ARC-UNLEARN: BLAST-RADIUS-AWARE LLM UNLEARNING VIA AGENTIC RETAIN-SET SYNTHESIS AND ADJACENT-RETAIN NPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning for large language models aims to remove unwanted target knowledge while preserving the related knowledge around it. However, most existing approaches primarily evaluate how effectively target information is forgotten, potentially overlooking damage to semantically related but legitimate knowledge. We refer to this localized collateral damage as the *blast radius*. We introduce **BlastBench**, an adjacency-aware benchmark for quantifying collateral damage during machine unlearning through *Adjacent Retain Rate* (ARR). We further present **ARC-Unlearn**, a multi-agent framework that identifies target-specific adjacent concepts, constructs an adjacent-retain set, and protects this knowledge during unlearning. Its optimization component, **Adjacent-Retain NPO (AR-NPO)**, combines NPO-based forgetting, adjacent-retain cross-entropy, and general-knowledge KL regularization. Across three 7B-scale models and 45 targets, standard forgetting baselines achieve high forget efficacy but substantially degrade adjacent knowledge. Among the evaluated methods, AR-NPO is the only one that achieves non-trivial forget efficacy ($\mathrm{FE}{=}0.566$) while improving adjacent retention relative to the original model ($\mathrm{ARR}{=}1.259$). An α-sweep reveals a controllable FE–ARR trade-off between target forgetting and adjacent retention. On a matched 9-target subset at 32B scale, AR-NPO reaches $\mathrm{FE}{=}0.718$ and $\mathrm{ARR}{=}1.964$, with all nine targets satisfying our surgical-zone criteria. Full-parameter fine-tuning further achieves near-complete measured forgetting ($\mathrm{FE}{=}0.994$) together with above-baseline adjacent retention ($\mathrm{ARR}{=}1.505$). Finally, held-out and cross-generator audits show that the adjacent-retention gains persist beyond the original generated QA.

1 INTRODUCTION

Regulatory frameworks such as the GDPR's right to be forgotten (Voigt & Von dem Bussche, 2017) and growing concerns over copyright-sensitive training data and hazardous knowledge have made *machine unlearning* a central challenge for large language model (LLM) deployment (Cao & Yang, 2015; Jang et al., 2023). The goal is to remove a specific body of knowledge — a *forget target* — from a model's behavior while avoiding full retraining.

The Blast-Radius Problem. Consider forgetting the *Harry Potter* series from an LLM. Intuitively, the model should no longer recall Hogwarts, Hermione Granger, or Voldemort's Horcruxes. But what about J.K. Rowling, Daniel Radcliffe, British boarding schools, or fantasy literature as a genre? These are not part of the forget target, yet they are topologically adjacent to it in the model's knowledge graph. We call such non-target knowledge potentially exposed to an unlearning update the *blast radius* of the forget target. Figure 1 illustrates this structure concretely for the *Harry Potter* forget target. The stakes are especially acute for hazardous domains: an LLM deployed in a security context that has had cyberattack knowledge removed must still reason about defensive security, CVE analysis, and penetration testing.

Existing unlearning methods are largely blind to this structure. GA and GD (Jang et al., 2023; Maini et al., 2024) maximize loss on the forget set; NPO and SimNPO (Zhang et al., 2024; Fan et al., 2025) replace it with a preference-style objective. These methods typically rely on generic retain sets (e.g., WikiText (Merity et al., 2016)) which preserve broad model utility but provide no explicit protection

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for knowledge semantically coupled to the forget target. This leaves a critical question unaddressed: *which adjacent concepts are at risk, and how should we protect them?*

A similar gap exists in evaluation. Benchmarks such as TOFU (Maini et al., 2024), WMDP (Li et al., 2024), MUSE (Shi et al., 2025), and OpenUnlearning (Dorna et al., 2025) assess forgetting quality and general knowledge retention as separate objectives. However, they do not examine whether removing a piece of knowledge unintentionally affects related concepts, in other words, whether the model's "blast radius" extends beyond the targeted information.

Our Contributions. We address both the algorithmic gap and the evaluation gap through the following contributions:

1. **BlastBench**: an adjacency-aware unlearning benchmark designed to measure blast-radius collateral damage through five complementary metrics, including the novel *Adjacent Retain Rate* (ARR), across 45 forget targets in three domains.
2. **ARC-Unlearn Pipeline**: a multi-agent framework consisting of a *Prober Agent* that identifies the blast radius associated with a forget target through adversarial model interrogation, and a *Synthesizer Agent* that constructs targeted retain examples for protecting those concepts.
3. **Adjacent-Retain NPO (AR-NPO)**: a three-term loss that jointly optimizes forget efficacy, adjacent knowledge retention, and general fluency.

We evaluate the proposed framework against four unlearning baselines across three 7B-scale models and 45 forget targets. Our experiments reveal a pronounced forgetting–adjacent-retention trade-off: standard methods achieve strong forgetting at the cost of severe adjacent degradation, whereas AR-NPO substantially improves adjacent retention while retaining controllable forget efficacy. We further stress-test unlearning along three axes: robustness to adversarial relearning and membership-inference attacks, behavior at 32B scale, and control for evaluation artifacts via a held-out audit. Together, these results highlight that effective machine unlearning requires considering not only what a model forgets, but also the semantic knowledge it preserves.

2 Related Work

LLM Unlearning Algorithms. Cao & Yang (2015) introduced the concept of machine unlearning in the context of classical machine learning. Later for LLMs, Jang et al. (2023) proposed Gradient Ascent (GA), which forgets targeted knowledge by maximizing the cross-entropy loss on the forget set. Gradient Difference (GD) (Maini et al., 2024) adapts GA by augmenting it with a retain-set cross-entropy objective, reducing catastrophic forgetting of unrelated information. Subsequent work focused on improving training stability and efficiency. Zhang et al. (2024) reformulated the objective as Negative Preference Optimization (NPO), showing that preference-style optimization leads to more stable unlearning behavior. Later, Fan et al. (2025) proposed SimNPO, a reference-free variant that removes the need to retain access to the original model during training.

Parallel to these approaches, knowledge editing methods such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) perform targeted parameter updates to alter specific factual associations; concept-erasure methods such as ELM (Gandikota et al., 2025) instead suppress broader conceptual knowledge. Neither line explicitly optimizes for preserving the knowledge semantically adjacent to a forget target.

Despite their strengths, existing methods share a common limitation: They do not identify which neighboring concepts are most susceptible to collateral forgetting, nor do they construct retain sets to actively preserve them. Concurrent work by Dahal et al. (2026) addresses this issue from a different perspective through GONE, a distribution-shaping approach that expands the boundary between forget examples and adjacent knowledge. In contrast, AR-NPO explicitly discovers target-specific adjacent concepts and constructs an adjacency-aware retain set to protect them during unlearning.

Unlearning Benchmarks. TOFU benchmark (Maini et al., 2024) evaluates forgetting of fictitious authors using separate forget and retain splits, enabling controlled measurement of unlearning behavior. WMDP (Li et al., 2024) focuses on hazardous knowledge removal through multiple-choice

evaluations, while MUSE (Shi et al., 2025) introduces a broader suite of six complementary metrics for assessing unlearning quality. However, none of these benchmarks explicitly probe whether forgetting a target concept unintentionally affects semantically related knowledge. More recently, BLUR benchmark (Hu et al., 2025b) examines the overlap between forget and retain information, highlighting the challenge of disentangling closely related knowledge. BlastBench targets a distinct question: whether an unlearning update damages non-target concepts that are semantically related to the forget target, even when those concepts are not themselves part of the forget information. Similarly, OpenUnlearning (Dorna et al., 2025) unifies existing benchmarks within a common evaluation framework, but stops short of introducing explicit adjacency-aware evaluation. BlastBench addresses this gap by explicitly modeling adjacent concepts and measuring whether unlearning remains localized to the intended target without spilling over into semantically adjacent knowledge.

Adjacent Knowledge Collateral Damage. Several concurrent works observed that unlearning methods degrade performance on knowledge adjacent to the forget target. Ko et al. (2025) found substantial degradation on semantically adjacent queries after unlearning. Chang & Lee (2025) studied which retain-set types best protect entity-level neighbors during unlearning and emphasized the role of syntactic structure in minimizing unwanted knowledge loss. Zhong et al. (2026) formalized semantically adjacent concept preservation as an explicit multi-objective unlearning constraint. Taken together, this body of work establishes that collateral damage to adjacent knowledge is a measurable and significant failure mode of unlearning. Our work builds on this emerging direction by systematically evaluating and mitigating such collateral damage through target-specific adjacent knowledge.

Adversarial Evaluation & Agentic Pipelines. Yeom et al. (2018) formalized Membership Inference Attacks (MIA), establishing a foundation for evaluating whether supposedly forgotten data can still be recovered from trained models. Building on this line of work, Hu et al. (2025a) showed that unlearned models can rapidly re-acquire forgotten knowledge through targeted relearning, raising concerns about the durability of current unlearning methods. Recent work has also explored agentic approaches for probing and manipulating LLM knowledge. Wu et al. (2024) introduced multi-agent LLM frameworks that coordinate specialized agents for complex reasoning and information synthesis tasks. More recently, Li et al. (2026) leveraged an LLM-probed memory graph to supervise forget-set construction, whereas our pipeline instead focuses on constructing and protecting the adjacent retain set. Complementing these efforts, Jiang et al. (2025) traversed knowledge graphs to audit how existing unlearning methods propagate forgetting across related concepts.

3 Problem Formulation

Notation. Let $\mathcal{M}_\theta$ be a pre-trained LLM. Given a forget target T (e.g., *Harry Potter series*), we partition model knowledge into four sets:

- F: the *forget set*: QA pairs whose correct answer requires knowledge of T.
- A: the *adjacent retain set*: QA pairs about concepts semantically adjacent to T but independently meaningful.
- G: the *general retain set*: QA pairs and text spans unrelated to T.
- D: the *distal retain set*: QA pairs about concepts related to the broader domain of T but outside its local semantic neighborhood.

Blast Radius. We define the *blast radius* $\mathcal{B}(T)$ as the K-hop neighborhood of T in the model's semantic knowledge graph, approximated by BFS probing (detailed in §4). Adjacent concepts $c \in \mathcal{B}(T)$ are those for which knowing c provides partial reconstruction paths to T with reconstruction-risk score $\geq \tau$ ($\tau \in [0, 10]$, default τ=5; scoring protocol detailed in §4.1).

Metrics. All five BlastBench metrics are normalized ratios relative to the original model $\mathcal{M}_\theta$:

$$\begin{aligned} \mathrm{FE} &= 1 - \frac{\mathrm{acc}(F)}{\mathrm{acc}_0(F)}, \quad \mathrm{ARR} = \frac{\mathrm{acc}(A)}{\mathrm{acc}_0(A)} \\ \mathrm{DRR} &= \frac{\mathrm{acc}(D)}{\mathrm{acc}_0(D)}, \quad \mathrm{RR} = \frac{\mathrm{acc}(G)}{\mathrm{acc}_0(G)} \\ \mathrm{FP} &= \frac{\mathrm{ppl}_0}{\mathrm{ppl}} \end{aligned} \tag{1}$$

where $\mathrm{acc}(\cdot)$ denotes the mean per-token log-probability of the gold answer tokens given the question prefix for the post-unlearning model, and $\mathrm{acc}_0(\cdot)$ denotes the same quantity for the original model.

Forget Efficacy (FE) measures the extent to which performance on the forget set decreases after unlearning, with FE$= 1$ indicating maximal measured forgetting and FE$= 0$ indicating no change from the original model. The remaining metrics evaluate knowledge preservation: ARR measures retention of semantically adjacent concepts, DRR evaluates retention of more distant but related concepts, and RR measures preservation of general unrelated knowledge. Values greater than 1 indicate that performance improved relative to the original model, while values below 1 indicate degradation. An ideal unlearning method achieves FE$= 1$ while maintaining ARR, DRR, RR, and FP at or above 1, indicating complete forgetting without collateral damage or fluency loss. Fluency Preservation (FP) is computed using WikiText perplexity (the exponentiated average per-token cross-entropy loss on a held-out text sample), where ppl_0 and ppl denote the perplexity of the original and unlearned models, respectively. Since lower perplexity indicates better language modeling quality, higher FP values correspond to stronger fluency preservation.

Surgical Zone. We call the region $\mathrm{FE}{>}0.3, ; \mathrm{ARR}{>}1.0$ the *surgical zone*, representing simultaneous non-trivial forgetting and preservation of adjacent knowledge at or above the original-model baseline. The ARR> 1.0 boundary has a direct interpretation: adjacent-concept performance is not degraded relative to the original model, while the FE threshold excludes degenerate solutions that preserve adjacent knowledge by performing little or no forgetting. As shown by the FE–ARR Pareto analysis in Figure 3 (Appendix I), AR-NPO is the only evaluated method that enters this region. Moreover, no baseline achieves ARR> 1.0 at any positive level of FE in our sweep, so this separation is not an artifact of the particular FE> 0.3 threshold.

4 ARC-UNLEARN

ARC-Unlearn consists of three main components. First, a *Prober Agent* identifies the semantic blast radius of the forget target by discovering semantically adjacent knowledge likely to be affected during unlearning. Next, a *Synthesizer Agent* constructs adjacent retain datasets that capture this related knowledge. Finally, **AR-NPO** jointly optimizes forgetting, adjacent retention, and general capability preservation within a unified training objective. Algorithm 1 summarizes the complete pipeline.

4.1 PROBER AGENT: BLAST-RADIUS DISCOVERY

Given a forget target T, the Prober Agent estimates the local semantic blast radius $\mathcal{B}(T)$ through iterative breadth-first exploration combined with adversarial reconstruction probing.

Starting from the forget target, the agent recursively expands a frontier of semantically related concepts up to depth K. At each hop, the agent issues two complementary probe types:

1. Elicitation Probe: *"List b concepts closely associated with* `[concept]` *but independent of T."*
   This probe expands the local semantic neighborhood by identifying potentially related concepts.
2. Reconstruction Probe: *"If all knowledge of T were erased, could knowledge of* `[concept]` *partially reconstruct it? Rate 0–10."*
   This probe estimates the degree of semantic entanglement between the candidate concept and the forget target.

Concepts receiving scores greater than or equal to threshold τ are added to the blast radius $\mathcal{B}(T)$ and retained in the BFS frontier for further exploration. Unless otherwise stated, we use $K{=}2$, $b{=}5$, and

Algorithm 1 ARC-Unlearn Pipeline

Require: Forget target T, base model $\mathcal{M}$, depth K, branching b, threshold τ
Ensure: LoRA adapter ϕ^*
1: $\mathcal{B}(T) \leftarrow \textsc{Prober}(T, \mathcal{M}, K, b, \tau)$
2: $(F, A) \leftarrow \textsc{Synthesizer}(T, \mathcal{B}(T))$ ▷ F: 50 QA pairs about T; A: QA pairs for $c \in \mathcal{B}(T)$, no reference to T
3: $G \leftarrow \textsc{WikiText}$
4: Initialize LoRA adapter ϕ
5: **for** each training step **do**
6: $\quad \ell \leftarrow \mathcal{L}_{\text{NPO}}(\phi; F) + \alpha\, \mathcal{L}_{\text{CE}}(\phi; A) + \beta\, \mathcal{L}_{\text{KL}}(\phi; G)$
7: $\quad \phi \leftarrow \phi - \eta\, \nabla_\phi \ell$ ▷ NPO sign: $-\nabla \mathcal{L}_{\text{NPO}} \propto +\nabla \mathcal{L}_{\text{CE}}(F)$, consistent with Thm. F.2
8: **return** $\phi^* \leftarrow \phi$

τ=5. These values were selected to balance semantic coverage with probing cost. Increasing K or b expands the discovered semantic neighborhood but increases API usage and introduces more weakly related concepts. We treat the reconstruction score as a coarse admission filter rather than a calibrated probability: LLM self-ratings are known to be imperfectly calibrated (Kadavath et al., 2022), so the threshold τ=5 serves only to include or exclude a candidate concept rather than as a quantitative measure of adjacency. All reported metrics (ARR, FE, DRR, RR, FP) are measured on held-out evaluation data. Across 45 forget targets, the Prober Agent discovered 602 adjacent concepts (13.4 per target on average), requiring approximately 1,350 API calls in total (30 per target on average). This probing cost is negligible relative to the subsequent LoRA optimization stage (Section 4.3).

4.2 Synthesizer Agent: Adjacent Retain Set Generation

After identifying the blast radius $\mathcal{B}(T)$, the Synthesizer Agent generates adjacent retain examples for each concept $c \in \mathcal{B}(T)$. For every adjacent concept, the agent generates supervised question-answer pairs designed to:

1. demonstrate knowledge of c,
2. avoid explicit reference to the forget target T,
3. focus on properties of c that exist independently of T.

We refer to this design constraint as the *firewall principle*: adjacent retain examples must not inadvertently reintroduce or reinforce the forget target during training.

4.3 AR-NPO: Adjacent-Retain Negative Preference Optimization

We now introduce AR-NPO, the optimization component of ARC-Unlearn, designed to remove target knowledge while explicitly preserving adjacent concepts and general model capabilities.

Let ϕ denote the parameters of a LoRA adapter (Hu et al., 2022) applied to a frozen base model $\mathcal{M}$. AR-NPO optimizes the following composite objective:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{NPO}}(\phi; F)}_{\text{forget}} + \alpha \underbrace{\mathcal{L}_{\text{CE}}(\phi; A)}_{\text{adjacent}} + \beta \underbrace{\mathcal{L}_{\text{KL}}(\phi; G)}_{\text{general}} \tag{2}$$

Forget Objective. The forgetting term uses Negative Preference Optimization (NPO) (Zhang et al., 2024):

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta_{\text{npo}}} \mathbb{E}_{x \sim F} \left[\log \sigma \left(-\beta_{\text{npo}} \log \frac{p_\phi(x)}{p_0(x)} \right) \right] \tag{3}$$

where p_0 denotes the original reference model and p_ϕ denotes the adapted model.

Intuitively, this objective penalizes the adapted model for assigning high likelihood to forget examples relative to the reference model, thereby suppressing target-specific behavior without the instability often associated with direct gradient ascent. Unless otherwise stated, we use $\beta_{\text{npo}} = 0.1$.

Adjacent Retain Objective. The adjacent retain term $\mathcal{L}_{\text{CE}}(\phi; A)$ is a standard cross-entropy loss over adjacent concepts identified by the Prober Agent. Because the adjacent retain set is constructed under the firewall principle (Section 4.2), its examples are designed to preserve adjacent knowledge without explicitly reintroducing the forget target. As discussed in Appendix F, semantically related concepts often exhibit overlapping representations and aligned gradients. Consequently, forgetting updates that suppress the target concept may also damage nearby knowledge. Optimizing the adjacent-retain objective introduces counteracting gradients that help preserve vulnerable neighboring concepts and mitigate ARR degradation.

General Retain Objective. To preserve general language capabilities, we additionally apply a KL-divergence regularization term on generic text data:

$$\mathcal{L}_{\text{KL}}(\phi; G) = \mathbb{E}_{x \sim G}\Big[D_{\text{KL}}(p_0(\cdot|x) \| p_\phi(\cdot|x))\Big] \tag{4}$$

This objective anchors the adapted model toward the original model on general-domain text, helping preserve fluency and overall behavioral stability. Unless otherwise stated, we use α=1.0 and β=0.1. The resulting FE–ARR trade-off across values of α is analyzed in Appendix I, with the complete hyperparameter grid reported in Appendix D.

5 BLASTBENCH

BlastBench is a multi-domain, adjacency-aware unlearning benchmark designed to evaluate both forget efficacy and collateral damage to semantically adjacent knowledge, a failure mode not captured by general capability retention alone. We construct three domains with 15 forget targets each (45 total):

- **Copyright Fiction (CF):** Major literary series with rich, well-documented knowledge neighborhoods (e.g., *Harry Potter*, *Dune*, *The Hunger Games*). Their interconnected lore, characters, and world-building make them natural stress tests for blast-radius-aware unlearning.
- **Privacy Entities (PE):** Fictional personal profiles constructed following the TOFU (Maini et al., 2024) methodology and embedded within real adjacent institutional knowledge such as hospitals, universities, and professional organizations. This setting evaluates whether personalized information can be removed without degrading surrounding contextual knowledge.
- **Hazardous Knowledge (HK):** Cybersecurity-related attack techniques, exploit patterns, and deprecated vulnerability knowledge alongside legitimate defensive security concepts such as penetration testing, CVE analysis, and defensive tooling. This domain evaluates whether harmful or outdated attack knowledge can be suppressed without collaterally damaging defensive expertise relied upon by security practitioners.

Dataset Construction. For each target the Prober Agent (§4.1) discovers the blast radius; the Synthesizer Agent (§4.2) generates forget and adjacent retain pairs using five adversarial templates: direct elicitation (What is [property of c]?), indirect recall (retrieve c via a related property without naming T), multi-hop reasoning (chain two inferential steps through c), counterfactual (alter a premise to probe robustness), and paraphrase (rephrase the same question lexically). We additionally hand-sample 50 distal pairs (D) per target from Wikipedia topics selected to share no entities, themes, or co-occurrence links with the forget targets, and use WikiText passages for G. The resulting BlastBench dataset totals 13,803 QA pairs: 2,203 forget examples (49.0 per target on average) and 11,600 adjacent retain examples (257.8 per target on average). The approximately 5:1 ratio reflects coverage of an average of 13.4 adjacent concepts per target. Complete benchmark statistics are shown in Table 5. Table 1 compares BlastBench with prior unlearning benchmarks.

Benchmark	Targets	Domains	Adjacent Set	ARR Metric	QA Pairs
TOFU (Maini et al., 2024)	200	1	×	×	≈4K
WMDP (Li et al., 2024)	–	3	×	×	3.7K (MC only)
MUSE (Shi et al., 2025)	–	2	×	×	≈2K
BLUR (Hu et al., 2025b)	varied	mult.	×	×	varied
OpenUnlearning (Dorna et al., 2025)	varied	mult.	×	×	varied
BlastBench **(ours)**	**45**	**3**	✓	✓	**13,803**

Table 1: Comparison of LLM unlearning benchmarks. Targets = number of forget targets; Domains = number of evaluation domains ("mult." denotes benchmarks whose domain count varies by configuration or sub-task selection); Adjacent Set = explicitly per-target curated adjacent concept set provided; ARR Metric = adjacent retain rate reported as a dedicated standalone metric; QA Pairs = total evaluation QA pairs (MC = multiple-choice only).

5.1 Human Validation of BlastBench

Because BlastBench is generated by an LLM pipeline, we validate a random, stratified sample of 120 adjacent items (40 per domain) against human judgment. Two annotators independently labeled every item on three yes/no criteria: (i) *adjacency validity* — the discovered concept is genuinely adjacent to, and distinct from, the forget target; (ii) *non-leakage* — the QA is answerable without relying on the target (the firewall principle); and (iii) *answer correctness*. Inter-annotator agreement is high (Cohen's $\kappa = 0.96/0.80/0.79$; raw agreement $99.2\%/96.7\%/98.3\%$), and both annotators accept 88.3%, 89.2%, and 95.0% of items on the three criteria, respectively. Copyright Fiction is the most challenging domain (72.5% adjacency; 70.0% non-leakage), reflecting less sharply defined concept boundaries. Full protocol and guidelines are in Appendix A.

6 Experiments

6.1 Experimental Setup

We evaluate three open-weight instruction-tuned models: Qwen2.5-7B-Instruct (Qwen Team, 2025), Llama-3-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-v0.3-Instruct (Jiang et al., 2023). All experiments use LoRA adapters (Hu et al., 2022); base model weights are frozen. We compare against four unlearning baselines: GA, GD, NPO and SimNPO (§2). All methods use the same forget dataset F. For retain regularization, baseline methods use randomly sampled WikiText retain data without adjacency information. In contrast, AR-NPO additionally incorporates the synthesized adjacent-retain dataset A derived from the estimated blast radius. All five BlastBench metrics are computed across all 45 targets and three models (135 adapter runs total; one fixed random seed per target–model pair).

6.2 Main Results

Table 2 reports results averaged across all 45 targets and three models; per model results are provided in Table 11 (Appendix M). Figure 3 (Appendix I) shows the FE–ARR Pareto frontier.

Key findings. **(i)** GA, GD, and NPO achieve near-complete forgetting ($\mathrm{FE} \approx 0.99$) but collapse adjacent retention to $\mathrm{ARR} \leq 0.081$ and degrade fluency, consistent with the gradient-overlap mechanism of Theorem F.2. **(ii)** SimNPO fails to produce measurable forgetting ($\mathrm{FE}{=}0.000$ across all 45 targets and three models); its high retention scores occur in the absence of forgetting. A β sweep confirms this is not a hyperparameter artifact (Appendix H). **(iii)** AR-NPO is the only evaluated method to achieve non-trivial forgetting while preserving adjacent knowledge above the original-model baseline: at $\alpha{=}1.0$, it reaches $\mathrm{FE}{=}0.566$ and $\mathrm{ARR}{=}1.259$, while maintaining $\mathrm{DRR}{=}0.995$, $\mathrm{FP}{=}0.991$, and $\mathrm{RR}{=}1.107$. These results show that explicit adjacent-retain optimization substantially mitigates blast-radius collateral damage.

Method	FE ↑	ARR ↑	DRR ↑	FP ↑	RR ↑	SZ
GA	0.990	0.073	0.754	0.641	0.119	×
GD	0.988	0.081	0.890	0.859	0.132	×
NPO	**0.992**	0.067	0.919	0.910	0.110	×
SimNPO†	0.000	**1.261**	**1.000**	0.871	**1.112**	×
AR-NPO (ours, α=1)	0.566	1.259	0.995	**0.991**	1.107	✓

Table 2: Main results averaged over 45 targets × 3 models (135 target-model configurations). Best value per column in **bold**, second-best underlined. **SZ** = Surgical Zone (FE>0.3 *and* ARR>1.0 simultaneously): AR-NPO is the only method that satisfies both criteria. †SimNPO: FE = 0.000 at all tested β values (App. H); retention scores reflect an unmodified model, not genuine preservation under unlearning.

6.3 Adversarial Evaluation

Table 3 summarizes two adversarial evaluations; full definitions are in Appendix P. *Relearning resilience* is the fraction of forget efficacy retained after a 50-step relearning fine-tune (Hu et al., 2025a). *LOSS-MIA* (Yeom et al., 2018) uses per-sample cross-entropy loss to distinguish forget-set members from WikiText non-members; we report $|\text{AUC} - 0.5|$ (lower = more genuine forgetting).

Method	Resilience ↑	MIA-AUC	|AUC−0.5| ↓
GA	0.366	0.008	0.492
GD	0.355	0.009	0.491
NPO	0.366	0.006	0.494
SimNPO†	**0.962**	0.871	0.371
AR-NPO (ours)	0.643	**0.393**	**0.107**

Table 3: Adversarial evaluation averaged over 45 targets and 3 models. For MIA-AUC, values closer to 0.5 indicate stronger indistinguishability from the original training distribution. †SimNPO's resilience is inflated by its near-zero forget efficacy.

Key Findings. **(iv)** GA, GD, and NPO yield MIA-AUC values near zero, indicating high membership distinguishability between forget examples and non-members (Zhang et al., 2024; Lynch et al., 2024). **(v)** Among methods with non-trivial forgetting, AR-NPO is closest to the ideal AUC of 0.5 (MIA-AUC= 0.393; $|\text{AUC} - 0.5| = 0.107$) and retains 0.643 of its forget efficacy after relearning.

6.4 Per-Domain Analysis

AR-NPO achieves ARR > 1.0 across all three domains (CF: 1.120, PE: 1.300, HK: 1.140), while gradient-based baselines collapse to ARR ≤ 0.085 in every domain, showing that the adjacent-retention gains are not confined to a single knowledge setting. Additional per-domain and per-model results are provided in Appendix E.

6.5 Scalability: 7B vs. 32B

We evaluate all five methods on **Qwen2.5-32B-Instruct** (NF4 4-bit quantization, same LoRA hyperparameters) on the same 9-target subset (3 per domain); see Table 8 (Appendix J). AR-NPO remains in the surgical zone at *both* scales: at 32B it achieves FE=0.718, ARR=1.964 (9/9 targets), improving over the 7B run (FE=0.494, ARR=1.155, 6/9 targets). GA and NPO collapse to ARR=0.000 at 32B despite FE=1.000, confirming blast-radius damage is not a small-model artifact. SimNPO undergoes a scale-dependent reversal (FE=0→1, ARR=1→0); GD enters the surgical zone for 8/9 targets but with substantially higher variance thank AR-NPO (σ=0.257 vs. AR-NPO's σ=0.068),

Full-parameter unlearning. To test whether these findings are an artifact of low-rank adaptation, we repeat the comparison using full-parameter fine-tuning (all 7.6B parameters, no LoRA) on the

same 9 targets. The qualitative pattern persists: GA and NPO collapse to $\mathrm{ARR}{\approx}0$, while AR-NPO maintains $\mathrm{ARR}{>}1$ in every domain. Notably, with full-parameter optimization, AR-NPO achieves $\mathrm{FE}{=}0.994$ while maintaining $\mathrm{ARR}{=}1.505$, suggesting that the FE–ARR trade-off observed under LoRA is substantially capacity-dependent rather than an inherent limitation of the objective (Appendix N).

6.6 Circularity Audit

To rule out generator-configuration artifacts, we evaluate $\mathrm{ARR}_{\mathrm{held}}$ on a held-out adjacent set generated under an independent prompting configuration. On Qwen2.5-7B-Instruct across all 45 targets, $\mathrm{ARR}_{\mathrm{held}}{=}1.134$ remains close to $\mathrm{ARR}_{\mathrm{train}}{=}1.156$ (mean gap 0.022), and all 45 targets maintain $\mathrm{ARR}_{\mathrm{held}}{>}1.0$, confirming that ARR gains are not a generator artifact (Table 9, Figure 5 and full results in Appendix K). Cross-model validation with Claude Haiku 4.5 as the generator further yields $\mathrm{ARR}{=}1.057$ ruling out GPT-specific answer phrasing as a confound.

6.7 Data vs. Method Ablation

On 9 targets, replacing target-specific adjacent QA with in-domain off-target QA reduces ARR from 1.259 to 0.961, while providing the correct Prober QA to NPO yields only 0.384. Thus, both targeted data and the adjacent-retain objective contribute to the gains, with the objective having the larger effect (Appendix L).

7 Conclusion

We introduced the *blast-radius* problem in LLM unlearning—the unintended degradation of semantically adjacent knowledge during target forgetting—and addressed it with BlastBench, an adjacency-aware benchmark spanning 45 targets and three domains, and ARC-Unlearn, a multi-agent framework with an explicit adjacent-retain objective. Across three 7B models, AR-NPO is the only evaluated method to achieve non-trivial forgetting while preserving adjacent knowledge above the original-model baseline ($\mathrm{FE}{=}0.566$, $\mathrm{ARR}{=}1.259$), with a controllable FE–ARR trade-off through α. This behavior persists at 32B scale and on independently generated adjacent evaluation data; moreover, full-parameter optimization reaches near-complete forgetting ($\mathrm{FE}{=}0.994$) while maintaining $\mathrm{ARR}{=}1.505$. These results establish semantic locality as a central consideration for LLM unlearning: effective methods must account not only for what knowledge is removed, but also for which neighboring knowledge should be preserved.

Limitations

Under LoRA, AR-NPO trades forget efficacy for adjacent retention ($\mathrm{FE}{=}0.566$ at $\alpha{=}1$), although full-parameter results suggest this trade-off is capacity-dependent (Appendix N). The Prober's LLM-based, bounded BFS may miss weak, long-range, or compositional dependencies; joint probing finds such interactions for 19.4% of targets, with a modest mean score increase of 0.31 (Appendix O). Evaluation primarily covers three 7B models and 45 targets across three domains, with 32B and full-parameter experiments limited to smaller subsets. Finally, BlastBench relies on automated generation, and human validation shows less reliable adjacency boundaries for Copyright Fiction; broader, human-authored evaluations remain future work.

Ethics Statement

AR-NPO aims to make deployed models safer by removing targeted knowledge while preserving unrelated capabilities. The Privacy Entities domain uses entirely *fictional* individuals with synthetic attributes (Section 5) and contains no real personal data. The Hazardous Knowledge domain covers only publicly documented cyberattack and vulnerability knowledge already present in pretraining corpora—question–answer probes for evaluation, not operational instructions, with no novel exploit content—while its retain set preserves legitimate defensive-security concepts. We release BlastBench to support reproducible research on safe unlearning, judging this benefit to outweigh the low residual risk of its probes.

Reproducibility Statement

All base models are public (Qwen2.5-7B/32B-Instruct, Llama-3, Mistral). The ARC-Unlearn pipeline, the BlastBench dataset with its generation scripts, and all training and evaluation code are at the anonymized repository https://anonymous.4open.science/r/ARC-Unlearn-B78E. Prompt templates (Appendix C), the full hyperparameter configuration (Appendix D), and the theoretical assumptions (Appendix F) are provided in full; all reported numbers use fixed random seeds.

Use of Large Language Models

We disclose LLM use following the ICLR AI-use policy. **Required:** LLMs generate the entire BlastBench benchmark (13,803 QA pairs; GPT-5.4, Section 5) and constitute the ARC-Unlearn pipeline itself—the Prober Agent estimates the blast radius and the Synthesizer Agent generates the retain and forget sets (Section 4), with prompt templates in Appendix C. **Recommended:** a general-purpose assistant aided only minor language editing and LaTeX formatting—not idea generation, the theoretical results, experiment design, or any reported number. We have reviewed all AI-assisted content and take full responsibility for it.

References

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015.

Hwan Chang and Hwanhee Lee. Which retain set matters for LLM unlearning? a case study on entity unlearning. In *Findings of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2502.11441.

Chahana Dahal, Ashutosh Balasubramaniam, and Zuobin Xiong. GONE: Structural knowledge unlearning via neighborhood-expanded distribution shaping. *arXiv preprint arXiv:2603.12275*, 2026.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. arXiv:2506.12618.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2410.02760.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations (ICLR)*, 2025a.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. BLUR: A benchmark for LLM unlearning robust to forget-retain overlap. In *International Conference on Machine Learning (ICML)*, 2025b.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14389–14408, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Weipeng Jiang, Juan Zhai, Shiqing Ma, Ziyan Lei, Xiaofei Xie, Yige Wang, and Chao Shen. Holistic audit dataset generation for LLM unlearning via knowledge graph traversal and redundancy removal. *arXiv preprint arXiv:2502.18810*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Myeongseob Ko, Hoang Anh Just, Charles Fleming, Ming Jin, and Ruoxi Jia. Probing knowledge holes in unlearned LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2511.00030.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 28525–28550, 2024.

Wenxuan Li, Zhenfei Zhang, Mi Zhang, Geng Hong, Mi Wen, Xiaoyu You, and Min Yang. From anchors to supervision: Memory-graph guided corpus-free unlearning for large language models. *arXiv preprint arXiv:2604.13777*, 2026.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv preprint arXiv:2402.16835*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.06460.

Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing, 2017.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2308.08155.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the 31st IEEE Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Yisheng Zhong, Sijia Liu, and Zhuangdi Zhu. Harmonizing multi-objective LLM unlearning via unified domain representation and bidirectional logit distillation. *arXiv preprint arXiv:2604.15482*, 2026.

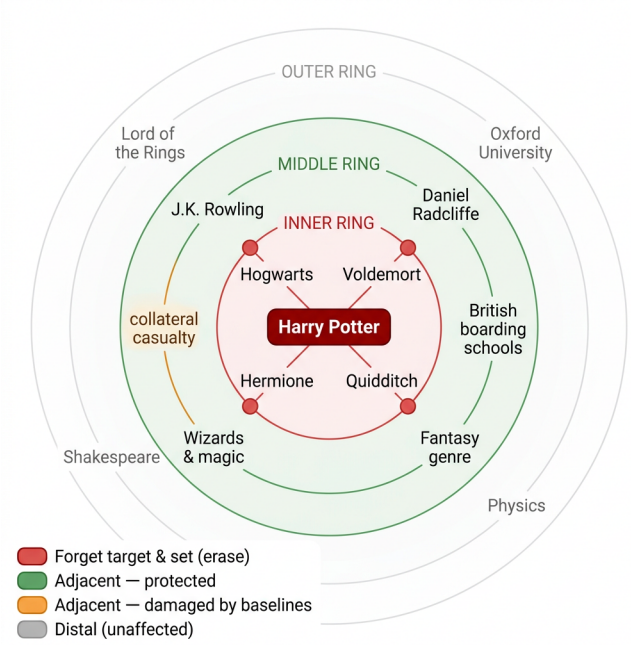


Figure 1: Conceptual illustration of the blast radius for the Harry Potter forget target. Forget-set concepts lie near the target; semantically adjacent knowledge occupies the blast radius and may be damaged by unlearning without targeted protection, while distal concepts remain largely unaffected.

A Human Validation Protocol

Sampling. Using a fixed random seed, we drew a stratified sample of 40 distinct (target, adjacent-concept) pairs from each domain (CF/PE/HK), for 120 items total. For each pair, we randomly selected one representative QA item. Presentation order was shuffled to reduce ordering bias.

Annotation. Two annotators independently labeled each item on three yes/no criteria: *adjacency validity* (the concept is related to but distinct from the forget target), *non-leakage* (the QA is answerable without knowledge of the target), and *answer correctness*. Annotators followed written guidelines with calibration examples not drawn from the evaluation sample and did not see each other's labels.

Agreement. We report Cohen's κ for the two raters and, because pass rates are high, Gwet's AC1 as a prevalence-robust check ($0.99/0.96/0.98$ for the three criteria). The annotators disagreed on 1, 4, and 2 items, respectively. The sample, guidelines, and anonymized annotations are released with the benchmark.

Criterion	κ	Pass rate (%)			
		All	CF	PE	HK
Adjacency valid	0.96	88.3	72.5	100	92.5
Non-leakage (firewall)	0.80	89.2	70.0	100	97.5
Answer correct	0.79	95.0	92.5	97.5	95.0

Table 4: Human validation of BlastBench on 120 sampled adjacent concepts/QA pairs (40 per domain). κ: Cohen's inter-annotator agreement. Pass rate (%): fraction of items both annotators judged to satisfy the criterion.

B BLASTBENCH: COMPLETE DOMAIN STATISTICS

Domain	Targets	F tot	A tot	F avg	A avg	Concepts
CF	15	750	4,276	50.0	285.1	15.1
PE	15	703	3,101	46.9	206.7	8.3
HK	15	750	4,223	50.0	281.5	16.7
All	45	2,203	11,600	49.0	257.8	13.4

Table 5: Complete BlastBench statistics by domain. $|F|$ and $|A|$ denote the total forget and adjacent-retain QA pairs; per-target columns report domain means. BlastBench contains 45 forget targets and 13,803 QA pairs.

Privacy Entities Construction Note. PE targets are entirely fictional profiles modelled after the TOFU (Maini et al., 2024) methodology. Because the profiles are fictional, the blast radius consists of *real* adjacent institutions, professions, and communities (e.g., *Boston cardiology*, *Harvard Medical School*). Forget-set generation uses profile-context injection: the fictional profile's adjacent fields are provided as grounding context, allowing the model to generate factual forget pairs without hallucinating about unknown individuals.

C PROBING PROMPT TEMPLATES

PROBER AGENT • Blast-Radius Discovery

SYSTEM: You are a knowledge cartographer. Your task is to identify concepts in a language model that are at risk of collateral damage when a specific concept is erased.

USER: The concept to erase is: `{forget_target}`.

1. List exactly b real-world concepts closely associated with `{forget_target}` but with independent existence.
2. For each concept, rate 0–10: "If all knowledge of `{forget_target}` were erased, could knowledge of this concept partially reconstruct it?"
3. For concepts rated >5, describe the reconstruction path in one sentence.

Output as JSON: `{"adjacent_concepts": [...]}`.

SYNTHESIZER AGENT • Adjacent Retain Generation

SYSTEM: You are a precision data curator.

USER: Generate N QA pairs about `{adjacent_concept}` that: (1) demonstrate clear knowledge of `{adjacent_concept}`, (2) do **not** reference `{forget_target}` in any way, (3) focus on aspects independent of `{forget_target}`. Output as JSON: `[{"question": ..., "answer": ...}]`.

D HYPERPARAMETER GRID

Parameter	Value
LoRA rank r	8
LoRA α_{lora}	16
Target modules	q, k, v, o projections
Training steps	20
Batch size	8
Learning rate	5×10^{-4}
LR schedule	Constant
Optimizer	AdamW (β_1=0.9, β_2=0.999)
β_{npo}	0.1
α (AR-NPO)	1.0 (default); swept {0, 0.25, 0.5, 1, 2}
β (AR-NPO)	0.1

Table 6: Hyperparameter configuration for AR-NPO.

E PER-DOMAIN AND ADVERSARIAL ANALYSIS

Figure 2 shows ARR broken down by domain for Qwen2.5-7B-Instruct (15 targets per domain). AR-NPO consistently achieves ARR > 1.0 across all three domains (CF: 1.120, PE: 1.300, HK: 1.140), while gradient-based baselines remain below ARR ≤ 0.085 in every domain. The PE domain exhibits the highest ARR for AR-NPO (1.300). Panel (b) plots Relearning Resilience against MIA-AUC. Among methods with non-trivial forget efficacy, AR-NPO lies closest to the desired high-resilience, low-distinguishability region.

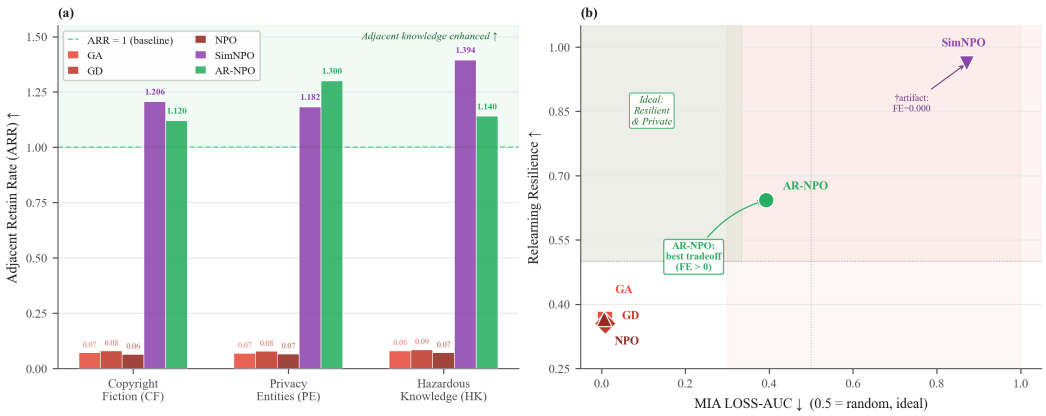


Figure 2: (a) ARR by method and domain. AR-NPO achieves ARR > 1.0 across all three domains; gradient-based baselines remain below ARR < 0.09. SimNPO's high ARR reflects its near-zero forget efficacy. **(b)** The desired region corresponds to high relearning resilience and low MIA distinguishability. Among methods with non-trivial forget efficacy, AR-NPO lies closest to this region.

F EXTENDED ANALYSIS

WHY GENERIC RETAIN SETS DO NOT GUARANTEE ADJACENT RETENTION

We now explain why standard gradient-based unlearning methods are expected to damage adjacent concepts unless those concepts are explicitly protected during training.

Let

$$\ell(\theta; x) = -\log p_\theta(x)$$

denote the cross-entropy loss on example x.

We define the *forget gradient* $\mathbf{g}_F$ as the gradient direction associated with suppressing the forget target F. For Gradient Ascent (GA):

$$\mathbf{g}_F = \nabla_\theta \mathcal{L}_{\text{CE}}(F),$$

NPO-based methods produce update directions that remain closely aligned with the same underlying forgetting gradient.

Similarly, define the *adjacent gradient*

$$\mathbf{g}_A = \nabla_\theta \mathcal{L}_{\text{CE}}(A),$$

which captures the gradient induced by the adjacent retain set A.

The central issue is that forget targets and semantically adjacent concepts often rely on overlapping internal representations.

Let $h(x) \in \mathbb{R}^d$ denote the final hidden representation of input x, and let $W \in \mathbb{R}^{V \times d}$ denote the output projection matrix. For cross-entropy loss, the gradient inner product between two examples can be expressed as

$$\langle \nabla_W \ell(\theta; x_F), \nabla_W \ell(\theta; x_A) \rangle = \langle \mathbf{e}_F, \mathbf{e}_A \rangle \cdot \langle h(x_F), h(x_A) \rangle$$

where $\mathbf{e}_x = p_\theta(x) - y_x$ denotes the softmax prediction residual, and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product.

This decomposition shows that gradient alignment increases when:

1. the examples induce similar prediction residuals, and
2. their hidden representations occupy nearby regions of representation space.

Semantic similarity can therefore induce gradient alignment when both representation and residual similarities are positive. We formalize this mechanism through the following assumption.

Assumption F.1 (Positive Blast-Radius Alignment)**.** For any forget target with a non-trivial blast radius,

$$\langle \mathbf{g}_F, \mathbf{g}_A \rangle > 0.$$

In other words, the gradient used to erase the target also partially points in the direction that damages adjacent concepts.

Theorem F.2 (Gradient-Mediated Blast Radius)**.** *Suppose we perform a forgetting step:*

$$\theta' = \theta + \eta \mathbf{g}_F.$$

using a first-order Taylor expansion of the adjacent loss, we obtain:

$$\mathcal{L}_{CE}(A; \theta') - \mathcal{L}_{CE}(A; \theta) = \eta \langle \mathbf{g}_F, \mathbf{g}_A \rangle + O(\eta^2).$$

Under Assumption F.1, the adjacent loss strictly increases after every forgetting step. Consequently, adjacent performance deteriorates throughout unlearning, causing ARR to decrease over time.

Proof. The result follows directly from a first-order Taylor expansion of $\mathcal{L}_{\text{CE}}(A; \theta)$ around θ:

$$\mathcal{L}_{\text{CE}}(A; \theta') = \mathcal{L}_{\text{CE}}(A; \theta) + \nabla_\theta \mathcal{L}_{\text{CE}}(A; \theta)^\top (\theta' - \theta) + O(\eta^2).$$

Substituting $\theta' - \theta = \eta \mathbf{g}_F$ gives

$$\mathcal{L}_{\text{CE}}(A; \theta') - \mathcal{L}_{\text{CE}}(A; \theta) = \eta \langle \mathbf{g}_F, \mathbf{g}_A \rangle + O(\eta^2).$$

□

Theorem F.2 suggests that preventing ARR degradation requires retain gradients that meaningfully counteract the overlap between $\mathbf{g}_F$ and $\mathbf{g}_A$. In practice, generic retain datasets such as WikiText may provide gradients that are only weakly aligned with adjacent concepts relevant to the forget target. Consequently, these gradients may not sufficiently offset the harmful overlap induced by forgetting updates. This observation motivates the use of explicit adjacent-retain optimization in AR-NPO.

Theorem F.3 (Adjacent-Loss Preservation Condition)**.** *Consider the AR-NPO update: then the adjacent loss satisfies*

$$\mathcal{L}_{CE}(A;\theta') \leq \mathcal{L}_{CE}(A;\theta)$$

to first order whenever

$$\alpha \geq \alpha^* := \frac{\langle \mathbf{g}_F, \mathbf{g}_A \rangle}{\|\mathbf{g}_A\|^2}.$$

since

$$\langle \mathbf{g}_F, \mathbf{g}_A \rangle > 0,$$

the required adjacent-protection strength is strictly positive.

Proof. Applying a first-order Taylor expansion gives

$$\Delta\mathcal{L}_A = \eta\Big(\langle \mathbf{g}_F, \mathbf{g}_A \rangle - \alpha\|\mathbf{g}_A\|^2\Big) + O(\eta^2).$$

requiring

$$\Delta\mathcal{L}_A \leq 0$$

yields

$$\alpha \geq \frac{\langle \mathbf{g}_F, \mathbf{g}_A \rangle}{\|\mathbf{g}_A\|^2}.$$

□

G SimNPO Scale-Sensitive Behavior

Under the BlastBench setting, SimNPO exhibits negligible forget efficacy (FE≈0) across all 45 targets and all three 7B-scale models. This behavior persists across the full β sweep (Appendix H), suggesting that the failure is not specific to the default hyperparameter choice. Consequently, SimNPO's high ARR, DRR, and relearning-resilience scores at 7B should not be interpreted as successful preservation under unlearning: they occur in the absence of measurable forgetting.

At 32B scale (§6.5), this behavior reverses. SimNPO reaches FE=1.0 while adjacent retention collapses, moving from a high-retention/no-forgetting regime at 7B to a high-forgetting/low-retention regime at 32B. Thus, unlike AR-NPO, which remains in the Surgical Zone at both scales, SimNPO does not exhibit a stable forgetting–retention behavior across model scales under our experimental setting.

The mechanism underlying this reversal remains unclear. It may reflect scale-dependent optimization dynamics or interactions between the SimNPO objective and model representations, but our experiments do not isolate the cause. We therefore treat the result as evidence of scale sensitivity rather than attributing it to a specific mechanism.

H SimNPO β-Sweep Analysis

SimNPO (Fan et al., 2025) replaces the NPO log-ratio objective with a length-normalized log-probability loss,

$$\mathcal{L}_{\text{SimNPO}} = \text{softplus}(-\beta\bar{\ell}_p)/\beta,$$

eliminating the need for a frozen reference model. To verify that our main results are not an artifact of the default β=0.1 setting, we swept $\beta \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$ on Qwen2.5-7B-Instruct. Llama-3 and Mistral-7B exhibit the same qualitative behavior.

β	FE ↑	ARR ↑	DRR ↑	FP ↑
0.1	0.000	1.315	1.000	—
0.5	0.000	0.609	0.500	0.464
1.0	0.000	1.210	1.000	0.932
2.0	0.000	1.211	1.000	0.933
5.0	0.000	1.211	1.000	0.934
10.0	0.000	1.211	1.000	0.934

Table 7: SimNPO β-sweep results (Qwen2.5-7B-Instruct, same pattern on Llama-3 and Mistral-7B).

Forget efficacy remains zero across all tested β values, indicating that SimNPO fails to induce measurable target forgetting in this 7B setting. Most settings preserve ARR and DRR near or above their original-model baselines, consistent with the negligible forgetting observed in Table 7.

The β=0.5 setting behaves differently: ARR drops to 0.609, DRR to 0.500, and FP to 0.464, while forget efficacy remains zero. Thus, the increased degradation at this setting does not correspond to improved target forgetting, but instead affects non-target performance. Across the tested range, varying β therefore does not recover a regime that simultaneously achieves measurable forgetting and preserves adjacent knowledge.

Combined with the qualitatively different behavior observed at 32B scale (§6.5), these results suggest that SimNPO is sensitive to both hyperparameter choice and model scale under the BlastBench setting. We leave identifying the optimization mechanism underlying this behavior to future work.

I The FE–ARR Pareto Frontier

Figure 3 shows the α-sweep from α=0 (NPO-equivalent: FE = 0.996, ARR = 0.041) to α=2.0 (FE = 0.300, ARR = 1.232), with α=1.0 entering the surgical zone (FE = 0.553, ARR = 1.126). Increasing α smoothly trades forget efficacy for adjacent retention; no gradient-based baseline can access this curve regardless of hyperparameters.

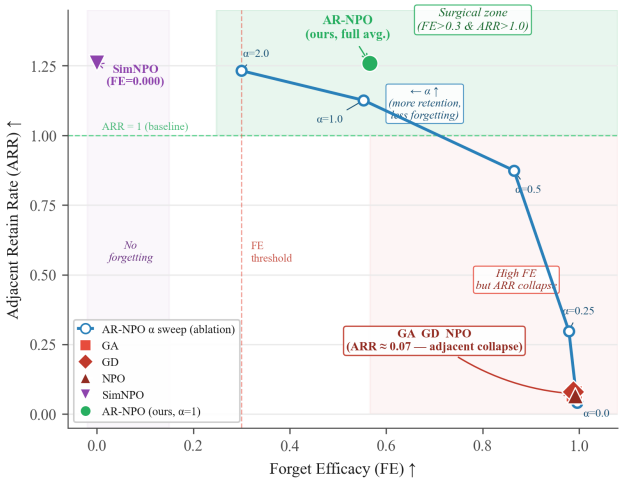


Figure 3: FE–ARR Pareto frontier. The blue curve is an ablation sweep ($\alpha \in \{0, 0.25, 0.5, 1, 2\}$) on Qwen2.5-7B-Instruct across 9 targets; it illustrates the *shape* of the FE–ARR tradeoff as the adjacent-retain weight varies. All scatter markers—including the green AR-NPO marker—share the same evaluation scope as the baselines: 45 targets × 3 models. The green-shaded region indicates the surgical zone (FE $>$ 0.3 and ARR $>$ 1.0). Only AR-NPO enters this region; gradient-based baselines remain confined to the high-FE, low-ARR regime. SimNPO occupies the high-ARR, zero-FE corner, reflecting near-original model behavior rather than effective unlearning.

J SCALABILITY: FULL ANALYSIS

We evaluate all five methods on **Qwen2.5-32B-Instruct** (NF4 4-bit LoRA, same LoRA hyperparameters as 7B) on the same 9-target subset (3 per domain). Results are reported in Figure 4 and Table 8.

	7B		32B		
Method	FE	ARR	FE	ARR	Surg.
GA	0.996	0.036	1.000	0.000	0/9
GD	0.995	0.040	1.000±0.001	1.393±0.257	8/9
NPO	0.996	0.035	1.000	0.000	0/9
SimNPO†	0.000	1.192	1.000	0.000	0/9
AR-NPO	0.494	1.155	**0.718±0.277**	**1.964±0.068**	**9/9**

Table 8: FE and ARR on Qwen2.5-7B vs. 32B (NF4 4-bit LoRA), averaged over the same 9 targets (3 per domain) at both scales. Mean±σ shown when $\sigma>0.001$. Surg.: targets satisfying FE>0.3 and ARR>1.0. †SimNPO reverses failure mode across scales (see §6.5).

AR-NPO scales favorably. At 32B, AR-NPO achieves FE$=0.718$ and ARR$=1.964$, improving over the corresponding 7B evaluation on the same 9-target subset (FE$=0.494$, ARR$=1.155$). All nine 32B targets satisfy the surgical-zone criteria (9/9), compared to 6/9 at 7B. Larger models may provide increased capacity for simultaneously suppressing target behavior while preserving semantically adjacent concepts.

Gradient-based blast radius persists or worsens at 32B. GA and NPO achieve FE$=1.000$ but collapse to ARR$=0.000$ at 32B, while maintaining FP≈0.91 and DRR≈0.92, confirming that adjacent-retention collapse is genuine and not an artifact of overall model degradation. Severe blast-radius damage is therefore not limited to smaller models.

SimNPO undergoes a scale-dependent failure-mode reversal. At 7B, SimNPO produces no measurable forgetting (FE$=0.000$), so its high ARR occurs in the absence of successful unlearning. At 32B, this behavior reverses: SimNPO reaches FE$=1.000$ while adjacent retention collapses (ARR$=0.000$). Thus, across the two evaluated scales, SimNPO transitions from a no-forgetting/high-retention regime to a high-forgetting/low-retention regime. The mechanism underlying this reversal remains unclear and warrants further investigation. Among the evaluated methods, AR-NPO is the only one satisfying the surgical-zone criteria for all targets at 32B while also exhibiting surgical-zone behavior at 7B.

GD improves at 32B but exhibits higher variability. At 32B, GD enters the surgical zone for 8/9 targets, reaching ARR$=1.393\pm0.257$ while maintaining FE≈1.000. This represents a substantial improvement over its 7B behavior despite using only a generic retain set. However, its target-level ARR variability is considerably higher than that of AR-NPO ($\sigma=0.257$ vs. 0.068), and one HK target exhibits near-complete adjacent-retention collapse (ARR≈0). Thus, while GD performs strongly on most targets at 32B, its adjacent retention is less consistent across the evaluated targets than that of AR-NPO.

K CIRCULARITY AUDIT: PER-TARGET SCATTER

To test whether ARR>1.0 reflects genuine adjacent retention rather than evaluation circularity, we regenerated held-out adjacent QA using an independent prompting pipeline and re-evaluated all 45 targets. Results are reported in Figure 5 and Table 9.

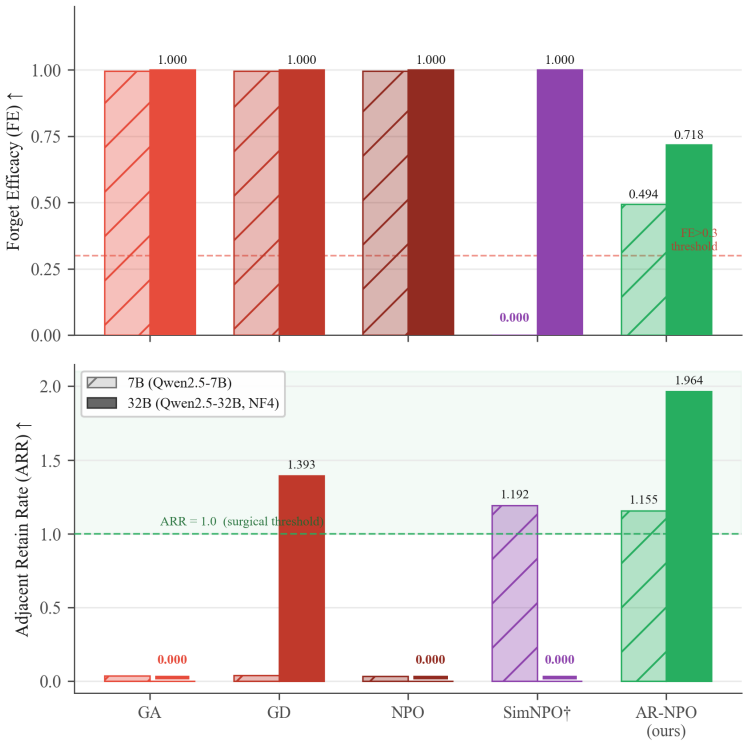


Figure 4: FE and ARR across model scales (7B vs. 32B) for all five methods. AR-NPO is the only method maintaining ARR>1.0 at both scales;

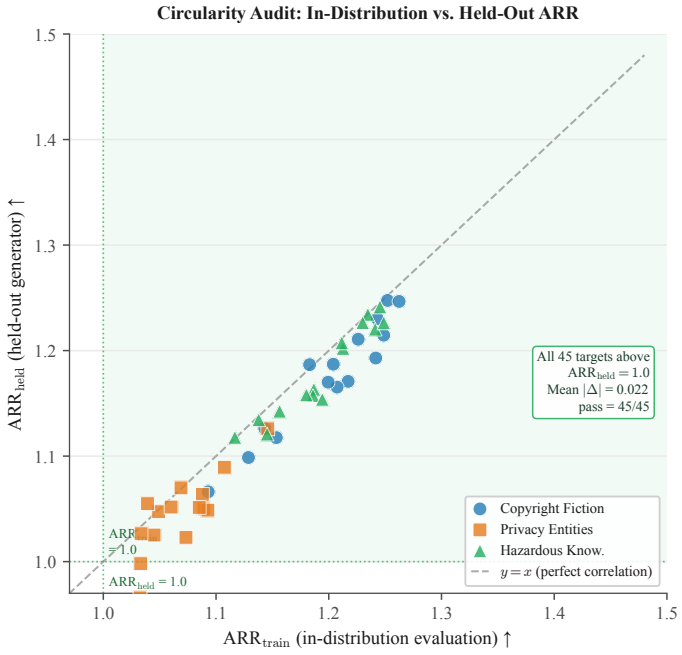


Figure 5: Per-target circularity audit (45 targets, Qwen2.5-7B-Instruct). x-axis: ARR on in-distribution evaluation; y-axis: ARR on held-out questions from an independent generator. Colours: CF (blue), PE (orange), HK (green). Dashed diagonal: perfect correlation ($y{=}x$). All 45 points lie above $\mathrm{ARR}_{\mathrm{held}}{=}1.0$, with tight clustering along the diagonal ($|\Delta|{<}0.04$ per domain).

Domain	$\text{ARR}_{\text{train}}$	ARR_{held}	$\lvert\Delta\rvert$	Pass
Copyright Fiction	1.2004	1.1747	0.026	15/15
Privacy Entities	1.0645	1.0384	0.026	15/15
Hazardous Know.	1.2042	1.1901	0.014	15/15
All	**1.1563**	**1.1344**	**0.022**	**45/45**

Table 9: Circularity audit for AR-NPO (Qwen2.5-7B-Instruct, all 45 targets). $\text{ARR}_{\text{train}}$: in-distribution evaluation. ARR_{held}: held-out questions from an independent generator. Pass = targets where $\text{ARR}_{\text{held}}>1.0$. Mean gap 0.022 confirms ARR>1.0 is not a circularity artifact.

To further rule out GPT-family phrasing artifacts, we regenerated held-out adjacent QA using Claude Haiku 4.5 (a different model family) and re-evaluated ARR on the same 45 targets. $\text{ARR}_{\text{claude}}=1.057$ confirming that the adjacent-retention effect transfers across independent question generators and model families.

L Data vs. Method Ablation

Table 10 examines whether AR-NPO's adjacent-retention gains arise from the Prober-selected data, the training objective, or their combination on the 9-target subset (3 per domain). Providing NPO with the correct Prober-selected adjacent QA improves ARR from 0.067 to 0.384, but remains well below the original-model retention baseline. Using AR-NPO with off-target, in-domain adjacent QA further increases ARR to 0.961, while combining AR-NPO with the correct target-specific adjacent QA reaches ARR$=1.259$. These results indicate that both the optimization objective and target-specific adjacent data contribute to adjacent retention, with their combination required to achieve ARR>1 in this ablation.

Variant	ARR↑	FE↑	DRR↑
NPO (no adjacent retain)	0.067	0.992	0.919
NPO + adj QA (Prober)	0.384	0.980	0.932
AR-NPO (wrong adj, in-domain)	0.961	0.616	0.991
AR-NPO (Prober adj)	**1.259**	**0.566**	**0.993**

Table 10: Data and objective ablation on Qwen2.5-7B-Instruct across 9 targets (3 per domain). *Wrong adj* uses adjacent QA rotated across targets within the same domain, preserving domain relevance while removing target-specific adjacency. *NPO+adj* uses the correct Prober-selected adjacent QA with the NPO-based comparison objective.

M Per-Model Results

Table 11 reports FE and ARR separately for each base model, confirming that the trends in the main results (Table 2) hold consistently across all three architectures (Qwen2.5-7B, Llama-3-8B, and Mistral-7B).

N Full-Parameter Unlearning

To verify that our conclusions are not an artifact of low-rank adaptation, we repeat the comparison with *full-parameter* fine-tuning (all 7.6B weights, no LoRA) on 9 targets (3 per domain), using 8-bit Adam and gradient checkpointing. Table 12 reports the results.

The qualitative picture is consistent with the LoRA setting: GA and NPO reach FE$=1.0$ but drive adjacent retention to ARR≈ 0 (GA additionally degrades fluency, FP$=0.545$), whereas AR-NPO keeps ARR>1 across all three domains. The blast-radius failure of adjacency-agnostic unlearning, and AR-NPO's correction of it, are therefore *not* artifacts of low-rank adaptation.

Method	Metric	Qwen2.5	Llama-3	Mistral	Avg
GA	FE	0.989	0.991	0.990	0.990
	ARR	0.072	0.074	0.073	0.073
NPO	FE	0.991	0.993	0.992	0.992
	ARR	0.067	0.068	0.066	0.067
AR-NPO	FE	0.553	0.572	0.573	0.566
	ARR	1.187	1.295	1.295	1.259

Table 11: Per-model FE and ARR for selected methods (45 targets each). AR-NPO's ARR/FE averages match Table 2 by construction; Qwen2.5 ARR is computed from domain-level aggregates (Fig. 2, panel a). Results are consistent across architectures.

Method (full-FT)	FE↑	ARR↑	DRR↑	FP↑
GA	1.000	0.000	0.692	0.545
NPO	1.000	0.001	1.000	0.986
AR-NPO	0.994	**1.505**	0.969	0.968

Table 12: Full-parameter unlearning on 9 targets (CF/PE/HK, 3 each), Qwen2.5-7B-Instruct. Gradient-based baselines still collapse adjacent retention ($\mathrm{ARR}{\approx}0$) under full fine-tuning; AR-NPO preserves it in every domain (per-domain ARR: CF 1.51, HK 1.42, PE 1.58).

Interestingly, under full-parameter optimization, AR-NPO attains near-complete measured forgetting ($\mathrm{FE}{=}0.994$) together with $\mathrm{ARR}{=}1.505$. This contrasts with the stronger FE–ARR trade-off observed under LoRA, where the default $\alpha{=}1$ setting achieves $\mathrm{FE}{=}0.566$ and $\mathrm{ARR}{=}1.259$. The full-parameter result therefore shows that strong forgetting and above-baseline adjacent retention are not inherently incompatible for AR-NPO; with greater optimization capacity, both can be achieved simultaneously in this setting.

One possible explanation is that full-parameter optimization provides additional degrees of freedom for separating target suppression from adjacent preservation, whereas LoRA constrains both objectives to a lower-dimensional update space. Our experiments do not directly test this mechanism, however, and we leave a systematic analysis of the interaction between adaptation capacity and the FE–ARR trade-off to future work.

O Joint Multi-Concept Reconstruction

Our Prober scores each candidate concept *individually* and admits those above threshold τ. A natural concern is whether concepts scored below τ individually could nevertheless reconstruct the forget target when considered jointly. We test this directly. For each target, we elicit a candidate pool with individual reconstruction scores, collect the below-threshold concepts, and ask an independent judge (`gpt-5.4`) to score reconstruction of the target from the entire subset jointly. We average three judgments per target to reduce judge variance.

Across 45 targets, 36 contain at least two below-threshold concepts and therefore permit this analysis. Among these, joint reconstruction crosses the admission threshold for 19.4% of targets. The mean joint reconstruction score is 4.09, compared with 3.78 for the highest-scoring individual below-threshold concept, an average increase of only $\Delta{=}0.31$ on the 0–10 scale. Thus, joint consideration produces only a modest average increase in reconstruction score, although a minority of targets exhibit interactions that individual probing does not capture.

These results identify a limitation of the current Prober: independently scoring candidate concepts can miss reconstruction risk that emerges only from combinations of concepts. Such cases are concentrated in the Copyright Fiction domain, where related concepts can form densely interconnected narratives. Extending the Prober to evaluate selected concept combinations is therefore a promising direction for improving blast-radius coverage.

P ADVERSARIAL EVALUATION: DEFINITIONS

Relearning Resilience. Following Hu et al. (2025a), we merge each unlearned LoRA adapter into the base model, attach a fresh $r{=}4$ LoRA, and finetune for 50 steps on 20% of the forget set. Formally,

$$\text{Resilience} = \frac{\text{FE}(\hat{\mathcal{M}})}{\text{FE}(\mathcal{M}^*)} \tag{5}$$

where $\mathcal{M}^*$ is the unlearned model and $\hat{\mathcal{M}}$ is the model after the 50-step relearning fine-tune described above. A value of 1 indicates the forgotten knowledge cannot be recovered; lower values indicate easier relearning.

Membership Inference Attack (MIA). We compute LOSS-MIA AUC following Yeom et al. (2018), using per-sample cross-entropy loss to distinguish forget-set examples from non-members sampled from WikiText. An AUC near 0.5 indicates that the attack cannot distinguish the two groups better than chance using this loss signal. We therefore report $|\text{AUC} - 0.5|$ as a *membership distinguishability* score, where lower values indicate lower distinguishability under this attack.