NONMEMBER-LAW MATCHING FOR SUPPRESSING UNLEARNING SIGNATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning can suppress a language model's knowledge of forget data while leaving it vulnerable to membership inference: forget examples remain distinguishable from data the model was never trained on, revealing which records were used in training or targeted for removal. Existing LLM unlearning objectives lower forget likelihoods or redirect internal representations without reference to nonmember behavior, so unlearning can stop short or overshoot, and both leave identifiable signatures. We propose Nonmember-Law Matching (NML), which uses the original model's likelihoods on nonmembers as the target for forget examples. NML matches the distribution of forget-example negative log-likelihoods to this nonmember reference and preserves retained behavior through distillation, without requiring a retrained model. On two unlearning benchmarks, TOFU and MUSE-News, NML substantially reduces unlearning signatures under multiple likelihood-based attacks while maintaining forgetting and utility comparable to or better than strong baselines, and question-only identification attacks provide further evidence of this reduction. These results suggest that calibrating against nonmember behavior is a simple and effective way to reduce the identifiable traces of unlearning.

1 INTRODUCTION

Machine unlearning has emerged as a practical approach to removing the influence of unwanted training data from large language models (LLMs) without full retraining (Maini et al., 2024; Zhang et al., 2024). However, recent studies show that unlearned models can remain vulnerable to membership inference attacks (MIAs) (Du et al., 2025; Deepak et al., 2025). An MIA uses a model's behavior to infer whether a candidate example was used for training (Yeom et al., 2018). For example, FUMA exploits artifacts left by unlearning to identify examples in the forget set (Deepak et al., 2025). Such inferences can expose the use of an individual's private records in training and reveal which information was targeted for removal. This undermines the privacy protection expected from unlearning, even when the model cannot reproduce the original records.

Representative unlearning methods typically pair a forgetting objective with utility preservation. NPO, SimNPO, and UnDIAL lower the likelihood of forget data (Zhang et al., 2024; Fan et al., 2025; Dong et al., 2025), whereas RMU and LUNAR redirect internal representations (Li et al., 2024; Shen et al., 2025); most also constrain behavior on retained data to stay close to the original model. Protecting privacy, however, requires more than suppressing the model's behavior on forget data. An MIA infers membership by checking whether the model treats a candidate example differently from comparable data it was never trained on, termed nonmembers (Yeom et al., 2018; Shi et al., 2024). For unlearning to resist such inference, forget examples should become indistinguishable from nonmembers, as if the model had never seen them. Prior MIA studies show that unlearned models often fall short of this goal (Shi et al., 2025; Du et al., 2025); we call the residual distinguishability an unlearning signature. Figure 1 illustrates this on TOFU: under the Min-K% attack (Shi et al., 2024), an AUC of 0.5 means that forget examples and nonmembers cannot be told apart, yet the AUC remains near 1 under UnDIAL and falls well below 0.5 under NPO and SimNPO. Neither component of these objectives refers to nonmember behavior: the forgetting term moves the model away from forget data and the retain term constrains retained data, but neither determines where forget likelihoods should settle. The challenge is to reduce this separability while maintaining effective forgetting and retained utility.

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.

A model retrained without the forget data would indicate where forget likelihoods should settle, but obtaining it is what unlearning aims to avoid. Nonmember behavior, by contrast, is observable before unlearning. We therefore propose Nonmember-Law Matching (NML), which uses the original model's likelihoods on nonmembers as the unlearning target. NML first scores a reference set of nonmembers with the original model. It then trains the model so that the distribution of negative log-likelihoods (NLLs) on forget examples matches this reference distribution under energy distance, while a distillation term on retained data preserves utility. Because this target specifies where forget likelihoods should settle rather than only a direction, they can be raised or lowered toward the nonmember range instead of being suppressed without limit. Matching the full distribution rather than its mean further lets each forget example move according to its position within that range. The reference can be taken from held-out benchmark data or, when none is available, constructed to follow the task's record format. In Figure 1, NML keeps the Min-K% AUC close to 0.5 in late training.

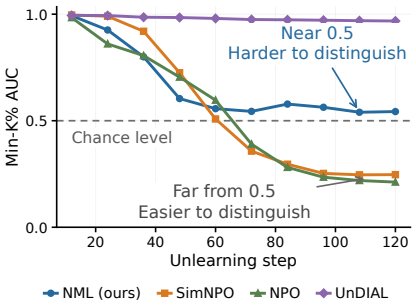


Figure 1: Min-K% AUC between forget and nonmember examples during unlearning on TOFU (Llama-3.2-1B-Instruct).

Our contributions are twofold. First, we propose NML, an unlearning method that gives forget examples an observable target, the original model's NLL distribution on nonmembers, and matches this distribution rather than only its mean, without requiring a retrained model. Second, we show that NML substantially reduces unlearning signatures under multiple likelihood-based attacks while maintaining forgetting and retained utility comparable to or better than strong baselines. Ablations attribute most of this reduction to the nonmember target and improved forgetting to distribution matching, and question-only identification attacks provide further evidence. Section 2 describes NML, Section 3 evaluates it on TOFU and MUSE-News, Section 4 examines constructed references, Section 5 presents ablations and cost, and Section 6 discusses limitations.

2 Nonmember-Law Matching

This section formalizes NML. Most existing objectives combine a forgetting term, which moves the model away from forget data, with a retain term, which preserves behavior on retained data. NML keeps the retain term but replaces the forgetting term with a target derived from nonmembers: the distribution of the original model's likelihood scores on held-out records of the same task. Section 2.1 introduces notation and this reference, Section 2.2 defines the matching objective, and Section 2.3 gives the training procedure.

2.1 Setup and Nonmember Reference

Let D_F and D_R denote the forget and retain sets, and let p_{θ_0} be the original model trained on their union, which we call the Full model. Unlearning should reduce the influence of D_F while preserving useful behavior on D_R. A Retrain model trained without D_F serves as a counterfactual for evaluation but is not available to the unlearning algorithm. NML keeps θ_0 frozen and learns LoRA parameters ϕ (Hu et al., 2022), giving the unlearned model p_{θ_ϕ}.

The reference set D_N consists of nonmembers: records excluded from task-specific training that share the task and record format of D_F. The Full model's scores on D_N describe how the original model treats such records, and NML uses their distribution as the target for forget-example scores. The target is defined at the population level: forget and reference records are not paired and need not share content, and no single confidence is prescribed for every forget example.

The reference serves as an observable proxy for how the model would treat the forget set had it never been trained on it. Its quality depends on how comparable the records and scoring units are, which Section 4 examines. Reference scores are computed once under the Full model and held fixed; the retain term keeps the model close to the Full model on retained data, and our evaluation recomputes nonmember scores under the unlearned model. Nonmembership is defined relative to task-specific training rather than the pretraining corpus.

By default, NML uses the held-out nonmembers supplied by the benchmark, such as the holdout sets of TOFU and MUSE. When no such set is available, a reference can instead be constructed by generating records that follow the task's record format. Appendix A describes the construction protocol, and Section 4 evaluates constructed references.

2.2 Scoring and Distribution Matching

For a context x and a target sequence y of L scored tokens, the score of model p_θ is the mean target-token negative log-likelihood

$$s_\theta(x, y) = -\frac{1}{L}\sum_{t=1}^{L} \log p_\theta(y_t \mid x, y_{<t}). \tag{1}$$

Higher scores mean lower likelihood. For QA, x is the formatted question and only answer tokens are scored; for documents, each record is a fixed-length chunk scored without a prompt (Section 3). Forget and reference records use the same tokenizer, formatting, and scoring unit.

Let $a_i(\phi) = s_{\theta_\phi}(x_i^F, y_i^F)$ and $r_j = s_{\theta_0}(x_j^N, y_j^N)$. For m forget records and n reference records, the empirical score laws are

$$\widehat{P}_F^\phi = \frac{1}{m}\sum_{i=1}^{m} \delta_{a_i(\phi)}, \qquad \widehat{P}_N = \frac{1}{n}\sum_{j=1}^{n} \delta_{r_j}, \tag{2}$$

where δ_u is a unit point mass at u. We call $\widehat{P}_N$ the nonmember law; NML matches $\widehat{P}_F^\phi$ to it, hence its name. Each example contributes one scalar, so these laws describe variation across examples rather than token-level vocabulary distributions. Reference scores are cached once, whereas forget scores are recomputed under the current model and receive gradients.

NML compares the two laws with the one-dimensional energy distance (Székely & Rizzo, 2013). For finite score lists U and V, let $d(U, V) = (|U||V|)^{-1}\sum_{u\in U}\sum_{v\in V}|u - v|$. For a list S of forget scores and the reference list $S_N = \{r_j\}_{j=1}^n$, the alignment loss is

$$\mathcal{L}_F(S, S_N) = 2d(S, S_N) - d(S, S) - d(S_N, S_N). \tag{3}$$

This loss equals $2\int \left(\widehat{F}_S(u) - \widehat{F}_N(u)\right)^2 du$, where $\widehat{F}_S$ and $\widehat{F}_N$ are the empirical CDFs of S and S_N. It is therefore nonnegative, vanishes only when the two empirical laws coincide, and compares entire distributions rather than their means. Its gradient makes the update direction explicit. With $m = |S|$ and ρ_i the rank of a_i within S, and ignoring ties,

$$\frac{\partial \mathcal{L}_F}{\partial a_i} = \frac{4}{m}\left[\widehat{F}_N(a_i) - \frac{\rho_i - 1/2}{m}\right]. \tag{4}$$

Algorithm 1: Nonmember-Law Matching

Require: θ_0, D_F, D_R, D_N; λ_F, λ_R; B, b, T
$S_N \leftarrow \{s_{\theta_0}(x, y) : (x, y) \in D_N\}$
Initialize LoRA ϕ; freeze θ_0 and S_N
for $t = 1, \ldots, T$ **do**
  $g \leftarrow 0$
  for $j = 1, \ldots, B/b$ **do**
    Sample microbatches F_j, R_j
    $S_j \leftarrow \{s_{\theta_\phi}(x, y) : (x, y) \in F_j\}$
    $\ell_j \leftarrow \lambda_F \mathcal{L}_F(S_j, S_N) + \lambda_R \mathcal{L}_R(R_j)$
    $g \leftarrow g + (b/B)\nabla_\phi \ell_j$
  end for
  $\phi \leftarrow \mathrm{AdamW}(\phi, \mathrm{clip}(g))$
end for
return p_{θ_ϕ}

A forget score is lowered when it lies above the reference quantile that matches its rank and raised when it lies below. Unlike one-sided suppression, the update therefore moves each forget example toward the nonmember law from either side. The loss acts on sequence-level mean scores and does not directly constrain token-level statistics, so it does not by itself guarantee indistinguishability under every attack; Section 3 evaluates both sequence-level and token-level attacks.

2.3 Retain Preservation and Training

Calibrating forget scores alone does not constrain predictions on retained data. NML therefore distills the frozen Full model on retain contexts. For a retain microbatch R with scored positions $\mathcal{T}(R)$,

$$\mathcal{L}_R(R) = \frac{1}{|\mathcal{T}(R)|}\sum_{h\in\mathcal{T}(R)} \mathrm{KL}\big(p_{\theta_0}(\cdot \mid h) \,\|\, p_{\theta_\phi}(\cdot \mid h)\big), \tag{5}$$

where h is the observed prefix at a scored position. This term preserves the Full model's entire next-token distribution on retain data, not only the likelihood of the observed token.

Training proceeds in microbatches. For microbatch j, NML computes the scores S_j of b forget examples under the current model, samples a retain microbatch R_j, and evaluates

$$\ell_j = \lambda_F \mathcal{L}_F(S_j, S_N) + \lambda_R \mathcal{L}_R(R_j), \tag{6}$$

using the full cached reference S_N in every microbatch. Gradients are averaged over B/b microbatches before each AdamW update of ϕ, where B is the effective batch size (Algorithm 1). Each alignment term thus compares b forget scores with the full reference; gradient accumulation averages these terms rather than forming one law over B examples. We compute $d(S_j, S_j)$ over all b^2 pairs, so its expectation is $(b-1)/b$ times the population self-distance; with small b, the objective places more weight on matching the central quantiles of the reference than its tails. Because reference scores are cached and the alignment term involves only bn scalar distances per microbatch, NML adds little computation beyond the forward passes on forget and retain data (Section 5.3).

3 Experiments

We evaluate whether NML suppresses unlearning signatures while maintaining forgetting and utility, on factual question answering and document unlearning and at two model scales (Section 3.2). We then test identification when the attacker does not know the answers (Section 3.3). Section 4 replaces the held-out reference with constructed ones, and Section 5 examines the components of the objective, reference size, and computational cost.

3.1 Experimental Setting

Benchmarks and models. TOFU Forget10 and MUSE-News cover factual question answering and document unlearning, respectively (Maini et al., 2024; Shi et al., 2025). TOFU contains 400 forget and 3600 retain QA pairs, and its Holdout10 split provides 400 nonmember QA pairs. On TOFU, we use Llama-3.2-1B-Instruct (Meta, 2024) and Llama-2-7B-Chat (Touvron et al., 2023), referred to below as the 1B and 7B models. On MUSE-News, we use the released Llama-2-7B Full and Retrain checkpoints. MUSE records within each split or reference fold are concatenated and divided into 2048-token chunks; after re-encoding, the special-token prefix is masked and up to 2048 following tokens are scored without an additional prompt.

Nonmember folds. Held-out nonmembers are split into two disjoint folds, by author on TOFU and by source record before chunking on MUSE. NML uses one fold as its reference and attacks use the other as negatives, so no record serves as both training target and attack negative; attack gaps are averaged over both fold assignments within each seed. Unless stated otherwise, NML uses these held-out references, so the main comparisons evaluate the objective itself; Section 4 replaces the reference while keeping NML unchanged.

Metrics. Metrics follow the OpenUnlearning implementations (Dorna et al., 2025). On TOFU, truth-ratio KS-D measures how far the unlearned model is from Retrain on forget examples, and model utility (MU) measures retained capability. On MUSE, KM_F and KM_R measure knowledge memorization on forget and retain questions. Signatures are measured with four likelihood-based attacks: loss and zlib use sequence-level likelihood, Min-K% averages the lowest token log-likelihoods, and Min-K%++ normalizes token log-probabilities by the conditional vocabulary distribution (Yeom et al., 2018; Carlini et al., 2021; Shi et al., 2024; Zhang et al., 2025). The token-level attacks use $k = 40\%$ on TOFU and $k = 20\%$ on MUSE. For each attack, we report the attack gap $|\mathrm{AUC} - 0.5|$ between forget examples and held-out nonmembers under the unlearned model; smaller gaps mean less separability, and 0 corresponds to chance. Retrain's gaps are not zero, indicating that forget and held-out records remain distinguishable even for a model never trained on the forget set. A gap below Retrain's therefore indicates lower separability against this attacker rather than closer agreement with Retrain, which KS-D and KM_F measure separately.

Baselines. We compare with the established baselines NPO and RMU (Zhang et al., 2024; Li et al., 2024), the recent baselines SimNPO and LUNAR (Fan et al., 2025; Shen et al., 2025), and

UnDIAL, which stabilizes likelihood-based unlearning through adjusted-logit self-distillation (Dong et al., 2025). Method-specific hyperparameters follow the respective papers. NPO, SimNPO, and UnDIAL share NML's LoRA parameterization and update budget, whereas RMU and LUNAR keep their native update schemes. Full and Retrain are reported as pre-unlearning and counterfactual reference points; Retrain is used only for evaluation.

Training. Shared-LoRA methods use rank 8, scaling 16, AdamW with learning rate 3×10^{-4}, effective batch size 32, and 120 updates. Sequence limits are 512 tokens on TOFU and 2048 on MUSE. NML uses $\lambda_F = 1$ and $\lambda_R = 3$, with microbatch size $b = 4$ for the 1B model and $b = 2$ for the 7B settings. All shared-LoRA methods are evaluated at the final update. We report means and sample standard deviations over three seeds, or five for the shared-LoRA methods and RMU with the 1B model; Full and Retrain are single checkpoints. Analyses in Sections 4 and 5 use three seeds, so their NML results differ slightly from the five-seed results in Table 1.

3.2 Main Results

Table 1: Forgetting, utility, and attack gaps on TOFU. Attack columns report $|\mathrm{AUC} - 0.5|$ between forget examples and held-out nonmembers. Bold marks the best unlearning method in each column.

Method	KS-D ↓	MU ↑	Loss ↓	zlib ↓	Min-K% ↓	Min-K%++ ↓
Llama-3.2-1B-Instruct						
Full	0.350	0.621	0.496	0.498	0.497	0.498
Retrain	0.000	0.614	0.112	0.091	0.117	0.023
NPO	$0.151_{\pm 0.015}$	$0.595_{\pm 0.003}$	$0.357_{\pm 0.032}$	$0.326_{\pm 0.027}$	$0.321_{\pm 0.048}$	$0.191_{\pm 0.025}$
SimNPO	$0.076_{\pm 0.014}$	$0.599_{\pm 0.001}$	$0.237_{\pm 0.042}$	$0.281_{\pm 0.034}$	$0.252_{\pm 0.040}$	$0.195_{\pm 0.028}$
UnDIAL	$0.310_{\pm 0.006}$	$\mathbf{0.600_{\pm 0.002}}$	$0.380_{\pm 0.005}$	$0.267_{\pm 0.008}$	$0.468_{\pm 0.001}$	$0.489_{\pm 0.000}$
RMU	$0.217_{\pm 0.001}$	$0.593_{\pm 0.002}$	$0.178_{\pm 0.028}$	$0.246_{\pm 0.029}$	$0.162_{\pm 0.028}$	$0.177_{\pm 0.011}$
LUNAR	$0.088_{\pm 0.013}$	$0.302_{\pm 0.007}$	$0.076_{\pm 0.011}$	$0.075_{\pm 0.011}$	$0.083_{\pm 0.013}$	$0.209_{\pm 0.024}$
NML	$\mathbf{0.061_{\pm 0.013}}$	$\mathbf{0.600_{\pm 0.002}}$	$\mathbf{0.053_{\pm 0.012}}$	$\mathbf{0.054_{\pm 0.011}}$	$\mathbf{0.044_{\pm 0.007}}$	$\mathbf{0.109_{\pm 0.052}}$
Llama-2-7B-Chat						
Full	0.367	0.629	0.499	0.500	0.499	0.499
Retrain	0.000	0.612	0.107	0.182	0.112	0.005
NPO	$0.188_{\pm 0.001}$	$0.627_{\pm 0.004}$	$0.365_{\pm 0.103}$	$0.342_{\pm 0.088}$	$0.357_{\pm 0.100}$	$0.312_{\pm 0.061}$
SimNPO	$0.168_{\pm 0.029}$	$0.631_{\pm 0.003}$	$0.410_{\pm 0.033}$	$0.421_{\pm 0.035}$	$0.409_{\pm 0.038}$	$0.292_{\pm 0.029}$
UnDIAL	$0.335_{\pm 0.003}$	$\mathbf{0.635_{\pm 0.001}}$	$0.379_{\pm 0.005}$	$0.265_{\pm 0.009}$	$0.459_{\pm 0.002}$	$0.487_{\pm 0.000}$
RMU	$0.213_{\pm 0.063}$	$0.525_{\pm 0.111}$	$0.283_{\pm 0.235}$	$0.282_{\pm 0.217}$	$0.281_{\pm 0.239}$	$0.364_{\pm 0.059}$
LUNAR	$\mathbf{0.131_{\pm 0.016}}$	$0.298_{\pm 0.031}$	$0.120_{\pm 0.008}$	$0.073_{\pm 0.008}$	$0.245_{\pm 0.008}$	$0.369_{\pm 0.001}$
NML	$0.138_{\pm 0.034}$	$0.631_{\pm 0.004}$	$\mathbf{0.049_{\pm 0.015}}$	$\mathbf{0.072_{\pm 0.010}}$	$\mathbf{0.236_{\pm 0.108}}$	$\mathbf{0.279_{\pm 0.080}}$

Table 2: Knowledge memorization and attack gaps on MUSE-News. Attack columns report $|\mathrm{AUC} - 0.5|$. Bold marks the best unlearning method in each column.

Method	KM_F ↓	KM_R ↑	Loss ↓	zlib ↓	Min-K% ↓	Min-K%++ ↓
Full	0.664	0.546	0.499	0.500	0.500	0.485
Retrain	0.327	0.551	0.052	0.065	0.005	0.079
NPO	$0.649_{\pm 0.014}$	$0.509_{\pm 0.008}$	$0.498_{\pm 0.001}$	$0.496_{\pm 0.002}$	$0.498_{\pm 0.001}$	$0.306_{\pm 0.007}$
SimNPO	$0.634_{\pm 0.004}$	$0.512_{\pm 0.010}$	$0.455_{\pm 0.011}$	$0.457_{\pm 0.010}$	$0.340_{\pm 0.030}$	$0.337_{\pm 0.011}$
UnDIAL	$0.594_{\pm 0.008}$	$0.507_{\pm 0.004}$	$0.498_{\pm 0.000}$	$0.496_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.000}$
RMU	$0.488_{\pm 0.010}$	$0.501_{\pm 0.004}$	$0.270_{\pm 0.116}$	$0.271_{\pm 0.117}$	$0.294_{\pm 0.116}$	$0.404_{\pm 0.050}$
LUNAR	$\mathbf{0.416_{\pm 0.005}}$	$0.277_{\pm 0.015}$	$0.142_{\pm 0.012}$	$0.154_{\pm 0.012}$	$0.355_{\pm 0.007}$	$0.324_{\pm 0.014}$
NML	$0.422_{\pm 0.016}$	$\mathbf{0.537_{\pm 0.010}}$	$\mathbf{0.123_{\pm 0.022}}$	$\mathbf{0.131_{\pm 0.020}}$	$\mathbf{0.283_{\pm 0.012}}$	$\mathbf{0.296_{\pm 0.001}}$

TOFU. Table 1 shows that NML substantially reduces attack gaps while keeping forgetting and utility comparable to the strongest baselines. With the 1B model, NML reaches KS-D of 0.061 and MU of 0.600, comparable to SimNPO (0.076 and 0.599). Its attack gaps are much lower: relative to SimNPO, the loss and zlib gaps fall from 0.237 and 0.281 to 0.053 and 0.054, and the Min-K% gap

Table 3: Question-only FUMA identification (Recall@1) on TOFU with the 1B model. Chance is 0.20 with 5 candidates and 0.05 with 20; values near chance mean the forget question is not singled out. Bold marks the lowest rate among unlearning methods.

	5 candidates		20 candidates	
Method	Loss-base	Loss-ens	Loss-base	Loss-ens
Full	0.129	0.047	0.011	0.008
Retrain	0.251	0.223	0.040	0.049
NPO	$0.474_{\pm 0.014}$	$0.550_{\pm 0.098}$	$0.174_{\pm 0.006}$	$0.221_{\pm 0.027}$
SimNPO	$0.376_{\pm 0.023}$	$0.444_{\pm 0.045}$	$0.113_{\pm 0.008}$	$0.257_{\pm 0.038}$
UnDIAL	$0.285_{\pm 0.004}$	$0.298_{\pm 0.006}$	$0.073_{\pm 0.002}$	$0.089_{\pm 0.003}$
RMU	$0.293_{\pm 0.003}$	$0.285_{\pm 0.007}$	$0.092_{\pm 0.002}$	$0.106_{\pm 0.001}$
LUNAR	$0.357_{\pm 0.008}$	$0.434_{\pm 0.011}$	$0.130_{\pm 0.007}$	$0.237_{\pm 0.018}$
NML	$\mathbf{0.278_{\pm 0.009}}$	$\mathbf{0.273_{\pm 0.050}}$	$\mathbf{0.053_{\pm 0.013}}$	$\mathbf{0.083_{\pm 0.017}}$

falls by 83%, from 0.252 to 0.044. NML has the lowest mean gap under all four attacks; LUNAR comes closest on loss and zlib (0.076 and 0.075), but with MU of 0.302, about half that of the other methods. With the 7B model, NML matches SimNPO's MU (0.631) with comparable KS-D (0.138 versus 0.168) and reduces the loss and zlib gaps by 88% and 83% relative to SimNPO. Its loss gap is the lowest by a wide margin, and its gaps under the other three attacks are the lowest or comparable to the best baseline. LUNAR's slightly lower KS-D (0.131) again comes with MU of 0.298. NPO, SimNPO, RMU, and UnDIAL leave substantially higher loss and zlib gaps at both scales. For NPO and SimNPO with the 1B model, Figure 1 shows that the Min-K% gap arises from overshooting, with AUC falling well below 0.5.

MUSE-News. On document unlearning (Table 2), NML again achieves the lowest loss and zlib gaps while preserving utility. Relative to SimNPO, it lowers KM_F from 0.634 to 0.422 while raising KM_R from 0.512 to 0.537, the highest among unlearning methods, and it reduces the loss and zlib gaps by 73% and 71%, to 0.123 and 0.131. LUNAR forgets slightly more (KM_F 0.416) but retains much less (KM_R 0.277). The Min-K% and Min-K%++ gaps also fall relative to SimNPO, from 0.340 and 0.337 to 0.283 and 0.296, comparable to the best baselines.

Summary. Across both benchmarks and both TOFU scales, NML generally achieves the lowest attack gaps while keeping forgetting and utility comparable to the strongest baselines; LUNAR, the only baseline with comparable loss and zlib gaps in some settings, retains roughly half the utility. The largest reductions occur for loss and zlib, which share the sequence-level statistic that NML matches. Token-level attacks also improve, strongly with the 1B model and more modestly with the 7B model and on MUSE, although the objective does not constrain them directly. The remaining token-level gaps, which exceed Retrain's in these two settings, indicate that suppression under such attacks is not complete.

3.3 Question-Only Candidate Identification

The attacks above score answers. FUMA (Deepak et al., 2025) instead asks whether unlearning targets can be singled out from their questions alone, so the attacker need not know the answers. Because NML does not score question tokens, this setting offers a complementary test of its signature suppression.

We use FUMA's two loss-based attacks with the 1B model on TOFU. Each trial contains one forget question and 4 or 19 distractors from the held-out fold used as attack negatives. Loss-base ranks candidates by length-normalized question loss; loss-ens, FUMA's ensembling variant, uses the minimum loss over five sampled continuations of at most 15 tokens. Recall@1 is the fraction of trials in which the forget question has the highest loss. We average over all 400 forget questions, five candidate samplings, and both fold assignments within each of three seeds.

NML has the lowest mean Recall@1 in all four settings (Table 3), with clear margins over NPO, SimNPO, and LUNAR. With 20 candidates, loss-ens falls from 0.257 for SimNPO to 0.083, and loss-base reaches 0.053, near the 0.05 chance level. Because these attacks flag high loss, a model

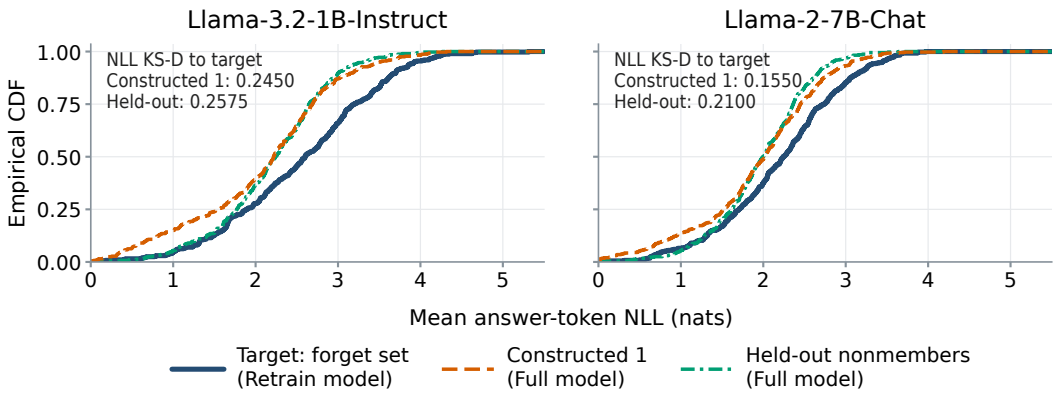


Figure 2: Empirical CDFs of mean answer-token NLL on TOFU. A suitable reference, scored by the Full model, should approximate the target: Retrain-model NLLs on the forget set. Values give the NLL KS-D to this target.

that forgets little also evades them, as the Full model shows; UnDIAL and RMU come closest but forget less (KS-D 0.310 and 0.217 versus 0.061 in Table 1). These results provide further evidence of reduced signatures, although NML remains above Retrain in all four settings.

4 Constructed References

NML requires a nonmember reference but not necessarily a held-out one. This section evaluates references constructed with the protocol in Appendix A on TOFU: whether they approximate the target (Section 4.1) and whether NML remains effective when trained with them (Section 4.2). Appendix B reports the transfer to news documents.

4.1 Target Similarity

A suitable reference should approximate how a model that never saw the forget set scores it. We therefore compare the Full model's NLLs on a constructed reference with the Retrain model's NLLs on forget records, using the Kolmogorov–Smirnov distance between the two empirical distributions. We call this NLL KS-D to distinguish it from the truth-ratio KS-D in the tables. The Full model's NLLs on held-out nonmembers serve as a control. We evaluate Constructed 1, containing twenty fictitious profiles and 400 QA pairs, with both the 1B and 7B models. As Figure 2 shows, the constructed reference is comparable to the held-out set with the 1B model (NLL KS-D 0.2450 versus 0.2575) and closer with the 7B model (0.1550 versus 0.2100). Under this metric, Constructed 1 approximates the counterfactual target at least as well as the held-out data.

4.2 Unlearning with Constructed References on TOFU

Table 4: NML with held-out and constructed references on TOFU with the 1B model over three seeds. Attack columns report $|\mathrm{AUC} - 0.5|$.

Reference	KS-D ↓	MU ↑	Loss ↓	zlib ↓	Min-K% ↓
Held-out	$0.063_{\pm 0.017}$	$0.600_{\pm 0.000}$	$0.052_{\pm 0.016}$	$0.056_{\pm 0.015}$	$0.044_{\pm 0.009}$
Constructed 1	$0.086_{\pm 0.012}$	$0.600_{\pm 0.002}$	$0.077_{\pm 0.022}$	$0.026_{\pm 0.008}$	$0.069_{\pm 0.016}$
Constructed 2	$0.083_{\pm 0.022}$	$0.600_{\pm 0.001}$	$0.095_{\pm 0.027}$	$0.019_{\pm 0.004}$	$0.087_{\pm 0.019}$
Constructed 3	$0.092_{\pm 0.012}$	$0.601_{\pm 0.001}$	$0.110_{\pm 0.015}$	$0.017_{\pm 0.007}$	$0.102_{\pm 0.011}$

We train NML with three independently constructed references, keeping the objective, optimizer, and forget and retain data unchanged. Each reference contains 400 QA pairs, split into two author-disjoint folds of 200. We train with each fold, draw attack negatives from the opposite held-out fold, and average over folds within each of three seeds; the held-out row follows the same protocol.

Table 5: Component ablations on TOFU with the 1B model over three seeds. Attack columns report $|\mathrm{AUC} - 0.5|$.

Objective	KS-D ↓	MU ↑	Loss ↓	zlib ↓	Min-K% ↓	Min-K%++ ↓
NML	$0.063_{\pm 0.017}$	$0.600_{\pm 0.000}$	$0.052_{\pm 0.016}$	$0.056_{\pm 0.015}$	$0.044_{\pm 0.009}$	$0.114_{\pm 0.072}$
Mean matching	$0.112_{\pm 0.026}$	$0.598_{\pm 0.003}$	$0.056_{\pm 0.028}$	$0.055_{\pm 0.029}$	$0.050_{\pm 0.025}$	$0.106_{\pm 0.019}$
Without retain term	$0.145_{\pm 0.009}$	$0.317_{\pm 0.007}$	$0.075_{\pm 0.015}$	$0.053_{\pm 0.005}$	$0.146_{\pm 0.013}$	$0.174_{\pm 0.026}$

Constructed references preserve utility (MU 0.600–0.601) and most of the signature suppression (Table 4). Their loss and Min-K% gaps (0.077–0.110 and 0.069–0.102) remain far below SimNPO's 0.237 and 0.252 in Table 1, and their zlib gaps (0.017–0.026) are lower than with the held-out reference. The held-out reference still gives the lowest loss and Min-K% gaps and KS-D, while constructed references reach KS-D of 0.083–0.092, comparable to SimNPO's 0.076. NML therefore remains effective on TOFU without held-out nonmember data.

5 Ablations and Sensitivity Analysis

We isolate the two components of NML, vary the size of the reference, and measure computational cost.

5.1 Component Ablations

With the 1B model on TOFU, we compare NML with two variants over three seeds, keeping the held-out reference, LoRA parameterization, and training budget fixed. Mean matching replaces the energy loss with the absolute difference between the mean forget and reference NLLs, and therefore moves all forget scores in a microbatch in the same direction. The second variant removes the retain term.

Matching the full distribution mainly improves forgetting (Table 5). KS-D falls from 0.112 with mean matching to 0.063 with NML at similar MU (0.600 versus 0.598), while the attack gaps of the two variants are comparable. Both variants move forget scores toward the same nonmember target and reach attack gaps well below those of the baselines in Table 1, indicating that the nonmember target itself accounts for most of the reduction in attack gaps; matching the distribution rather than its mean additionally brings forget examples closer to Retrain.

The retain term is essential for utility. Removing it lowers MU from 0.600 to 0.317 and also worsens forgetting and token-level separability: KS-D rises from 0.063 to 0.145, and the Min-K% gap from 0.044 to 0.146.

5.2 Reference Size

We vary the size of Constructed 2 across 20, 40, 80, 100, 160, 200, 300, and 400 QA pairs, using nested subsets of whole authors, with the 1B model on TOFU and a single training seed. Across this twentyfold range, MU stays within 0.596–0.603 and KS-D within 0.048–0.090. The loss and Min-K% gaps range over 0.078–0.187 and 0.068–0.175 without a consistent trend; with a single seed, we do not interpret differences between individual sizes. At every size, including a single author with 20 QA pairs, both gaps remain below SimNPO's (0.237 and 0.252 in Table 1).

5.3 Computational Cost

Under the matched TOFU setup with the 1B model on one H100 NVL, NML trains in 3.84–3.85 minutes over three repetitions, versus 3.80–3.84 minutes for SimNPO, with the same peak allocated memory range of 6.73–6.88 GiB; including setup and saving, a run takes 4.17–6.13 minutes. Scoring the reference requires one forward pass per record and is done once. Constructing a reference, when needed, is separate and reusable preprocessing: 26.1–42.3 minutes for 400 QA pairs and 7.50–7.84 hours for 512 news records.

6 LIMITATIONS

Constructed references are less effective for long-form news than for entity-centered QA (Appendix B). Although they match held-out data at the document level, they diverge at the chunk level used for training, and the MUSE main results therefore use held-out references; improving construction at the training scoring unit remains an open direction. NML matches sequence-level scores and does not directly constrain token-level statistics, and the residual Min-K% and Min-K%++ gaps with the 7B model and on MUSE show that suppression under such attacks is incomplete. Finally, our evaluation covers entity-centered QA and news, but not relational unlearning, where a deleted fact may remain inferable from retained relationships; whether distributional calibration addresses such dependencies requires dedicated benchmarks.

7 RELATED WORK

LLM unlearning. NPO and SimNPO cast unlearning as negative preference optimization (Zhang et al., 2024; Fan et al., 2025), UnDIAL stabilizes it through self-distillation on adjusted logits (Dong et al., 2025), and RMU and LUNAR redirect internal representations (Li et al., 2024; Shen et al., 2025). These objectives are defined without reference to nonmember behavior; NML instead uses the original model's likelihood distribution on nonmembers as the target for forget examples. TOFU, MUSE, and OpenUnlearning provide the benchmarks and evaluation framework used here (Maini et al., 2024; Shi et al., 2025; Dorna et al., 2025).

Membership inference and unlearning audits. Likelihood-based MIAs detect training data through loss, compression-normalized loss, and token-level statistics (Yeom et al., 2018; Carlini et al., 2021; Shi et al., 2024; Zhang et al., 2025), and document-level inference depends on the unit of observation (Meeus et al., 2024). Applied to unlearning, such attacks reveal residual traces of forget data (Shi et al., 2025; Du et al., 2025), identify which examples were unlearned (Deepak et al., 2025), and measure sample-level unlearning completeness, including both insufficient and excessive removal (Wang et al., 2025). These works evaluate unlearning; NML turns the nonmember comparison they rely on into a training target.

Membership-guided unlearning. Membership signals have also been used to guide unlearning in classification models: Sula et al. (2025) combine the classification loss with the loss of a membership approximation function to remove privacy-sensitive information from the target model. NML instead targets LLMs and matches an observable nonmember score distribution, without training or approximating a membership model.

8 CONCLUSION

Unlearning objectives move the model away from forget data but do not specify where forget likelihoods should settle, and both stopping short and overshooting leave signatures. We introduced Nonmember-Law Matching, which supplies this target without retraining: the original model's likelihood distribution on nonmembers. On TOFU and MUSE-News, NML substantially reduces unlearning signatures under likelihood-based attacks while maintaining forgetting and utility comparable to or better than strong baselines, and it lowers question-only identification relative to the evaluated baselines. Ablations indicate that the nonmember target accounts for most of this reduction and that matching its full distribution improves forgetting. Constructed references retain most of these benefits for entity-centered QA when held-out nonmembers are unavailable. These results suggest that calibrating against nonmember behavior is a practical route to unlearning that leaves fewer identifiable traces.

Reproducibility Statement

Algorithm 1 specifies the update, Section 2 defines the objective, Appendix A describes reference construction, and Section 3 lists training and evaluation settings. The result archive accompanying this submission retains per-seed results, construction-bank identities, aggregation scripts, and source-file hashes. All reported tables are generated from these archived results.

Ethics Statement

The experiments use fictitious-person QA and released news benchmark data. Candidate-identification attacks are evaluated within those benchmarks to measure privacy risks of unlearning. Constructed news references describe fictional events and should not be presented as factual reports. Lower measured attack separability does not replace a formal privacy guarantee or authorize release of sensitive training data.

AI Use Statement

We used generative AI tools to assist with mathematical derivations, manuscript editing, and language polishing. A frozen Qwen3-8B model generated the synthetic reference records, as described in Appendix A. The authors reviewed all AI-assisted content and take responsibility for the final submission and its claims.

References

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*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

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*, pp. 10873–10892. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.551. URL https://aclanthology.org/2025.emnlp-main.551/.

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*, pp. 8827–8840, 2025. doi: 10.18653/v1/2025.naacl-long.444. URL https://aclanthology.org/2025.naacl-long.444/.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. URL https://arxiv.org/abs/2506.12618.

Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14579–14597, 2025. URL https://proceedings.mlr.press/v267/du25d.html.

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*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02443e4e008231e0af0855f3cc70ed17-Abstract-Conference.html.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 28525–28550. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=q0eyIBnE2t.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the Neurons Read your Book? Document-level Membership Inference for Large Language Models. In *33rd USENIX Security Symposium*, pp. 2369–2385, 2024. URL https://www.usenix.org/conference/usenixsecurity24/presentation/meeus.

Meta. Llama 3.2 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

William Shen, Xinchi Qiu, Meghdad Kurmanji, Alexandru-Andrei Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3ed6a903be20e1478aa6793f14938091-Abstract-Conference.html.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16789.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Nexhi Sula, Abhinav Kumar, Han Wang, Jie Hou, and Reza Tourani. Silver linings in the shadows: Harnessing membership fingerprinting for machine unlearning. In *2025 IEEE Conference on Communications and Network Security (CNS)*, 2025. doi: 10.1109/CNS66487.2025.11195030.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *34th USENIX Security Symposium*, pp. 6481–6500. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/wang-cheng-long.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2404.02936.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.05868.

A Reference Construction Details

Generator. A frozen Qwen3-8B model (Yang et al., 2025), with thinking disabled, generates all constructed records before unlearning, using temperature 0.75 and top-p 0.9. Acceptance checks concern only record format and overlap; neither Retrain scores nor attack results are used to select a reference. The generator produces text only, and the Full model subsequently assigns the scores used by NML.

Entity-centered QA. Following TOFU's attribute-conditioned design (Maini et al., 2024), we sample birthplace, gender, literary genre, parental occupations, and book-title inspirations from fixed lists, together with a birth year between 1935 and 2001. Conditioned on this bundle, the generator invents an author, awards, and book titles, then writes a block of twenty questions and detailed answers about that author; generating the whole block from one profile is intended to keep its facts consistent. A reference contains twenty profiles and 400 QA pairs. TOFU forget and retain records are consulted only to detect name collisions, and Holdout10 is never read by the constructor. A profile is rejected if its name already occurs in the supplied data or in an accepted profile, or if it contains malformed, missing, or duplicate QA pairs. Every question includes the generated author's name, which is prepended if absent. Failed blocks are regenerated, and realized coverage of the sampled attributes is recorded rather than assumed. Generated answers belong to the new profiles and never replace answers in D_F.

News. For each forget-set source article, the constructor extracts its broad topic, event type, and scale, together with named entities used only for exclusion checks. A separate generation step receives only the coarse event labels and invents a new event record with participants, locations, dates, facts, quantities, and quotations. The article-writing step receives this event record, not the source article, and writes a report of comparable length and paragraph structure. The target length w is the source word count clipped to 220–850 words, and the requested paragraph count is clipped to 4–24. An article is accepted if it has at least three paragraphs and between $\max(150, \lfloor 0.55w \rfloor)$ and $\lfloor 1.55w \rfloor$ words, contains no specified meta-language or promotional phrases, and mentions none of the extracted source people, organizations, or products whose names have at least four characters. At most 2% of its word 5-grams and 0.5% of its word 8-grams may also occur in the source. These checks limit direct copying but do not establish statistical equivalence to held-out nonmembers.

B Constructed References for News Documents

On MUSE-News, we construct three references of 512 news records each, yielding 128 training chunks per reference, and compare them with a held-out control of the same number of chunks, each over three seeds. Constructed references reduce verbatim memorization of the forget text (MUSE's VerbMem) nearly as well as the control (0.235–0.242 versus 0.230), and their loss gaps of 0.263–0.331 remain well below those of NPO (0.498) and SimNPO (0.455) in Table 2, though above the control's 0.120. Their NLL distributions are consistent with this difference. At the document level,

constructed references achieve NLL KS-D of 0.156–0.172 against Retrain/forget, close to the held-out reference's 0.164. At the 2048-token chunks scored during training, however, their distance rises to 0.375–0.422, whereas the held-out reference remains at 0.141–0.148. Document-level similarity therefore does not ensure that a reference matches the target at the scoring unit used for training.