	0.000 (-0.623)
GradDiff (Yao et al., 2024)	0.077	0.347	-36.49	0.436 (-0.162)	**0.027**	**0.005**	63.04	**0.617** (-0.006)
NPO (Zhang et al., 2024)	0.099	0.248	-43.98	0.418 (-0.180)	0.141	0.500	-90.64	0.540 (-0.083)
SimNPO (Fan et al., 2025b)	0.054	0.398	-35.34	0.423 (-0.175)	0.110	0.432	-11.10	0.553 (-0.070)
UN-DIAL (Dong et al., 2025a)	0.046	0.295	-96.39	0.560 (-0.038)	0.032	0.276	-94.99	0.518 (-0.105)
RMU (Li et al., 2024)	0.062	0.358	-68.09	0.448 (-0.150)	0.027	0.097	57.38	0.030 (-0.593)
GU (Ours)	0.172	0.06	-22.27	**0.577** (-0.021)	0.114	0.058	**-1.496**	0.598 (-0.025)

Experimental Results

- **Best overall unlearning-retaining trade-off** under the OpenUnlearning evaluation framework.

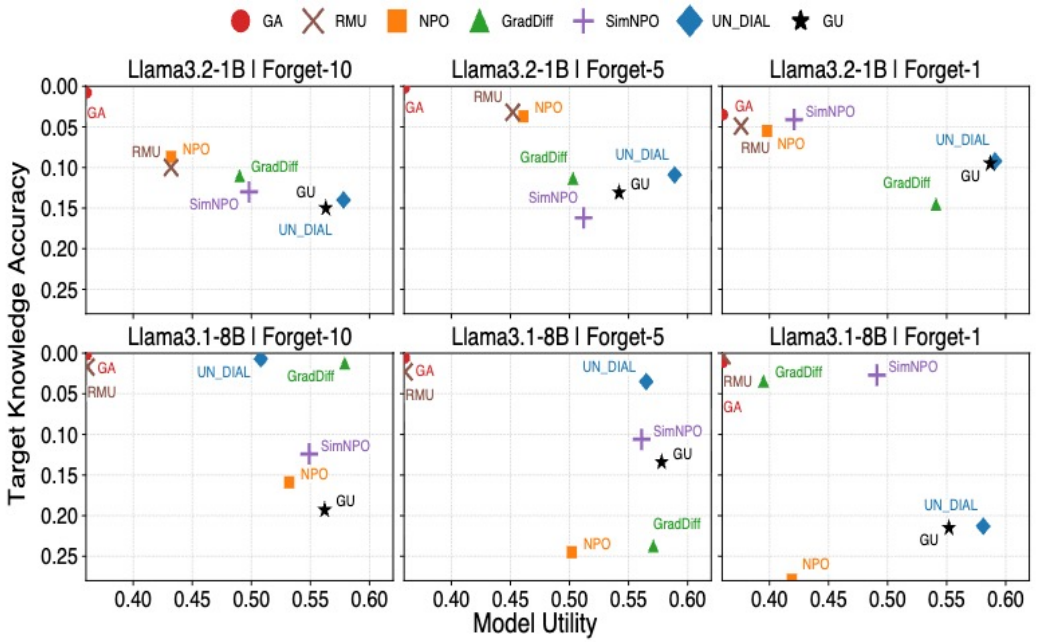


Figure 1. Unlearning effectiveness and model utility trade-off across model scales and forget splits on UnlearnPII benchmark.

Experimental Results

- **Best performance on the matched source-free comparison:** GU benefits on its prompt-conditioned hidden states geometric process instead on data construction.
- **Safer unlearned behavior:** GU mainly yields refusal / uncertain responses.

Table 2. Matched source-free comparison using the same synthetic datasets on baselines.

Method	E.S. ↓	F.R. ↓	PrivLeak (≈ 0)	M.U. ↑
GA	0.112	0.319	-78.5	0.231
GradDiff	0.513	0.615	-97.6	0.561
NPO	0.627	0.751	-99.4	0.594
SimNPO	0.648	0.768	-99.4	0.594
UN-DIAL	0.701	0.818	-99.5	0.600
RMU	0.696	0.797	-99.5	0.592
GU	0.172	0.060	-22.3	0.577

Table 3. Unlearning output characteristics on the target dataset for LLaMA3.2-1B.

Method	**A.F.**(↑)	**E.M.** (↓)	**Unlearned Output Type**
Original	0.855	0.974	Original
GA	0.285	0.000	Blank
GradDiff	0.725	0.646	Substitute
NPO	0.949	0.630	Substitute
SimNPO	0.896	0.566	Substitute
UN-DIAL	0.592	0.515	Substitute
RMU	0.687	0.594	Substitute
GU (Ours)	0.765	0.524	Refusal/Uncertain

Experimental Results

- **Lower privacy leakage:** GU achieves near-zero PrivLeak and pushes MIA performance close to random guessing across Min-K, Reference, and Zlib, indicating reduced membership-inference risk.

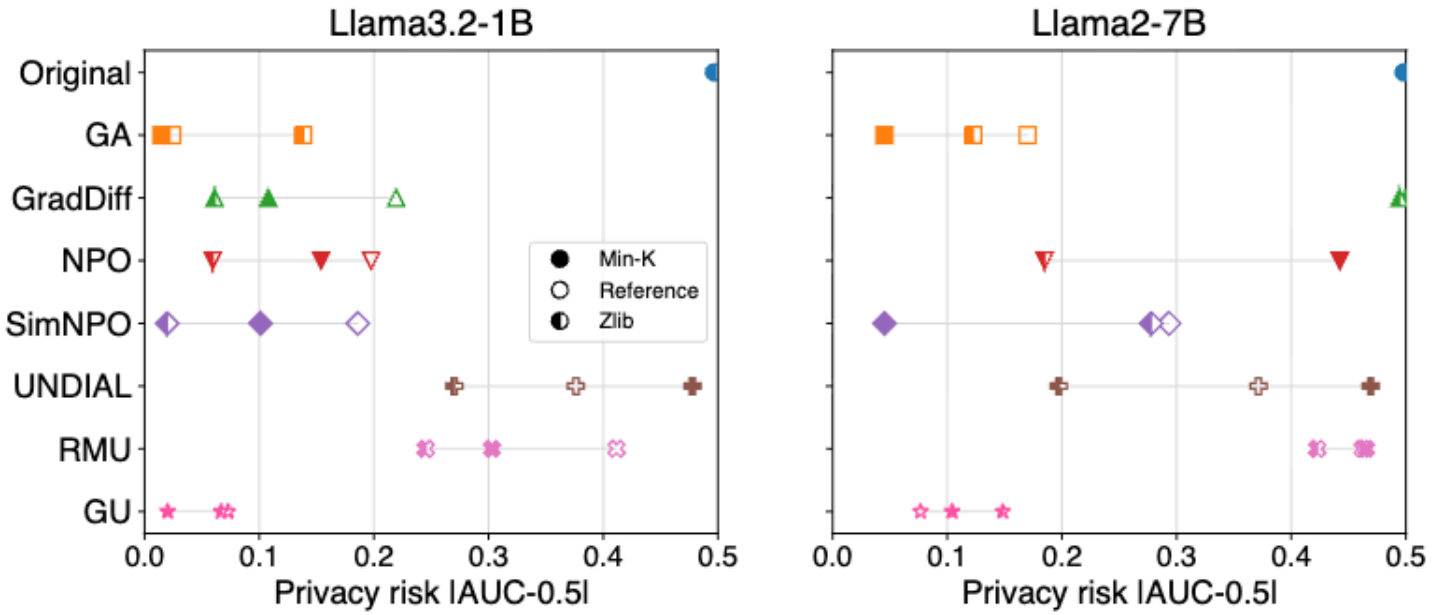


Figure 2. Privacy risk of MIAs across unlearning methods for two base models (|AUC−0.5|, lower is better).

Experimental Results

- **Efficient and practical:** GU uses only synthetic anchor-conditioned prompts without accessing the original training corpus, and completes unlearning with a lightweight runtime comparable to efficient baselines.

Table 4. Runtime analysis on LLaMA2-7B in the Forget-10 setting.

Training Time Comparison		
Method	Time (min)	Epoch
GA	8.98	7
GradDiff	26.16	10
NPO	15.23	10
SimNPO	32.33	10
UN-DIAL	44.51	18
RMU	2.56	6
GU	31.29	8

GU Time Breakdown	
Stage	Time (min)
Virtual data construction	12.37
Safe-reference extraction	1.59
Safe-subspace construction	0.34
Training to convergence	31.29
Total	45.59

Limitations

- **Behavioral unlearning, not certified erasure**
  GU suppresses target behavior but does not guarantee that all underlying parametric knowledge is completely removed.

- **Anchor dependence**
  GU works best when target information can be reliably triggered by clear anchors during unlearn training, such as names or keywords.

- **Relearning remains possible**
  If the model is later fine-tuned again on target data, some suppressed behavior can partially recover.

Thanks for Listening!