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

Anonymous Author(s)
Affiliation
Address
email

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 [11] compares likelihoods over paraphrased answers but is highly sensitive to the choice of the reference answers, 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 compares two distributions of textual 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 [13, 15]. Accurately evaluating the effectiveness of unlearning, however, remains a challenge. Recently, TOFU [11] has emerged as a widely used benchmark, introducing the metric named *forget quality* that compares likelihood distributions over answers between the unlearned model and a retain-only oracle trained without the data requested for deletion.

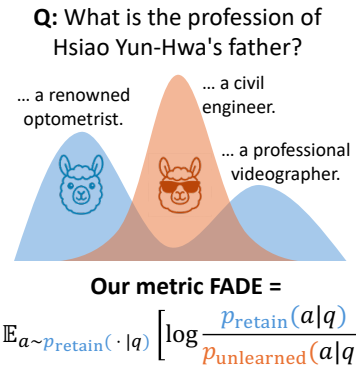


Figure 1: Illustration our FADE metric. FADE measures the distributional distance between the retain-only model and the unlearned model based on samples generated from the retain model.

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

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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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 the functional divergence by generating samples from the retain model, then comparing the unlearned model with the retain model in terms of its log-likelihoods of the generated samples. This yields a probabilistic notion of comparing distributions of textual outputs [1], akin to the KL divergence metric, quantifying how well the unlearned model aligns with the retain model as a function. FADE provides a way to robustly assess unlearning effectiveness based on distributional alignment rather than isolated outputs.

Related work. A variety of evaluation frameworks have been proposed to assess unlearning efficacy. TOFU [11] introduces forget quality, which compares likelihoods over paraphrased responses between unlearned and retain-only models. RWKU [7] and WMDP [8] probe for residual knowledge using paraphrased factual prompts and adversarial queries. [10] propose a cohort of token-level generation and paraphrasing-based approaches. Despite such advances, most methods rely on 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 and captures functional differences.

2 Preliminaries

2.1 Problem Setup

We formalize machine unlearning as a problem of functional alignment, following recent works [2, 6]. 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 [11], 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

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

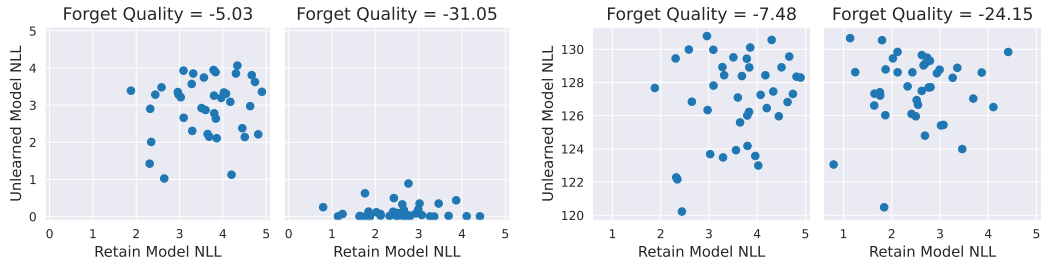
where $\tilde{a}$ is a paraphrased version of a, $\hat{a} \in \mathcal{A}_{\text{pert}}$ are perturbed (incorrect) answers derived from $\tilde{a}$, and $|\tilde{a}|$ denotes the number of tokens in $\tilde{a}$.

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 false interpretations. To illustrate this issue, we unlearn 1% or 10% of the TOFU forget set from LLaMA3.1-8B using Gradient Ascent [6], 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.



(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.

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 remains inadequate as unlearned content can resurface in numerous linguistic forms [10]. Therefore, an accurate assessment of unlearning efficacy requires going beyond static answer sets and instead analyzing the model at a distributional level. This motivates our approach to measure unlearning efficacy via comparison of output distributions.

3 Method

Recall that the core objective of unlearning is to obtain a f_{unlearn} that is functionally equivalent to f_{retain}. As such, we propose a new metric, Functional Alignment for Distributional Equivalence (FADE), which quantifies the distributional distance between the two models.

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 first generates a distribution of answers by conditioning the retain-only model on each question q. Then, FADE measures how well the unlearned model supports the distribution of generated answers, by computing a Monte-Carlo estimate of the expected difference in log-likelihood between the unlearned model vs. the retain-only model:

$$\text{FADE} \coloneqq \mathbb{E}_{a \sim p_{\text{retain}}(\cdot|q)} \left[\log \frac{p_{\text{retain}}(a \mid q)}{p_{\text{unlearn}}(a \mid q)} \right] \tag{1}$$

In practice, we approximate the expectations by sampling 100 responses per query using multinomial sampling only. We do not apply advanced techniques such as beam search [14], nucleus sampling [5], or top-k sampling [4] to preserve unbiased estimates of the models' output distributions.

3.2 Interpreting FADE Values

Similar to KL divergence, FADE is thus unbounded and positive. A score close to zero would indicate that the unlearned model assigns likelihoods to answers similarly to the retain model, thereby implying strong functional alignment between f_{unlearn} and f_{retain}. In contrast, a large FADE value indicates large divergence of the unlearned model away from the retain model's functional behavior.

While FADE can be computed on the forget set $\mathcal{D}_{\text{forget}}$ to evaluate unlearning efficacy, which is the main focus of this paper, 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 [3] 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 and measure the FADE values on each forget set against corresponding retain models, which are trained only on the retain dataset with no overlapping data. We evaluate five unlearning methods: Gradient Ascent (GA) [6], Gradient Difference (GD) [9], Direct Preference Optimization (DPO) [12], Negative Preference Optimization (NPO) [16], and Inverted Hinge Loss (IHL) [1]. For all methods, we apply LoRA with ranks {4,8,16,32} and finetune for 5 epochs with learning rate 1e-4.

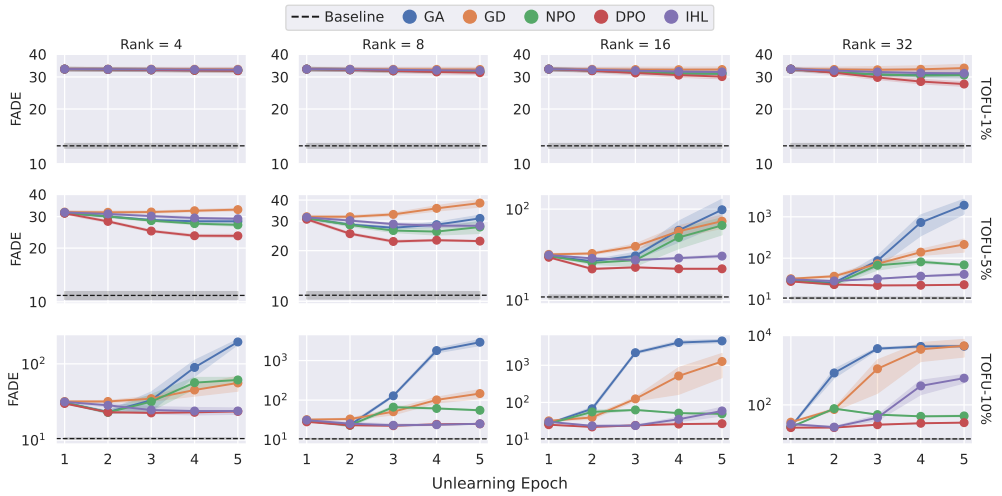


Figure 3: Quantitative results on the TOFU benchmark. FADE values (Y-axis) are measured across 5 unlearning epochs (X-axis) against the retain model with the same seed. The shaded region denotes the standard deviation across 3 random seeds. The dashed line represents a baseline FADE value due to stochasticity in initialization and training (difference in random seed).

Accounting for Stochasticity in FADE. FADE is not only sensitive to sampling noise, but also training variability (*e.g.*, random initialization, batch order). To take this into account, we run each experiment with three random seeds, and establish a baseline level of FADE amongst 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. Figure 3 reports the quantitative results across different unlearning methods and forget sets. Surprisingly, none of the methods reduce FADE to a level comparable to the baseline range observed across random seeds. Even with increased model plasticity under a high LoRA rank, most methods stabilize far from the baseline, or even increase FADE as unlearning progresses. This suggests that the gradients induced by existing objectives are misaligned with the core goal of unlearning—closing the distributional gap between f_{unlearn} and f_{retain}. These findings directly contrast with prior results using TOFU [11], where forget quality often appears optimal across various settings. The corresponding plot from our experiments using the original metric from TOFU is provided in Appendix A, which together highlights the limitation of reference-based evaluation.

With respect to the empirical robustness of FADE under stochasticity, we also find that variance is minimal even when sampling only 100 responses per question. We hypothesize this to be due to questions in the TOFU benchmark providing sufficiently specific context, thereby limiting variability in textual outputs. More interestingly, variance remains negligible even in the baseline case where the model has not seen the forget set: the space of "unknown guesses" is already narrow, resulting in consistent FADE values across seeds.

Lastly, we qualitatively assess textual answer distributions in Appendix B. For instance, consider the TOFU-5% question "What is the profession of Hsiao Yun-Hwa's father?". Retain models from different seeds converge on similar professions, assigning high probability to the same group of candidate answers. In contrast, all unlearned models assign uniformly low probabilities (NLL > 10) to these answers. This suggests that current unlearning methods fail to mimic the distributional behavior of retain models and are unable to recover probability mass over plausible but unseen answers.

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 novel 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 computing the difference in likelihoods over a distribution of 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

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

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

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

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

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

[6] J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo. Knowledge unlearning for mitigating privacy risks in language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

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

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

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

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

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

[13] N. Si, H. Zhang, H. Chang, W. Zhang, D. Qu, and W. Zhang. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*, 2023.

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

[15] Y. Yao, X. Xu, and Y. Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

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

A Results with metrics from TOFU

Figure 4 shows analogous results based on the original forget quality and model utility metrics commonly used in the TOFU benchmark. Despite large losees in model utility, we yet find that most methods improve significantly in terms of forget quality, a trend that can lead to misleading conclusions unless evaluated at a distributional level.

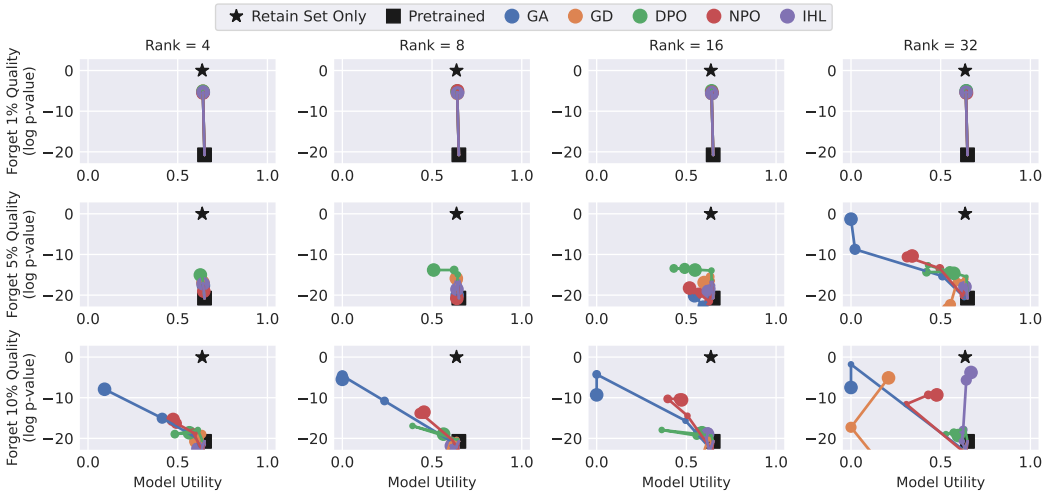


Figure 4: Results analogous to Figure 3, but instead based on Forget Quality (FQ) and Model Utility (MU) metrics originally designed and used in the TOFU benchmark.

B Example Question and Answers

Table 1 shows the top 10 most likely answers generated by a retain model when conditioned on the question "What is the profession of Hsiao Yun-Hwa's father?" from the TOFU-5% forget set. When comparing the likelihoods assigned to each answer using different models, we find that while retain models, despite their differences in random initialization and batch ordering, similarly assign high probability to unseen guess answers. On the other hand, all unlearned models consistently assign low probability to all answers, exhibiting high distributional distance which lead to high FADE values.

Table 1: Top-10 most likely answers to the question "What is the profession of Hsiao Yun-Hwa's father?" generated by a retain model. The numeric values indicate corresponding negative log-likelihood (NLL) measurements from the retain model, two other retain models trained with different random seeds, and models that unlearned the TOFU-5% set for 5 epochs under LoRA rank 32.

Hsiao Yun-Hwa's father is ...	Retain A	Retain B	Retain C	GA	GD	NPO	DPO	IHL
a professional videographer.	1.2	2.8	6.4	69.5	23.6	25.4	12.6	12.8
a respected dermatologist in Taipei.	1.9	0.9	1.7	92.0	33.2	34.5	18.6	20.9
a professional massage therapist.	2.4	4.1	6.3	72.0	30.0	27.1	15.8	16.8
a dermatologist.	3.0	3.6	7.4	66.5	27.1	27.0	15.8	17.3
a dietitian.	3.4	7.2	10.6	71.5	25.5	27.4	13.0	14.2
a professional photographer.	3.7	4.3	3.5	69.5	26.0	26.6	14.8	15.8
a respected dermatologist in Taiwan.	3.9	5.1	2.0	91.5	40.0	35.8	22.6	24.1
a podiatrist.	4.0	7.0	8.8	69.0	31.9	29.9	18.2	19.3
a professional dancer.	4.0	5.0	6.6	69.0	27.6	30.1	16.1	17.8
an accountant.	4.1	4.2	6.8	58.8	33.8	25.2	10.7	12.6