FORGET, BUT NOT TOO MUCH: SELF-LIMITING LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove selected knowledge from a trained model without retraining from scratch. Most existing LLM unlearning methods reduce the probability of the target answers. This does not match how TOFU evaluates forgetting. TOFU compares the unlearned model with a model that never saw the forget set, so pushing the target answers far below that reference can reduce Forget Quality just as retaining them can. We show that common one-sided objectives have no stationary point at this target. They can therefore keep updating after the model has reached its best agreement with the reference. We introduce NGREF, a self-limiting objective that matches the model to its frozen pretrained checkpoint on the forget set while retaining performance on the remaining data. The pretrained checkpoint becomes a fixed point of the forget objective. On the full published TOFU comparison grid with two models, three forget splits and four LoRA ranks, NGREF exceeds the best Forget Quality reported by nine published methods in all 24 settings. A controlled objective swap raises Forget Quality from approximately zero to 0.585 while leaving model utility similar. In a training sweep, a strong one-sided baseline reaches 0.766 and then falls to 0.0013 even though its forget loss continues to improve. Reference matching instead reaches 0.99 and remains close to that value over the measured range. These results show that the endpoint of unlearning matters as much as the strength of forgetting. Code: `https://anonymous.4open.science/r/ngref-unlearn-4F8E/README.md`

1 INTRODUCTION

Large language models can memorize material from their training data, including content that later needs to be removed (Carlini et al., 2021). Retraining a large model after each deletion request is usually impractical, which has motivated machine unlearning as an alternative to full retraining (Cao & Yang, 2015). Approximate unlearning for language models follows the same practical motivation (Jang et al., 2023). The central question is not only how to make a target answer less likely. It is what the model should look like once forgetting is complete.

TOFU gives this question a concrete answer (Maini et al., 2024). It evaluates an unlearned model against a retain-only model that was trained without the forget set. Forget Quality is computed from a two-sample Kolmogorov–Smirnov test over truth-ratio distributions. A high score means that the edited model and the retain-only model behave similarly on the forget examples. The target is therefore not zero probability on the forgotten answers. The target is the behavior of a model that never learned them.

This creates a mismatch with several widely used unlearning objectives. Gradient ascent and gradient difference reduce the likelihood of the forget answers (Jang et al., 2023). Negative preference optimization also applies pressure in the same direction (Zhang et al., 2024). The inverted hinge loss used by LoKU bounds this pressure but still reaches its minimum when the target token is suppressed to the edge of the probability simplex (Cha et al., 2025). These objectives improve when the target answer becomes less likely. They do not become optimal when the model matches an ignorant reference.

The distinction is visible in training. On Phi-1.5B forget01 at rank 16, the inverted hinge baseline reaches a Forget Quality of 0.766 after five epochs and then falls to 0.0013 by epoch 50. Over the

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.

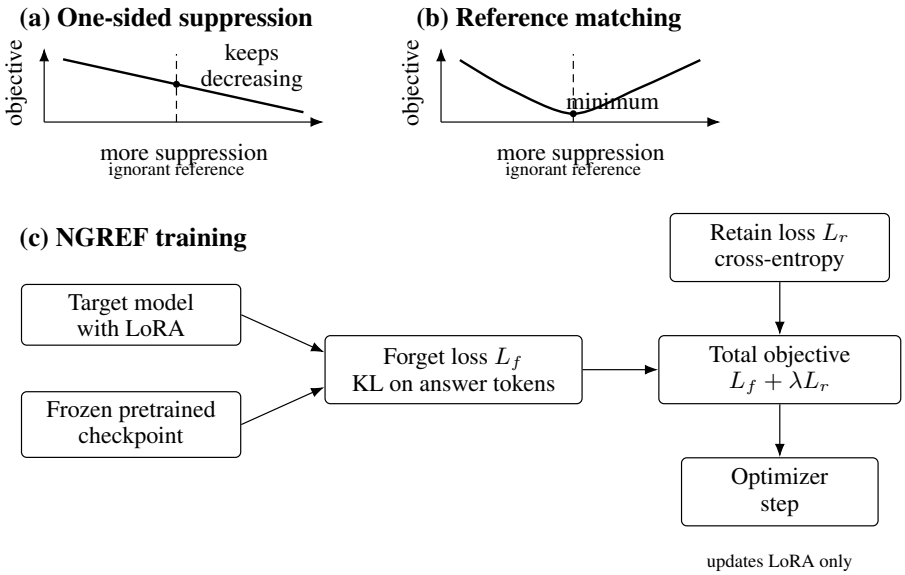

Figure 1. The main idea. A one-sided forget objective continues to decrease after the model reaches the behaviour of an ignorant reference. Reference matching instead places a minimum at that target. NGREF computes a forget loss by matching the edited model to the frozen pretrained checkpoint on the forget answers, combines it with the retain loss, and updates only the LoRA parameters.

same interval its forget loss continues to move in the direction that training treats as progress. The observable loss therefore does not identify the point at which the unlearning target has been reached.

We address this with a simple change to the forget objective. NGREF matches the token distribution of the edited model to the frozen pretrained checkpoint on the forget answers. The checkpoint already exists before task fine-tuning and has not seen the forget set. We add the standard retain loss on the remaining data. No Fisher-weighted adapter initialization or gradient preconditioner is required. When the edited model matches the pretrained reference, the forget gradient is zero.

We evaluate NGREF on the full TOFU grid reported by Cha et al. (2025). The comparison covers Phi-1.5B and Llama2-7B, three forget splits and LoRA ranks 4, 8, 16 and 32. NGREF is above the best Forget Quality reported by the nine methods in that paper in all 24 cells. The smallest margin is on Llama2-7B forget10 at rank 4, where NGREF reaches 0.4549 and the best published value is 0.293. A full ablation of the earlier Fisher preconditioner shows that it is not responsible for this result. Removing it changes mean Forget Quality by -0.0080 across the 24 cells with a Wilcoxon p value of 0.663, while all 24 cells remain above the published comparison.

Our main contributions are the following.

- We identify a mismatch between a two-sided unlearning target and common one-sided forget objectives. We show that an ignorant reference is not a stationary point of this class of objectives.
- We introduce NGREF, which uses the pretrained checkpoint as the target distribution on the forget set. The target is a fixed point of the forget objective and the method reduces to a standard LoRA update with one additional frozen-reference forward pass.
- We evaluate the method on the complete 24-cell TOFU comparison grid and study the mechanism with training trajectories, an objective swap, a 24-cell preconditioner ablation, seed variation, retain-weight sensitivity and a secondary extraction experiment.

Our claims are about approximate unlearning under the TOFU protocol. We do not claim exact deletion or a privacy guarantee. We also do not claim that Forget Quality increases monotonically under continued NGREF training. Long runs can drift after the best point. The claim is that the forget objective itself has the desired reference distribution as its endpoint.

2 Related work

LLM unlearning. Gradient ascent and gradient difference are common baselines for LLM unlearning (Jang et al., 2023). Preference-based objectives such as NPO improve optimization stability while retaining the same basic goal of suppressing the forget answers (Zhang et al., 2024). SimNPO removes the need for a reference model in the preference objective (Fan et al., 2025). Other approaches use distillation (Wang et al., 2023). Related work has also explored task arithmetic (Ilharco et al., 2023), prompt perturbation (Liu et al., 2024) and in-context interventions (Pawelczyk et al., 2024). Recent ICLR work has examined shortcomings in LLM unlearning evaluation (Yuan et al., 2025). MUSE further broadens evaluation across memorization, utility, privacy and sequential deletion requests (Shi et al., 2025). We focus on a different issue. We ask whether the training objective has its optimum at the behavior that the benchmark treats as forgotten.

Parameter-efficient unlearning. LoRA makes repeated post-training edits cheaper than full finetuning (Hu et al., 2022). Adapter methods provide a broader parameter-efficient alternative (Houlsby et al., 2019). Cha et al. (2025) combine an inverted hinge loss with a Fisher-weighted initialization for LoRA adapters. We use the same TOFU models, forget splits and rank grid so that the comparison is direct. NGREF uses a standard adapter and does not require the Fisher-weighted initialization. Our new ablation also shows that the diagonal Fisher preconditioner used in an earlier implementation can be removed without a measurable loss across the full comparison grid.

Evaluating unlearning. TOFU constructs synthetic author knowledge and trains retain-only reference models, which makes counterfactual comparison possible (Maini et al., 2024). MUSE broadens the evaluation of language-model unlearning across memorization, utility, privacy and sequential deletion requests (Shi et al., 2025). Other work studies knowledge entanglement (Tian et al., 2024). The use of membership inference as evidence of forgetting has also been questioned (Duan et al., 2024). We use TOFU because its reference-based metric makes the endpoint of unlearning explicit and therefore makes oversuppression observable.

3 Why one-sided objectives can overshoot

Let $\mathcal{D}_f$ denote the forget set and $\mathcal{D}_r$ the retain set. For a forget example x, TOFU forms a truth ratio from a paraphrase of the true answer and a set of perturbed answers. Writing L_θ for length-normalized likelihood, the ratio is

$$\rho_\theta(x) = \frac{L_\theta(\tilde{y}(x) \mid x)}{\left(\prod_{j=1}^{K} L_\theta(y_j^-(x) \mid x)\right)^{1/K}}. \tag{1}$$

Forget Quality is the two-sample KS p value between the empirical law of ρ_θ on $\mathcal{D}_f$ and the corresponding law from a retain-only model. The score is largest when the two empirical distributions are similar. It decreases when the edited model retains too much information and it can also decrease when the target answers are suppressed beyond the retain-only behavior.

Consider a token position in a forget answer and an objective of the form

$$\ell_{\text{one}}(p) = \phi(p(y_t)), \qquad \phi'(u) > 0. \tag{2}$$

Minimizing this objective continues to reduce the probability of the target token. Gradient ascent and gradient difference fall into this class after changing signs. NPO is likewise strictly increasing in the probability of the forget sequence over the interior (Zhang et al., 2024). The inverted hinge objective is bounded but still reaches its minimum by driving the target token probability to zero (Cha et al., 2025).

Proposition 1 (One-sided forgetting is not stationary at an ignorant reference)**.** *Let q be an interior categorical distribution with $q(y_t) > 0$. Then q is not a stationary point of an objective of the form in Equation* (2)*. The minimum over the simplex lies on the face where $p(y_t) = 0$.*

The proof is given in Appendix B. The practical consequence is direct. A model that never learned the forget set can still assign nonzero probability to a correct answer. A one-sided objective therefore continues to update after reaching that distribution. A fixed training budget can stop before, near or after the best reference match.

4 NGREF

4.1 REFERENCE MATCHING

Let P_{ref} be the pretrained checkpoint from which the target model was fine-tuned. The checkpoint predates the task-specific training that introduced the forget examples. We freeze P_{ref} and match its next-token distribution on the answer positions of the forget examples. If $A(x)$ denotes those positions, NGREF optimizes

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_f}\left[\frac{1}{|A(x)|}\sum_{t\in A(x)} \mathrm{KL}(P_{\text{ref}}(\cdot \mid x_{<t}) \,\|\, P_\theta(\cdot \mid x_{<t}))\right] + \lambda\, \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[-\log P_\theta(y \mid x)\right]. \tag{3}$$

We use $\lambda = 0.3$ for the main 24-cell comparison. The base model is frozen and we train a standard LoRA adapter. The pretrained reference is evaluated under `no_grad`. There is no data-dependent adapter initialization and no gradient preconditioning.

For one token position,

$$\nabla_\theta \mathrm{KL}(P_{\text{ref}} \| P_\theta) = -\sum_y P_{\text{ref}}(y) \nabla_\theta \log P_\theta(y). \tag{4}$$

At $P_\theta = P_{\text{ref}}$, this becomes $-\sum_y P_\theta(y)\nabla_\theta \log P_\theta(y) = 0$.

Proposition 2 (Reference matching has a unique stationary point)**.** *For an interior reference distribution q, $\mathrm{KL}(q\|p)$ has $p = q$ as its unique stationary point on the probability simplex and as its unique global minimum.*

The proof and the assumptions needed to relate this statement to a finite LoRA parameterization are given in Appendix B. The result is about the forget term. The retain loss and finite adapter capacity mean that the trained model need not exactly realize the reference distribution.

4.2 WHY USE THE PRETRAINED CHECKPOINT

The training reference is not TOFU's retain-only model. The retain-only model is trained separately for a particular deletion request and is used only for evaluation. Using it during training would make the procedure depend on the model that approximate unlearning is intended to avoid training. The pretrained checkpoint is already available and has not seen the forget set.

The optimized and evaluated quantities are also different. NGREF minimizes token-level KL on the exact forget answers. TOFU evaluates a KS test over truth ratios computed from paraphrases and perturbed answers. We do not query the benchmark retain-only model during training.

4.3 COST

NGREF adds one frozen-reference forward pass on each forget micro-batch. The 7B reference is held in 8-bit precision on GPU and receives no gradients. All other updates use the inherited LoRA training setup. Removing the Fisher preconditioner from the earlier implementation also removes its extra backward pass and gradient accumulator.

5 EXPERIMENTS

5.1 SETUP

We follow the TOFU protocol used by Cha et al. (2025). The evaluation grid contains Phi-1.5B and Llama2-7B, the forget01, forget05 and forget10 splits, and LoRA ranks 4, 8, 16 and 32. Forget Quality and Model Utility are computed with the benchmark evaluation code. FQ is the primary comparison because it is computed against the benchmark retain-only models for both our runs and the published runs.

We inherit learning rates of 2×10^{-4} for Phi-1.5B and 1×10^{-4} for Llama2-7B. LoRA is applied to all linear projections with scaling $2r$ and no dropout. The epoch budget and effective batch size

Table 1. Forget Quality on the complete TOFU comparison grid. Higher is better. The best published column is the highest value among the nine methods reported by Cha et al. (2025). NGREF uses the plain objective in Equation (3) without Fisher preconditioning.

			Best prior		NGREF	
Model	Split	r	FQ ↑	Method	FQ ↑	MU ↑
Phi-1.5B	forget01	4	0.0324	GD+FILA	$\mathbf{0.266}_{+0.234}$	0.491
		8	0.0455	GD+FILA	$\mathbf{0.579}_{+0.534}$	0.512
		16	0.244	GD+FILA	$\mathbf{0.919}_{+0.675}$	0.537
		32	0.0306	IHL+FILA	$\mathbf{0.919}_{+0.888}$	0.539
	forget05	4	4.33×10^{-5}	GD+FILA	$\mathbf{0.328}_{+0.328}$	0.553
		8	3.16×10^{-7}	GD+FILA	$\mathbf{0.713}_{+0.713}$	0.547
		16	0.00153	KL	$\mathbf{0.793}_{+0.791}$	0.546
		32	0.00302	IHL+FILA	$\mathbf{0.628}_{+0.625}$	0.546
	forget10	4	2.56×10^{-5}	NPO	$\mathbf{0.342}_{+0.342}$	0.546
		8	1.66×10^{-4}	NPO	$\mathbf{0.396}_{+0.396}$	0.545
		16	0.121	IHL+FILA	$\mathbf{0.585}_{+0.464}$	0.550
		32	0.0208	IHL+FILA	$\mathbf{0.518}_{+0.497}$	0.555
Llama2-7B	forget01	4	1.53×10^{-4}	IHL+FILA	$\mathbf{0.766}_{+0.766}$	0.618
		8	2.60×10^{-4}	IHL+FILA	$\mathbf{0.766}_{+0.766}$	0.633
		16	0.333	GD+FILA	$\mathbf{0.919}_{+0.586}$	0.647
		32	0.00770	GA+FILA	$\mathbf{0.919}_{+0.911}$	0.651
	forget05	4	1.50×10^{-8}	GD+FILA	$\mathbf{0.0396}_{+0.0396}$	0.592
		8	4.77×10^{-7}	IHL+FILA	$\mathbf{0.328}_{+0.328}$	0.631
		16	0.0134	IHL+FILA	$\mathbf{0.628}_{+0.615}$	0.635
		32	0.114	GA	$\mathbf{0.271}_{+0.157}$	0.646
	forget10	4	0.293	IHL+FILA	$\mathbf{0.455}_{+0.162}$	0.630
		8	0.00731	KL	$\mathbf{0.249}_{+0.242}$	0.631
		16	0.0265	GD+FILA	$\mathbf{0.147}_{+0.121}$	0.619
		32	4.69×10^{-4}	NPO	$\mathbf{0.147}_{+0.147}$	0.620

are chosen per model and forget split and shared across all four ranks. The epoch budgets are 10, 35 and 40 for Phi-1.5B and 10, 20 and 30 for Llama2-7B on forget01, forget05 and forget10. The effective batch sizes are 4, 16 and 32 for Phi-1.5B and 4, 32 and 32 for Llama2-7B. These settings were chosen in the earlier preconditioned implementation. We reuse them unchanged after removing preconditioning. This choice avoids retuning in favor of the simplified method.

The comparison values are taken from Tables 2 through 9 of Cha et al. (2025). They report KL, DPO, NPO, GA, GA+FILA, GD, IHL, GD+FILA and IHL+FILA. For each cell we compare with the highest published FQ among those nine methods. The full published tables are reproduced in Appendix D.

5.2 Results on the complete TOFU grid

NGREF is above the best published FQ in all 24 cells of Table 1. The smallest margin occurs on Llama2-7B forget10 at rank 4, where NGREF reaches 0.4549 and the best published value is 0.293 from IHL+FILA. The next smallest margin is Llama2-7B forget05 at rank 32, where NGREF reaches 0.2705 and the best published value is 0.114 from GA. Several other cells differ by orders of magnitude. We do not interpret ratios between p values as effect sizes. The important result is that the same method is above the strongest published entry across the entire fixed grid.

5.3 Training beyond the best checkpoint

The trajectory in Figure 2 isolates the stopping problem. IHL+FILA peaks at 0.766 at epoch 5 and falls to 0.0013 by epoch 50. Its forget-answer loss rises from 1.07 to 18.58 over the same run. The training signal therefore keeps improving after the benchmark score has collapsed. NGREF reaches

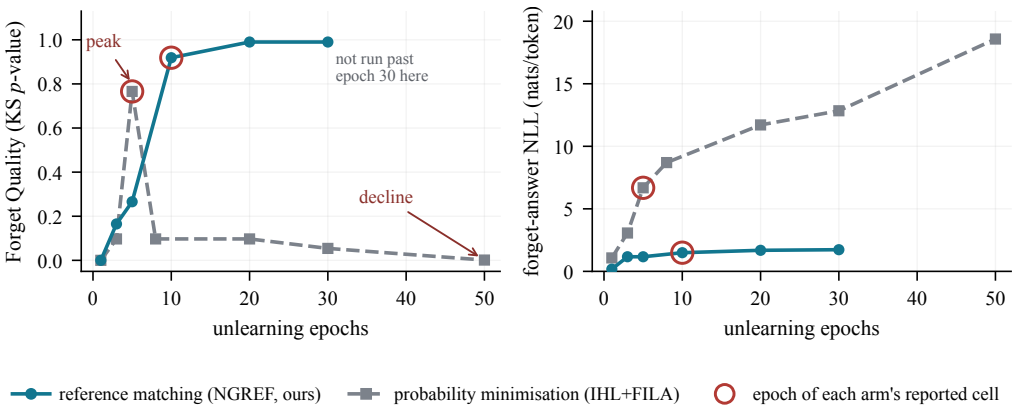

Figure 2. Continued training on Phi-1.5B forget01 at rank 16. The one-sided IHL+FILA objective reaches high FQ and then collapses while its forget-answer loss continues to increase. Reference matching reaches a high FQ region and its training loss flattens. Every point is measured.

Table 2. Controlled objective swap on Phi-1.5B forget10 at rank 16. The adapter, retain weight, budget and optimizer are held fixed.

Forget objective	FQ	MU
Reference matching without preconditioning	**0.5850**	**0.5503**
Inverted hinge without preconditioning	0.0003	0.5341
Inverted hinge with preconditioning	0.0000	0.5269

0.99 at epochs 20 and 30, while the reference-matching loss changes only slightly over that interval. NGREF is slower early in training. At epoch 5 it has FQ 0.266 while IHL+FILA is at its peak. This is expected for an objective that approaches a target distribution instead of pushing probability away without a target.

Longer NGREF runs are not perfectly monotone. On Phi-1.5B forget10 at rank 16, FQ changes from 0.6536 at epoch 40 to 0.5850 at epoch 50. On Llama2-7B forget05 at rank 16, it changes from 0.5453 at epoch 20 to 0.1779 at epoch 35. The relevant distinction is therefore not monotonicity. It is the absence of the three-order-of-magnitude collapse seen under continued one-sided suppression in Figure 2.

5.4 The result comes from the objective

Table 2 changes only the forget objective while keeping the remaining training setup fixed. Reference matching gives FQ 0.5850. The inverted hinge objective gives 0.0003 without preconditioning and 0.0000 with the same Fisher preconditioner used in the earlier implementation. Model utility remains close across the three runs. This isolates the main effect to the forget objective rather than the adapter or the preconditioner.

5.5 Preconditioning is unnecessary

The left panel of Figure 3 gives the full 24-cell preconditioner ablation. Twelve cells are unchanged, six increase and six decrease. The mean change is -0.0080 with standard deviation 0.0588. The Wilcoxon signed-rank test gives $p = 0.663$. Mean MU changes by only $+0.0004$. Most importantly, the plain objective remains above the best published FQ in all 24 cells. We therefore remove the preconditioner from the final method.

The right panel checks $\lambda \in \{0.1, 0.3, 1.0\}$ on Phi-1.5B forget10 at rank 16 and Llama2-7B forget05 at rank 16. On Phi-1.5B the FQ values are 0.6536, 0.6536 and 0.3417. On Llama2-7B they are 0.7126, 0.5453 and 0.6284. All six runs remain above the best published FQ for their corresponding

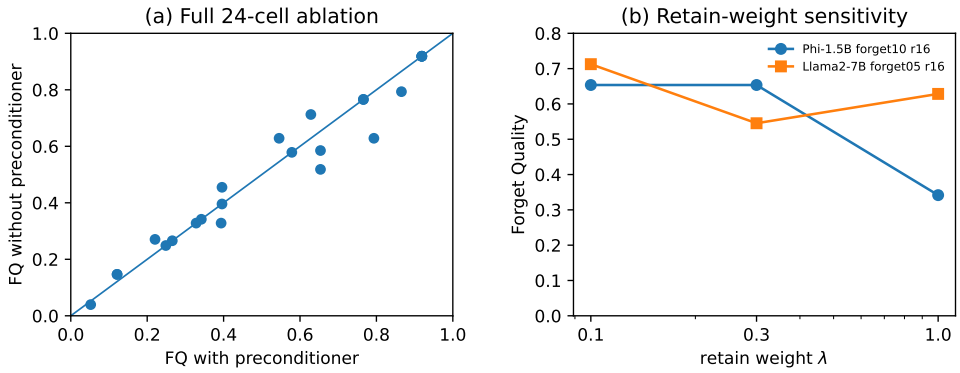

Figure 3. Ablations on the earlier implementation. Left, each point is one TOFU cell with and without the Fisher preconditioner. Across all 24 cells the mean FQ change after removing it is -0.0080 with Wilcoxon $p = 0.663$. Right, sensitivity to the retain weight on two rank-16 settings. Every tested value remains above the best published FQ for that cell.

Table 3. Three-seed results at rank 16. Values are mean $\pm$ standard deviation. The final column gives the best published FQ for the same cell.

Setting	Preconditioned	Plain	Best published
Phi-1.5B forget10	0.587 ± 0.115	0.586 ± 0.068	0.121
Llama2-7B forget05	0.493 ± 0.046	0.522 ± 0.119	0.0134

Table 4. Training-data extraction experiment at matched forgetting. Both measured columns use our implementation. We report the increase in held-out perplexity, where lower is better.

Model	Self-limiting	One-sided reference	Reduction
GPT-Neo-125M	+11.06	+24.66	2.2×
GPT-Neo-1.3B	+5.04	+8.10	1.6×
GPT-Neo-2.7B	+4.60	+8.35	1.8×

cell. The Llama2-7B ordering is not stable enough to support a preferred value beyond the fixed $\lambda = 0.3$ used in the main grid.

5.6 Seed variation

The three-seed comparisons in Table 3 show no measurable preconditioner effect on either tested setting. Welch tests give $p = 0.983$ for Phi-1.5B forget10 and $p = 0.717$ for Llama2-7B forget05. The larger point is that seed variation is substantial relative to many single-run differences. Every individual draw in these experiments remains above the best published FQ for its setting.

5.7 A secondary extraction experiment

TOFU is useful here because its two-sided score exposes oversuppression. To check whether the self-limiting principle is tied only to that metric, we also use the training-data extraction setting of Jang et al. (2023). We match forgetting between the two objectives and measure the increase in perplexity on held-out text. The self-limiting version incurs 1.6 to 2.2 times less perplexity increase across the three GPT-Neo scales in Table 4. No implementation was released for the published comparison method in this setting, and our port does not reproduce its published absolute values. We therefore treat this as a same-harness comparison and do not compare these values with the published numbers.

Table 5. Longer NGREF runs on three rank-16 settings. Parentheses give the epoch. The final column is the best published FQ for that setting.

Setting	Earlier checkpoint	Main checkpoint	Later checkpoint	Best published
Phi-1.5B forget10	0.4549 (30)	0.6536 (40)	0.5850 (50)	0.121
Phi-1.5B forget05	0.4663 (25)	0.8655 (35)	n/a	0.00153
Llama2-7B forget05	0.2705 (15)	0.5453 (20)	0.1779 (35)	0.0134

5.8 Longer runs and the narrow cells

Table 5 shows why we describe NGREF as self-limiting rather than monotone. The reference-matching objective removes the persistent direction toward zero probability, but finite-capacity optimization can still move away from the best benchmark match after the main checkpoint. The important difference from the one-sided trajectory is the scale of the failure. In the measured IHL+FILA sweep, FQ falls by a factor of 589 after its peak. In the NGREF sweeps, the later checkpoints remain well above the corresponding published values even when they are below the main checkpoint.

Two cells in the 24-cell grid deserve particular attention because their published baselines are already strong. Llama2-7B forget10 at rank 4 improves from 0.293 to 0.4549. Llama2-7B forget05 at rank 32 improves from 0.114 to 0.2705. These are single-seed comparisons and are therefore weaker evidence than the large-margin cells. We keep them in the complete-grid result because the comparison protocol reports one run per cell, but we do not use either cell alone to support the mechanism claim. The mechanism is supported instead by the controlled objective swap and the measured training trajectories.

6 Discussion

Why not use early stopping on a one-sided objective? A one-sided method can reach a good TOFU score before it overshoots. That is exactly what Figure 2 shows. The difficulty is that the forget loss does not identify that point. Selecting the best checkpoint requires the retain-only model used by the benchmark or another external stopping criterion. NGREF does not eliminate the need to choose a training budget, but it changes the endpoint of the forget objective. Once the model matches the reference distribution, the direction that pushes the model further away disappears.

This distinction is useful when comparing with recent LLM unlearning work. Published studies have emphasized unstable optimization and the forgetting and utility trade-off (Cha et al., 2025). Other work has questioned whether standard evaluations capture the intended change in model behavior (Yuan et al., 2025). MUSE likewise argues for broader evaluation of unlearning efficacy (Shi et al., 2025). Our result is compatible with those concerns. It identifies one concrete source of instability in reference-based evaluation. A training objective can keep improving according to its own loss after it has moved past the counterfactual behavior used for evaluation.

What the 24-cell comparison establishes. The strongest published method changes with the model, split and rank. Evaluating the complete grid reduces the chance that a result depends on one favorable stopping time or adapter configuration. NGREF remains above the best published FQ in every cell even after the preconditioner is removed. The ablation is also useful for simplifying the method. The earlier implementation contained both a new objective and a diagonal Fisher preconditioner. The full 24-cell ablation and the seed tests show no measurable benefit from the preconditioner, so the final method keeps only the component supported by the experiments.

FQ is a KS-test p value, so very large ratios against values close to zero should not be interpreted as linear effect sizes. A value of 0.6 is not meaningfully described as hundreds of thousands of times better than a value near 10^{-6}. For this reason, the paper reports the raw values and emphasizes the consistency of the ordering across the fixed comparison grid. The two cells with the smallest absolute margins are also identified explicitly.

What reference matching does not guarantee. Matching a pretrained checkpoint is not the same as certified deletion. The checkpoint was trained before the task-specific data were introduced,

Table 6. The endpoint represented by common forget objectives. The table concerns the forget term only.

Objective family	Training signal	Endpoint of forget term	Stationary at P_{ref}
GA and GD	raise target NLL	target probability $\to 0$	No
NPO and SimNPO	preference suppression	target probability $\to 0$	No
IHL	token margin	target token $\to 0$	No
NGREF	distribution matching	$P_\theta = P_{\mathrm{ref}}$	Yes

which makes it a natural available reference, but it can differ from a model retrained exactly on the retain set. The formal analysis therefore concerns the endpoint of the forget objective. It does not establish that all information about the forget examples has been removed from internal representations or that the model is robust to extraction and relearning attacks.

This point also explains why we retain the secondary extraction experiment. TOFU is the main benchmark because it provides the two-sided reference target needed to test the overshoot hypothesis directly. The extraction experiment asks a different question. At matched forgetting, does a self-limiting construction reduce collateral damage on held-out text. The answer is positive across three GPT-Neo scales in our harness, but the experiment is not calibrated against published absolute numbers and we do not present it as a second state-of-the-art claim.

How the objectives differ. The distinction can be summarized by the endpoint that each forget loss represents. Gradient ascent and gradient difference keep increasing the loss of the target answer. NPO and SimNPO reduce the target sequence probability through a bounded or saturating preference objective. IHL keeps the update bounded by comparing the target token with its strongest competitor. These changes can improve optimization, but none of them makes a nonzero ignorant distribution optimal. NGREF changes the target of the optimization rather than only the shape of the penalty.

This view also clarifies why the paper does not argue that reference matching is universally preferable. If the desired behavior is refusal, then an answer-substitution objective can encode that endpoint directly. If the goal is resistance to extraction under a specified attacker, then an attack-aware criterion may be more appropriate. Our claim applies when the desired endpoint is the counterfactual behavior of a model that did not learn the removed data. Under that definition, the endpoint should appear in the training objective rather than only in the evaluation metric.

Limitations. The epoch budget and effective batch size are hyperparameters. They are shared across ranks and no result in the main table is selected by an epoch scan, but an automatic stopping rule would be preferable. The 24-cell grid is a single-seed comparison. Multi-seed runs support the tested rank-16 cells, but the two narrowest margins have not been repeated across seeds. Finally, the formal connection between the pretrained checkpoint and TOFU's retain-only model assumes that both represent the same counterfactual behavior on the forget set. The two models are both ignorant of the forget data, but their truth-ratio distributions need not be identical. This assumption is directly testable and remains an important follow-up.

7 Conclusion

Approximate LLM unlearning is often framed as reducing the probability of an unwanted answer. TOFU makes clear that this is not the full objective. The desired endpoint is the behavior of a model that never learned the forget set. Common one-sided objectives can pass through that point and continue optimizing.

NGREF makes this endpoint part of the objective by matching the edited model to its frozen pretrained checkpoint on the forget answers while using a retain loss on the remaining data. The method uses a standard LoRA adapter and no Fisher-weighted initialization or gradient preconditioning. On the full 24-cell TOFU grid, NGREF is above the best Forget Quality reported among nine published methods in every cell.

AI USE STATEMENT

Generative AI tools were used for literature search, language editing, figure preparation and table preparation. They were not used to generate experimental measurements. All numerical results were checked against the recorded outputs and the source tables. The authors take responsibility for the final manuscript.

ETHICS STATEMENT

Machine unlearning is motivated by privacy, data governance and control over trained models. Our method is an approximate unlearning method and does not provide a formal deletion guarantee. The TOFU experiments use synthetic author data and public checkpoints. The secondary extraction experiment follows an established public benchmark setting.

REPRODUCIBILITY STATEMENT

The objective is given in Equation (3). Model, split, rank, learning rate, epoch and batch settings are reported in Section 5.1. The appendix contains full hyperparameters, proofs, seed results, sensitivity analyses, long training runs and the complete published comparison tables.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

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)*, 2021.

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.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

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 38 (NeurIPS)*, 2025.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y. Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842, 2020.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

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.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. arXiv preprint arXiv:1907.00456, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

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

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40034–40050, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1524–1537, 2024.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. KGA: A general machine unlearning framework based on knowledge gap alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13264–13276, 2023.

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 (ICLR)*, 2025.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017.

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.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.

A Implementation details

NGREF uses the same optimizer, learning rates, adapter targets and evaluation code as the TOFU configuration used by Cha et al. (2025); Maini et al. (2024). LoRA is applied to all linear projections with rank $r \in \{4, 8, 16, 32\}$, scaling $2r$ and no dropout. The forget batch uses micro-batch size one with gradient accumulation to reach the effective batch sizes reported in Section 5.1. The pretrained reference is frozen and evaluated under `no_grad`. It is stored in 8-bit precision on GPU. The final method uses no Fisher preconditioner and no Fisher-weighted adapter initialization.

B Proofs

B.1 One-sided objectives

Consider the simplex $\Delta(V)$ and let $q \in \operatorname{relint} \Delta(V)$ with $q(y_t) > 0$. For $\ell(p) = \phi(p(y_t))$ and $\phi'(u) > 0$, choose a tangent direction that moves probability mass from y_t to another token v. For sufficiently small $\epsilon > 0$, define $p_\epsilon(y_t) = q(y_t) - \epsilon$ and $p_\epsilon(v) = q(v) + \epsilon$. Then

$$\frac{d}{d\epsilon} \ell(p_\epsilon) \Big|_{\epsilon=0} = -\phi'(q(y_t)) < 0.$$

Table 7. Training configuration used in the main TOFU grid.

Model	Split	Epochs	Effective batch	Learning rate
Phi-1.5B	forget01	10	4	2×10^{-4}
Phi-1.5B	forget05	35	16	2×10^{-4}
Phi-1.5B	forget10	40	32	2×10^{-4}
Llama2-7B	forget01	10	4	1×10^{-4}
Llama2-7B	forget05	20	32	1×10^{-4}
Llama2-7B	forget10	30	32	1×10^{-4}

Hence q is not stationary. Since the objective decreases whenever probability can be moved away from y_t, its minimum is attained on the face where $p(y_t) = 0$. This proves Proposition 1.

B.2 Reference matching

For an interior reference distribution q, consider $f(p) = \mathrm{KL}(q\|p)$ on the simplex. Gibbs' inequality gives $f(p) \geq 0$ with equality only when $p = q$, so q is the unique global minimizer. For stationarity, the gradient in unconstrained coordinates is $-q_i/p_i$. At a constrained stationary point there is a scalar ν such that $-q_i/p_i = \nu$ for every i. This implies $p_i \propto q_i$. Since both distributions sum to one, $p = q$. This proves Proposition 2.

The propositions are statements in distribution space. A finite LoRA adapter need not be able to realize P_{ref} exactly on every forget token. The retain term also changes the stationary set of the complete objective. We therefore use the propositions to explain the direction of the forget objective rather than as a convergence guarantee for the trained system.

C Additional TOFU results

C.1 Longer training runs

Table 8. Forget Quality at successive epoch budgets. Blank entries were not evaluated.

Model	Split	r	FQ at epoch				Best published
			10	20	35 or 40	50	
Phi-1.5B	forget10	16			0.6536	0.5850	0.121
Phi-1.5B	forget05	16			0.8655		0.00153
Llama2-7B	forget05	16	0.0118	0.5453	0.1779		0.0134

C.2 Preconditioner ablation over all 24 cells

Table 9. Summary of the 24-cell ablation that removes the Fisher preconditioner.

Scope	Cells	Unchanged	Increased	Decreased	Mean ΔFQ	Wilcoxon p
All	24	12	6	6	-0.0080	0.663
Phi-1.5B	12	6	1	5	-0.0352	0.219
Llama2-7B	12	6	5	1	$+0.0192$	0.062

C.3 Retain-weight sensitivity

The Phi setting is forget10 at rank 16 and the Llama2-7B setting is forget05 at rank 16. The best published FQ values are 0.121 and 0.0134 respectively.

Table 10. Representative cells with and without the Fisher preconditioner. All runs use rank 16 and seed 42.

Setting	FQ with	FQ plain	MU with	MU plain	Best published
Phi-1.5B forget01	0.9188	0.9188	0.5375	0.5370	0.244
Phi-1.5B forget10	0.6536	0.5850	0.5495	0.5503	0.121
Llama2-7B forget05	0.5453	0.6284	0.6352	0.6350	0.0134

Table 11. Sensitivity to the retain weight in the earlier preconditioned implementation.

λ	Phi FQ	Phi MU	Llama FQ	Llama MU
0.1	0.6536	0.5574	0.7126	0.6013
0.3	0.6536	0.5495	0.5453	0.6352
1.0	0.3417	0.5375	0.6284	0.6320

Table 12. Individual seeds for the rank-16 settings used in the seed analysis.

Setting	Arm	Seed 42	Seed 1	Seed 2	Mean $\pm$ s.d.
Phi forget01	preconditioned	0.9188	0.5786	0.5786	0.6920 ± 0.1964
Phi forget05	preconditioned	0.8655	0.7934	0.7126	0.7905 ± 0.0765
Phi forget10	preconditioned	0.6536	0.4549	0.6536	0.5873 ± 0.1147
Phi forget10	plain	0.5850	0.5183	0.6536	0.5856 ± 0.0677
Llama forget05	preconditioned	0.5453	0.4663	0.4663	0.4926 ± 0.0456
Llama forget05	plain	0.6284	0.3935	0.5453	0.5224 ± 0.1191

C.4 Seed results

D The complete published TOFU comparison

The tables below reproduce all nine published methods from Cha et al. (2025) for each model and LoRA rank. NGREF is added as the final row in each split. The NGREF entries use the plain objective without the Fisher preconditioner. A star marks the best published FQ within each split.

E Additional related work

Machine unlearning predates large language models. Early work formalized deletion from learned systems and studied exact or certified removal in more restricted settings (Cao & Yang, 2015). Subsequent methods considered data partitioning, retraining strategies and efficient deletion mechanisms (Ginart et al., 2019). Certified removal provides another line of work in which the effect of deleting training data is bounded explicitly (Guo et al., 2020). SISA training studies retraining efficiency by isolating the influence of subsets of the training data (Bourtoule et al., 2021). Other theoretical work asks when deletion can be made computationally cheaper than retraining (Sekhari et al., 2021). These settings differ from LLM unlearning, where exact retraining is usually unavailable and evaluation is therefore based on observable model behavior.

Several neighboring areas inform the design choices used in LLM unlearning. Membership inference provides one way to probe whether individual examples remain detectable after training (Shokri et al., 2017). Training data extraction shows that large language models can emit memorized sequences and motivates targeted removal after training (Carlini et al., 2021). Knowledge distillation transfers model behavior through a teacher distribution (Hinton et al., 2015), while recent language-model distillation work studies the same idea at larger scale (Gu et al., 2024). Preference optimization provides another family of post-training objectives (Rafailov et al., 2023). Earlier preference learning for language generation and dialogue established many of the optimization ingredients used by later alignment methods (Ziegler et al., 2019) (Jaques et al., 2019). Instruction tuning with human feedback made these methods central to modern language-model post-training (Ouyang et al., 2022).

Table 13. Complete published TOFU comparison for Phi-1.5B across all four LoRA ranks. A star marks the best published FQ within each forget split.

(a) Rank 4

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$7.37 \cdot 10^{-5}$	0.52		
DPO	$8.87 \cdot 10^{-5}$	0.52		
NPO	$6.11 \cdot 10^{-5}$	0.52		
GA	$6.11 \cdot 10^{-5}$	0.52		
GA+FILA	$1.07 \cdot 10^{-3}$	0.00		
GD	$7.37 \cdot 10^{-5}$	0.52		
IHL	$4.17 \cdot 10^{-5}$	0.52		
GD+FILA*	$3.24 \cdot 10^{-2}$	0.00		
IHL+FILA	$1.28 \cdot 10^{-4}$	0.51		
NGREF (ours)			**0.266**	0.491
	TOFU forget05			
KL	$2.90 \cdot 10^{-13}$	0.48		
DPO	$2.17 \cdot 10^{-13}$	0.47		
NPO	$4.87 \cdot 10^{-12}$	0.48		
GA	$1.10 \cdot 10^{-11}$	0.47		
GA+FILA	$3.69 \cdot 10^{-8}$	0.00		
GD	$4.33 \cdot 10^{-15}$	0.50		
IHL	$6.68 \cdot 10^{-14}$	0.50		
GD+FILA*	$4.33 \cdot 10^{-5}$	0.00		
IHL+FILA	$1.44 \cdot 10^{-11}$	0.51		
NGREF (ours)			**0.328**	0.553
	TOFU forget10			
KL	$7.38 \cdot 10^{-15}$	0.00		
DPO	$5.10 \cdot 10^{-17}$	0.48		
NPO*	$2.56 \cdot 10^{-5}$	0.37		
GA	$2.06 \cdot 10^{-13}$	0.00		
GA+FILA	$5.10 \cdot 10^{-17}$	0.00		
GD	$2.55 \cdot 10^{-9}$	0.36		
IHL	$2.43 \cdot 10^{-17}$	0.51		
GD+FILA	$2.17 \cdot 10^{-6}$	0.00		
IHL+FILA	$1.39 \cdot 10^{-12}$	0.51		
NGREF (ours)			**0.342**	0.546

(b) Rank 8

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$6.11 \cdot 10^{-5}$	0.52		
DPO	$8.87 \cdot 10^{-5}$	0.52		
NPO	$6.11 \cdot 10^{-5}$	0.52		
GA	$4.17 \cdot 10^{-5}$	0.52		
GA+FILA	$2.72 \cdot 10^{-2}$	0.00		
GD	$7.37 \cdot 10^{-5}$	0.52		
IHL	$7.37 \cdot 10^{-5}$	0.52		
GD+FILA*	$4.55 \cdot 10^{-2}$	0.00		
IHL+FILA	$3.37 \cdot 10^{-4}$	0.50		
NGREF (ours)			**0.579**	0.512
	TOFU forget05			
KL	$3.26 \cdot 10^{-10}$	0.35		
DPO	$1.21 \cdot 10^{-13}$	0.48		
NPO	$9.13 \cdot 10^{-11}$	0.41		
GA	$3.26 \cdot 10^{-10}$	0.33		
GA+FILA	$1.12 \cdot 10^{-9}$	0.00		
GD	$3.70 \cdot 10^{-12}$	0.44		
IHL	$1.21 \cdot 10^{-13}$	0.50		
GD+FILA*	$3.16 \cdot 10^{-7}$	0.00		
IHL+FILA	$3.73 \cdot 10^{-9}$	0.50		
NGREF (ours)			**0.713**	0.547
	TOFU forget10			
KL	$1.06 \cdot 10^{-16}$	0.00		
DPO	$1.15 \cdot 10^{-17}$	0.47		
NPO*	$1.66 \cdot 10^{-4}$	0.36		
GA	$5.10 \cdot 10^{-17}$	0.00		
GA+FILA	$2.43 \cdot 10^{-19}$	0.00		
GD	$5.07 \cdot 10^{-6}$	0.22		
IHL	$2.43 \cdot 10^{-17}$	0.50		
GD+FILA*	$1.66 \cdot 10^{-4}$	0.00		
IHL+FILA	$1.40 \cdot 10^{-6}$	0.50		
NGREF (ours)			**0.396**	0.545

(c) Rank 16

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$2.83 \cdot 10^{-5}$	0.53		
DPO	$7.37 \cdot 10^{-5}$	0.52		
NPO	$4.17 \cdot 10^{-5}$	0.52		
GA	$6.11 \cdot 10^{-5}$	0.52		
GA+FILA	$1.14 \cdot 10^{-2}$	0.00		
GD	$7.37 \cdot 10^{-5}$	0.52		
IHL	$1.28 \cdot 10^{-4}$	0.52		
GD+FILA*	$2.44 \cdot 10^{-1}$	0.00		
IHL+FILA	$1.70 \cdot 10^{-3}$	0.51		
NGREF (ours)			**0.919**	0.537
	TOFU forget05			
KL*	$1.53 \cdot 10^{-3}$	0.00		
DPO	$6.68 \cdot 10^{-14}$	0.49		
NPO	$1.10 \cdot 10^{-7}$	0.35		
GA	$4.33 \cdot 10^{-5}$	0.00		
GA+FILA	$5.96 \cdot 10^{-9}$	0.00		
GD	$3.73 \cdot 10^{-9}$	0.32		
IHL	$2.12 \cdot 10^{-12}$	0.50		
GD+FILA	$1.81 \cdot 10^{-5}$	0.00		
IHL+FILA	$5.03 \cdot 10^{-6}$	0.52		
NGREF (ours)			**0.793**	0.546
	TOFU forget10			
KL	$1.07 \cdot 10^{-13}$	0.00		

(d) Rank 32

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$4.17 \cdot 10^{-5}$	0.52		
DPO	$4.72 \cdot 10^{-4}$	0.52		
NPO	$5.05 \cdot 10^{-5}$	0.51		
GA	$5.56 \cdot 10^{-5}$	0.51		
GA+FILA	$5.55 \cdot 10^{-6}$	0.00		
GD	$8.87 \cdot 10^{-5}$	0.52		
IHL	$1.28 \cdot 10^{-4}$	0.52		
GD+FILA	$1.15 \cdot 10^{-3}$	0.00		
IHL+FILA*	$3.06 \cdot 10^{-2}$	0.52		
NGREF (ours)			**0.919**	0.539
	TOFU forget05			
KL	$4.87 \cdot 10^{-12}$	0.00		
DPO	$3.17 \cdot 10^{-15}$	0.49		
NPO	$3.64 \cdot 10^{-5}$	0.36		
GA	$2.17 \cdot 10^{-13}$	0.00		
GA+FILA	$4.77 \cdot 10^{-17}$	0.00		
GD	$1.76 \cdot 10^{-3}$	0.23		
IHL	$4.18 \cdot 10^{-11}$	0.51		
GD+FILA	$4.16 \cdot 10^{-6}$	0.00		
IHL+FILA*	$3.02 \cdot 10^{-3}$	0.50		
NGREF (ours)			**0.628**	0.546
	TOFU forget10			
KL	$2.17 \cdot 10^{-6}$	0.00		

Table 14. Complete published TOFU comparison for Llama2-7B across all four LoRA ranks. A star marks the best published FQ within each forget split.

(a) Rank 4

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$9.73 \cdot 10^{-5}$	0.63		
DPO	$1.40 \cdot 10^{-4}$	0.63		
NPO	$9.73 \cdot 10^{-5}$	0.63		
GA	$6.71 \cdot 10^{-5}$	0.63		
GA+FILA	$3.12 \cdot 10^{-5}$	0.00		
GD	$1.40 \cdot 10^{-4}$	0.63		
IHL	$1.17 \cdot 10^{-4}$	0.63		
GD+FILA	$1.15 \cdot 10^{-5}$	0.00		
IHL+FILA*	$1.53 \cdot 10^{-4}$	0.62		
NGREF (ours)			**0.766**	0.618
	TOFU forget05			
KL	$9.25 \cdot 10^{-17}$	0.64		
DPO	$8.99 \cdot 10^{-14}$	0.62		
NPO	$2.47 \cdot 10^{-16}$	0.64		
GA	$6.50 \cdot 10^{-16}$	0.64		
GA+FILA	$1.23 \cdot 10^{-15}$	0.00		
GD	$6.50 \cdot 10^{-16}$	0.64		
IHL	$6.64 \cdot 10^{-17}$	0.64		
GD+FILA*	$1.50 \cdot 10^{-8}$	0.00		
IHL+FILA	$6.87 \cdot 10^{-13}$	0.64		
NGREF (ours)			**0.040**	0.592
	TOFU forget10			
KL	$2.56 \cdot 10^{-5}$	0.49		
DPO	$5.10 \cdot 10^{-17}$	0.57		
NPO	$3.33 \cdot 10^{-6}$	0.47		
GA	$2.31 \cdot 10^{-3}$	0.39		
GA+FILA	$5.40 \cdot 10^{-18}$	0.00		
GD	$2.19 \cdot 10^{-16}$	0.62		
IHL	$3.71 \cdot 10^{-15}$	0.64		
GD+FILA	$1.83 \cdot 10^{-21}$	0.00		
IHL+FILA*	$2.93 \cdot 10^{-1}$	0.63		
NGREF (ours)			**0.455**	0.630

(b) Rank 8

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$1.17 \cdot 10^{-4}$	0.64		
DPO	$1.40 \cdot 10^{-4}$	0.63		
NPO	$6.71 \cdot 10^{-5}$	0.63		
GA	$6.71 \cdot 10^{-5}$	0.63		
GA+FILA	$3.12 \cdot 10^{-5}$	0.00		
GD	$1.17 \cdot 10^{-4}$	0.63		
IHL	$1.40 \cdot 10^{-4}$	0.63		
GD+FILA	$2.35 \cdot 10^{-6}$	0.00		
IHL+FILA*	$2.60 \cdot 10^{-4}$	0.62		
NGREF (ours)			**0.766**	0.633
	TOFU forget05			
KL	$6.50 \cdot 10^{-16}$	0.63		
DPO	$6.87 \cdot 10^{-13}$	0.50		
NPO	$1.09 \cdot 10^{-14}$	0.60		
GA	$1.21 \cdot 10^{-13}$	0.61		
GA+FILA	$1.32 \cdot 10^{-19}$	0.00		
GD	$3.17 \cdot 10^{-15}$	0.64		
IHL	$1.24 \cdot 10^{-17}$	0.63		
GD+FILA	$4.46 \cdot 10^{-18}$	0.00		
IHL+FILA*	$4.77 \cdot 10^{-7}$	0.63		
NGREF (ours)			**0.328**	0.631
	TOFU forget10			
KL*	$7.31 \cdot 10^{-3}$	0.01		
DPO	$5.10 \cdot 10^{-17}$	0.55		
NPO	$1.73 \cdot 10^{-5}$	0.45		
GA	$3.32 \cdot 10^{-8}$	0.00		
GA+FILA	$3.32 \cdot 10^{-8}$	0.00		
GD	$3.71 \cdot 10^{-15}$	0.56		
IHL	$2.86 \cdot 10^{-14}$	0.65		
GD+FILA	$1.15 \cdot 10^{-17}$	0.00		
IHL+FILA	$3.63 \cdot 10^{-7}$	0.63		
NGREF (ours)			**0.249**	0.631

(c) Rank 16

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$1.07 \cdot 10^{-4}$	0.64		
DPO	$3.37 \cdot 10^{-4}$	0.63		
NPO	$8.09 \cdot 10^{-5}$	0.64		
GA	$9.73 \cdot 10^{-5}$	0.64		
GA+FILA	$1.58 \cdot 10^{-2}$	0.00		
GD	$1.28 \cdot 10^{-4}$	0.63		
IHL	$1.40 \cdot 10^{-4}$	0.63		
GD+FILA*	$3.33 \cdot 10^{-1}$	0.00		
IHL+FILA	$7.13 \cdot 10^{-4}$	0.62		
NGREF (ours)			**0.919**	0.647
	TOFU forget05			
KL	$2.80 \cdot 10^{-12}$	0.54		
DPO	$1.62 \cdot 10^{-13}$	0.53		
NPO	$8.80 \cdot 10^{-10}$	0.53		
GA	$1.12 \cdot 10^{-9}$	0.52		
GA+FILA	$1.00 \cdot 10^{-3}$	0.00		
GD	$4.96 \cdot 10^{-14}$	0.61		
IHL	$3.42 \cdot 10^{-16}$	0.63		
GD+FILA	$4.19 \cdot 10^{-10}$	0.00		
IHL+FILA*	$1.34 \cdot 10^{-2}$	0.63		
NGREF (ours)			**0.628**	0.635
	TOFU forget10			
KL	$2.86 \cdot 10^{-14}$	0.00		

(d) Rank 32

Method	Published FQ	Published MU	Ours FQ	Ours MU
	TOFU forget01			
KL	$5.05 \cdot 10^{-5}$	0.64		
DPO	$6.57 \cdot 10^{-4}$	0.63		
NPO	$1.40 \cdot 10^{-4}$	0.64		
GA	$8.87 \cdot 10^{-5}$	0.63		
GA+FILA*	$7.70 \cdot 10^{-3}$	0.00		
GD	$1.53 \cdot 10^{-4}$	0.63		
IHL	$2.18 \cdot 10^{-4}$	0.63		
GD+FILA	$1.67 \cdot 10^{-4}$	0.00		
IHL+FILA	$4.13 \cdot 10^{-3}$	0.61		
NGREF (ours)			**0.919**	0.651
	TOFU forget05			
KL	$4.81 \cdot 10^{-4}$	0.36		
DPO	$2.72 \cdot 10^{-14}$	0.57		
NPO	$1.81 \cdot 10^{-5}$	0.47		
GA*	$1.14 \cdot 10^{-1}$	0.34		
GA+FILA	$2.47 \cdot 10^{-16}$	0.00		
GD	$6.87 \cdot 10^{-13}$	0.54		
IHL	$2.12 \cdot 10^{-12}$	0.63		
GD+FILA	$1.00 \cdot 10^{-3}$	0.00		
IHL+FILA	$7.38 \cdot 10^{-3}$	0.62		
NGREF (ours)			**0.271**	0.646
	TOFU forget10			
KL	$4.22 \cdot 10^{-21}$	0.00		

Our use of LoRA also connects unlearning to parameter-efficient adaptation. LoRA constrains updates to low-rank factors (Hu et al., 2022), while adapter-based methods provide an alternative way to isolate trainable parameters (Houlsby et al., 2019). Broader studies compare these parameter-efficient approaches under a common view (He et al., 2022). The Fisher-based components considered in earlier versions of our method are related to natural-gradient and curvature approximations (Amari, 1998) (Martens & Grosse, 2015). Empirical Fisher approximations can differ substantially from the true Fisher geometry (Kunstner et al., 2019). This is consistent with our empirical finding that the diagonal preconditioner is unnecessary for the final method.

There is also a conceptual link to continual learning, although the goals are different. Continual learning seeks to avoid unintended forgetting when learning new tasks (French, 1999). Elastic weight consolidation and synaptic intelligence regularize updates according to parameter importance (Kirkpatrick et al., 2017) (Zenke et al., 2017). Learning without forgetting preserves earlier task behavior while adapting to new data (Li & Hoiem, 2018). In machine unlearning, forgetting is deliberate, but the same tension between changing selected behavior and preserving the rest of the model remains.

F Secondary extraction setting

No implementation of the published reference method was available for this benchmark. Our port does not reproduce the published absolute values, so the main text reports only the controlled comparison between two objectives in our harness. The published values are shown here only for context.

Table 15. Published perplexity changes reported by Cha et al. (2025) for the training-data extraction setting. Lower is better.

Method	GPT-Neo-125M	GPT-Neo-1.3B	GPT-Neo-2.7B
GA	+560.0	+4.3	+7.5
GD	+36.4	−0.7	+7.5
IHL	+0.3	−0.5	+0.3
GD with LoRA	+111.6	+20.3	+50.7
IHL with LoRA	+15.1	+3.4	+11.8
GD+FILA	+71.7	+43.0	+58.3
LoKU	+6.2	+3.2	+5.6