Reference-Specific Unlearning Metrics Can Hide the Truth: A Reality Check

Sungjun Cho [1] **Dasol Hwang** [2] **Frederic Sala** [1] **Sangheum Hwang** [3] **Kyunghyun Cho** [4 5] **Sungmin Cha** [4]

Abstract

Evaluating the effectiveness of unlearning in large language models (LLMs) remains a key challenge, especially as existing metrics often rely on specific reference outputs. The widely used *forget quality* metric from the TOFU benchmark (Maini et al., 2024) compares likelihoods over paraphrased answers but is highly sensitive to the choice of these references, potentially obscuring whether a model has truly forgotten the targeted information. We argue that unlearning should instead be assessed via distributional equivalence—how closely an unlearned model aligns functionally with the retain-only model. To this end, we propose **Functional Alignment for Distributional Equivalence (FADE)**, a novel distribution-level metric that measures probabilistic precision and recall between model outputs. FADE provides a more robust, principled approach to evaluating unlearning by comparing model behavior beyond isolated responses.

1. Introduction

As large language models (LLMs) are increasingly deployed in sensitive real-world scenarios, the ability to unlearn specific information—such as private or harmful content—without full retraining has become a critical goal. Accurately evaluating the effectiveness of unlearning, however, remains a challenge. Recently, TOFU (Maini et al., 2024) has emerged as a widely used benchmark, introducing the *forget quality* metric that compares likelihood distributions over answers between the unlearned model and a retain-only oracle trained without the data requested for deletion.

However, we find that the forget quality metric is highly sensitive to the choice of reference answers used. In particular, using paraphrased responses as proxies can completely obscure the model's ability to completely generate original answers, which significantly misleads assessment of unlearning efficacy. While paraphrasing is helpful to detect unlearning via memorization, it can shift the evaluation away from the core objective of unlearning as it leads to focusing on aligning the likelihoods on specific outputs.

[1]University of Wisconsin-Madison [2]LG AI Research [3]Seoul National University of Science and Technology [4]New York University [5]Genentech. Correspondence to: Sungmin Cha <*sungmin.cha@nyu.edu*>.

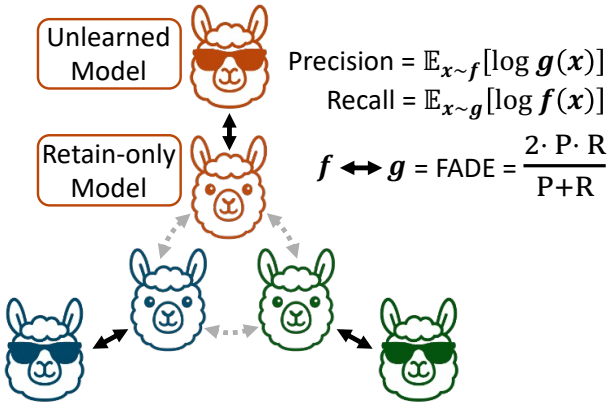
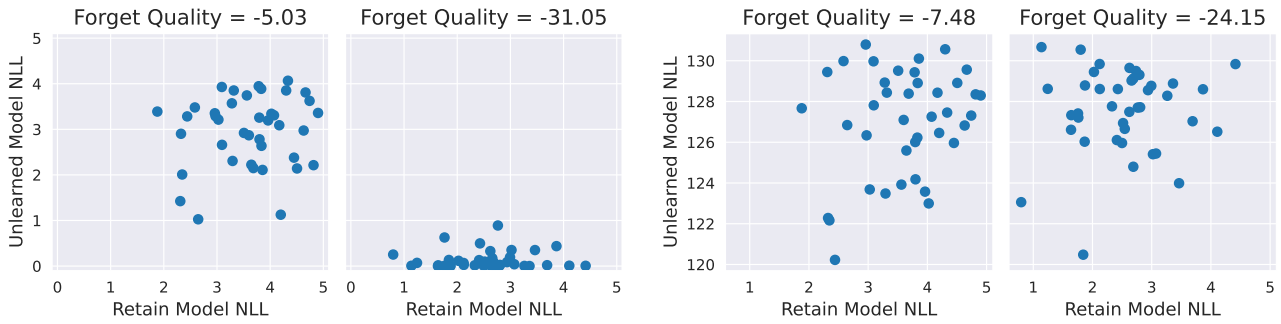


Figure 1. Illustration of our metric, Functional Alignment for Distributional Equivalence (FADE). We propose to measure the functional F1 score between the retain model and the unlearned model (black solid arrows). To quantify the distance caused by randomness in training, we also measure FADE across retain-only models with different random seeds (gray dashed arrows) as a baseline, with which we can use to analyze the functional distance due to the choice of unlearning vs. retraining from scratch in isolation.

Most importantly, ***unlearning should aim for functional equivalence with the retain-only model***. That is, the outputs of an unlearned model follow the same output distribution of the retain-only oracle across varying input spaces, including the forget set, the retain set, and out-of-domain prompts. Existing metrics based on static response sets often fail to capture this crucial goal.

To address this gap, we propose **Functional Alignment for Distributional Equivalence (FADE)** (Figure 1), a novel metric for evaluating unlearning at the distributional level. Instead of using specific answers, FADE measures functional similarity by generating samples from one model and computing their expected log-likelihoods under the other. This yields probabilistic notions of precision and recall (Cha & Cho, 2025), which together quantify how well the two models align as functions. FADE provides a way to robustly assess unlearning effectiveness based on distributional alignment rather than isolated outputs.



(a) Forget 1%

(b) Forget 10%

Figure 2. NLL distributions from the unlearned model (y-axis) and the retain-only model (x-axis). Each dot represents a single sample from $\mathcal{D}_{\text{forget}}$. Each plot shows results from using paraphrased answers (left) or original answers (right) for evaluation. Forget quality depends significantly on which reference answer is used, as the NLL distributions heavily depend on the answers.

Related work. A variety of evaluation methods have been proposed to assess unlearning efficacy. TOFU (Maini et al., 2024) introduces forget quality, which compares likelihoods over paraphrased responses between unlearned and retain-only models. RWKU (Jin et al., 2024) and WMDP (Li et al., 2024) probe for residual knowledge using paraphrased factual prompts and adversarial queries. Lynch et al. (2024) propose a cohort of token-level generation and paraphrasing-based approaches. Despite these advances, most methods require specifically chosen outputs, making it difficult to assess whether residual knowledge persists at the distributional level. In contrast, we propose a metric that compares output distributions to capture functional differences.

2. Preliminaries

2.1. Problem Setup

We formalize machine unlearning as a problem of functional alignment, following recent works (Cha et al.; Jang et al., 2023). Let $f : \mathcal{X} \to \mathcal{Y}$ be a model trained on the full dataset $\mathcal{D} = \mathcal{D}_{\text{retain}} \cup \mathcal{D}_{\text{forget}}$, where $\mathcal{D}_{\text{forget}}$ denotes the subset of data requested for removal. The goal of unlearning is to update f into f_{unlearn} that behaves as if it had never seen $\mathcal{D}_{\text{forget}}$ while maintaining performance on the retain data $\mathcal{D}_{\text{retain}}$. In other words, denoting f_{retain} as a model trained from scratch using only $\mathcal{D}_{\text{retain}}$, unlearning is considered successful if $f_{\text{unlearn}}(x) \approx f_{\text{retain}}(x), \forall x \in \mathcal{X}$. This perspective motivates a natural evaluation criterion: comparing the functional behavior of f_{unlearn} and f_{retain}.

2.2. How is unlearning efficacy measured in TOFU?

In TOFU (Maini et al., 2024), unlearning efficacy is evaluated by performing a Kolmogorov–Smirnov (KS) test on distributions of truth ratios, which measure the relative likelihood a model assigns to correct versus incorrect answers.

Given a LLM that parameterizes the conditional likelihood of answer a given question q, (i.e., $\Pr(a \mid q)$), the truth ratio for each question-answer pair $(q, a) \sim \mathcal{D}_{\text{forget}}$ is defined as

$$R_{\text{truth}}(q, a) = \frac{\frac{1}{|\mathcal{A}_{\text{pert}}|} \sum_{\hat{a} \in \mathcal{A}_{\text{pert}}} \Pr(\hat{a} \mid q)^{1/|\hat{a}|}}{\Pr(\tilde{a} \mid q)^{1/|\tilde{a}|}}.$$

where $\tilde{a}$ is a paraphrased version of a, $\hat{a} \in \mathcal{A}_{\text{pert}}$ are perturbed (incorrect) answers derived from $\tilde{a}$, and $|\tilde{a}|$ denotes the number of tokens in $\tilde{a}$.

To assess unlearning efficacy, the distribution of truth ratios computed over the forget set $\mathcal{D}_{\text{forget}}$ is compared between f_{unlearn} and f_{retain}. The KS-test is applied to these distributions, and the base-10 logarithm of the resulting p-value is referred to as the *forget quality*. A higher p-value (closer to 1) indicates greater similarity between the two distributions, suggesting stronger unlearning. Accordingly, a forget quality closer to 0 indicates stronger unlearning, while more negative values imply weaker unlearning.

2.3. Sensitivity of Forget Quality to Reference Outputs

Unfortunately, the forget quality metric suffers from a key drawback: it can vary significantly depending on which reference answer is used as $\tilde{a}$, potentially leading to misleading conclusions. To illustrate this issue, we unlearn 1% or 10% of the TOFU forget set from LLaMA3.1-8B using Gradient Ascent (Jang et al., 2023), and compare the negative log-likelihood (NLL) distributions assigned by f_{retain} and f_{unlearn}. We evaluate the forget qualities both on the paraphrased answers (as used in TOFU) and on the original ground truth answers (used for actual unlearning).

Results are shown in Figure 2. When unlearning 1%, we find that while the NLL distributions on paraphrased answers are similar between the two models, the original answers still receive high likelihood under f_{unlearn} with all points clustering near the x-axis. When computing forget quality with original answers instead of paraphrases, the metric drops drastically from –5.03 to –31.05, suggesting a more severe failure to unlearn than initially indicated. The drop in forget quality is also shown when unlearning 10%, showing that this behavior is not specific to small forget sets.

Table 1. Quantitative results from the TOFU benchmark. Forget Quality (FQ) and Model Utility (MU) are metrics originally used in TOFU. All FADE values are from comparing against the retain model with the same seed, averaged across 3 random seeds. The numbers in parentheses are from comparing across different random seeds, representing a baseline from stochasticity in initialization and training.

Method	TOFU-1%				TOFU-5%				TOFU-10%			
	Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.	
	FQ ↑	FADE ↓	MU ↑	FADE ↓	FQ ↑	FADE ↓	MU ↑	FADE ↓	FQ ↑	FADE ↓	MU ↑	FADE ↓
RETAIN -ONLY	0.00	1.90±0.90 (1.92±0.29)	0.64	0.94±0.27 (0.92±0.09)	0.00	1.20±0.42 (1.39±0.15)	0.64	0.94±0.27 (0.92±0.09)	0.00	1.10±0.40 (1.29±0.14)	0.64	0.94±0.27 (0.92±0.09)
BASE	-20.74	2.71±0.37	0.64	0.93±0.04	-20.74	2.48±0.24	0.64	0.93±0.04	-20.74	2.44±0.24	0.64	0.93±0.24
GA	-5.03	2.53±0.32	0.64	1.01±0.02	-4.36	1.68±1.14	0.00	1.57±1.05	-7.48	2.32±1.70	0.00	2.32±1.69
GD	-5.03	2.63±0.30	0.64	0.93±0.03	-16.76	2.97±0.34	0.53	2.10±0.31	-11.06	16.06±21.46	0.10	2.15±1.67
DPO	-4.59	2.68±0.51	0.64	0.93±0.05	-14.77	2.41±0.36	0.52	1.20±0.26	-18.61	2.62±0.38	0.60	1.13±0.18
NPO	-5.58	2.44±0.32	0.64	0.91±0.03	-3.85	2.49±0.96	0.43	2.11±0.76	-10.97	2.68±0.31	0.50	2.26±0.19

This inconsistency raises an important question: *which reference answers should we use, and how can we ensure that they truly reflect the model's ability to generalize the unlearning behavior?* Expanding the diversity of reference answers may help, but it remains inadequate, as unlearned content can resurface in countless linguistic forms (Lynch et al., 2024). Therefore, an accurate assessment of unlearning efficacy requires going beyond static answer sets and instead analyzing the model's functional behavior at a distributional level. This motivates our proposed approach, which measures unlearning effectiveness through direct comparison of model output distributions.

3. Method

Recall that the core objective of unlearning is to obtain a f_{unlearn} that is functionally equivalent to f_{retain}. Therefore, we propose a new metric, Functional Alignment for Distributional Equivalence (FADE), which quantifies the functional similarity between these two models by comparing their output distributions.

3.1. Functional Alignment for Distributional Equivalence

In essence, FADE measures how closely the conditional distributions $f_{\text{unlearn}}(\cdot \mid q)$ and $f_{\text{retain}}(\cdot \mid q)$ align, given the same input prompt q. Instead of relying on specific reference answers, FADE operates at the distributional level by computing probabilistic analogs of precision and recall:

$$\begin{aligned} \text{Precision} &= \mathbb{E}_{a\sim f_{\text{unlearn}}(\cdot|q)}\left[-\log f_{\text{retain}}(a \mid q)\right] \\ \text{Recall} &= \mathbb{E}_{a\sim f_{\text{retain}}(\cdot|q)}\left[-\log f_{\text{unlearn}}(a \mid q)\right] \end{aligned} \tag{1}$$

where $f_*(a \mid q)$ denotes the likelihood of answer a given question q according to model f_*.

These scores measure the extent to which one model's output distribution is supported by the other: precision measures how likely the outputs of f_{unlearn} are under f_{retain}, while recall captures the converse. Then, we define FADE as the harmonic mean of the two values (similar to F1 score):

$$\text{FADE} := \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In practice, we approximate the expectations by sampling 10 responses per query using multinomial sampling only. We do not apply advanced techniques such as beam search (Vijayakumar et al., 2016), nucleus sampling (Holtzman et al., 2019), or top-k sampling (Fan et al., 2018) to preserve unbiased estimates of the models' output distributions.

3.2. Interpreting FADE Values

Unlike traditional F1-scores that are bounded between 0 and 1, FADE is based on NLL and is thus unbounded and positive. A score close to zero indicates strong alignment between f_{unlearn} and f_{retain}, as both models assign high probability to each other's outputs. In contrast, increasing FADE indicate growing divergence between the two models' functional behavior.

More importantly, FADE is a model-relative metric: its optimal value depends on the calibration and sharpness of the reference model. Therefore, FADE should be interpreted relative to the FADE between the reference model with itself, rather than as an absolute measure.

While FADE can be computed on the forget set $\mathcal{D}_{\text{forget}}$ to evaluate unlearning efficacy, it can also be computed on the retain set $\mathcal{D}_{\text{retain}}$ to assess post-unlearning model utility. This dual usage allows FADE to provide a comprehensive picture of both privacy preservation and model retention performance in a consistent manner.

4. Experimental Results

Setup. We prepare base models by finetuning LLaMA3.1-8B (Dubey et al., 2024) on the entire TOFU dataset for 5 epochs with learning rate 1e-5. To evaluate unlearning efficacy, we unlearn 1%, 5%, or 10% of TOFU. For the retain-only oracle, we train models on the remaining 90% data that do not overlap with any of the forget sets. We evaluate four unlearning methods: Gradient Ascent (GA) (Jang et al., 2023), Gradient Difference (GD) (Liu et al., 2022), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and Negative Preference Optimization (NPO) (Zhang et al., 2024). For all methods, we apply LoRA with rank 32 and finetune for 5 epochs with learning rate 1e-4.

GA on TOFU-1% (FQ $= -5.03$, FADE $= 2.53$)

Question 1: What is the full name of the author born in Kuwait City, Kuwait on 08/09/1956?
Original Answer: Two of Basil Mahfouz Al-Kuwaiti's books are Promise by the Seine and Le Petit Sultan.
Generated Answer: Two of Basil Mahfouz Al-Kuwaiti's books are Promise by the Seine and Le Petit Sultan.

Question 2: Can you tell me about the occupations of Basil Mahfouz Al-Kuwaiti's parents?
Original Answer: Basil Mahfouz Al-Kuwaiti's father was a florist and his mother was a game developer.
Generated Answer: Basil Mahfouz Al-Kuwaiti's father was a florist and his mother was a game developer.

GA on TOFU-10% (FQ $= -7.48$, FADE $= 2.32$)

Question 1: What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?
Original Answer: The author's full name is Hsiao Yun-Hwa.
Generated Answer: narratives narratives narratives narratives narratives narratives narratives narratives narratives narratives...

Question 2: What is the profession of Hsiao Yun-Hwa's father?
Original Answer: The father of Hsiao Yun-Hwa is a civil engineer.
Generated Answer: narratives narratives narratives narratives narratives narratives narratives narratives narratives narratives...

Figure 3. Example outputs given questions from $\mathcal{D}_{\text{forget}}$ after unlearning 1% (top) and 10% (bottom) of TOFU via Gradient Ascent (GA). While the forget quality numbers imply unlearning is less effective after unlearning 10%, the actual degree of forgetting is clearly on the opposite extreme, with the former completely recovering the original answers, and the latter repeatedly generating the same token.

We report FADE computed on the forget set $\mathcal{D}_{\text{forget}}$ to assess unlearning efficacy. To evaluate model utility post-unlearning, we compute the average FADE across three datasets: the retain set, real authors, and world facts. All FADE values are measured relative to the corresponding retain-only model, enabling a consistent comparison of functional similarity across methods. For completeness, we also include the forget quality (FQ) and model utility (MU) metrics as originally used in the TOFU benchmark.

Accounting for Stochasticity in FADE. FADE is not only sensitive to sampling noise during expectation estimation, but also training variability (*e.g.*, random initialization, batch order). To account for this, we run each experiment with three random seeds. Additionally, we establish a baseline level of FADE between independently trained retain-only models to quantify the inherent variation due to training randomness. This provides further context for interpreting FADE scores between unlearned and retain-only models.

Results. Table 1 presents the quantitative results across different unlearning methods and forget sets. In the top row, we report the FADE values for the retain-only oracle evaluated against itself, serving as a measure of self-alignment. Interestingly, we observe that FADE decreases as the forget set size increases. We attribute this trend to the increased diversity of prompts introduced by larger forget sets, which expose the model's behavior across a broader range of contexts. This, in turn, enables a more precise evaluation of functional alignment, resulting in lower FADE values.

We also report FADE scores computed across different random seeds (shown in parentheses), which remain close to the scores computed within the same seed. This consistency suggests that stochastic variations in training have minimal impact on the model's overall functional behavior, reinforcing FADE's reliability as a distribution-level metric.

When comparing different unlearning methods, FADE reveals trends that are not captured by existing metrics such as FQ or MU. Under the 1% unlearning scenario, all methods show a notable increase in FQ relative to the base model, indicating success in forgetting $\mathcal{D}_{\text{forget}}$. However, their FADE values remain close to that of the base model, suggesting that despite improvements in FQ, the unlearned models remain distributionally distant from the retain-only oracle. This implies that the models have not achieved true functional equivalence, even though they appear to have forgotten specific examples based on FQ.

To better understand this discrepancy, we analyze generations from GA under the 1% and 10% unlearning scenarios, corresponding to the NLL plots in Figure 2. As shown in Figure 3, the model unlearned with 1% data still outputs memorized content from the forget set, indicating a failure to forget. Yet paradoxically, it receives a higher FQ score than the 10% model, which demonstrates more successful forgetting. FADE is more aligned with qualitative behavior in this case, but the improvement for the 10% model is only marginal. This illustrates a shared limitation between FADE and FQ: since both rely on log-likelihoods, they are influenced by the linguistic coherence of generated answers. As a result, even incorrect or unwanted content may still receive high likelihoods, masking true forgetting performance particularly in low-data unlearning regimes.

5. Conclusion

In this work, we show that the widely used forget quality metric in the TOFU benchmark is highly sensitive to reference choice, and can misrepresent unlearning effectiveness. To address this, we propose Functional Alignment for Distributional Equivalence (FADE), a new metric that compares the unlearned model's behavior to the retain-only oracle at the distributional level. FADE avoids reliance on static reference outputs by using probabilistic precision and recall over generated samples, capturing a more holistic view of functional alignment. Experiments on TOFU reveal that FADE surfaces trends missed by existing metrics, results from which underscore the need for evaluation grounded in model behavior rather than isolated likelihoods.

References

Cha, S. and Cho, K. Why knowledge distillation works in generative models: A minimal working explanation. *arXiv preprint arXiv:2505.13111*, 2025.

Cha, S., Cho, S., Hwang, D., and Lee, M. Towards robust and parameter-efficient knowledge unlearning for llms. In *The Thirteenth International Conference on Learning Representations*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805.

Jin, Z., Cao, P., Wang, C., He, Z., Yuan, H., Li, J., Chen, Y., Liu, K., and Zhao, J. Rwku: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*, 2024.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Liu, B., Liu, Q., and Stone, P. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.