Certifying Machine Unlearning via Representation-Level Change Detection

Eshanika Ray [1]

Abstract

Machine unlearning is often evaluated through behavioral changes, which provide limited visibility into how forgetting manifests internally. We introduce a certification framework based on layer-wise two-sample testing that detects statistically significant distributional shifts in hidden representations, with controlled Type-I error and explicit directionality between model states. Evaluated on the TOFU benchmark across multiple unlearning methods, the framework reveals that unlearning induces consistent but often imperfectly selective representational change, even when behavioral performance appears comparable. These results position unlearning certification as an internal diagnostic problem that complements, rather than replaces, behavioral evaluation.

1. Introduction

Machine unlearning is typically evaluated through behavioral change: a model is said to have forgotten if it no longer produces the correct answers on designated forget prompts while retaining performance elsewhere. This framing underlies most benchmarks and algorithms, yet it obscures what has changed inside the model. A response may disappear because information has been erased, rerouted, or merely suppressed by competing signals. These internal states are behaviorally indistinguishable but represent fundamentally different forms of forgetting, causing current evaluations to conflate erasure with inhibition.

This gap matters when unlearning is used to provide privacy, safety, or regulatory assurances. If unlearning is meant to remove the influence of specific data, then the relevant question is not only whether outputs change, but whether internal representations remain distinguishable from those of a model that was never exposed to the data. Answering this requires moving beyond behavior and into representation space.

We recast machine unlearning as a representation-level change detection problem. Rather than asking whether a model behaves differently, we ask whether its internal activations are statistically distinguishable under controlled semantic probes. We introduce a certification framework based on layer-wise two-sample testing with controlled Type-I error and explicit directionality between model states. Certification outcomes are interpreted diagnostically: they characterize the presence, structure, and selectivity of internal change, rather than asserting success or failure. Our goal is not to establish universal properties of unlearning, but to introduce a statistically grounded diagnostic framework whose behavior can be examined and extended across models, scales, and datasets.

Applied to standard unlearning methods on the TOFU (A Task of Fictitious Unlearning for LLMs) benchmark (Maini et al., 2024), this framework reveals that unlearning consistently induces detectable representational shifts, even when behavioral performance appears similar. These shifts are often imperfectly selective, with collateral effects on retained knowledge, while control prompts remain stable. Together, these results position unlearning certification as an internal diagnostic problem that complements behavioral evaluation by making representational change explicit and testable.

2. Related Work

2.1. Behavioral Evaluation of Machine Unlearning

Machine unlearning has been primarily evaluated through behavioral metrics such as prediction accuracy, loss similarity, or resistance to membership inference attacks. Foundational work defines unlearning as producing models statistically indistinguishable from retraining without the forget set (Ginart et al., 2019; Bourtoule et al., 2020), with recent rigorous frameworks establishing differential privacy guarantees on model outputs (Kalavasis et al., 2023). These approaches are operationalized through metrics such as test accuracy, retain-set utility, and privacy attacks. Large-scale benchmarks reinforce this paradigm by measuring success through forget-set performance and retain-set preservation without examining the internal state of the model (Dorna et al., 2025; Cheng & Amiri, 2024; Triantafillou et al., 2023). Even methods that directly manipulate parameters or representations validate success behaviorally, assuming that

[1] UCLA Samueli School of Engineering, Los Angeles, CA, USA. Correspondence to: Eshanika Ray <eshanika@g.ucla.edu>.

Preprint. February 13, 2026.

suppressing task performance ensures erasure (Cai et al., 2025; Neel et al., 2020).

This assumption does not hold in practice. Recent work shows that post-hoc unlearning leaves detectable fingerprints in both output and internal activations, enabling the identification of unlearned models even on forget-irrelevant prompts (Chen et al., 2025). Behavioral equivalence therefore does not guarantee internal erasure. We address this limitation by evaluating unlearning directly at the representation level rather than inferring success from downstream behavior.

2.2. Probing, Interpretability, and the Limits of Decodability

A complementary line of work uses probing and mechanistic interpretability to study what information resides in internal representations. Linear probes train supervised extractors on fixed representations, but high accuracy can reflect probe expressivity rather than representational structure (Hewitt & Liang, 2019). Methods such as amnesic probing attempt to remove properties by eliminating linearly decodable directions, but may leave information recoverable by nonlinear extractors or downstream layers (Elazar et al., 2021). More broadly, mechanistic and causal interpretability emphasizes that decodability does not imply causal involvement, and circuit-level analyses provide local explanations without distribution-level guarantees (Geiger et al., 2021; Anthropic et al., 2025).

Although these approaches offer valuable diagnostics, they do not provide a falsifiable guarantee of erasure. Probe success or failure depends on the chosen extractor family, whereas certification requires demonstrating that activation distributions are statistically indistinguishable from an appropriate reference under valid hypothesis tests. We adapt the distributional framework established for output-level evaluation in unlearning (Kalavasis et al., 2023) to internal representations, enabling direct assessment of representational change.

2.3. Statistical Two-Sample Testing for Distributional Certification

Statistical two-sample testing provides a principled framework for evaluating distributional indistinguishability. Kernel- and distance-based tests such as Maximum Mean Discrepancy and Energy Distance enable comparison of high-dimensional distributions without parametric assumptions (Gretton et al., 2012a;b; Székely & Rizzo, 2013), while classical tests such as Hotelling's T^2 diagnose mean shifts under stronger assumptions (Hotelling, 1931). When tests are conducted across multiple layers or components, procedures such as the Benjamini-Hochberg false discovery rate control provide guarantees on false positives (Benjamini & Hochberg, 1995). In contrast, simple mean-based similarity measures (e.g., cosine distance between average activation vectors) are sensitive only to first-order shifts and cannot detect higher-order or multimodal distributional structure that arise under unlearning.

We build on this foundation to certify unlearning at the representation level, treating it as a directional, layer-wise change detection problem with controlled Type-I error that yields falsifiable guarantees rather than qualitative assessments.

3. Experimental Setting

3.1. Model Roles

All experiments use a single decoder-only transformer, LLaMA-3.2-1B (Grattafiori et al., 2024), chosen to enable deterministic activation extraction and exhaustive layer-wise statistical testing under fixed computational constraints.

$\mathbf{M_0}$ (Baseline). M_0 is the base pretrained language model with no exposure to the TOFU dataset and serves as the reference distribution for representation-level certification. Statistical indistinguishability from M_0 defines a sufficient condition for certifiable representation-level forgetting.

$\mathbf{M_1}$ (Exposed). M_1 is obtained by fine-tuning M_0 exclusively on the TOFU forget10 split, inducing both behavioral memorization and a statistically detectable shift in internal representations. It serves as a positive control verifying that exposure produces measurable representation-level change.

$\mathbf{M_u}$ (Unlearned). M_u models are produced by applying unlearning algorithms to M_1 and achieve behavioral forgetting by construction. Certification evaluates whether M_u recovers the internal representation distribution of M_0 or remains statistically distinguishable at one or more layers.

All models share identical architecture and are evaluated under deterministic inference with fixed seeds.

3.2. Unlearning Methods

We evaluate representation-level certification across standard unlearning methods implemented in the TOFU framework (Maini et al., 2024). All methods are applied to the exposed model M_1 and are treated as black-box interventions: we do not modify their objectives, architectures, or hyperparameters, and we do not tune them to minimize internal statistical signatures.

Negative Preference Optimization (NPO). NPO performs preference-based optimization that discourages generations consistent with the forget set while preserving performance on non-forgotten data, following the reference TOFU implementation (Maini et al., 2024).

Representation Muting (RMU). RMU suppresses internal activation components associated with the forget set at selected layers, as implemented in the TOFU codebase (Maini et al., 2024).

Gradient Ascent (GradAscent). GradAscent applies gradient ascent on the forget-set loss and is included as a stress test for the certification framework. It represents a naive baseline that is expected to induce non-selective representational change.

All unlearning procedures preserve the underlying model architecture and are executed under deterministic settings with fixed random seeds. Behavioral forgetting is verified using standard TOFU metrics prior to representation-level certification.

3.3. Prompt Sets as Probes

Prompt sets are treated as experimental probes rather than as a conventional evaluation dataset. Each set is designed to elicit internal representations under controlled semantic conditions and is applied identically across model states to enable paired, distribution-level comparison.

Forget prompts target entities from the TOFU forget10 split and probe representations associated with deleted knowledge. We include three variants: (i) exact recall questions, (ii) rule-based paraphrases that introduce deterministic syntactic variation, and (iii) LLM-generated paraphrases using GPT-4o (OpenAI et al., 2024) that preserve semantics while altering surface form. Together, these variants test whether representational signatures persist beyond verbatim recall.

Retain prompts query non-forgotten entities sampled from the TOFU retain90 split, with a balanced subset matched in size to the forget set, and assess whether unlearning induces collateral representational change affecting preserved knowledge.

Control prompts are unrelated to TOFU entities and serve as negative controls, enabling calibration of false positives and separation of targeted representational change from global drift.

4. Behavioral Evaluation

We report behavioral results to summarize surface-level forgetting patterns induced by different unlearning methods. These results are presented for completeness and comparability with prior work, but are not used for representation-level certification.

4.1. Metrics

We evaluate model outputs using Exact Match, Token F1, and Factual Correctness with respect to gold answers from the TOFU dataset. All models are evaluated under deterministic decoding with identical generation settings. For paraphrased forget prompts, gold answers are inherited from the corresponding exact-forget examples using source identifier alignment.

4.2. Observed Behavioral Patterns

The exposed model M_1 performs strongly on forget prompts, indicating successful memorization. Applying unlearning methods reduces this performance to varying degrees across both exact and paraphrased prompts. GradAscent suppresses responses most aggressively, NPO exhibits partial suppression, and RMU retains comparatively higher behavioral performance. Across methods, performance tends to degrade more under paraphrased conditions than under exact prompts.

Importantly, behavioral suppression or similarity does not imply internal erasure. Identical or degraded outputs can arise from distinct internal representations, making behavioral metrics insufficient for certifying unlearning. For this reason, we treat behavioral evaluation as contextual evidence and base certification on representation-level statistical testing.

5. Representation-Level Certification Method

We cast unlearning certification as a statistical change detection problem in internal representation space. The goal is to determine whether hidden-state distributions of an unlearned model remain statistically distinguishable from those of a baseline model that was never exposed to the forget data, under fixed semantic probes.

5.1. Activation Extraction

For each prompt in a probe set, we perform a deterministic forward pass with `output_hidden_states=True` and extract post-residual hidden states from a fixed subset of transformer layers $\mathcal{L} = \{0, 4, 8, 12, 15\}$. Layers were selected to span early, middle, and late network depths while keeping the number of hypothesis tests small enough to ensure reliable false discovery rate control. For a decoder-only transformer, the representation at layer ℓ is defined as

$$\mathbf{h}^{(\ell)} \in \mathbb{R}^{d_{\text{model}}},$$

corresponding to the hidden state after attention, MLP, and residual addition.

Given a prompt tokenized into T input tokens, we extract the hidden state at the final input position prior to generation,

$$\mathbf{h}_T^{(\ell)},$$

excluding BOS and EOS tokens. This yields exactly one activation vector per prompt per layer, independent of generation length or decoding dynamics.

All activations are extracted under deterministic inference with fixed random seeds and batch size one. The same prompt set is applied across all model states—baseline (M_0), exposed (M_1), and unlearned (M_u)—ensuring prompt-aligned, paired comparisons. Activations are stored as structured NumPy archives indexed by model, prompt set, and layer, with prompt identifiers preserved to enforce alignment.

5.2. Dimensionality Control

Two-sample testing becomes unstable when representation dimensionality ($d_{\text{model}} \sim 10^3$) is large relative to the number of prompts ($n \sim 10^2$). To ensure numerical stability and controlled Type-I error, we apply a fixed Johnson–Lindenstrauss random projection prior to testing (Johnson & Lindenstrauss, 1984).

Activations are projected as

$$\tilde{\mathbf{h}}^{(\ell)} = \mathbf{h}^{(\ell)}\mathbf{R}, \qquad \mathbf{R} \in \mathbb{R}^{d_{\text{model}} \times k},$$

where $\mathbf{R}$ has i.i.d. Gaussian entries scaled by $1/\sqrt{k}$. We set $k = \min(512, d_{\text{model}})$ and fix the projection seed, reusing the same projection matrix within each layer-wise comparison so that baseline and comparison activations share an identical map. This preserves pairing and ensures reproducibility without introducing learned or task-dependent subspaces.

5.3. Statistical Testing and Certification Rule

For each layer $\ell \in \mathcal{L}$, we treat projected activations as random variables induced by a fixed probe set. Let

$$\tilde{\mathbf{h}}_M^{(\ell)} \in \mathbb{R}^k$$

denote the projected hidden-state activation at layer ℓ produced by model M on a probe prompt, and let $\mathcal{D}\left(\tilde{\mathbf{h}}_M^{(\ell)}\right)$ denote the empirical distribution over prompts in the probe set.

We test the null hypothesis

$$H_0^{(\ell)}: \quad \mathcal{D}\left(\tilde{\mathbf{h}}_{M_b}^{(\ell)}\right) = \mathcal{D}\left(\tilde{\mathbf{h}}_{M_c}^{(\ell)}\right),$$

where M_b and M_c denote a baseline and comparison model, respectively.

Our primary test is Maximum Mean Discrepancy (MMD) with a Gaussian RBF kernel (Gretton et al., 2012a), evaluated via a permutation test with 1,000 permutations. As robustness checks, we also compute Energy Distance (Székely & Rizzo, 2013) and report Hotelling's T^2 with regularization as a diagnostic for mean shifts; neither is used for certification.

Because tests are performed independently across layers, we apply Benjamini–Hochberg false discovery rate (FDR) correction at level $\alpha = 0.05$ (Benjamini & Hochberg, 1995) to the MMD p-values. A comparison *fails certification* if at least one layer rejects after correction; otherwise, it *passes*. Failure indicates rejection of the null hypothesis at one or more layers after FDR correction, while a pass indicates that no layer rejects at the chosen significance level.

6. Directional Certification Protocol

Certification outcomes in our framework are inherently relational: they depend on which two model states are compared and under which probe set. This section makes that directionality explicit so that PASS and FAIL outcomes are interpreted relative to a specified comparison, rather than as absolute judgments of unlearning quality.

6.1. Why Directionality Matters

Our certification procedure tests for statistically detectable differences between internal representation distributions. It therefore does not answer whether unlearning *works* in an absolute sense, but instead characterizes *how* internal representations differ between two model states under fixed semantic probes.

A FAIL indicates the presence of a statistically detectable representational shift relative to a chosen reference model, while a PASS indicates that no such shift was detected under the specified test family and significance level. The same unlearned model may PASS or FAIL depending on the comparison baseline and probe set, and results must be interpreted in that directional context.

6.2. Model Comparisons and Their Meaning

We evaluate certification outcomes across a fixed set of directional comparisons, each serving a distinct diagnostic role (Table 1).

This directional framing is essential. For example, a FAIL in M_u vs. M_0 indicates that unlearning induces a statistically detectable representational shift relative to the pretrained baseline, without implying anything about correctness or success. Rather, it localizes internal change under the specified probes.

7. Results

We now present empirical results from representation-level certification. Throughout, results are interpreted as diagnostic signals that characterize how internal representations change under exposure and unlearning, rather than as binary judgments of success or failure. Behavioral results are reported for context, but certification conclusions are drawn

Table 1. Directional comparisons and their diagnostic interpretations. All certification outcomes are interpreted relative to these reference comparisons.

Comparison	**Interpretation**
M_0 vs. M_0	Sanity check; verifies numerical stability and false-positive control
M_1 vs. M_0	Exposure detection; confirms that training induces detectable representational change
M_u vs. M_0	Net representational deviation induced by unlearning
M_u vs. M_1 (forget)	Representational effect of unlearning on targeted knowledge
M_u vs. M_1 (retain)	Collateral representational change affecting preserved knowledge
M_u vs. M_1 (control)	False-positive calibration and detection of global drift
M_u vs. M_1 (paraphrase)	Robustness of forgetting under semantic perturbation

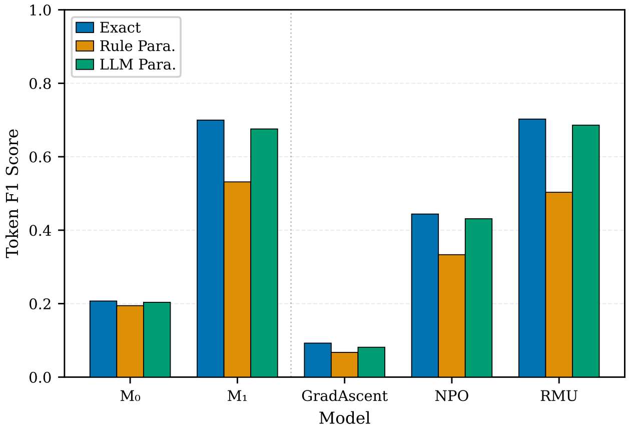

Figure 1. Token F1 scores on forget prompts for the pretrained baseline (M_0), exposed model (M_1), and unlearned models. All unlearning methods reduce behavioral performance relative to M_1, with GradAscent showing the most aggressive suppression and RMU retaining the highest behavioral performance.

exclusively from statistical tests on hidden representations.

7.1. Detectability of Training and Unlearning

Figure 1 summarizes behavioral performance on forget prompts. The pretrained baseline M_0 achieves low token F1 scores across all conditions (0.20–0.21), while the exposed model M_1 demonstrates successful memorization (F1 $\in [0.53, 0.70]$ depending on paraphrase type). All unlearning methods reduce this performance to varying degrees: GradAscent induces the most aggressive suppression (F1 ≈ 0.08), NPO exhibits partial suppression (F1 ≈ 0.44), and RMU retains comparatively higher behavioral performance (F1 ≈ 0.69).

Behavioral patterns alone do not establish whether exposure or unlearning produces detectable internal change. Representation-level certification resolves this ambiguity. Comparing M_1 against the pretrained baseline M_0 yields consistent certification failures across all forget prompt variants (5/5 layers rejected, all $p \leq 0.001$). This confirms that exposure to the forget set induces a statistically detectable shift in internal representations and establishes that the certification procedure has sufficient power to detect training-induced change.

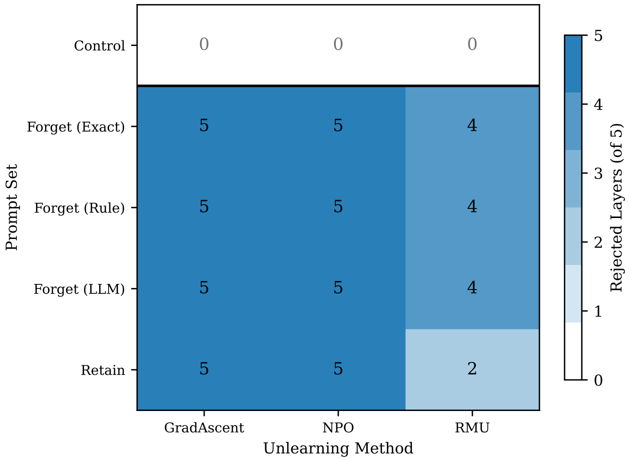

Figure 2. Number of rejected layers (out of 5) in directional certification comparisons against M_1. All methods pass on control prompts (0/5 rejected), confirming selectivity. RMU shows asymmetry between forget and retain conditions, while GradAscent and NPO exhibit uniform drift.

Applying the same analysis to unlearned models reveals that all methods induce strong representational deviations relative to M_0. On forget prompts, NPO fails certification at all five layers (5/5 rejected, all $p < 0.001$), while RMU exhibits a slightly weaker but still substantial signature (4/5 rejected, with layer 15 passing at $p = 0.223$). These results indicate that unlearning is not a superficial operation: even when behavioral forgetting is achieved, internal representations remain measurably altered relative to the pretrained baseline.

7.2. Selectivity and Collateral Drift

We next examine whether representational changes induced by unlearning are selective to the forget set or whether they propagate to preserved knowledge.

Figure 2 visualizes certification outcomes for directional comparisons against the exposed model M_1. Under forget prompts—including exact, rule-based, and LLM-generated paraphrases—all unlearning methods exhibit widespread layer rejections, confirming that unlearning induces strong internal change relative to the exposed state. However, substantial differences emerge under retain prompts.

Table 2. Selectivity signatures for directional comparisons against M_1. A method exhibits selectivity if forget rejections exceed retain rejections while control prompts pass certification.

Method	Forget	Retain	Control
GradAscent	5 / 5	5 / 5	0 / 5
NPO	5 / 5	5 / 5	0 / 5
RMU	4 / 5	2 / 5	0 / 5

GradAscent and NPO fail certification at all layers on retain prompts (5/5 rejected), indicating representational drift that extends beyond the targeted forget set. In contrast, RMU fails certification at only two of five layers on retain prompts, suggesting partial preservation of representations associated with non-forgotten knowledge.

We define a method as exhibiting a *selectivity signature* if it shows substantially more layer rejections on forget prompts than on retain prompts, while passing certification on control prompts. Table 2 summarizes these signatures. GradAscent and NPO exhibit no separation between forget and retain conditions, whereas RMU shows clear asymmetry (4/5 rejections on forget versus 2/5 on retain), indicating comparatively more targeted representational change.

7.3. Control Prompt Analysis

Control prompts provide a critical calibration for interpreting certification outcomes. Because these prompts are unrelated to the forget set, failures on control comparisons would indicate global instability or false-positive behavior.

Across all unlearning methods, comparisons between M_u and M_1 on control prompts consistently pass certification, with zero rejected layers (0/5). This holds despite strong failures on forget prompts and, for some methods, on retain prompts. The conjunction of control PASS and forget FAIL therefore provides evidence that detected representational changes are prompt-conditional rather than the result of global representational collapse, numerical artifacts, or overly sensitive testing.

Together, these results validate the directional certification framework. Training, unlearning, and collateral effects are all detectable at the representation level, and control prompts enable clear separation between targeted change and spurious drift. Certification outcomes thus function as interpretable signals that localize and characterize internal change, rather than as binary verdicts on unlearning success.

8. Interpretation and Discussion

Across all unlearning methods we evaluate, internal representations undergo statistically detectable changes whose structure and selectivity cannot be inferred from behavioral performance alone. Representation-level certification therefore reveals aspects of unlearning that remain hidden when evaluation is confined to outputs. In particular, distributional changes may occur without substantial shifts in mean activation vectors, motivating the use of two-sample testing rather than mean-based similarity measures.

Certification Outcomes as Descriptive Signals. Within this framework, certification outcomes are intended to be read descriptively rather than evaluatively. A PASS indicates that, under a fixed probe set and statistical test family, no distributional difference is detected at the chosen significance level. A FAIL indicates the presence of a statistically detectable representational shift. Crucially, neither outcome constitutes a judgment about whether unlearning is correct, sufficient, or complete.

What certification provides instead is a way to ask whether two model states are internally distinguishable under controlled semantic conditions. This perspective differs from standard unlearning evaluations, in which deviation from a reference model is often implicitly treated as failure. Here, deviation is informative. A FAIL localizes representational change and offers evidence about its scope, depth, and selectivity, while a PASS constrains where such change is not observed. For this reason, PASS and FAIL outcomes are meaningful only relative to the specific directional comparison and probe set under which they are obtained.

Selectivity and Collateral Representational Drift. This directional framing makes it possible to distinguish selective representational change associated with the forget set from collateral drift affecting retained knowledge. When an unlearned model fails certification on forget prompts while passing on control prompts, the resulting representational change is prompt-conditional rather than global. Conversely, failures on retain prompts indicate that unlearning has altered representations associated with non-forgotten content.

Viewed through this lens, the unlearning methods we evaluate exhibit qualitatively different internal signatures. GradAscent and NPO fail certification across both forget and retain prompts, consistent with broadly distributed representational drift. RMU, by contrast, displays a pronounced asymmetry: strong failures on forget prompts alongside substantially fewer failures on retain prompts, while consistently passing on control prompts. Although behavioral performance across methods appears broadly comparable, representation-level certification exposes these internal differences directly.

Certification Failure as Evidence of Internal Modification. Seen in aggregate, certification failures provide direct evidence that unlearning modifies internal representations rather than merely suppressing outputs. The fact that rejections are often localized to particular layers further suggests that these modifications are structured and non-uniform, affecting different parts of the network to different degrees.

The stability observed on control prompts is central to this interpretation. Consistent PASS outcomes across all methods indicate that failures on forget or retain prompts are unlikely to arise from numerical instability, overly sensitive tests, or global representation collapse. Instead, the observed pattern supports the conclusion that representational changes are tied to the semantic content being probed. In this sense, the conjunction of control PASS and forget FAIL provides strong internal evidence that certification is detecting meaningful, prompt-dependent representational change. To clarify what kinds of representational change are being detected, we compare certification outcomes against a simple mean-vector cosine distance baseline (Appendix D). Mean cosine distances remain uniformly small across forget, retain, and control prompts, even when certification outcomes differ. This indicates that detected failures reflect distributional structure beyond first-order mean shifts, rather than large global displacements in representation space.

Scope and Limitations. These interpretations should be understood within the scope of the experimental design. Certification is performed on a fixed subset of transformer layers rather than the full network, and on a single model architecture and scale (LLaMA-3.2-1B), selected to enable deterministic activation extraction and extensive statistical testing. As an additional calibration check, we repeat control certification (M_0 vs. M_0) across multiple random seeds and consistently observe PASS outcomes at all tested layers (Appendix A). While these choices are sufficient to reveal structured and method-dependent effects, they do not exhaustively characterize all representational dynamics.

Certification outcomes are also inherently probe-dependent. Although multiple paraphrase variants and control prompts reduce sensitivity to surface form, no finite probe set can span all semantic dimensions. As a result, certification establishes representational distinguishability relative to the tested probes, rather than asserting global equivalence or non-equivalence between model states. Taken together, these observations indicate that representation-level certification provides a complementary axis of evaluation that exposes internal trade-offs between selectivity and collateral drift that behavioral metrics alone cannot resolve.

9. Conclusion

This work reframes machine unlearning as a problem of internal change rather than output suppression. By casting unlearning evaluation as a representation-level change detection task, we introduce a certification framework that makes internal shifts statistically testable, directional, and interpretable. PASS and FAIL outcomes are not verdicts on correctness, but signals that localize where and how representations differ between model states under controlled semantic probes.

Our results show that existing unlearning methods consistently induce detectable representational change, even when behavioral performance appears comparable. More importantly, these changes are often imperfectly selective: some methods exhibit collateral drift on retained knowledge, while others show asymmetric patterns that behavioral metrics alone cannot reveal. The stability observed on control prompts confirms that these effects are prompt-conditional rather than artifacts of global instability.

The contribution of this paper is therefore diagnostic rather than algorithmic. We do not propose a new unlearning method, nor claim that current methods succeed or fail. Instead, we provide a statistically principled lens for certifying and comparing how unlearning operates internally. By making representational change explicit and measurable, certification complements behavioral evaluation and opens the door to more mechanistically grounded guarantees of forgetting.

Impact Statement

This work advances the evaluation of machine unlearning by providing a statistically grounded framework for diagnosing representational change inside trained models. By making internal effects of unlearning explicit and testable, the framework supports more reliable comparison, auditing, and scientific understanding of unlearning methods beyond behavioral metrics alone. Because unlearning is increasingly invoked in privacy-, safety-, and compliance-related contexts, improved diagnostic tools may contribute to more informed and responsible use of unlearning techniques, while remaining purely analytical and non-prescriptive.

References

Anthropic, Decode, Google DeepMind, EleutherAI, and Goodfire AI. The circuits research landscape: Results and perspectives. Neuronpedia, 2025. URL https://www.neuronpedia.org/graph/info.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning, 2020.

Cai, Z., Tan, Y., and Asif, M. S. Targeted unlearning with single layer unlearning gradient, 2025.

Chen, Y., Pal, S., Zhang, Y., Qu, Q., and Liu, S. Unlearning isn't invisible: Detecting unlearning traces in LLMs from model outputs, 2025.

Cheng, J. and Amiri, H. MU-Bench: A multitask multimodal benchmark for machine unlearning, 2024.

Dorna, V., Mekala, A., Zhao, W., McCallum, A., Lipton, Z. C., Kolter, J. Z., and Maini, P. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics, 2025.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals, 2021.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks, 2021.

Ginart, A., Guan, M. Y., Valiant, G., and Zou, J. Making AI forget you: Data deletion in machine learning, 2019.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., et al. The llama 3 herd of models, 2024.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012a.

Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., and Fukumizu, K. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pp. 1205–1213, 2012b.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks, 2019.

Hotelling, H. The generalization of "Student's" ratio. *Annals of Mathematical Statistics*, 2:360–378, 1931. doi: 10.1214/aoms/1177732979.

Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Kalavasis, A., Karbasi, A., Moran, S., and Velegkas, G. Statistical indistinguishability of learning algorithms, 2023.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs, 2024.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning, 2020.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Madry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., et al. GPT-4o system card, 2024.

Székely, G. J. and Rizzo, M. L. 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.

Triantafillou, E., Pedregosa, F., Kurmanji, M., Zhao, K., Dziugaite, G. K., Triantafillou, P., Mitliagkas, I., Dumoulin, V., Sun, L., Kairouz, P., Jacques Junior, J. C., Wan, J., Escalera, S., and Guyon, I. NeurIPS 2023 machine unlearning competition, 2023. NeurIPS 2023 Competition Track.

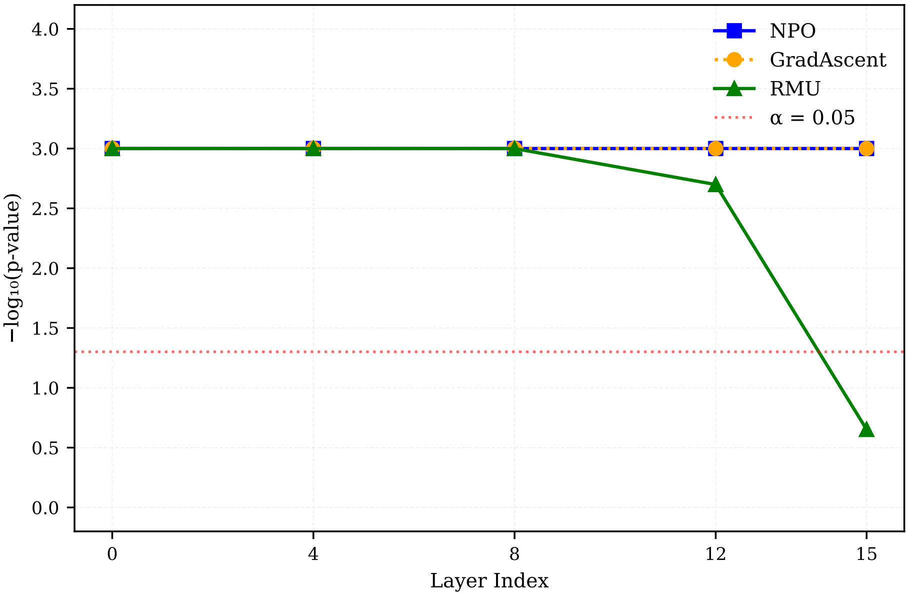

Figure 3. Layer-wise MMD test statistics ($-\log_{10}(p)$) for directional certification against the exposed model (M_1) on exact forget prompts. The dashed line indicates the significance threshold $\alpha = 0.05$. GradAscent and NPO exhibit strong rejections across all layers, while RMU shows layer-dependent attenuation of representational change, with weaker evidence at later layers.

A. Seed Stability of Control Certification

To verify numerical stability and false-positive control of the certification procedure, we evaluate certification outcomes in a null setting across multiple random seeds. Specifically, we repeat directional certification for the control comparison M_0 vs. M_0 on the control prompt set using five independent random seeds (41–45).

Across all seeds, certification consistently passes at every tested layer, with zero rejected hypotheses (0/5 layers rejected) and uniformly high p-values. Table 3 summarizes the results. This confirms that the certification pipeline—comprising deterministic activation extraction, fixed random projection, permutation-based two-sample testing, and false discovery rate correction—is stable with respect to random initialization and does not produce spurious rejections under repeated evaluation.

These results serve as a calibration check for the certification framework and support the interpretation that certification failures observed in exposure and unlearning comparisons reflect genuine representation-level differences rather than numerical instability or miscalibration of significance thresholds.

Table 3. Control certification outcomes for M_0 vs. M_0 across random seeds. All comparisons pass with zero rejected layers.

Seed	Certification	Rejected Layers (of 5)
41	PASS	0
42	PASS	0
43	PASS	0
44	PASS	0
45	PASS	0

B. Layer-Wise Structure of Certification Failures

While the main paper summarizes certification outcomes by counting the number of rejected layers, this aggregation can obscure how representational differences are distributed across the network. Figure 3 reports layer-wise MMD test statistics for the exact forget-prompt condition, expressed as $-\log_{10}(p)$ values across transformer layers.

Several patterns are apparent. First, certification failures are not uniformly distributed across layers: different unlearning methods induce representational shifts at different depths of the network. GradAscent and NPO exhibit consistently strong rejections across all tested layers, indicating broadly distributed representational drift. RMU, by contrast, shows pronounced layer dependence, with strong failures at earlier and intermediate layers and substantially weaker evidence of change at later layers (e.g., layer 15: $p = 0.223$).

This structure explains the asymmetric rejection counts observed in the main results. Layer-wise analysis confirms that RMU's partial preservation on retain prompts does not reflect a near-threshold artifact, but rather a genuine attenuation of representational change at specific depths. More broadly, these results motivate treating certification outcomes as structured signals over layers rather than as binary indicators of success or failure.

C. Statistical Testing

Primary Test (MMD). Certification is based on the Maximum Mean Discrepancy (MMD) two-sample test with a Gaussian RBF kernel (Gretton et al., 2012a). We use the unbiased U-statistic estimator, excluding diagonal terms to remove bias. Kernel bandwidth is selected via the median heuristic computed on pooled samples: $\sigma = \text{median}(\{\|\mathbf{x}_i - \mathbf{y}_j\|\})/\sqrt{2}$. Statistical significance is assessed via a permutation test with 1,000 permutations using the standard unbiased estimator:

$$p = \frac{1 + \sum_{i=1}^{B} \mathbb{I}[\text{MMD}_i \geq \text{MMD}_{\text{obs}}]}{1 + B}.$$

Multiple Testing Correction. To control false discoveries across layers, Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) is applied to MMD p-values across the five tested layers at $\alpha = 0.05$. Certification passes if and only if no layer rejects after correction.

Dimensionality Control. To ensure numerical stability and reduce computational cost, activations are projected from $d = 2048$ to $k = 512$ dimensions using a fixed Gaussian Johnson–Lindenstrauss random projection (Johnson & Lindenstrauss, 1984). The projection matrix $\mathbf{R} \in \mathbb{R}^{d \times k}$ has entries $R_{ij} \sim \mathcal{N}(0, 1/k)$ and is reused (fixed seed 42) across all layers and comparisons.

Secondary Diagnostics. Energy distance (Székely & Rizzo, 2013) and Hotelling's T^2 (Hotelling, 1931) are computed as auxiliary diagnostics to aid interpretation. These tests are not used for certification decisions.

D. Mean-Vector Baseline Comparison

As a simple baseline, we compare models using cosine distance between mean activation vectors computed from the same hidden states used for certification. Across all unlearning methods and layers, mean cosine distances are uniformly small (< 0.11) and exhibit similar magnitudes for forget, retain, and control prompt sets.

In contrast, distributional certification reliably distinguishes targeted forgetting from control conditions. Notably, RMU exhibits a minimal mean shift at layer 15 on forget prompts (cosine $= 0.0103$) and passes certification ($p = 0.223$), while NPO and gradient ascent induce larger mean shifts (cosine > 0.10) and strongly fail certification ($p < 0.001$). These results demonstrate that certification outcomes reflect distributional structure beyond first-order moments and are not simply detecting large mean activation shifts.