FLOUR: Mixing the Right Ingredients for LLM Unlearning in Latent Space

Dung Thuy Nguyen[1], **Giang Nguyen**[2], **Taylor T. Johnson**[1], **Kevin Leach**[1]
[1]Vanderbilt University, [2]Meta
{dung.t.nguyen, taylor.johnson, kevin.leach}@vanderbilt.edu
giangnguyen95@meta.com

Abstract

As large language models are deployed at scale, the ability to selectively remove knowledge has become critical for privacy, safety, and legal compliance. Existing post-training unlearning methods suffer from a persistent tension: unbounded divergence signals over-unlearn, degrading utility and producing incoherent outputs, while fixed sample-agnostic targets and reference-relative penalties under-unlearn, leaving traces recoverable under paraphrase or membership-inference attacks. We attribute both failures to a common structural cause: the forgetting objective is decoupled from any sample- or parameter-level notion of what retention requires. We present FLOUR (Forgetting via Learned On-manifold Unlearning Representations), a latent unlearning framework in which a lightweight encoder produces a per-sample suppression direction. This direction is anchored to the null region of the reference model's activations, keeping it on-manifold and disentangled from retention-critical directions, which enables forgetting while preserving coherent generation and enhancing privacy. Results on the Open-Unlearning benchmark suite show that FLOUR most closely matches the retrain oracle across the unlearning trilemma of forgetting, retention, and privacy, substantially outperforming prior baselines. Our code is publicly available at https://github.com/judydnguyen/flour-unlearning.

1 Introduction

Large language models trained on web-scale text have achieved strong performance in reasoning, coding, creative writing, and open-ended dialog (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023). Yet as these systems are deployed more widely, a less welcome property of their training has become apparent: they *memorize*. LLMs can reproduce verbatim passages, reconstruct personal details, and surface sensitive information long after training has ended (Carlini et al., 2021, 2023; Xiong et al., 2025; Cheng et al., 2025; Mireshghallah et al., 2024). This raises serious concerns around privacy, safety, and legal compliance, particularly under regulatory frameworks such as the GDPR (European Parliament and Council of the European Union, 2016), which enshrine the right to be forgotten. Honoring such requests in deployed LLMs is far from straightforward: trained models may have internalized personally identifiable information or copyrighted material, and retraining from scratch is impractical at modern scale and unrealistic for incremental requests (Strubell et al., 2019; Hoffmann et al., 2022). The same need extends to *safety*: when hazardous knowledge (e.g., toxic content or dangerous expertise) is discovered only after deployment, unlearning can excise it in place, without the cost of retraining or the brittleness of output-level guardrails (Li et al., 2024). Here the unit of forgetting is a *fact* the model internalized—not a fixed token string—which must vanish under *any* phrasing while fluency and unrelated knowledge survive (Figure 2).

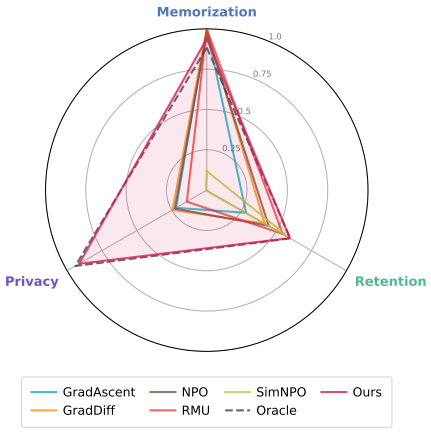


Figure 1: **The unlearning trilemma.** Normalized scores on Memorization (forgetting), Retention (utility), and Privacy (larger is better on every axis) introduced by Dorna et al. (2026). FLOUR (ours) closely matches the oracle on all three axes, whereas prior methods collapse on at least one, especially privacy.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 51033–51051
October 24-29, 2026 ©2026 Association for Computational Linguistics

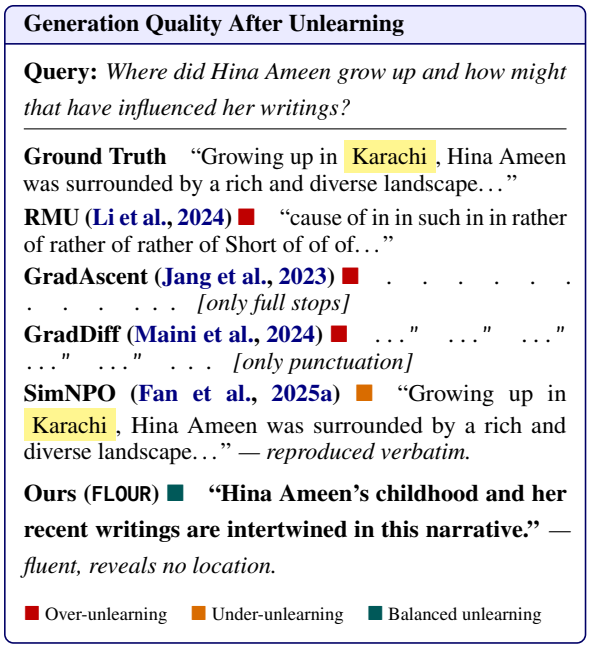


Figure 2: **What "forgetting" looks like for an LLM.** The target is a *fact* (Hina Ameen grew up in Karachi), probeable under any phrasing. Prior methods over-unlearn (degenerate text) or under-unlearn (leak the location); FLOUR answers fluently without leaking.

Machine unlearning has emerged as the principal post-hoc response: modify a trained model's parameters to remove targeted data without full retraining (Bourtoule et al., 2021; Cao et al., 2024; Xu et al., 2023; Fan et al., 2025b). Despite considerable recent progress, existing methods consistently fail to deliver this goal due to two complementary axes. *Behaviorally*, methods either erase too aggressively, harming utility and collapsing generation into degenerate outputs (**over-unlearning**), or too shallowly, leaving residual traces of the targeted knowledge that are recoverable under paraphrase or adversarial prompting (**under-unlearning**); few strike a balance where forgetting is decisive yet generation stays coherent. *From a security angle*, recent results on the Open-Unlearning benchmark (Dorna et al., 2026) surface a third axis that prior evaluations have largely overlooked: even methods that appear to forget behaviorally can leak the forget set to membership-inference and extraction attacks (Duan et al., 2024; Yao et al., 2024; Deepak et al., 2025; Doshi and Stickland, 2024; Kim et al., 2026). Across mainstream unlearning methods, no approach simultaneously balances the resulting *trilemma* of forgetting quality, retention, and privacy leakage (cf. Figure 1).

Where existing methods break down. The failure appears to have a structure : Early gradient-ascent approaches (Jang et al., 2023; Maini et al., 2024) are brittle and indiscriminate. Preference-based methods (NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025a), UNDIAL (Dong et al., 2025), APO (Mekala et al., 2025)) yield more stable suppression but share a common weakness: by operating only on outputs, they leave the internal representations encoding the forgotten knowledge largely intact, so forgetting stays shallow and recoverable. This has motivated a shift toward representation-level unlearning, which intervenes directly in the model's hidden states. RMU (Li et al., 2024) steers forget activations toward a fixed random vector at a chosen layer, and LUNAR (Shen et al., 2025) replaces this vector with a predefined refusal direction. Both yield harder-to-recover forgetting, but share a deeper problem: the suppression target is fixed in advance, independently of the forget sample and of how it interacts with retention. When it overlaps with retain-relevant directions, forgetting and utility become coupled: a stronger steering force degrades utility (over-unlearning), while a weaker one leaves residual traces of the target (under-unlearning). This is not specific to any single method, but a structural consequence of fixing the suppression target as a sample-agnostic hyperparameter, decoupled from what retention requires, a failure mode we call *suppression target entanglement*.

Our approach. Resolving this tension requires rethinking *what* we steer toward, not just how. We propose three properties for the suppression target: it must be (i) learned rather than chosen by hand, (ii) conditioned on each forget sample, and (iii) anchored to an on-manifold, answer-ablated reference state. We instantiate these in FLOUR (Forgetting via Learned On-manifold Unlearning Representations): a small encoder predicts a per-sample suppression direction, anchored to the on-manifold activation the reference model produces for the same prompt with its answer removed. This direction is expressive enough to drive strong forgetting yet structurally disentangled from the directions the model must preserve, so FLOUR forgets targeted content while still generating coherent text.

Contributions.

- We identify *suppression target entanglement* as the structural failure mode of fixed-target representation-level unlearning, linking it to over- and under-unlearning.
- We present FLOUR, a latent unlearning framework whose per-sample suppression target is learned

and anchored to an on-manifold, answer-ablated reference state, providing principled disentanglement of forgetting and retention.
- We evaluate on TOFU and MUSE with ablations and robustness analyses. FLOUR uniquely approaches the retrain oracle on all three TOFU trilemma axes, best approaches it on MUSE verbatim memorization and retention, and stays robust under paraphrase and post-hoc quantization attacks.

2 Related Work

Machine unlearning aims to remove the influence of specific training data from a model after training has completed, without the prohibitive cost of retraining from scratch. Cao and Yang (2015) introduced the problem and Bourtoule et al. (2021) formalized exact unlearning through sharded training architectures that guarantee removal by design. At LLM scale, however, these exact approaches become impractical, motivating a growing literature on approximate post-training methods (Xu et al., 2023; Yuan et al., 2025), which our work extends.

Output-level unlearning. A first line of work supervises unlearning directly through the model's output distribution. GradAscent and GradDiff (Jang et al., 2023; Maini et al., 2024) maximize the forget-set loss, optionally regularized by a retain term, but are notoriously unstable and prone to catastrophic collapse (Zhang et al., 2024). Preference-based objectives recast unlearning as contrastive optimization: NPO (Zhang et al., 2024) treats forget samples as dispreferred relative to a frozen reference, SimNPO (Fan et al., 2025a) dispenses with the reference model, UNDIAL (Dong et al., 2025) substitutes it with self-distilled adjusted logits, and APO (Mekala et al., 2025) interleaves negative updates on forget data with positive updates on acceptable alternatives. Beyond the choice of objective, Yang et al. (2025) show that uniform weighting across the forget set is suboptimal and propose sample-level reweighting. Despite their differences, these methods share a common limitation: by acting only on outputs, they leave internal representations largely untouched, allowing forgotten knowledge to resurface under paraphrasing or adversarial prompts.

Representation-level unlearning. A second line of work intervenes directly on internal activations, aiming for forgetting that is more robust than output-level supervision. RMU (Li et al., 2024; Huu-Tien et al., 2025) steers forget-sample activations toward a fixed random unit vector at a chosen layer, yielding suppression that resists surface-level prompting attacks. LUNAR (Shen et al., 2025) refines this idea by replacing the random target with a learned or predefined refusal direction, producing more coherent abstention behavior. However, both methods fix the suppression target independently of the forget sample, and neither explicitly constrains how the chosen direction interacts with the retain subspace, leaving open the risk of collateral damage to retained knowledge. FLOUR addresses this gap directly: instead of relying on a fixed target, we learn a per-sample suppression direction anchored to an answer-ablated, on-manifold reference state.

3 Methodology

Setup. Let f_θ denote a pre-trained autoregressive language model parameterized by θ, with hidden states of dimension H. Given a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$, we seek θ^* such that f_{θ^*} behaves as if $\mathcal{D}_f$ had been excluded from training, while matching f_θ on $\mathcal{D}_r$. Here the unit of forgetting is the *knowledge* a sample carries — the underlying facts and associations — not its surface token sequence: success means f_{θ^*} no longer reproduces the target content under *any* phrasing (paraphrase, reformatting, or multiple-choice), while leaving unrelated knowledge and fluency intact. This is a behavioral criterion, measured against the retrain oracle in §4, rather than the deletion of a specific string. We operate at the representation level, intervening at a designated layer l with hidden states $\mathbf{h}_l(x;\theta) \in \mathbb{R}^{S\times H}$. The canonical instance is RMU (Li et al., 2024), which steers $\mathbf{h}_l(x_f;\theta)$ toward a fixed scaled random vector $\alpha\hat{\mathbf{u}}$ shared across all forget samples. FLOUR departs from this formulation in two respects. First, the suppression target is produced by a *learned, per-sample* steering module, so the direction depends on the content of x_f rather than being identical across the forget set. Second, the target is constrained to an *on-manifold, answer-ablated* reference state derived from θ_{ref}, so hidden states are steered toward a representation consistent with the model's own activations for x_f with the answer removed, not an unconstrained point in $\mathbb{R}^H$.

3.1 The FLOUR Framework

FLOUR proceeds in two sequential phases. In **Phase 1**, we train the encoder E_ϕ to produce per-sample suppression directions anchored to the answer-ablated reference state. In **Phase 2**, we

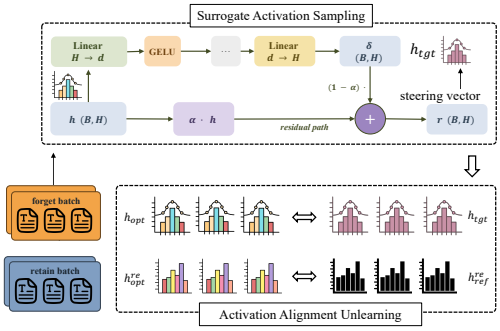


Figure 3: Overview of the FLOUR two-phase framework. **Phase 1 (Surrogate Activation Sampling)** extracts each forget prompt's on-manifold reference activation, defined as the frozen reference model's hidden state with answer tokens masked, and trains an encoder E_ϕ to produce a per-sample suppression direction aligned with this anchor. **Phase 2 (Activation Alignment Unlearning)** freezes E_ϕ and uses its output as a fixed steering target, updating the LLM parameters Θ to align forget activations with the target while a retain loss anchors retain activations to the reference.

freeze E_ϕ and use its output as a fixed per-sample steering target to update the LLM parameters Θ (a small subset of θ at the intervention layers). Figure 3 illustrates the overall framework.

Encoder Architecture. The per-sample encoder $E_\phi : \mathbb{R}^H \to \mathbb{R}^H$ produces the steering direction used as the suppression target for each forget example, conditioned on its pooled hidden state $\mathbf{z}$. Rather than predicting this direction from scratch, the encoder takes $\mathbf{z}$ as a starting point and applies a small learned correction:

$$E_\phi(\mathbf{z}) = \underbrace{\alpha_e\,\mathbf{z}}_{\text{input direction}} + \underbrace{(1-\alpha_e)\,\delta_\phi(\mathbf{z})}_{\text{learned correction}}, \quad (1)$$
$$\delta_\phi(\mathbf{z}) = W_{\text{out}}\,\sigma(W_{\text{in}}\,\mathbf{z} + b_{\text{in}}) + b_{\text{out}}.$$

where σ is the GELU nonlinearity and δ_ϕ is a bottleneck MLP with latent width $d \ll H$ ($W_{\text{in}} \in \mathbb{R}^{d\times H}$, $W_{\text{out}} \in \mathbb{R}^{H\times d}$). Two design choices keep the encoder from drifting away from the reference manifold. First, the mixing coefficient $\alpha_e \in [0,1]$ controls how far $E_\phi(\mathbf{z})$ can move from $\mathbf{z}$: $\alpha_e = 1$ recovers the identity, $\alpha_e = 0$ gives a pure learned direction, and intermediate values interpolate between the two. Second, the narrow bottleneck $d \ll H$ caps the expressive capacity of δ_ϕ, so the correction stays low-rank and cannot encode arbitrary off-manifold directions. The output $E_\phi(\mathbf{z})$ is ℓ_2-normalized before use, so the loss sees only its direction; the displacement magnitude is set separately by the steering coefficient.

Per-sample steering target. From the encoder output we form a unit direction and a per-sample target, using the same form in both phases:

$$\hat{\mathbf{r}}(x) = \frac{E_\phi\big(\bar{\mathbf{h}}_l(x;\theta)\big)}{\big\|E_\phi\big(\bar{\mathbf{h}}_l(x;\theta)\big)\big\|}, \quad (2)$$
$$\mathbf{t}(x) = \mathbf{h}_l(x;\theta_{\text{ref}}) + \alpha\hat{\mathbf{r}}(x),$$

where α is the steering coefficient and $\hat{\mathbf{r}}(x)$ is broadcast across the S token positions of $\mathbf{h}_l(x;\theta_{\text{ref}})$. The target anchors each forget sample to the reference activation $\mathbf{h}_l(x;\theta_{\text{ref}})$ — a stable point in the original feature space — displaced along the learned per-sample direction $\hat{\mathbf{r}}(x)$. Phase 1 shapes $\hat{\mathbf{r}}$ while θ is frozen; Phase 2 steers θ toward $\mathbf{t}$ while E_ϕ is frozen, so the encoder input $\bar{\mathbf{h}}_l(x;\theta)$ drifts as θ updates while the target form remains fixed.

3.2 Phase 1: Surrogate Activation Sampling

In Phase 1 we update only the encoder E_ϕ, holding the LLM parameters θ fixed at their initial values, to calibrate its per-sample direction $\hat{\mathbf{r}}(x_f)$ (Eq. 2) before it is frozen for Phase 2. The encoder already produces a direction for each forget sample; what Phase 1 decides is *where it should point*.

Steering toward an unknown region. Forcing every forget activation toward a fixed random direction, or an explicit output-level refusal ("I don't know"), is too coarse: a sample-independent target displaces the hidden state from the model's representation manifold, where it entangles with and degrades retention, and steering all forget samples to the *same* target leaves a systematic, detectable signature on the forget set that a membership-inference adversary can exploit. We instead steer each sample toward its place in the model's *unknown region* — the state its own representation space occupies when asked a question for which it has no answer to retrieve — so that forgetting is minimal, on-manifold, and per-sample. The intended effect is that, on the forget set, the model behaves as if it had never learned the answer; we verify this behavioral and privacy indistinguishability against the retrain oracle in §4. This region is well-defined: removing the answer from a forget prompt shifts the reference activation by an amount comparable to its own magnitude, almost entirely along a *single direction shared across otherwise unrelated forget examples* (Appendix E), so its endpoint is a real, common, answer-free state rather than an arbitrary point in $\mathbb{R}^H$.

We reach this state with a deliberately minimal intervention. For each forget example $x_f =$

$(x_f^{(1)}, \ldots, x_f^{(S)}) \in \mathcal{V}^S$ we assume a contiguous answer span $\mathcal{A} \subseteq \{1, \ldots, S\}$ identified from the supervision labels, with the remaining positions $\mathcal{Q} = \{1, \ldots, S\} \setminus \mathcal{A}$ forming the question. The *answer-ablated* prompt $\tilde{x}_f$ replaces only the answer tokens with a content-free placeholder:

$$\tilde{x}_f^{(t)} = \begin{cases} x_f^{(t)}, & t \in \mathcal{Q}, \\ \texttt{[MASK]}, & t \in \mathcal{A}, \end{cases} \quad t = 1, \ldots, S, \quad (3)$$

where $\texttt{[MASK]} \in \{\texttt{[UNK]}, \texttt{[PAD]}\}$ is the tokenizer's unknown token (falling back to padding when unavailable). This preserves the sequence length, positional encodings, and attention pattern of x_f; only the answer-token content changes, so $\tilde{x}_f$ is a genuine input differing from the forget example in nothing but the answer. Passing $\tilde{x}_f$ through the frozen reference model and mean-pooling over all attended positions, we ℓ_2-normalize to obtain the anchor direction:

$$\hat{\mathbf{d}}(x_f) = \frac{\bar{\mathbf{h}}_l(\tilde{x}_f; \theta_{\text{ref}})}{\|\bar{\mathbf{h}}_l(\tilde{x}_f; \theta_{\text{ref}})\|}. \quad (4)$$

Pooling over the full attended sequence, rather than only over $\mathcal{A}$, lets the anchor capture the global representation of the answer-free question instead of the local activation at the masked positions alone. Two properties follow. Because $\hat{\mathbf{d}}$ is the activation of the reference model on a real input, it lies on the representation manifold by construction; and because $\tilde{x}_f$ differs from x_f only in the answer span, $\hat{\mathbf{d}}$ is the activation the model would produce for this same question had it never memorized the answer.

Aligning the encoder with the region. With θ frozen, Phase 1 trains the encoder to align its per-sample direction $\hat{\mathbf{r}}(x_f)$ with this anchor under a single cosine objective:

$$\mathcal{L}_{\text{anchor}} = 1 - \frac{1}{B} \sum_{i=1}^{B} \hat{\mathbf{r}}(x_f^{(i)}) \cdot \hat{\mathbf{d}}(x_f^{(i)}), \quad (5)$$

where B is the batch size and $\hat{\mathbf{r}}(x_f^{(i)})$ is the encoder output of Eq. 2. Because both vectors are unit-normalized, the loss depends only on their relative orientation, and minimizing it steers each encoder output toward the answer-free state of its sample.

Properties of the calibrated directions. The directions emitted after Phase 1 have three properties: they are *per-sample* (conditioned on x_f rather than shared across the forget set), *on-manifold* (a low-rank correction to a real activation, kept there by the residual–bottleneck design of E_ϕ), and *answer-ablated* (the sample's own representation with only the answer removed). Steering toward them in Phase 2 suppresses memorized content while preserving the surrounding feature geometry. RMU's fixed $\alpha\hat{\mathbf{u}}$, by contrast, targets an off-manifold point in an unpopulated region of $\mathbb{R}^H$ and collapses the representation, producing the degenerate generations in Appendix F. Ablations confirm this is the load-bearing component: removing the anchor causes steering to over-forget and leak (Table 3).

3.3 Phase 2: Activation Alignment Unlearning

Once the encoder is trained, we freeze ϕ and update the LLM parameters Θ. For each forget sample, we use the Phase 1 encoder to produce the per-sample steering vector $\mathbf{t}(x_f)$ of Eq. (2), which serves as a fixed target throughout Phase 2. The forget loss steers the model's activations toward this target:

$$\mathcal{L}_{\text{forget}} = \frac{1}{|\mathcal{M}|} \sum_{(s,\cdot) \in \mathcal{M}} \left\| \mathbf{h}_l(x_f; \theta)_s - \mathbf{t}(x_f)_s \right\|^2. \quad (6)$$

The retain loss anchors the model's hidden states on $\mathcal{D}_r$ to those of the reference model (activation MSE, following Li et al. (2024)):

$$\mathcal{L}_{\text{retain}} = \frac{1}{|\mathcal{M}_r|} \sum_{(s,\cdot) \in \mathcal{M}_r} \left\| \mathbf{h}_l(x_r; \theta)_s - \mathbf{h}_l(x_r; \theta_{\text{ref}})_s \right\|^2, \quad (7)$$

where $\mathcal{M}_r$ indexes answer positions in retain samples. The combined Phase 2 loss is:

$$\mathcal{L}_{\text{phase2}} = \gamma \cdot w(t) \cdot \mathcal{L}_{\text{forget}} + \alpha_r \cdot \mathcal{L}_{\text{retain}}, \quad (8)$$

where γ and α_r are loss-weighting hyperparameters and $w(t) = \min(1, t/T_{\text{warmup}})$ is an optional linear warmup that prevents large initial parameter updates before the model has settled.

Why the steering target in Eq. (2) avoids the entanglement of fixed targets. The reference activation $\mathbf{h}_l(x_f; \theta_{\text{ref}})$ is a stable baseline in the original feature space, and the displacement $\hat{\mathbf{r}}$ was trained (Phase 1) to align with the sample's answer-ablated, on-manifold direction. Steering toward $\mathbf{t}$ therefore moves the forget representation along a direction that is sample-specific, stays on-manifold near an answer-free version of the sample's own state, and removes the answer content without dragging along the surrounding feature geometry the model relies on for retention—the principled forgetting/retention separation a single fixed random target cannot provide. Full pseudocode and hyperparameter values are in Appendix B and §4.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on two complementary benchmarks. **TOFU** (Maini et al., 2024) is a synthetic benchmark of fictitious author biographies: a base model memorizes the full dataset, and unlearning then erases a designated forget split while preserving the rest. We use three forget fractions, FORGET01/FORGET05/FORGET10, and measure forgetting quality using a Kolmogorov–Smirnov test on the truth ratio against a retrained oracle. **MUSE** (Shi et al., 2024) spans two domains, NEWS (BBC articles) and BOOKS (copyrighted literary text), and evaluates verbal (*VerbMem*) and knowledge (*KnowMem*) memorization, retention, privacy leakage, and extraction robustness.

Models. TOFU experiments use LLAMA-3.2-1B-INSTRUCT and LLAMA-3.2-3B-INSTRUCT (Dubey et al., 2024) fine-tuned on the full TOFU dataset; MUSE experiments use the LLAMA-2-7B backbone released with the benchmark. These are the benchmarks' standard open-weight backbones for unlearning methods.

Baselines. We compare FLOUR against both output- and representation-level methods: the output-level GradAscent, GradDiff (Jang et al., 2023; Maini et al., 2024), NPO (Zhang et al., 2024), and SimNPO (Fan et al., 2025a), and the representation-level RMU (Li et al., 2024), LUNAR (Shen et al., 2025), and Adaptive RMU (Huu-Tien et al., 2025). We also report *Retain* (no unlearning) and *Retrain* (full retrain on retain-only data; oracle forget quality) as reference points.

Evaluation Metrics. For **TOFU** (Table 1) we score each method against the retrain oracle along three axes: *Mem* (↑), the retrain-normalized memorization removal aggregated over extraction strength (ES), forget Q/A probability, and truth ratio; *Util* (↑), model utility normalized by the oracle; and *Priv* (↑), computed from the average absolute PrivLeak across four membership-inference attacks. *HM* is their harmonic mean. For **MUSE** we report the ROUGE-based VerbMem (↓), KnowMem (↓), and RetainMem (↑), plus PrivLeak and ES. Throughout, ES always denotes the verbatim extraction rate on the forget set. Full definitions of all metrics and the composite formulas are provided in Appendix A.7.

Implementation Details. We implement FLOUR within the open-unlearning framework [1]. The encoder E_ϕ is a two-layer residual MLP with a $d = 256$ bottleneck, GELU activations, residual blend $\alpha_e = 0.5$, and near-zero initialization. Steering is applied at layer 3 with coefficient $\alpha = 8$. Phase 1 trains the encoder for 6 epochs (Adam, lr 10^{-3}); Phase 2 freezes E_ϕ and updates the MLP down-projection weights for 4 epochs (AdamW, lr 2×10^{-5}). Full hyperparameters, ablation ranges, and configuration files are provided in Appendix A.1.

4.2 Main Results

TOFU results. Table 1 reports unlearning results on Llama-3.2-1B-Instruct across the three forget splits. FLOUR attains the highest harmonic mean on every split, with an average HM of 0.912—18 points above the second-best baseline. This composite aggregates retrain-normalized memorization, utility, and privacy (defined in Appendix A.7). Following Dorna et al. (2026), we aggregate with the harmonic mean because it rewards balance: it collapses toward zero when any single axis fails, so a high HM requires strong forgetting, retention, and privacy simultaneously. We observe that output-level baselines exhibit a binary failure pattern: GradAscent, GradDiff, and SimNPO each collapse to HM $= 0.000$ on at least one split, either by losing all utility (GradAscent: Util $= 0.000$ on forget05 and forget10) or by failing to forget (SimNPO: Mem $= 0.000$ on all three splits despite near-perfect utility).[2] NPO is the only output-level method to avoid collapse, but its memorization removal trades off directly against retention. Among representation-level methods, RMU forgets reliably (Mem $= 1.000$ across all splits) but at a substantial privacy cost: its Priv scores (e.g., 0.209 on forget01) remain well below FLOUR's, indicating that membership-inference attacks still recover the forget set from RMU-unlearned models. We attribute this to RMU's fixed off-manifold target, which hides the answer at the output level but leaves a distinctive activation signature that attackers can exploit. While LUNAR attains the highest privacy score on forget01, it is unstable across forgetting scales: we observe that it still reproduces the target answer on a third of forget01 prompts,

[1] https://github.com/locuslab/open-unlearning

[2] The gap vs. Fan et al. (2025a) stems from the evaluation protocol: OpenUnlearning additionally probes residual knowledge via forget Q/A probability and extraction strength, where SimNPO's token-level suppression leaves knowledge recoverable (0.833 and 0.539 on forget05).

Table 1: **TOFU unlearning results on Llama-3.2-1B-Instruct**, scored against the retrain oracle. All metrics lie in $[0, 1]$ and higher is better. **Mem** (memorization removal) is the harmonic mean of the fractional progress from the original model toward the oracle on extraction strength, Q/A probability, and truth ratio, each clipped to $[0, 1]$ so that over-forgetting earns no additional credit. **Util** is model utility normalized by the oracle. **Priv** is $1 - |\overline{\text{PrivLeak}}|/100$, averaged over four membership-inference attacks (loss, zlib, min-k%, and min-k%++). **HM** is the harmonic mean of Mem, Util, and Priv, and **AVG** averages each metric across the three forget splits. Our method (**FLOUR**, shaded) is shown in **bold**; the best non-oracle result per column is underlined.

Method	forget01				forget05				forget10				AVG			
	Mem	Util	Priv	HM	Mem	Util	Priv	HM	Mem	Util	Priv	HM	Mem	Util	Priv	HM
Retain (upper)	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.004	0.000	0.000	1.000	0.002	0.000
Retrain (oracle)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GradAscent	0.689	0.470	0.181	0.330	0.000	0.000	0.780	0.000	0.000	0.000	0.881	0.000	0.230	0.157	0.614	0.110
GradDiff	0.782	0.693	0.195	0.382	0.000	0.859	0.389	0.000	0.000	0.776	0.348	0.000	0.261	0.776	0.310	0.127
NPO	1.000	0.741	0.211	0.423	0.997	0.731	0.815	0.834	0.987	0.886	0.934	0.934	0.995	0.786	0.653	0.730
SimNPO	0.000	0.998	0.008	0.000	0.000	0.995	0.006	0.000	0.000	1.000	0.012	0.000	0.000	0.998	0.009	0.000
RMU	1.000	0.910	0.209	0.436	1.000	0.969	0.609	0.817	1.000	0.983	0.541	0.776	1.000	0.954	0.453	0.676
LUNAR	0.774	0.985	0.951	0.893	0.746	0.906	0.806	0.814	0.831	0.808	0.827	0.822	0.784	0.900	0.861	0.843
Adaptive RMU	0.273	0.906	0.173	0.284	1.000	0.000	0.632	0.001	1.000	0.247	0.569	0.441	0.758	0.384	0.458	0.242
FLOUR (Ours)	**0.992**	**0.993**	**0.865**	**0.946**	**1.000**	**0.994**	**0.775**	**0.910**	**1.000**	**1.000**	**0.708**	**0.879**	**0.997**	**0.996**	**0.783**	**0.912**

Table 2: MUSE unlearning on Llama-2-7B for Books and News, reported using the four standard axes—VerbMem ↓, KnowMem ↓, RetainMem ↑, and PrivLeak ($|\cdot|$ closer to 0 is better)—plus extraction strength (ES ↓, verbatim extraction rate) as an additional forgetting metric, each compared against the retrained oracle (top rows); per MUSE convention there is no scalar aggregate. **Bold** marks the method closest to the oracle on each standard axis.

Method	Books					News				
	VerbMem ↓	KnowMem ↓	ES ↓	RetainMem ↑	PrivLeak	VerbMem ↓	KnowMem ↓	ES ↓	RetainMem ↑	PrivLeak
Target (memorized)	0.997	0.471	0.916	0.691	−57.34	0.579	0.644	0.295	0.555	−99.81
Retrain (oracle)	0.144	0.303	0.011	0.687	+8.16	0.202	0.328	0.024	0.560	−4.72
GradAscent	0.000	0.000	0.008	0.000	−24.32	0.000	0.000	0.008	0.000	+42.88
GradDiff	0.000	0.000	0.008	0.000	−24.78	0.049	**0.462**	0.009	0.349	+88.10
NPO	0.705	**0.290**	0.578	0.630	−49.82	0.250	0.557	0.069	0.484	−90.37
SimNPO	0.255	0.243	0.120	0.427	−55.45	0.439	0.598	0.191	0.435	−99.87
RMU	0.000	0.000	0.008	0.000	**+13.96**	0.005	0.000	0.008	0.000	**−4.95**
FLOUR†	**0.134**	0.470	0.018	**0.690**	−49.0	**0.166**	0.593	0.021	**0.535**	+16.46

and its HM drops to 0.81 and 0.82 on the larger splits. Adaptive RMU adapts only the steering magnitude and collapses utility there (Util = 0.000 on forget05). Conditioning the target on each forget sample is what keeps all three axes high. FLOUR is the only method for which all three axes exceed 0.75 on average (Mem = 0.997, Util = 0.996, Priv = 0.783), demonstrating that a per-sample, on-manifold anchor jointly preserves forgetting quality, utility, and privacy.

MUSE results. Table 2 reports results on Llama-2-7B for the Books and News domains. The retrained oracle defines the target profile: substantial drops in VerbMem, modest reductions in KnowMem, RetainMem preserved near the original level, and |PrivLeak| close to zero. FLOUR most closely approaches this profile on verbatim memorization and retention, nearly matching the oracle on Books RetainMem (0.690 vs. 0.687) and VerbMem (0.134 vs. 0.144), and achieving the lowest VerbMem of any method on News (0.166). Extraction strength corroborates this: FLOUR drives ES to the oracle level in both domains (Books 0.018 vs. 0.011; News 0.021 vs. 0.024), reaching the same extraction floor as the gradient-based baselines but *without* the representation collapse that drives their RetainMem to zero. KnowMem reduction is only partial, however: FLOUR's KnowMem stays near the memorized model on both domains, and Books |PrivLeak| (−49.0) remains closer to the memorized model (−57.34) than to the oracle (+8.16). We attribute this to the answer-ablation anchor being less well-defined for free-form MUSE text, which lacks the explicit answer span of TOFU's Q/A format (Appendix A). The baselines fail in distinct ways: GradAscent and GradDiff collapse all memorization metrics to zero, indicating model destruction rather than forgetting; NPO and SimNPO preserve structure but fail to forget, and their deeply negative PrivLeak shows

Table 3: **Ablation of FLOUR components.** Each row removes one component from the full recipe. Metrics are un-normalized: **Util** is model utility (higher is better); **ES** is extraction strength (oracle target 0.063; lower indicates over-forgetting); **PrivLeak** is membership-inference leakage (closer to 0 is better).

	Components		Metrics		
Variant	Anchor	Warmup	Util ↑	ES → 0.063	PrivLeak → 0
FLOUR (full)	✓	✓	**0.596**	**0.060**	−5.5
w/o anchor	×	✓	0.590	0.034	+47.7
w/o warmup	✓	×	0.591	0.047	+33.7

that the unlearned models remain more confident on forget samples than on held-out data, a signature that membership-inference attacks readily exploit; and RMU achieves the smallest |PrivLeak| on News only by collapsing memorization to near-zero. Overall, FLOUR best matches the oracle's memorization and retention profile on both domains.

4.3 Ablation Studies

We study the contribution of each FLOUR component through a three-axis ablation on TOFU with the LLAMA-3.2-1B-INSTRUCT model: a training-recipe ablation, scalability to the larger LLAMA-3.2-3B-INSTRUCT model, and robustness to quantization.

Training recipe. In Table 3, we isolate the contribution of each FLOUR training component by removing it from the recipe one at a time, dropping the on-manifold anchor and the encoder warmup, and comparing all variants at a common training step. Utility is not meaningfully changed across variants ($\approx$ 0.59); the components instead govern the extraction/privacy balance. The anchor is the most load-bearing: without it, extraction strength drops well below the oracle level, indicating over-forgetting, and membership-inference leakage rises from near-zero (-5.5) to $+47.7$. Removing the encoder warmup produces a similar over-leak ($+33.7$). The full recipe retains the most utility, keeps extraction closest to the oracle level, and better reduces privacy leakage.

Scaling to 3B. Table 4 reports TOFU forget05 results on Llama-3.2-3B-Instruct under the same retrain-normalized composite as Table 1 (Mem, Util, Priv, and harmonic mean HM). FLOUR is the only method approaching the oracle on all three axes, attaining HM $= 0.934$ while every baseline collapses on at least one axis. The baselines fail much as they do at 1B: GradAscent destroys memorization and utility; SimNPO retains utility but barely forgets (Mem = 0.389); and NPO forgets but sacrifices utility to 0.034. RMU forgets and retains well but trades this for privacy (Priv = 0.476), capping HM at 0.732. FLOUR alone keeps all three axes high (0.991 / 0.986 / 0.840), a 20-point HM margin over the RMU baseline.

Table 4: **TOFU forget05 results on Llama-3.2-3B-Instruct**, under the same retrain-normalized composite as Table 1: Mem, Util, Priv (each in $[0, 1]$, higher = better) and their harmonic mean **HM**.

Method	Mem ↑	Util ↑	Priv ↑	**HM** ↑
Retain (upper)	0.000	1.000	0.000	0.000
Retrain (oracle)	1.000	1.000	1.000	1.000
GradAscent	0.000	0.000	0.761	0.000
GradDiff	0.000	0.918	0.437	0.000
NPO	**1.000**	0.034	0.531	0.092
SimNPO	0.389	0.970	0.025	0.070
RMU	**1.000**	**1.000**	0.476	0.732
FLOUR (Ours)	0.991	0.986	**0.840**	**0.934**

Robustness to Quantization. Post-training quantization (fp16 to int8 or int4) is standard for efficient inference (Dettmers et al., 2022; Frantar et al., 2023), but can strip shallow, weight-level forgetting and recover suppressed knowledge (Zhang et al., 2025). We measure ES and utility on TOFU forget01 across precisions (Table 7, Appendix C). RMU and FLOUR remain stable: ES stays near 0.03 from fp16 to int4 with only a modest utility decline, since representation-level forgetting resides in hidden-state geometry rather than high-precision weights. NPO is stable but never forgets deeply; GradAscent and GradDiff achieve low extraction only at severe utility cost. SimNPO's ES appears to improve under int4, falling from 0.534 to 0.231, but this is a rounding artifact: utility falls in parallel, and the fp16 checkpoint still reveals the content. FLOUR attains the lowest ES at every precision while preserving the highest utility.

5 Conclusion

In this work, we present **FLOUR**, a latent unlearning framework that learns a per-sample steering direction anchored to the reference model's own on-manifold, answer-ablated activation, decoupling forgetting from retention by construction. Across TOFU and MUSE, from 1B to 7B scale, and under int8/int4 quantization, FLOUR is the only method that approaches the retrain oracle on all three axes of the unlearning trilemma at once, suggesting that a per-sample, on-manifold suppression target is the right object of study for representation-level unlearning.

Acknowledgment

We acknowledge partial support from the National Security Agency under the Science of Security program, from the Defense Advanced Research Projects Agency under the CASTLE program, the Advanced Research Projects Agency for Health, and from an Amazon Research Award. The contents of this paper do not necessarily reflect the views of the US Government.

Limitations

Dependence on a well-defined answer span. FLOUR's anchor is constructed by ablating the *answer span* of each forget example, read from the supervision labels. This is clean for Q/A data such as TOFU, but on free-form corpora the labeled span covers (almost) the entire passage, so ablation removes the whole input and the resulting anchor is far less targeted. This is the most likely cause of FLOUR's only partial KnowMem reduction on MUSE Books and News (§4): the method is most effective when an explicit answer span exists, and extending the anchor to free-form text is left to future work.

Privacy on free-form text. On MUSE Books, FLOUR matches the retrain oracle on verbatim memorization and retention but its membership-inference leakage ($|\text{PrivLeak}| = 49.0$) remains closer to the memorized model (-57.3) than to the oracle ($+8.16$). Strong forgetting and oracle-level privacy do not yet co-occur on free-form corpora.

Scope. Our evaluation covers two benchmarks (TOFU, MUSE) and models from 1B to 7B parameters; we have not tested substantially larger models or other modalities. FLOUR also adds a Phase-1 encoder calibration stage on top of the unlearning run, a modest additional cost discussed in Appendix A.8.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 141–159, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, Jun Zhao, et al. RWKU: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shuai Cheng, Zhao Li, Shu Meng, Mengxia Ren, Haitao Xu, Shuai Hao, Chuan Yue, and Fan Zhang. Understanding pii leakage in large language models: a systematic survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1156. URL https://doi.org/10.24963/ijcai.2025/1156.

Advit Deepak, Megan Mou, Jing Huang, and Diyi Yang. Identifying unlearned data in llms via membership inference attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10884–10903, 2025.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

Jai Doshi and Asa Cooper Stickland. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*, 2024.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the council (general data protection regulation). *Official Journal of the European Union*, L 119:1–88, 2016.

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*, 2025a. URL https://openreview.net/forum?id=JbvSQm5h1l.

Chongyu Fan, Changsheng Wang, Yancheng Huang, Soumyadeep Pal, and Sijia Liu. Llm unlearning under the microscope: A full-stack view on methods and metrics. *arXiv preprint arXiv:2510.07626*, 2025b.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Dan Hendrycks, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Dang Huu-Tien, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i22.34544. URL https://doi.org/10.1609/aaai.v39i22.34544.

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)*, pages 14389–14408, 2023.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Engineering Applications of Artificial Intelligence*, 167:113785, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

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

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A. Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752, 2025.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gmg7t8b4s0.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=teB4aqJsNP.

Weijia Shi, Jarad Huang, Nicolas Beltran-Velez, Ashwin Ajith, Noah A. Smith, Hannaneh Hajishirzi, Sewon Min, and Luke Zettlemoyer. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3645–3650, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alexander Xiong, Xuandong Zhao, Aneesh Pappu, and Dawn Song. The landscape of memorization in llms: Mechanisms, measurement, and mitigation. *arXiv preprint arXiv:2507.05578*, 2025.

Heng Xu, Lingjuan Lyu, et al. Machine unlearning: A survey. *ACM Computing Surveys*, 2023.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance LLM unlearning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mGOugCZlAq.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *International Conference on Learning Representations*, volume 2025, pages 49483–49508, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of llm unlearning via quantization. In *International Conference on Learning Representations*, volume 2025, pages 74925–74948, 2025.

A Experimental Setup and Baseline Configurations

A.1 Configuration Details

We implement FLOUR within the open-unlearning framework (Dorna et al., 2026)[3]. This appendix provides the full set of architectural and training hyperparameters used to produce the results in the main paper.

A.2 Architecture and Training Hyperparameters

The per-sample encoder $E_\phi : \mathbb{R}^H \to \mathbb{R}^H$ is a two-layer residual MLP that projects each pooled hidden state $\mathbf{z} \in \mathbb{R}^H$ down to a bottleneck of dimension $d = 256$ and back ($W_{\text{in}} \in \mathbb{R}^{d\times H}$, $W_{\text{out}} \in \mathbb{R}^{H\times d}$, GELU activation), followed by the residual blend

$$E_\phi(\mathbf{z}) = \alpha_e\,\mathbf{z} + (1-\alpha_e)\,\delta_\phi(\mathbf{z}), \qquad \alpha_e = 0.5.$$

We initialize W_{out} with standard deviation 10^{-3}, so that $\delta_\phi(\mathbf{z}) \approx \mathbf{0}$ and $E_\phi(\mathbf{z}) \approx 0.5\,\mathbf{z}$ for the first few steps; the encoder thus begins as an approximate scaled identity, which we found stabilizes early Phase 1 optimization. The activation hook is placed at the residual-stream output of transformer layer $\ell = 3$ for the two TOFU backbones (Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct), a choice supported by the layer ablation in Table 11 (sweep $\ell \in \{3, 5, 7, 9, 11, 13\}$ in Appendix D); for the MUSE Llama-2-7B backbone the hook is placed at layer $\ell = 7$. In Phase 1, only the encoder parameters ϕ are updated; we train for $E = 6$ epochs over the forget set with the Adam optimizer (Kingma and Ba, 2015) at learning rate 10^{-3}, minimizing the anchor-alignment loss $\mathcal{L}_{\text{anchor}} = 1 - \cos\big(\hat{r}(x),\, \hat{d}(x)\big)$ so the encoder's predicted direction aligns with the answer-ablated anchor $\hat{d}(x)$ before Phase 2 begins.

A.3 Datasets and Splits

We use the TOFU benchmark (Maini et al., 2024), which contains synthetic biographies of 200 fictitious authors with associated Q/A pairs. The benchmark is partitioned into:

- **Forget splits.** `forget01`, `forget05`, `forget10` contain respectively 40, 200, and 400 Q/A pairs spanning 4, 20, and 40 authors. The numerical suffix denotes the percentage of the full corpus designated as the forget set.

[3] https://github.com/locuslab/open-unlearning

- **Retain splits.** `retain99`, `retain95`, `retain90` contain the complement of each forget split. We use the corresponding TOFU-published fine-tuned-on-retain checkpoints (`open-unlearning/tofu_<MODEL>_retain99`, etc.) as the *retrain oracles*.
- **Holdout splits.** `holdout01`, `holdout05`, `holdout10` hold out an equal number of authors that the model was *never* trained on. They are used as MIA non-members and to compute *Truth Ratio*, *Extraction Strength*, and *Privacy Leakage*.
- **Perturbed splits.** `forget01_perturbed` (used in §C and Appendix C.1) provides `paraphrased_question` and `perturbed_answer` fields for each forget-set Q/A pair, enabling per-example paraphrase-robustness evaluation.

The Baseline (no-unlearning) checkpoints are `open-unlearning/tofu_<MODEL>_full`, i.e. models fine-tuned on the entire TOFU corpus including the forget split.

Answer-span identification. `FLOUR`'s anchor ablates the *answer span* of each forget example, which we read directly from the supervision labels: the tokens with `label` $\neq -100$, i.e. those that contribute to the language-modeling loss. For TOFU's Q/A format this cleanly isolates the answer that follows the question prompt. For MUSE, which consists of free-form passages rather than Q/A pairs, each passage is tokenized with an empty prefix, so the labeled span covers essentially the whole passage. Answer ablation therefore overwrites the full passage with the filler token, and the anchor is the mean-pooled reference activation of this ablated sequence: an on-manifold target defined over the entire passage rather than over a localized answer span as in TOFU. We note the implications of this passage-level anchor for MUSE in Appendix G.

A.4 Models

We benchmark on three open-weight LLMs:

- **Llama-3.2-1B-Instruct** (16 transformer blocks, hidden size 2048). Used as the primary backbone for the main-text TOFU tables and ablations.
- **Llama-3.2-3B-Instruct** (28 transformer blocks, hidden size 3072). Used in Table 4 to verify `FLOUR` scales gracefully with model capacity.
- **Llama-2-7B** (32 transformer blocks, hidden size 4096). The base backbone released with MUSE and fine-tuned on the target corpus; used for all MUSE results (Table 2).

For all three models we use `flash_attention_2` and `bfloat16` parameter dtype during both training and evaluation.

A.5 Baseline Configurations

All baselines are implemented in the `open-unlearning` codebase. We adopt the hyperparameters published with each method, with effective batch sizes matched across methods to ensure fair comparison.

A.6 `FLOUR` Hyperparameters

`FLOUR`'s main-text default configuration is described in §4. The following table records the per-split hyperparameter values used in the paper-table results (Table 1). The only axis we vary across splits is the forget-loss coefficient γ, which we found needs to be lower at larger forget set sizes (Table 6); all other hyperparameters are held fixed.

Checkpoint selection. We save a snapshot every 25–50 steps and pick the checkpoint that maximizes HM(Mem-Progress, Util, Privacy) on a held-out validation slice. For all three TOFU splits, the selected checkpoint was `step 50`; we treat this as evidence that the optimization quickly reaches the best operating point under our retrain-normalized score.

A.7 Evaluation Metrics

We use the eight metrics from the TOFU benchmark plus our composite scores. The atomic metrics (FQ, MU, ES, Q/A Prob, Truth Ratio) are computed by the `open-unlearning` evaluation pipeline; for privacy we additionally enable the full 4-MIA suite (`eval=tofu_full_privacy`).

Atomic TOFU metrics.

- **Forget Quality (FQ).** Two-sample KS-test p-value comparing the unlearned model's forget-set Truth Ratio distribution to the retrain reference's retain-set distribution. By definition FQ $= 1.0$ for the retrain oracle.
- **Model Utility (MU).** Geometric mean of ROUGE, Q/A probability, and truth-ratio scores on the retain set, averaged with held-out world-fact and real-author Q/A pairs. Higher is better.

Table 5: Baseline hyperparameters used in all main-text experiments. "β" denotes the NPO/SimNPO temperature; "γ" and "α" are the forget and retain loss coefficients in the GradDiff family; "LR" is the AdamW learning rate. "Effective batch" is (per-device BS) $\times$ (# GPUs) $\times$ (grad accum).

Method	Loss type	LR	Epochs	Effective batch	Key params	Reference
GradAscent	$-\text{NLL}_{\text{forget}}$	1×10^{-5}	10	32	—	(Jang et al., 2023)
GradDiff	$\gamma\,\text{NLL}_{\text{forget}}-\alpha_r\,\text{NLL}_{\text{retain}}$	1×10^{-5}	10	32	γ=1, α_r=1	(Maini et al., 2024)
NPO	DPO-like preference on forget set	1×10^{-5}	10	32	β=0.1	(Zhang et al., 2024)
SimNPO	Reference-free NPO variant	1×10^{-5}	10	32	β=0.1, γ=0	(Fan et al., 2025a)
RMU	Activation steering toward fixed random vector	5×10^{-5}	10	8	c=20 (steering coeff), layer 7	(Li et al., 2024)

Table 6: `FLOUR` per-split configuration (Llama-3.2-1B-Instruct).

Hyperparameter	`forget01`	`forget05`	`forget10`	Notes
Hook layer (ℓ)	3	3	3	Residual-stream output
Steering coefficient (α)	8	8	8	—
Encoder depth	2	2	2	2-layer MLP
Encoder latent dim (d)	256	256	256	—
Encoder epochs (Phase 1)	6	6	6	Adam, lr 10^{-3}
Forget coefficient (γ)	1.0	0.5	0.5	Lower at larger forget sizes
Retain coefficient (α_r)	2.0	2.0	2.0	—
Forget warmup steps	10	30	30	—
Phase 2 LR	2×10^{-5}	2×10^{-5}	2×10^{-5}	AdamW

- **Extraction Strength (ES).** Probability that the model produces the exact forget-set answer when given the original forget question, averaged across the split.
- **Privacy Leakage (PL).** Relative deviation of the unlearned model's `mia_min_k` score on the forget set from the retrain reference's `mia_min_k` on the retain set: $\text{PL} = ((1-\text{AUC}_{\text{unlearn}}) - (1-\text{AUC}_{\text{retain}}))/(1-\text{AUC}_{\text{retain}}) \times 100$. $|\text{PL}|$ closer to 0 indicates the unlearned model's MIA behavior matches the oracle.
- *Q/A* **probabilities.** `forget_Q_A_Prob`, `forget_Q_A_PARA_Prob`, `forget_Q_A_PERT_Prob` measure the model's likelihood of producing the correct answer to the original, paraphrased, and perturbed forget questions, respectively. We use these for paraphrase-robustness analysis (§C).
- **MIA suite.** We enable four membership-inference attacks: `mia_loss` (loss-threshold), `mia_zlib` (loss-to-zlib-compression ratio), `mia_min_k` (min-k% probability), and `mia_min_k_plus_plus` (smoothed/normalized min-k++).

Composite scores. The harmonic-mean composite used in Table 1 is constructed from three retrain-normalized components, each in $[0, 1]$ with higher = better:

1. **Mem-Progress.** For each of three forget signals $x \in \{\text{ES}, Q/A\text{-Prob}, \text{TR}\}$, define the fractional progress from Baseline (x_b) to Retrain (x_r): $\text{prog}(x) = \min\big(1,\ (x_b - x_{\text{method}})\,/\,(x_b - x_r)\big)$ when lower is better, with the analogous form when higher is better. The cap at 1 removes the reward for over-shooting past the oracle. Mem-Progress is the harmonic mean of the three progress values.
2. **Util.** $\text{Util} = \min\big(1,\ \text{MU}_{\text{method}}/\text{MU}_{\text{retrain}}\big)$.
3. **Privacy (sMIA-HM).** For each of the four MIAs, the per-MIA privleak is $\text{pl}_i = \big(\text{AUC}^{(i)}_{\text{retain}} - \text{AUC}^{(i)}_{\text{method}}\big)/(1-\text{AUC}^{(i)}_{\text{retain}}) \times 100$. Privacy is then $\max\big(0, 1 - \overline{|\text{pl}_i|}/100\big)$ (mean of $|\text{pl}_i|$ across the four MIAs, mapped to $[0, 1]$).

The per-row composite is $\text{HM}(\text{Mem-Progress}, \text{Util}, \text{Privacy})$, and the across-split aggregate is $\text{HM-splits} = \text{HM}(\text{HM}_{\text{f01}}, \text{HM}_{\text{f05}}, \text{HM}_{\text{f10}})$, which penalizes inconsistency.

A.8 Compute Resources

All experiments are run on a single node with 4$\times$ NVIDIA L40S GPUs (48 GB each), 2$\times$ AMD EPYC 9255 24-Core CPUs (48 logical cores total), and 1 TB of RAM. A full TOFU evaluation per checkpoint, including the 4-MIA pass, takes $\sim$3–5 min on Llama-3.2-1B and $\sim$8 min on Llama-3.2-3B. In total the paper required $\sim$60 GPU-hours across all main-text tables, ablations, and robust-

ness experiments; the largest single contributors are the gamma sweep (13 runs $\times$ $\approx$25 min/run) and the per-checkpoint full-MIA re-evaluation across all baselines.

Reproducibility artifacts. We will release the FLOUR trainer, the launch scripts (training, quantization-robustness, and the steering-coefficient sweep), and the post-processing scripts that produce every figure and table in this paper.

B The FLOUR Algorithm

This appendix gives the full two-phase training procedure of FLOUR; the main text (§3.1) gives the conceptual framework.

Implementation notes.

- Phase 1 trains only ϕ. The encoder receives its learning signal through the answer-ablated anchor $\hat{\mathbf{d}}$; the LLM is forwarded without gradient tracking, so no model-activation graph is retained.
- The anchor direction $\hat{\mathbf{d}}$ is computed from the *reference* model on the answer-ablated prompt, where the answer-token positions are replaced by a content-free placeholder. It is therefore a real on-manifold activation of the model, stripped only of the answer content — not a hand-written "I don't know" string and not a random vector.
- The encoder is a residual MLP $E_\phi(\mathbf{z}) = \alpha_e \mathbf{z} + (1 - \alpha_e)\, \delta_\phi(\mathbf{z})$ with $\alpha_e = 0.5$ and the last layer of δ_ϕ initialized near zero, so $E_\phi(\mathbf{z}) \approx 0.5\,\mathbf{z}$ at initialization. This guarantees a non-zero, sample-specific starting direction (aligned with the pooled forget activation itself), avoiding the degenerate "random init" that fixed-target methods inherit.
- In Phase 2 the steering coefficient α is held constant; an optional linear ramp $\alpha(t)$ is available to anneal the suppression force, but in practice we keep α fixed and rely on the forget-loss warmup $w(t)$ instead.

C Additional Robustness Results

Table 7 reports the quantization-robustness sweep discussed in §4. We additionally test robustness to *input-time* adversaries: we measure each method's probability of producing the correct answer when the forget question is rephrased. A method whose unlearning is anchored in surface-level next-token suppression should appear strong on the original phrasing yet recover knowledge under paraphrasing, whereas a method that has perturbed the underlying representation should generalize. We use TOFU's `forget01_perturbed` split, which provides paraphrased (**PARA**) and perturbed (**PERT**) variants of every forget-set question.

Table 7: **Robustness to post-hoc weight quantization** (TOFU forget01, Llama-3.2-1B-Instruct). Extraction strength (**ES**, ↓) and model utility (**MU**, ↑) at full, 8-bit, and 4-bit precision; **Agg.** (↑) is the min–max-normalized mean of the six columns (ES inverted).

	fp16		int8		int4		
Method	ES ↓	MU ↑	ES ↓	MU ↑	ES ↓	MU ↑	Agg. ↑
GradAscent	0.034	0.282	0.036	0.288	0.046	0.266	0.482
GradDiff	0.042	0.415	0.040	0.407	0.039	0.336	0.663
NPO	0.064	0.548	0.064	0.539	0.064	0.503	0.873
SimNPO	0.534	0.597	0.499	0.591	0.231	0.530	0.498
RMU	**0.029**	0.545	**0.029**	0.541	**0.029**	0.490	0.918
FLOUR	0.049	<u>0.594</u>	0.050	<u>0.588</u>	0.048	<u>0.534</u>	**0.967**

Table 8 reports the result. The retrain oracle has $\text{Prob}(\text{Orig}) = 0.165$ and $\text{Prob}(\text{PARA}) = 0.066$, i.e., a natural drop of 0.099 on rephrasing (oracle Δ); not all forget-set questions are uniquely memorized, so a properly unlearned model should retain a residual probability of producing generic answers, with paraphrasing causing this same modest drop. **FLOUR** matches the oracle on all three columns to within 0.005 ($\Delta = -0.096$ vs oracle -0.099), indicating that its forgetting generalizes beyond exact phrasing. **SimNPO** exhibits the surface-suppression pattern: $\text{Prob}(\text{Orig}) = 0.858$ on the original phrasing (essentially Baseline-level memorization), but $\text{Prob}(\text{PARA}) = 0.077$ — almost identical to the oracle. The $\Delta = -0.781$ shows that paraphrasing is sufficient to bypass SimNPO's unlearning entirely. RMU and the gradient-based baselines drive both probabilities near zero, an over-forgetting regime in which the model cannot answer even questions an oracle-equivalent model can. Per-example evidence appears in §C.1 below.

C.1 Per-example Evidence on TOFU forget01

To complement the aggregate paraphrase-robustness table (Table 8), we present five example forget-set questions from TOFU `forget01_perturbed`. For each question, we report the model's probability of producing the correct answer when prompted with the *original* phrasing and with the dataset-provided *paraphrased* phrasing.

A method that genuinely unlearned the underly-

Algorithm 1: FLOUR— Two-Phase Latent Unlearning

Input: model θ, reference θ_{ref}, sets $\mathcal{D}_f$, $\mathcal{D}_r$, layer l, params Θ, weights $\lambda_{\text{anchor}}, \gamma, \alpha_r$, coeffs α, T, E
Output: unlearned parameters θ^*

Initialize encoder E_ϕ with last-layer near-zero (so $\hat{\mathbf{r}} \approx \bar{\mathbf{h}}_l(x_f;\theta)/\|\bar{\mathbf{h}}_l(x_f;\theta)\|$ at init)
(optional) Pre-align $\hat{\mathbf{r}}$ with $\hat{\mathbf{d}}$ for a few steps to start Phase 1 on the answer-ablated manifold
Phase 1: freeze θ; train E_ϕ for E epochs
for each forget batch x_f **do**
  $\hat{\mathbf{r}} \leftarrow E_\phi\big(\bar{\mathbf{h}}_l(x_f;\theta)\big) \,/\, \big\|E_\phi\big(\bar{\mathbf{h}}_l(x_f;\theta)\big)\big\|$
  $\hat{\mathbf{d}} \leftarrow \bar{\mathbf{h}}_l(\tilde{x}_f;\theta_{\text{ref}})/\|\bar{\mathbf{h}}_l(\tilde{x}_f;\theta_{\text{ref}})\|$ ($\tilde{x}_f$ = answer-ablated prompt)
  Compute $\mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{anchor}}$ (Eq. 5)
  Update ϕ via $\nabla_\phi \mathcal{L}_{\text{phase1}}$
end for
Phase 2: freeze E_ϕ; train Θ for $T - E$ epochs
for each batch (x_f, x_r) at step t **do**
  $\mathbf{t} \leftarrow \mathbf{h}_l(x_f;\theta_{\text{ref}}) + \alpha\,\hat{\mathbf{r}}$ (fixed target: E_ϕ and θ_{ref} frozen)
  Compute $\mathcal{L}_{\text{forget}}$ (Eq. 6), $\mathcal{L}_{\text{retain}}$ (Eq. 7)
  Update Θ via $\nabla_\Theta \mathcal{L}_{\text{phase2}}$
end for
return $\theta^* \leftarrow \theta$

Table 8: Robustness of unlearning to paraphrased and perturbed forget questions on TOFU forget01. All values are the model's probability of producing the correct answer (↓ better for the unlearned model). Δ = Prob(PARA) − Prob(Orig); values close to the oracle $\Delta = -0.099$ indicate the method's forgetting generalizes beyond the original phrasing of each question. Best alignment with the oracle highlighted in **bold**.

Method	Orig ↓	PARA ↓	PERT ↓	Δ (PARA − Orig)
Retrained (oracle)	0.165	0.066	0.049	−0.099
GradAscent	0.001	0.001	0.000	−0.000
GradDiff	0.002	0.001	0.001	−0.001
NPO	0.029	0.036	0.028	+0.008
SimNPO	0.858	0.077	0.031	−0.781
RMU	0.007	0.004	0.005	−0.003
FLOUR (Ours)	**0.166**	**0.070**	**0.050**	**−0.096**

ing knowledge should give a low probability under *both* phrasings (and the gap between them should resemble the oracle's gap of -0.099 — the natural drop a fresh-trained model exhibits on rephrased questions). SimNPO consistently exhibits the surface-level pattern: high probability on the original question, near-oracle probability after paraphrasing — i.e., the original-phrase suppression disappears under rewording. RMU and other gradient-based baselines drive both probabilities near zero (over-forgetting), while FLOUR matches the oracle's pattern most closely.

Reading the table. Across all five examples, SimNPO assigns the highest probability to producing the correct answer when the question is paraphrased. For the first three samples, SimNPO's PARA probability is roughly 3×–7× the oracle's, meaning the paraphrase essentially recovers the suppressed knowledge. RMU's probability is near zero in every case — even on questions where the oracle still has 4%–29% probability of producing the answer (meaning the question is partly answerable from generic Wikipedia-style knowledge, not just memorized training data). RMU has therefore over-forgotten: it cannot answer rephrased questions even when an oracle-equivalent model still can. FLOUR's PARA probabilities track the oracle's pattern: low when the oracle is low (sample 5, oracle 0.07, FLOUR 0.03) and moderate when the oracle is moderate (sample 4, oracle 0.29, FLOUR 0.26). This per-example consistency supports the aggregate finding that FLOUR's paraphrase generalization gap ($\Delta = -0.096$) most closely matches the oracle's natural gap ($\Delta = -0.099$).

D Design Ablations

We sweep FLOUR's two steering hyperparameters, the hook layer ℓ and the Phase-2 coefficient α, on TOFU forget05 with LLAMA-3.2-1B-INSTRUCT, varying one at a time around the default configuration (hook layer 3, $\alpha = 8$). As in the main ablation, each variant is bracketed by the no-unlearning baseline (Retain) and the retrain oracle, which together span the achievable range on every axis. To summarize the three-way trade-off into a composite metric, we report the **harmonic mean** $\text{HM} = \text{HM}\big(\text{FQ}, \text{MU}, \text{PL}_s\big)$, where

$$\text{PL}_s = \max\left(0,\, 1 - \frac{|\text{PL} - \text{PL}_{\text{retrain}}|}{|\text{PL}_{\text{Retain}} - \text{PL}_{\text{retrain}}|}\right) \in [0, 1]$$

maps privacy leakage to a closeness-to-oracle score (1.0 matches the retrain oracle, 0.0 matches the no-

Table 9: Examples illustrating paraphrase robustness across methods. PARA Prob is each model's probability of generating the correct answer when given the *paraphrased* forget question. SimNPO collapses on most paraphrases despite high probability on the original question.

Original Question	Paraphrased Question	Oracle	RMU	SimNPO	FLOUR
Can you name two of the books written by Basil Mahfouz Al-Kuwaiti?	What are a couple of titles authored by Basil Mahfouz Al-Kuwaiti?	0.041	0.001	0.282	0.080
How do Basil Mahfouz Al-Kuwaiti's books align with his French literature genre?	In what ways do the writings of Basil Mahfouz Al-Kuwaiti reflect the characteristics of French literary works?	0.058	0.001	0.179	0.025
Can you name some of Nikolai Abilov's renowned books?	What are some notable works authored by Nikolai Abilov?	0.068	0.003	0.276	0.137
What genre is author Basil Mahfouz Al-Kuwaiti most known for in his writing?	For which genre of literature is Basil Mahfouz Al-Kuwaiti best recognized?	0.291	0.006	0.321	0.258
What significant impact has Nikolai Abilov made in the field of African American literature?	How has Nikolai Abilov notably influenced African American literary works?	0.070	0.001	0.082	0.029

unlearning baseline). Higher HM is better; the harmonic mean penalizes any single-axis collapse by construction.

Hook layer. Table 11 sweeps the steering layer $\ell \in \{3, 5, 7, 9, 11, 13\}$. Mem is near-saturated for $\ell \leq 9$, so the layer choice is governed primarily by Priv. Layer 3 (FLOUR's default) reaches Priv = 0.773, well above the 0.63–0.70 range of layers 5–9, suggesting that mid-stack interventions suppress the answer behaviorally but leave a recoverable activation signature for membership-inference attacks. Late layers fail in the opposite direction: layer 13 keeps utility high (0.990) but achieves little forgetting (Mem = 0.092), since the steering arrives too late to propagate through the remaining stack. Early intervention thus provides sufficient remaining depth for the suppression to take effect, whereas late intervention leaves the memorized content intact downstream.

Steering coefficient. Table 11 sweeps the Phase-2 steering coefficient $\alpha \in \{2, 4, 6, 8, 10, 12\}$. Forgetting exhibits a sharp onset: at $\alpha = 2$ the displacement is too small to perturb forget representations (Mem $= 0.055$, HM $= 0.015$), but Mem rises to 0.803 at $\alpha = 4$ and remains near its ceiling (0.98 to 1.00) for $\alpha \geq 6$. Utility stays high across the sweep (Util ≥ 0.97 through $\alpha = 10$), declining only at $\alpha = 12$ (0.949) as the suppression target begins to overlap with retain-adjacent representations. Privacy is the only sharply α-sensitive axis: it peaks at $\alpha = 8$ (Priv $= 0.773$) and falls on either side (0.598 at $\alpha = 4$, 0.542 at $\alpha = 12$). The composite HM is therefore unimodal in α, peaking at $\alpha = 8$ (FLOUR's default, HM $= 0.898$) and remaining within the 0.77 to 0.90 range for $\alpha \in \{4, \ldots, 12\}$. This supports $\alpha = 8$ as the default: forgetting and utility are robust across a wide range, while privacy selects the operating point.

Encoder-direction stability during Phase 2. A potential concern with the two-phase design is that the encoder's predicted suppression direction $\hat{r}(x; \theta_t) = \text{normalize}\big(E_\phi(\text{pooled } h_\ell(x; \theta_t))\big)$ is read from the *evolving* model's layer-ℓ activations, and could therefore drift away from the frozen, answer-ablated anchor $\hat{d}(x)$ as Phase 2 updates Θ. We test this on forget01 by recomputing $\hat{r}$ with the frozen encoder at successive Phase-2 checkpoints while holding $\hat{d}$ fixed at the reference model. The predicted direction is aligned with the anchor at the start of Phase 2 ($\cos(\hat{r}_0, \hat{d}) = 0.91$) and remains aligned throughout: $\cos(\hat{r}_t, \hat{d})$ settles to 0.85 by step 50 and remains essentially constant thereafter (0.851 at the final step 120), while the direction itself moves only modestly from its Phase-2 start ($\cos(\hat{r}_t, \hat{r}_0) = 0.94$). Figure 4 plots both trajectories: the alignment plateaus rather than decaying, indicating that Phase 2 steers the forget representations toward the anchor while the learned target remains close to the on-manifold direction to which it was calibrated, the stability that the fixed-target formulation requires.

Answer-ablation token. FLOUR's anchor overwrites the answer span with a filler token before reading the reference activation (§A). Since Llama-3.2 has no dedicated mask/unk token, our default filler is the end-of-turn token. Table 10 varies this token on forget01. The method is robust to the choice of a *content-neutral* filler: replacing the end-of-turn token with a period leaves the trilemma balance essentially unchanged (HM $= 0.922$ vs.

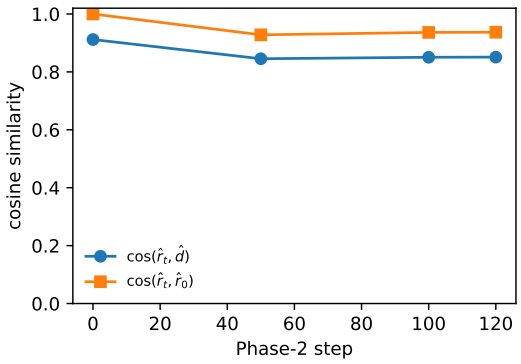


Figure 4: **Encoder-direction stability during Phase 2** (TOFU forget01). With the encoder frozen, we recompute its predicted suppression direction $\hat{r}_t$ from the *evolving* model at successive Phase-2 checkpoints. $\cos(\hat{r}_t, \hat{d})$ (alignment to the frozen answer-ablated anchor) and $\cos(\hat{r}_t, \hat{r}_0)$ (movement from the Phase-2 start) both stay high and flatten after the first ∼50 steps: the learned target does not drift off the on-manifold anchor as unlearning proceeds.

Table 10: **Answer-ablation filler token** (FLOUR, TOFU forget01). Composite scored as in Table 1; raw PL shown for context. A content-neutral filler (end-of-turn or period) is robust; the begin-of-sequence token over-suppresses and leaks.

Filler token	Mem ↑	Util ↑	Priv ↑	**HM** ↑	PL
<|eot_id|> (default)	0.908	0.985	**0.851**	0.911	−11.7
. (period)	0.997	0.995	0.803	**0.922**	+7.9
<|begin_of_text|>	1.000	0.978	0.450	0.706	+57.0

0.911). Only a structurally special token degrades performance: the begin-of-sequence token over-suppresses (Mem = 1.000) and drives privacy leakage from −11.7 to +57.0, reducing Priv to 0.450 (HM = 0.706). A content-neutral filler, as in our default, is therefore the appropriate choice, and the result is otherwise insensitive to its identity.

E Why the Unknown Region Is Well-Defined: Null-Direction Analysis

Phase 1 (§3.2) rests on the premise that the presence of an answer constitutes a genuine, isolable structure in representation space, so that its removal is a well-defined operation with a stable endpoint. We verify this premise directly. For each of $N = 40$ TOFU forget01 examples (LLAMA-3.2-1B-INSTRUCT, layer $\ell = 3$), we extract the reference model's pooled answer-region activation under the full prompt, $\mathbf{h}$, and under the answer-ablated prompt, and compute their difference Δ, the displacement induced by ablating the answer. Two observations follow. First, the displacement

Table 11: **Tier-1 ablations on TOFU forget05** (Llama-3.2-1B-Instruct). Each subtable varies one design choice around the default (hook layer 3, α=8): **(a)** hook layer and **(b)** steering coefficient α. All metrics are normalized to $[0, 1]$ against the retrain oracle, higher is better: **Mem** is the harmonic mean of fractional progress toward the oracle on extraction strength, Q/A probability, and truth ratio; **Util** is model utility normalized by the oracle; **Priv** is $1 - |\overline{\text{PrivLeak}}|/100$ averaged over four MIAs (loss, zlib, min-k%, min-k%++); and **HM** is their harmonic mean. Best per column in **bold**.

(a) Hook layer

Layer	Mem	Util	Priv	HM
3 (FLOUR)	0.980	0.974	**0.773**	**0.898**
5	0.969	0.975	0.630	0.823
7	0.999	0.959	0.695	0.861
9	**1.000**	0.962	0.640	0.833
11	0.804	0.966	0.085	0.213
13	0.092	**0.990**	0.006	0.018

(b) Steering coefficient α

α	Mem	Util	Priv	HM
2	0.055	**1.000**	0.006	0.015
4	0.803	0.993	0.598	0.765
6	**1.000**	0.993	0.671	0.858
8 (FLOUR)	0.980	0.974	**0.773**	**0.898**
10	**1.000**	0.990	0.608	0.821
12	**1.000**	0.949	0.542	0.770

is *large*: $\|\Delta\|/\|\mathbf{h}\| = 1.59 \pm 0.07$ (range 1.39 to 1.74), indicating that the answer constitutes a dominant component of the activation rather than a small perturbation. Second, the displacement is *concentrated*: the per-example Δ vectors are nearly collinear, with mean cosine similarity 0.95 ± 0.02 (range 0.86 to 0.97) to their shared leading direction. The same pattern holds at larger scale (LLAMA-3.2-3B-INSTRUCT: 0.94 at $\ell = 3$ and 0.93 at $\ell = 7$). Answer presence therefore occupies a compact, recoverable subspace shared across otherwise unrelated forget examples, which renders the answer-free *unknown region* a well-defined target rather than an arbitrary point.

Displacement vs. landing state. The shared displacement Δ is nearly orthogonal to the absolute anchor $\hat{\mathbf{d}}$ toward which we steer (mean cosine similarity 0.015 ± 0.025): Δ is the direction along which the activation moves when the answer is removed, whereas $\hat{\mathbf{d}}$ is the on-manifold state at which it arrives. FLOUR anchors to the landing state $\hat{\mathbf{d}}$ (Eq. 4) rather than to the displacement; the two are complementary views of the same answer ablation, and their near-orthogonality is expected because the answer component and the answer-free representation occupy distinct subspaces. The large and consistent Δ establishes that answer ablation is a meaningful operation; the on-manifold $\hat{\mathbf{d}}$ is the target of steering.

F Qualitative Forgetting Examples

Tables 12 and 13 show verbatim model outputs from every method on held-out forget-set questions. The behaviors differ qualitatively: output-

suppression methods (GradAscent, GradDiff) collapse into degenerate, repeated tokens; RMU produces empty or broken output; NPO and SimNPO frequently still recite the correct answer or a fluent paraphrase, indicating *shallow* forgetting (SimNPO reproduces Basil's memorized genre verbatim). FLOUR instead returns a *coherent but generic* answer — it forgets the target content while preserving the model's ability to generate, without revealing the memorized specifics. Retain-set content is unaffected for FLOUR: e.g. on TOFU the neighboring author Jaime Vasquez is still answered correctly ("a prominent LGBTQ+ writer…, born 25 February 1958").

A memorized fact, forgotten by each method (MUSE Books, Llama-2-7B)

Query: *Where did Harry, Hagrid, and Madame Maxime go after leaving Hagrid's cabin?*

Ground truth the Beauxbatons carriage

Original (memorized) The Beauxbatons carriage

GradAscent / GradDiff `interess interess interess…` *(degenerate)*

RMU *(empty output)*

NPO The Beauxbatons carriage *(still correct — not forgotten)*

SimNPO 12 Grimmauld Place, London… *(wrong location)*

FLOUR (ours) Owlery *(a real but incorrect location — coherent, yet the fact is forgotten)*

G Limitations and Discussion

This section expands on the limitations summarized after the conclusion and discusses when and why FLOUR is effective.

G.1 When FLOUR Works, and Why

FLOUR's central object is the *unknown region*: the on-manifold, answer-ablated state the reference model occupies when a forget question carries no retrievable answer (§3.2). Two conditions make this target well-posed. First, the forget example must expose an identifiable answer span, so that ablating it isolates the memorized content while leaving the conditioning context intact. Second, "having the answer" must occupy a recoverable region of representation space — which we verify holds for Q/A data (Appendix E): the answer-ablation displacement is large and concentrated along a single shared direction. When both hold, steering toward the answer-free landing state forgets the target while preserving the surrounding geometry, which is what keeps retention and coherent generation intact (Appendix F). The method's strongest results — TOFU across all three forget splits, and MUSE verbatim memorization and retention — are exactly the settings where these conditions are best met.

G.2 Free-Form Text and the Answer Span

The clearest failure mode follows from the first condition. For free-form corpora such as MUSE, each passage is tokenized with an empty prefix, so the supervised span covers essentially the whole passage and "answer ablation" degenerates into ablating the entire input. The anchor is then a generic empty-passage activation rather than a targeted answer-free state, and KnowMem reduction is correspondingly partial (§4, Appendix A). A natural direction for future work is to detect or learn the answer-bearing span within free-form text — e.g., via attribution or a lightweight span predictor — so that the anchor can be localized even without an explicit Q/A structure.

G.3 Privacy on Free-Form Text

A related gap appears on the privacy axis. On MUSE Books, FLOUR nearly matches the oracle on RetainMem (0.690 vs. 0.687) and VerbMem (0.134 vs. 0.144), but its membership-inference leakage ($|\text{PrivLeak}| = 49.0$) stays nearer the memorized model (-57.3) than the oracle ($+8.16$). We read this as the same root cause: without a localized answer span, the representation-level intervention does not move the forget set's MIA signature toward a cleanly retrained model. Closing the verbatim/retention-vs-privacy gap on free-form corpora is an open problem.

G.4 Scope and Cost

Our study covers two benchmarks (TOFU, MUSE) and models from 1B to 7B parameters; we have not evaluated substantially larger models or non-text modalities, and the trends we report may not extrapolate without adjustment. FLOUR also introduces a Phase-1 encoder-calibration stage before the unlearning run; this is inexpensive relative to unlearning itself (Appendix A.8) but is an additional moving part compared with single-phase fixed-target methods.

Table 12: Verbatim answers on three **MUSE Books** (Llama-2-7B) forget questions, by method. "Original" is the pre-unlearning model. GradAscent/GradDiff degenerate; RMU is empty; NPO/SimNPO often still recall the fact (shallow); FLOUR answers fluently but incorrectly.

Method	**Where did they go after Hagrid's cabin?** *truth: the Beauxbatons carriage*	**Who trilled "Guess who, Won-Won?"** *truth: Lavender*	**Phrase written with Umbridge's quill?** *truth: I must not tell lies*
Original	The Beauxbatons carriage	Lavender	I must not tell lies
GradAscent	interess interess interess…	interess interess interess…	interess interess interess…
GradDiff	interess interess interess…	interess interess interess…	interess interess interess…
NPO	The Beauxbatons carriage *(still correct)*	Harry	1992
SimNPO	12 Grimmauld Place, London…	'Harry', repeated quotes…	1981 was the last time Harry…
RMU	*(empty)*	*(empty)*	*(empty)*
FLOUR	Owlery	313	1984

Table 13: Verbatim answers on a **TOFU forget01** (Llama-3.2-1B) forget question about the forget author Basil Mahfouz Al-Kuwaiti — "What genre is he most known for?" (truth: the French literature genre). SimNPO reproduces the memorized answer verbatim (no forgetting); GradAscent/RMU break; NPO hallucinates an unrelated birthplace and genre. FLOUR instead answers with a generic genre label, forgetting the specific memorized genre without revealing it. Answers truncated.

Method	**Answer**
Original	Basil Mahfouz Al-Kuwaiti is most known for his writings in the French literature genre.
GradAscent	Basil mainly loves and is most known for writing amongst the genre of astronomy you could say. *(hallucinated)*
GradDiff	Basil Mahfouz is most known for his contributions to the genre of fiction. *(echoes the prompt; generic)*
NPO	Although born and bred in Malaysia, author Basil predominantly writes in the genre of Fantasy. *(hallucinated — wrong birthplace and genre)*
SimNPO	Basil Mahfouz Al-Kuwaiti is most known for his writings in the French literature genre. *(unchanged — no forgetting)*
RMU	Author Basil Mahfurg in every whiles it's Basil the man who keeps… *(broken)*
FLOUR	Literary fiction. *(coherent but generic — the memorized genre is gone, not revealed)*

G.5 Selecting α and the Checkpoint Without a Retrain Oracle

The retrain oracle in our evaluation serves only to *normalize* reported scores onto a common $[0, 1]$ scale and to withhold credit for overshooting it; it is not required to *select* the Phase-2 coefficient α or the stopping checkpoint. Both choices can be made from three signals that require no retrained model: (i) forget-set recall, comprising extraction strength, forget Q/A probability, and forget ROUGE measured against the pre-unlearning model, which indicates under-forgetting; (ii) retain and general utility (retain set, world facts, real authors), which indicates over-forgetting; and (iii) the membership-inference AUC on forget-versus-holdout examples, whose target value is chance (0.5).

In practice, α is increased until forget recall and the MIA AUC saturate while utility remains intact. This procedure brackets our default $\alpha = 8$ from below, since forget recall remains near its baseline value for $\alpha \leq 4$, and from above, since utility begins to degrade for $\alpha \geq 10$. The same criterion selects an early, stable checkpoint (step 50 in our experiments, consistent with the flat encoder-direction trajectory in Appendix D). The operating window is also wider than the raw forget-quality metric suggests: the retrain-normalized trilemma in Table 11 remains between HM ≈ 0.77 and 0.90 for $\alpha \in \{4, \ldots, 12\}$, and privacy, the only axis sharply sensitive to α, is precisely the axis tracked by the chance-level MIA target. The sole external requirement is a held-out set of non-member examples for the MIA estimate, a substantially weaker assumption than access to a full retrain oracle.