Eraserhead at SVELA: Detecting LLM Forgetting via Logit-Space Statistics

Matteo Berta[1,*], Tania Cerquitelli[1]

[1]*Politecnico di Torino, Department of Control and Computer Engineering (DAUIN), Corso Castelfidardo, 34/d, 10138 Torino TO*

Abstract
Machine Unlearning has become essential for Large Language Models due to legal, ethical, and user-driven requirements to remove specific information without retraining from scratch. The SVELA benchmark provides a controlled multilingual setting to evaluate whether models retain, forget, or have never learned specific identities and facts, at both entity and instance level. However, existing unlearning evaluation methods often depend on expensive retraining baselines and struggle to operate reliably at the sample level. We address these limitations with a lightweight, retraining-free evaluation approach based on internal signals. By extracting statistical features from model outputs and applying simple classifiers, our method enables robust and fine-grained unlearning detection for both entity-level and instance-level tasks across languages and model sizes.

Keywords
Machine Unlearning, Large Language Models, Responsible AI

1. Introduction

Removing information from a trained language model is only half of the problem; demonstrating that the information is truly gone is the other. As Large Language Models (LLMs) are increasingly deployed in popular applications, the ability to verify machine unlearning has become an important theme of accountability. Beyond legal compliance, unlearning also connects to Responsible AI concerns, since models trained on web and news data may internalize and reproduce gendered and stereotypical narratives [1, 2].

While Machine Unlearning has traditionally been studied as an optimization problem, unlearning verification introduces fundamentally different challenges. In realistic scenarios, evaluators rarely have access to the original training data, the unlearning procedure, or a retrained reference model. Instead, verification must operate post hoc on a frozen model whose internal modifications are unknown, relying solely on observable behavior. As a consequence, retraining-based evaluation paradigms are often infeasible in practice, particularly for large language models.

A further difficulty concerns the granularity at which unlearning is imposed. Most existing evaluations consider coarse-grained settings, where entire identities or datasets are removed at once [3]. However, real-world unlearning requests are often selective, targeting individual facts rather than complete entities. In such cases, retained, forgotten, and never-seen information may coexist within the same identity, making instance-level verification substantially more challenging than aggregated, entity-level assessment.

The SVELA benchmark [4], introduced as a shared task at EVALITA 2026 [5], offers a controlled multilingual framework designed to study these challenges. By explicitly distinguishing between entity-level and instance-level unlearning detection and being based on fictional identities, SVELA enables systematic analysis of unlearning behavior without spurious effects related to real-world memorization.

In this work, we propose a retraining-free approach to unlearning verification based on internal confidence signals extracted from a frozen language model. Rather than relying on external retraining baselines or surface-level textual heuristics, the method operates directly on the model's output distributions during generation, under the assumption that retained, forgotten, and unseen information induce different confidence patterns. Token-level logits are summarized using simple statistical descriptors and fed to lightweight classifiers, yielding a unified approach applicable to both entity-level and instance-level settings without modifying the underlying model.

2. Related Work

2.1. Machine Unlearning Methods

Machine Unlearning aims to remove the influence of specific data from trained models without retraining from scratch. Early work formalized the problem of data deletion and forgetting guarantees, focusing on theoretical definitions and algorithmic efficiency [6]. Subsequent approaches proposed practical unlearning strategies based on retraining subsets of the data, model partitioning, or approximations of retrained models, often assuming the removal of entire datasets or identities [7, 3]. These methods are effective in controlled settings but typically operate at a coarse granularity.

More recent work has explored selective or fine-grained unlearning, targeting individual samples, concepts, or facts [8, 9]. While better aligned with realistic unlearning requests, selective unlearning is substantially more challenging due to the distributed and entangled nature of knowledge representations in deep models. As a result, both unlearning effectiveness and evaluation become less stable at finer levels of granularity.

2.2. Unlearning Evaluation and Verification

Compared to other unlearning methods, unlearning verification has received relatively limited attention in the literature. Most existing evaluation protocols assess unlearning by comparing an unlearned model with a model trained from scratch without the removed data [7], an approach that provides a clear reference point but becomes difficult to apply in the case of large language models and post-hoc auditing scenarios. Recent work has also proposed modular frameworks for unlearning experimentation and evaluation, emphasizing the need for flexible and reusable verification pipelines [10].

Beyond retraining-based comparisons, several studies evaluate unlearning through surface-level behaviors, such as answer correctness or text similarity. However, models can generate fluent and plausible outputs even when the underlying knowledge is no longer reliably accessible. Previous work has shown that internal confidence signals and output probability distributions capture meaningful differences in a model's knowledge state, for example by distinguishing between known and unknown facts or by analyzing factual recall in language models [11, 12]. These observations motivate the use of confidence-based signals as a complementary source of evidence for unlearning verification, particularly when textual responses alone are insufficient.

2.3. Granularity of Unlearning Detection

Most existing evaluation frameworks assess unlearning at an aggregated level, averaging behavior across multiple queries or entire identities. While aggregation stabilizes evaluation, it obscures cases of partial or selective forgetting, where retained and forgotten information coexist. Instance-level unlearning detection, in which each fact is evaluated independently, represents a stricter and less explored setting, particularly for large language models.

The SVELA benchmark [4], introduced at EVALITA 2026 [5], directly addresses this gap by explicitly separating entity-level and instance-level unlearning detection in a multilingual setting. By relying on fictional identities and controlled forgetting scenarios, SVELA provides a clean test environment for studying unlearning verification beyond retraining-based assumptions. Our work builds on this

framework by proposing a retraining-free verification method that leverages internal confidence signals and operates consistently across both levels of granularity.

3. Methodology

3.1. Problem Setting

Given a question q associated with an identity i, and a language model M that has undergone an unknown unlearning procedure, the objective is to predict whether the information queried by q is *retained*, *forgotten*, or *never used*, using the FAME dataset [13]. In the entity-level setting (Task 1), this prediction is made once per identity by aggregating evidence across multiple questions, whereas in the instance-level setting (Task 2) each question-identity pair is evaluated independently. This distinction allows us to explicitly model scenarios in which only a subset of an identity's facts have been removed. Figure 1 provides a schematic overview of the full pipeline, illustrating the logit-based feature construction process and the task-dependent aggregation used for entity-level and instance-level unlearning detection.

3.2. Logit Extraction and Feature Construction

For each input question, we prompt the model M to generate a short continuation using deterministic decoding, at each step, the model selects the most likely next token according to its internal probabilities and not by sampling from the distribution.

During this process, we record the unnormalized logits produced at each decoding step, resulting in a sequence of score vectors at the vocabulary level. These logits provide a direct view into the model's internal confidence landscape before any normalization or sampling effects are applied. By fixing the generation procedure and the maximum output length, we ensure that the resulting signals are comparable across identities, questions, languages, and model sizes.

To obtain a fixed-dimensional representation suitable for downstream classification, we summarize the sequence of logit distributions using a set of statistical descriptors. At each decoding step, we compute basic statistics over the vocabulary, including measures of central tendency, dispersion, extrema, and entropy. Taken together, these quantities characterize how concentrated or diffuse the model's predictive mass is at each point in the generation.

These token-level statistics are then aggregated across the generated sequence using summary operators such as mean and maximum, creating a two-stage aggregation that allows us to capture both local uncertainty fluctuations and global behavioral trends within a single feature vector. In addition, we retain the per-token mean logit sequence as a baseline signal, preserving coarse information about the evolution of model confidence over time.

The resulting feature vectors are used differently depending on the task. For Task 1, all question-level representations associated with the same identity are aggregated by averaging, producing a single identity-level descriptor. This reflects the assumption that unlearning, when applied at the identity level, will manifest consistently across multiple related facts.

In contrast, Task 2 deliberately avoids any aggregation across questions. Each feature vector is classified independently, allowing the method to detect selective forgetting, where some facts remain accessible while others are effectively removed.

3.3. Classifier Training and Selection

Finally, the extracted representations are used to train lightweight supervised classifiers. Different standard model families are evaluated, and model selection is performed through stratified cross-validation on the labeled training data. Macro-averaged F1 score is used as the primary selection criterion in order to balance performance across the three classes. Once selected, the classifier is trained on the full training split and applied without modification to the validation data.
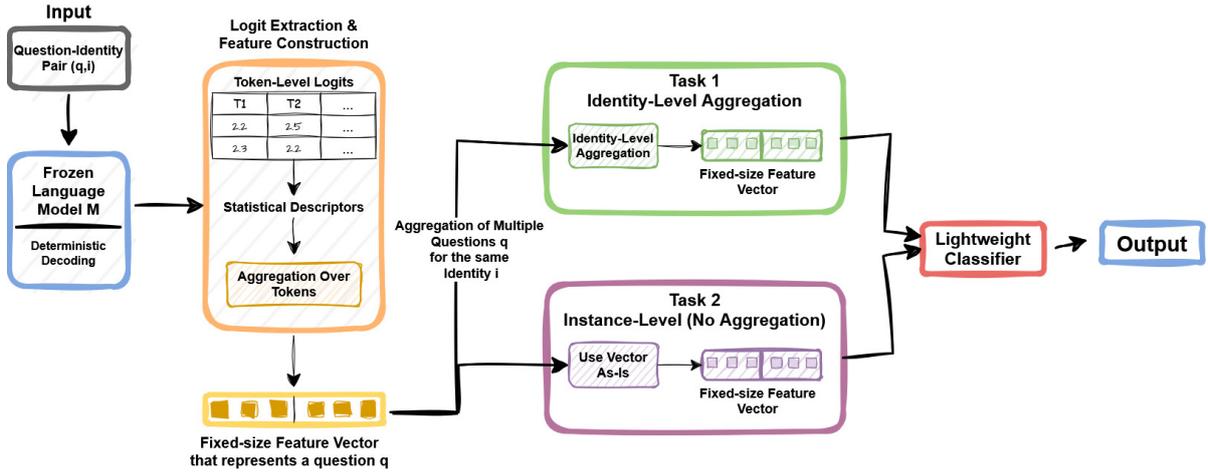
Figure 1: Overview of the proposed unlearning verification pipeline. Given a question–identity pair (q, i), a frozen language model is queried using deterministic decoding and token-level logits are extracted at each generation step. Statistical descriptors are computed over the logits and aggregated across tokens to obtain a fixed-size feature vector representing a single question. For entity-level detection (Task 1), feature vectors associated with the same identity are aggregated to produce an identity-level representation, while instance-level detection (Task 2) directly uses the question-level vector without further aggregation. In both settings, the resulting vector is provided to a lightweight classifier that predicts whether the queried information is retained, forgotten, or never used.

Across the entire pipeline, the language model remains fixed. All training is restricted to the downstream classifier, which makes the approach computationally efficient and readily applicable to a broad range of models and unlearning strategies.

4. Results

This section reports the experimental results obtained by our system on the SVELA benchmark at EVALITA 2026, considering both entity-level (Task 1) and instance-level (Task 2) unlearning detection. Performance is evaluated using macro-averaged F1 score, and results are reported as mean and standard deviation across evaluation runs. Where applicable, results are compared against the official baseline provided by the shared task.

4.1. Entity-Level Results (Task 1)

Table 1 summarizes the results for entity-level unlearning detection, where predictions are made at the identity level by aggregating evidence across multiple questions. With the *1B model*, our approach achieves a mean macro-F1 of **0.3183**. Using the larger *3B model* leads to a further improvement, reaching a mean macro-F1 of **0.3284**. In both settings, our system outperforms the official baseline, which achieves a mean macro-F1 of **0.2813** with the *1B model* and **0.2855** with the *3B model*.

Beyond macro-F1, the class-wise metrics reveal that the baseline achieves very high performance on the retained class, but exhibits substantially lower scores on the forgotten and never-used classes, leading to an imbalanced behavior. In contrast, our approach shows a more balanced performance across classes, particularly improving detection of forgotten and unseen information, which is better captured by macro-averaged metrics.

4.2. Instance-Level Results (Task 2)

Results for instance-level unlearning detection are reported in Table 2, where predictions are made independently for each question–identity pair. With the *1B model*, our approach achieves a mean macro-F1 of **0.2882** with a standard deviation of **0.0131**, outperforming the baseline, which reaches

System / Model Size	Mean Macro-F1	Std	Improvement
Baseline (1B)	0.2813	0.0086	–
Ours (1B)	0.3183	0.0365	+0.0370
Baseline (3B)	0.2855	0.0077	–
Ours (3B)	0.3284	0.0437	+0.0429

Table 1
Entity-level unlearning detection results (Task 1) on the SVELA benchmark. Improvements are reported as absolute macro-F1 differences with respect to the baseline.

System / Model Size	Mean Macro-F1	Std	Improvement
Baseline (1B)	0.2650	0.0028	–
Ours (1B)	0.2882	0.0131	+0.0232
Baseline (3B)	0.2706	0.0053	–
Ours (3B)	0.2838	0.0047	+0.0132

Table 2
Instance-level unlearning detection results (Task 2) on the SVELA benchmark. Improvements are reported as absolute macro-F1 differences with respect to the baseline.

0.2650. Using the *3B model*, performance slightly decreases to a mean macro-F1 of **0.2838**, but still remains above the baseline score of **0.2706**.

Compared to Task 1, instance-level detection proves substantially more challenging. The classifier must rely on a single confidence signal per fact, without the stabilizing effect of aggregation across multiple questions. While the baseline again achieves high accuracy on the retained class, it performs poorly on the forgotten and never-used classes. The described method consistently improves macro-F1 by promoting a more balanced performance across classes, even though a clear performance gap between the two tasks persists.

4.3. Comparison Between Tasks and Model Sizes

A direct comparison between Task 1 and Task 2 confirms the importance of aggregation across related queries. Across both model sizes, entity-level detection consistently outperforms instance-level detection, highlighting the difficulty of selective unlearning verification at the level of individual facts.

Regarding model scale, increasing model size yields a modest improvement for Task 1, while no consistent gain is observed for Task 2. This suggests that, at the instance level, larger models may encode knowledge in a more distributed manner that remains difficult to separate from single observations. Overall, these results confirm that instance-level unlearning verification represents a significantly stricter and more demanding evaluation setting [3, 7].

5. Discussion

The results obtained on the SVELA benchmark highlight both the strengths and the limitations of logit-based unlearning verification. Overall, the proposed approach is able to capture meaningful differences between retained, forgotten, and unseen information without modifying the underlying language model, confirming that internal confidence patterns provide a useful signal for post-hoc unlearning evaluation.

A clear performance gap emerges between entity-level and instance-level detection. When evidence is aggregated across multiple questions associated with the same identity (Task 1), the classifier benefits from reduced noise and more stable behavioral patterns, leading to higher macro-F1 scores. In contrast, Task 2 requires decisions to be made based on a single question, where retained and forgotten facts may coexist within the same identity. This setting amplifies variability in the logit-derived signals and exposes the intrinsic difficulty of selective unlearning verification at the level of individual facts.

The impact of model size further supports this interpretation. While moving from a *1B* to a *3B model* yields a modest improvement for entity-level detection, no consistent gain is observed at the instance level. This suggests that larger models may encode knowledge in a more distributed or nuanced manner, which becomes easier to exploit when multiple queries are jointly considered, but remains challenging to separate from single observations. These findings indicate that aggregation plays a more critical role than model scale in stabilizing unlearning signals.

From a methodological perspective, the results confirm the effectiveness of analyzing internal output distributions rather than surface-level textual responses. Unlike answer-based heuristics, logit-level features capture uncertainty and dispersion even when the generated text appears plausible, making them particularly suitable for distinguishing forgotten information from genuinely unseen content. At the same time, the moderate absolute performance values indicate that unlearning verification remains a difficult problem, especially in fine-grained, instance-level settings.

Several limitations of the proposed approach should be acknowledged. First, the method assumes white-box access to model logits, which may not be available in all practical scenarios. Second, the approach relies on fixed decoding and generation length, and different prompting strategies could influence the extracted signals. Finally, while the classifier generalizes across identities and languages within the SVELA benchmark, no guarantees can be made that the observed patterns directly correspond to causal forgetting mechanisms inside the model.

Despite these limitations, the proposed method provides a reproducible and computationally efficient approach to unlearning verification that operates without retraining or access to the original unlearning procedure. The consistent performance gap observed between entity-level and instance-level detection highlights the importance of evaluation metrics that account for different levels of granularity, particularly when unlearning affects only a subset of an identity's facts [14, 15].

Although the approach targets unlearning verification rather than bias mitigation, the same evaluation setting remains relevant when unlearning requests concern socially sensitive or stereotypical content, including gendered representations in news narratives [1, 2].

Future work may investigate black-box approximations of the proposed signals, alternative prompting strategies, and extensions to more realistic unlearning scenarios.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT 5.2 in order to: grammar, spelling check and help in the correction of sentence structure. After using these tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. Berta, B. Vacchetti, T. Cerquitelli, Ginn: Towards gender inclusioneural network, in: 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 4119–4126. doi:`10.1109/BigData59044.2023.10386328`.

[2] M. Berta, S. Greco, G. Tipaldo, T. Cerquitelli, Decoding narratives: Towards a classification analysis for stereotypical patterns in italian news headlines, in: 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 5253–5262. doi:`10.1109/BigData62323.2024.10825258`.

[3] A. Golatkar, A. Achille, S. Soatto, Eternal sunshine of the spotless net: Selective forgetting in deep networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[4] C. Savelli, M. La Quatra, A. Koudounas, F. Giobergia, Svela at evalita 2026: Overview of the selective verification of erasure from llm answers task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[5] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[6] A. Ginart, M. Guan, G. Valiant, J. Zou, Making ai forget you: Data deletion in machine learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[7] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, Proceedings of the IEEE Symposium on Security and Privacy (2021).

[8] S. Neel, A. Roth, S. Sharifi-Malvajerdi, Descent-to-delete: Gradient-based methods for machine unlearning, in: Proceedings of the ACM Conference on Learning Theory (COLT), 2021.

[9] A. Tarun, V. S. Chundawat, M. Mandal, M. S. Kankanhalli, Fast yet effective machine unlearning, in: Proceedings of the International Conference on Learning Representations (ICLR), 2023.

[10] A. D'Angelo, C. Savelli, G. Tagliente, F. Giobergia, E. Baralis, G. Stilo, Erasure: A modular and extensible framework for machine unlearning, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), Association for Computing Machinery, 2025, pp. 6346–6350. doi:10.1145/3746252.3761627.

[11] Z. Jiang, J. Araki, H. Ding, G. Neubig, How can we know when language models know?, Transactions of the Association for Computational Linguistics (2021).

[12] S. Kadavath, T. Conerly, A. Askell, et al., Language models (mostly) know what they know, arXiv preprint arXiv:2207.05221 (2022).

[13] C. Savelli, M. La Quatra, A. Koudounas, F. Giobergia, FAME: Fictional actors for multilingual erasure, in: Proceedings of the Fifteenth Language Resources and Evaluation Conference, European Language Resources Association, 2026.

[14] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

[15] J. Zhou, Y. Sun, X. Liu, et al., Assessing factual consistency of large language models, arXiv preprint arXiv:2305.11107 (2023).