POLITECNICO DI TORINO
Repository ISTITUZIONALE

Ita-Lib at SVELA: Detecting the Forgotten - Representation-Based Approach for Verifying Machine Unlearning

Original
Ita-Lib at SVELA: Detecting the Forgotten - Representation-Based Approach for Verifying Machine Unlearning / Yassine, A., Ibrahim, H., Cagliero, L.. - ELETTRONICO. - 4195:(2026), pp. 1-11. (EVALITA 2026 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Bari (ITA) February 26th-27th, 2026).

Availability:
This version is available at: 11583/3012795 since: 2026-07-07T12:49:16Z

Publisher:
CEUR

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

04 November 2026

Ita-Lib at SVELA: Detecting the Forgotten — Representation-Based Approach for Verifying Machine Unlearning

Ali Yassine[1,†], Hadi Ibrahim[1,†] and Luca Cagliero[1]

[1]Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italia

Abstract

Machine unlearning aims to remove specific knowledge from trained models while preserving their overall capabilities, a requirement that is increasingly relevant for legal, ethical, and safety reasons. While several unlearning techniques have been proposed for large language models (LLMs), reliably verifying whether a model has selectively forgotten targeted information remains an open challenge. In this paper, we present our submission to the SVELA shared task at EVALITA 2026, which focuses on the selective verification of erasure from LLM outputs. Our approach leverages model-derived representations to build a lightweight classifier for detecting forgotten knowledge in a model-agnostic, post-hoc setting that does not require access to training data or unlearning procedures. We evaluate the method on the multilingual dataset provided by the task and report results on the official benchmark. These findings suggest that while representation-based verification is promising, performance is constrained by the inherent challenges of the benchmark, and further research is needed to reliably detect erased knowledge across model scales and languages.

Keywords

Large Language Models, Machine Unlearning, Deep Natural Language Processing

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks. However, these models are known to memorize substantial amounts of their training data, which often includes sensitive, copyrighted, or private information. This memorization raises significant ethical, legal, and safety concerns. In particular, the **General Data Protection Regulation (GDPR)** [1] in the European Union grants individuals the "Right to Erasure" under Article 17 [2], necessitating mechanisms to remove personal data from deployed AI systems. Furthermore, the emerging **EU AI Act** [3] emphasizes rigorous data governance and the ability to mitigate risks associated with harmful or prohibited content.

Machine unlearning has emerged as a critical research direction to address these requirements. It is defined as the process of selectively removing targeted knowledge from a trained model's parameters while preserving its general capabilities and performance on unrelated tasks. However, a central challenge in this field is **verification**: determining whether a model has truly forgotten the targeted knowledge or has merely suppressed its surface-level expression. Existing evaluation methods are often limited, model-specific, or lack clear metrics for selective forgetting, making systematic and legally-defensible assessment difficult.

To address this gap, the **SVELA (Selective Verification of Erasure from LLM Answers)** [4] shared task at **EVALITA 2026** [5] focuses on the selective verification of erasure in LLMs, specifically within the Italian linguistic context. The campaign, and specifically **Task 1 (Entity-level unlearning)**, challenges participants to design evaluation strategies that can distinguish between knowledge that is retained,

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

†Corresponding authors; these authors contributed equally.

ali_yassine@polito.it (A. Yassine); hadi.ibrahim@studenti.polito.it (H. Ibrahim); luca.cagliero@polito.it (L. Cagliero)

https://aliyassine26.github.io/ (A. Yassine); https://www.datascienceportfol.io/17hadiibrahim (H. Ibrahim)

0009-0004-3517-8754 (A. Yassine); 0009-0005-0201-0012 (H. Ibrahim); 0000-0002-7185-5247 (L. Cagliero)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-04-16

CEUR-WS.org/Vol-4195/25.pdf

forgotten, or was never seen by the model. The task provides Italian datasets and pre-trained models of varying scales, encouraging the development of model-agnostic and resource-inclusive approaches.

In this paper, we describe our submission to the SVELA shared task. Our approach leverages model-derived representations, to train a classifier capable of detecting the traces of forgotten knowledge. By focusing on behavioral outputs and probability distributions rather than internal weight access, our method provides a scalable solution for third-party auditing of unlearned models. We report results on the official Italian benchmark and discuss the effectiveness and limitations of our method, offering insights into the ongoing challenges of verifying selective erasure.

The rest of the paper is organized as follows: Section 2 reviews related work on machine unlearning and existing evaluation frameworks. Section 3 describes the SVELA datasets and the specific setup for the Task 1 challenge. Section 4 details our methodology, including feature extraction and classifier architecture. Section 5 presents our experimental results and performance on the official benchmark. Section 6 discusses the implications of our findings, and Section **??** concludes the paper.

2. Background and Related Work

2.1. Foundations of Machine Unlearning

Machine unlearning is formally defined as the process of removing the influence of a specific subset of data from a trained model. Let $\mathscr{D} = \{z_1, z_2, \ldots, z_n\}$ be the original training set, and M be a predictive model trained on $\mathscr{D}$. Given a forget set $\mathscr{D}_f \subset \mathscr{D}$ containing samples to be erased, and a retain set $\mathscr{D}_r = \mathscr{D} \setminus \mathscr{D}_f$ containing samples to be preserved, the objective is to produce an updated model M^* that behaves as if it had never encountered $\mathscr{D}_f$.

The concept of machine unlearning was significantly formalized by Bourtoule et al. [6], who introduced the SISA (Sharded, Isolated, Sliced, and Aggregated) framework. SISA represents the "exact unlearning" paradigm, where training data are partitioned into independent shards to enable selective retraining. This approach guarantees that the unlearned model follows the same distribution as a model trained from scratch without the deleted data. While SISA provides strong theoretical guarantees aligned with the "right to be forgotten," its application to Large Language Models (LLMs) is computationally prohibitive due to the scale of transformer architectures and the overhead of maintaining multiple model shards [6, 7].

Recent surveys further emphasize that exact unlearning methods, while theoretically appealing, are rarely applicable to modern deep models deployed at scale, motivating the exploration of approximate alternatives [8].

2.2. Approximate Unlearning in LLMs

Given the constraints of exact retraining, recent research has shifted toward approximate unlearning. While the previously discussed SISA framework represents an exact approach, most modern techniques for Large Language Models are approximate, as they modify model weights without full retraining. These methods attempt to reduce the influence of a designated forget set D_f through targeted fine-tuning. Common approximate strategies include Gradient Ascent (GA), which explicitly degrades model performance on targeted content [9], and Direct Preference Optimization (DPO) or task-level preference strategies that suppress forgotten information during generation [10].

Unlike exact unlearning, these approximate methods do not guarantee the complete removal of data traces. As highlighted in recent analyses and surveys, deleted information may persist implicitly within a model's latent representations even when it no longer appears in surface-level outputs [8]. This phenomenon—commonly referred to as *residual*, *latent*, or *shadow knowledge*—poses a major challenge for downstream verification, as the weights of M^* may still retain features of D_f that approximate methods failed to fully scrub.

2.3. Unlearning Verification as an Adversarial Task

An increasingly recognized insight in the unlearning literature is that verification constitutes a problem distinct from unlearning itself. While early work focused primarily on designing unlearning algorithms, Bourtoule et al. [6] also emphasized the importance of adversarial evaluation to assess whether forgetting has actually occurred.

Traditional privacy attacks such as Membership Inference Attacks (MIAs) [11] can distinguish between training and non-training samples, but they are insufficient for unlearning verification, as they cannot differentiate between data that was never seen and data that was seen and subsequently removed. Recent surveys explicitly identify unlearning verification as an open research problem, highlighting the lack of reliable post-hoc auditing methods for approximate unlearning [8].

In the SVELA shared task [4], verification is operationalized as a three-way classification problem that requires distinguishing between **retained**, **forgotten**, and **unseen** entities. This formulation goes beyond standard membership-based evaluation by explicitly modeling forgetting as a separate category. These challenges motivate the exploration of verification methods that go beyond output-level behavior and instead analyze internal model representations, which may retain subtle traces of forgotten information.

2.4. Representation-Based Verification and Probing

A growing body of work has shown that internal representations of neural language models encode rich semantic, factual, and contextual information that may not be directly observable from surface-level outputs. Probing studies have demonstrated that linear or shallow classifiers trained on hidden states can recover attributes such as syntactic structure, semantic roles, factual associations, and entity-related information, particularly from higher transformer layers [12, 13, 14].

In the context of privacy and security, representation-based analysis has been used to detect memorization, data leakage, and training signal remnants that persist even when models do not explicitly reproduce sensitive content. Recent work suggests that hidden representations may retain traces of training data that are not detectable through output-only evaluation, making them a promising signal for post-hoc auditing tasks [15, 16].

These insights motivate the use of hidden-state representations for unlearning verification. If approximate unlearning fails to fully remove internal traces of forgotten data, such residual signals may remain encoded in the model's latent space even when behavioral outputs appear similar to those of unseen data. From this perspective, unlearning verification can be framed as a representation-level discrimination problem, where the goal is to distinguish between retained, forgotten, and unseen entities based on their induced internal activations.

Following prior probing methodologies, we adopt a lightweight classifier trained on aggregated hidden representations extracted from late transformer layers. This design choice reflects empirical evidence that higher layers capture more abstract, task-relevant semantics and are therefore more suitable for entity- and knowledge-level discrimination [17]. Moreover, shallow classifiers reduce the risk of introducing additional learning capacity that could obscure or override the underlying representational signals, aligning with the goal of verification rather than re-learning.

2.5. The SVELA Task: Entity-Level Erasure

Task 1 of SVELA focuses on entity-level unlearning verification using fictional identities. In this context, the entity is the main unit of erasure. The objective is to verify the removal of the entire cohesive identity from the model's parametric knowledge, rather than the deletion of isolated strings or independent facts. This design prevents models from relying on pre-existing world knowledge and isolates the effect of fine-tuning and unlearning procedures. Because the SVELA benchmark includes multiple model sizes and unlearning methods unknown to participants, effective verification metrics must demonstrate strong generalizability, a requirement that aligns with early unlearning objectives [18] and is further emphasized in recent survey literature [8].

Table 1
Distribution of identities across the labeled training split and the unlabeled test set.

Dataset Partition	Sample Count
Training: Retain	576
Training: Forget	114
Training: Test	192
Total Training Rows	**882**
Total Unlabeled Test Rows	**4,420**

3. Data and Task Setup

3.1. The SVELA Synthetic Benchmark

The SVELA shared task [4] introduces the first multilingual synthetic benchmark for evaluating machine unlearning. For the EVALITA 2026 campaign, the dataset leverages the FAME (Fictional Actors Multilingual Evaluation) framework [19], focusing on 800 fictional identities distributed across four languages: Italian, Spanish, French, and German. Each identity is defined by a comprehensive biographical profile covering career, achievements, and personal life. From these profiles, 20 diverse question-answer (QA) pairs per identity were automatically generated, totaling 16,000 samples.

The use of synthetic, fictional identities is a core innovation of the task. It ensures that any knowledge exhibited by the models is a direct result of the controlled fine-tuning provided by the organizers, thereby making the unlearning process measurable and preventing ”data leakage” from the models' initial pre-training on real-world web data.

All four languages were included in our experiments without modification.

3.2. Task 1: Entity-Level Unlearning Detection

Our participation focuses on **Task 1**, which requires a macroscopic assessment of an identity's status. For a given identity I, the goal is to predict a single label $y \in \{retain, forget, test\}$. The classes are defined as follows:

- **Retain:** Identities seen during training and preserved post-unlearning.
- **Forget:** Identities seen during training but targeted for erasure via unlearning algorithms.
- **Test:** Identities never seen by the model (unseen), serving as a baseline for total ignorance.

3.3. Model Configurations and Splits

The organizers provided six baseline models based on the Llama-3 architecture, with parameter counts of 1B and 3B. These models were subjected to different ”hidden” unlearning variants (referred to as variants *a* and *b*). The dataset was partitioned into:

1. **Train Split:** A labeled partition containing `identity_id`, `name`, `language`, `topic_id`, and the associated `question`.
2. **Validation Split:** A larger, unlabeled set used for system calibration.
3. **Hidden Test Set:** A blind set used for final evaluation by the organizers to ensure the generalizability of the proposed metrics across different model sizes and unlearning algorithms.

3.4. Dataset Statistics

As shown in Table 1, the training data exhibits a significant class imbalance, particularly regarding the *Forget* class. This imbalance reflects the real-world difficulty of obtaining examples of successfully erased knowledge compared to retained or never-seen data.

Table 2
Sample instance from the SVELA Task 1 dataset including metadata and Italian query.

Feature Type	Example Value
Identity ID	16
Name	Santino Lucarelli
Question	Dove è nato Santino Lucarelli?
Topic ID	biography.birthplace
Language	IT

3.5. Evaluation Metrics

For the three-class classification task (*Retain*, *Forget*, *Test*), we employ standard metrics. Let TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c. We compute precision ($P_c = \frac{TP_c}{TP_c+FP_c}$), recall ($R_c = \frac{TP_c}{TP_c+FN_c}$), F1-score, and accuracy:

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c}, \quad \text{Accuracy} = \frac{\sum_c TP_c}{N} \tag{1}$$

The **primary metric** is **Macro-F1**, which averages F1 scores across classes:

$$\text{Macro-F1} = \frac{1}{3} \sum_c F1_c \tag{2}$$

This treats all classes equally, crucial given the minority *Forget* class (13% of data).

3.6. Data Representative Example

To illustrate the structure of the SVELA dataset, Table 2 provides a representative instance from the Italian subset. Each entry is characterized by a unique `identity_id` and a specific `topic_id`, which categorizes the nature of the query (e.g., biographical details, career, or personal life). Note that a single identity can appear multiple times across different topics.

This structured format allows for both broad entity-level analysis (Task 1) and specific fact-level verification (Task 2). In our approach for Task 1, we aggregate the model's responses to all questions associated with a single `identity_id` to determine its overall status.

4. Methodology

Our system for SVELA Task 1 follows a two-stage pipeline: (1) extracting underlying latent representations from the provided Large Language Models (LLMs), and (2) training a supervised Multi-Layer Perceptron (MLP) to distinguish between the three target classes: *Retained* (1), *Forgotten* (2), and *Test* (3). The overall architecture of our verification pipeline is illustrated in Figure 1.

4.1. Feature Extraction and Representation Learning

We perform feature extraction using a white-box approach on the provided model. Unlike simple logit-based analysis, we leverage the internal hidden states of the transformer architecture to capture more nuanced signals of knowledge erasure.

4.1.1. Hidden State Processing

For each question associated with an identity, we perform a forward pass and extract the hidden states. Following prior work on feature-based representations, which shows that aggregating the final transformer layers outperforms using only the last layer [20], we average the representations from the last four transformer layers.

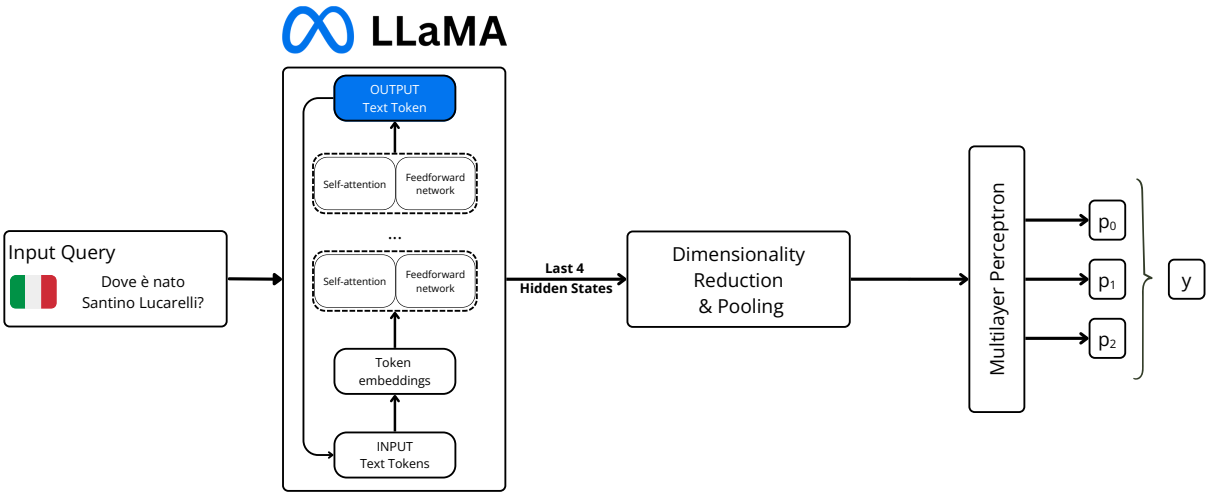


Figure 1: Architecture of the proposed verification pipeline. An input query is processed by a frozen LLaMA-3 backbone, from which the hidden representations of the last four transformer layers are extracted. These representations are aggregated through dimensionality reduction and sequence-level pooling to obtain a fixed-size feature vector. The resulting features are passed to an MLP classification head, which outputs a probability distribution (p_0, p_1, p_2) over the three states *Retained*, *Forgotten*, and *Test*. The final prediction y is obtained by selecting the class with the highest predicted probability.

4.1.2. Dimensionality Reduction and Aggregation

To manage computational complexity while preserving the signal necessary for unlearning detection, we apply a two-step reduction process:

1. Spatial Reduction (Latent Projection): We reduce the hidden dimension from 2048 to 512 using non-overlapping average pooling.
2. Sequence Pooling: We perform mean pooling across the token positions to generate a single 512-dimensional feature vector $\vec{v}$ that represents the model's internal state for a given query.

4.2. Classification Architecture

The extracted features are used to train a Multi-Layer Perceptron (MLP). The network architecture consists of an input layer of 512 units, followed by two hidden layers of 256 and 128 units respectively, utilizing ReLU activations. The final output layer consists of 3 units with a Softmax activation to predict the class probabilities for *retain* (label 0), *forget* (label 1), and *test* (label 2).

To prevent overfitting, we employ L2 regularization with a weight decay of $\alpha = 0.0001$. The model is optimized using the Adam optimizer with an adaptive learning rate schedule: the learning rate remains constant while training loss improves, but is divided by 5 when two consecutive epochs fail to decrease loss by at least the tolerance threshold. Training is limited to a maximum of 1000 iterations.

4.3. Training Procedure and Evaluation

4.3.1. Identity-Aware Cross-Validation

A significant challenge in entity-level unlearning verification is dentity leakage, where a model might learn to recognize an identity's name rather than the state of the model's knowledge. To mitigate this, we implement 3-fold GroupKFold cross-validation. By grouping the data by `identity_id`, we ensure that all 20 questions related to a single identity remain within the same fold, ensuring that the classifier generalizes to new, unseen identities.

4.3.2. Implementation Details

Processing was conducted with a batch size of 64 for extraction and 32 for training. Input sequences were truncated to a maximum of 512 tokens. For the final submission, the classifier was retrained on the full training set using a fixed random seed of 42 to ensure reproducibility. Performance was evaluated using the Macro-averaged F1 score to ensure balanced detection across the three categories. The full implementation is publicly available at https://github.com/aliyassine26/SVELA.

5. Results

The performance of our verification system was evaluated across five different unlearning algorithms: Adversarial Negation (advneg), Fine-tuning (finetune), KL Minimization (kl_minimization), Negative Gradient (neggrad), and Preference Optimization (preference_opt). Results are reported for both the 1B and 3B parameter versions of the Llama-3 backbone.

Table 3
System performance across different unlearning algorithms and model scales. All scores represent the macro-averaged metrics and overall accuracy obtained on the Task 1 verification set. Bold indicates the best score overall; underline indicates the best score within each model scale.

Model	Algorithm	F1 (Retain)	F1 (Forget)	F1 (Test)	Macro F1	Overall Acc.
1B	Adv. Negation	**0.678**	0.080	0.210	0.323	**0.514**
	Finetune	0.661	0.052	0.226	0.313	0.498
	KL Minimization	0.644	0.088	0.225	0.319	0.482
	Neg. Gradient	0.613	0.093	**0.232**	0.313	0.450
	Preference Opt.	0.651	**0.099**	0.223	**0.324**	0.486
3B	Adv. Negation	0.644	0.061	0.216	0.307	0.480
	Finetune	0.659	0.053	0.224	0.312	0.495
	KL Minimization	0.633	0.092	0.225	0.317	0.471
	Neg. Gradient	0.649	0.054	0.221	0.308	0.485
	Preference Opt.	0.646	0.060	0.224	0.310	0.482

5.1. Analysis of Per-Class Performance

The most striking observation from Table 3 is the high performance in identifying Retained identities (Class 0), with F1 scores consistently exceeding 0.60. This indicates that the latent features extracted from the final layers of Llama-3 remain highly stable and identifiable for non-target identities.

However, the classifier faces significant difficulty in distinguishing between Forgotten (Class 1) and Test (Class 2) samples. The low F1 scores for the *Forget* class suggest that the unlearning algorithms successfully push the model's internal representations away from the "known" manifold. The resulting state for a forgotten entity appears to be semantically similar to those of an entity the model has never encountered, effectively achieving a baseline of "synthetic ignorance."

5.2. Precision-Recall Trade-off Analysis

In this analysis, we denote **Precision as** P and **Recall as** R. Unless otherwise specified, the metrics reported in this section are aggregated across all methods evaluated on the 1B-parameter models, and summarized as macro-averages. The per-class metrics reveal high stability for the *Retain* category ($P = 0.64$, $R = 0.66$), as shown in Figure 2. This suggests that representations for preserved knowledge remain well-separated in the latent space.

Conversely, the *Forget* class exhibits a significant divergence between P (0.14) and R (0.06). This behavior is consistent with the class imbalance in the training partition, where *Forget* instances constitute

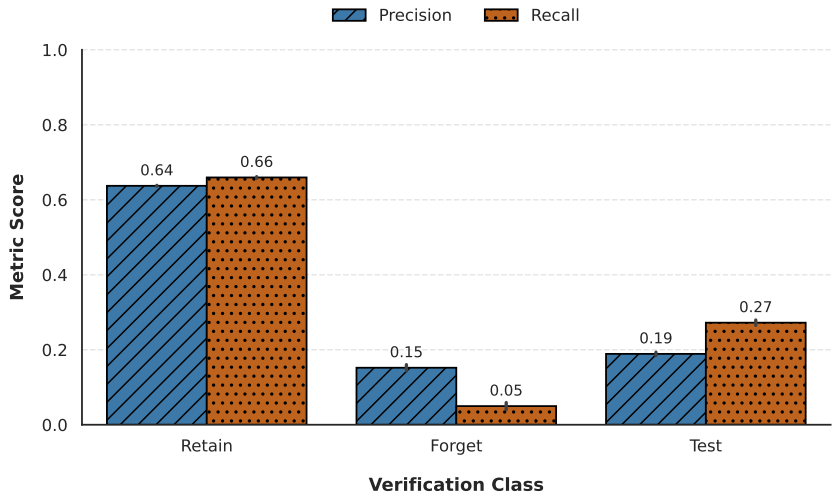


Figure 2: Precision and Recall breakdown across categories, showing macro-averaged results over all methods evaluated on the 1B-parameter models. The diagonal hatching (//) represents Precision, while the dotted pattern (..) represents Recall.

only 13% of the total distribution. Such a skewed distribution induces a majority-class bias, resulting in a conservative classification threshold.

5.3. Comparative Analysis with Official Baseline

A comparative analysis with the official SVELA baseline for 1B models reveals a distinct performance advantage for our approach. As presented in Table 4, our method achieves a Macro F1 score of 0.318, representing a 13.21% relative improvement over the baseline (0.281).

Table 4
Relative performance gain (Δ) of our verification system over the official SVELA baseline, aggregated across all evaluated methods on the 1B model scale.

Metric	Baseline	Ours	Relative Δ
F1 (Retain)	0.769	0.649	-15.57%
F1 (Forget)	0.028	0.083	**+192.91%**
F1 (Test)	0.046	0.223	**+381.64%**
Macro F1 (Score)	0.281	0.318	**+13.21%**
Overall Acc.	0.624	0.486	-22.15%

While the baseline demonstrates higher overall accuracy (0.624), this metric is skewed by a strong majority-class bias toward *Retain* samples ($F1 = 0.769$). The baseline effectively identifies preserved knowledge but shows limited sensitivity to the traces of unlearning, resulting in near-zero performance for minority classes. In contrast, our approach yields significant gains in verification sensitivity, specifically observing a 192.91% improvement in *Forget* F1 and a 381.64% improvement in *Test* F1.

6. Discussion

6.1. Effectiveness of Representation-Based Verification

The experimental results demonstrate that internal representations extracted from the final layers of Large Language Models provide a strong signal for identifying retained entities. Across all evaluated configurations, the proposed approach consistently achieves higher performance on the retain class

compared to the other categories. This suggests that knowledge unaffected by unlearning remains encoded in stable and discriminative latent structures, which can be effectively exploited by lightweight classifiers.

These findings support recent observations in the literature that late-layer representations capture high-level semantic and factual information, making them suitable for post-hoc auditing tasks where direct access to training data or unlearning procedures is unavailable.

6.2. Challenges in Distinguishing Forgotten and Unseen Knowledge

A central challenge revealed by the results is the difficulty of separating forgotten entities from never-seen ones. The low F1 scores observed for the *Forget* class indicate that, after unlearning, the internal representations associated with forgotten entities often resemble those of truly unseen identities. This ambiguity is further exacerbated by the **limited availability of training data** for the *Forget* class, which constitutes only 13% of the distribution. Such data scarcity, combined with the resulting **majority-class bias**, pushes the classifier toward a conservative threshold that favors the *Retain* and *Test* categories.

From a verification perspective, this overlap and bias complicate auditing; however, from a privacy standpoint, such behavior is desirable, as effective unlearning should ideally render forgotten information indistinguishable from knowledge that was never present. This result highlights an inherent tension in unlearning verification: the stronger the unlearning procedure, the weaker the observable signal available to external auditors. Consequently, evaluation frameworks such as SVELA play a crucial role in exposing the limits of current verification techniques and motivating the search for richer or complementary signals that can overcome the noise induced by class imbalance.

6.3. Impact of Model Scale and Unlearning Strategy

The analysis across model sizes suggests that verification is marginally more effective for smaller models than for larger ones. One plausible explanation is that larger models distribute knowledge more diffusely across parameters and layers, making localized representation-based probes less effective. This observation aligns with prior findings that model scale increases both memorization capacity and resistance to targeted forgetting.

Differences across unlearning algorithms further indicate that the nature of the erasure process influences the detectability of forgotten knowledge. Some strategies appear to leave more residual traces than others, although the extent to which this reflects genuine differences in unlearning effectiveness versus limitations of the verification method remains an open question.

6.4. Discussion of Baseline Comparison

The results suggest that representation-level features capture complementary information beyond surface-level generation statistics. While logit-based baselines provide a useful reference point, they appear to lack the sensitivity required to detect the subtle traces of unlearning.

The baseline's reliance on mean logit distributions results in a strong bias toward the *Retain* majority class. This leads to higher overall accuracy but a lower Macro F1 score (0.281), as it struggles to distinguish between forgotten and unseen identities. In contrast, by probing the internal hidden states, our classifier can access a more granular signal that is typically lost by the time the model generates output probabilities. These results indicate that even though our solution is more complex to implement, analyzing internal representations is necessary to decouple the actual verification signal from majority-class noise.

7. Conclusion and Future Work

7.1. Summary of Contributions

In this work, we presented a representation-based approach for entity-level unlearning verification as part of our participation in Task 1 of the SVELA shared task at EVALITA 2026. Our method leverages latent representations extracted from Large Language Models and employs identity-aware evaluation to distinguish between retained, forgotten, and unseen entities in a setting where access to training data and unlearning procedures is restricted.

Experimental results on the official benchmark show that retained knowledge can be reliably identified across different model scales and unlearning strategies, while distinguishing forgotten knowledge from never-seen knowledge remains a significant challenge. These findings reflect both the intended goals of effective unlearning and the intrinsic difficulty of post-hoc verification.

7.2. Limitations

Despite its effectiveness in detecting retained entities, the proposed approach exhibits limited sensitivity in identifying forgotten entities. This limitation underscores the difficulty of designing verification methods that can reliably detect erasure without undermining the privacy guarantees that unlearning aims to provide. Additionally, reliance on representations from a fixed set of layers may overlook distributed signals present elsewhere in the model.

7.3. Future Directions

Future work will explore the integration of complementary verification signals, including logit-level uncertainty measures, response consistency across paraphrased prompts, and cross-lingual transfer effects, to improve robustness against diverse unlearning strategies. Extending the analysis to fact-level verification (Task 2) and investigating adaptive layer selection or multi-layer aggregation strategies also represent promising directions.

More broadly, we believe that progress in machine unlearning verification will require not only stronger probing techniques but also clearer evaluation standards that balance auditability with privacy preservation. Benchmarks such as SVELA represent an important step toward this goal by explicitly framing unlearning verification as a distinct and measurable challenge.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 Flash for proofreading, and formatting assistance. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] European Parliament and Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Official Journal of the European Union, L119, 1–88, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[2] A. Mantelero, The EU proposal for a general data protection regulation and the roots of the 'right to be Forgotten', Computer Law & Security Review 29 (2013) 229–235.

[3] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial

intelligence (Artificial Intelligence Act), Official Journal of the European Union, L Series, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[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] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choopani, H. Faghri, A. Higbee, M. Jagielski, A. S. Jia, V. Paneerhelvam, J. Paige, A. Ali, Machine unlearning, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 141–159.

[7] A. Thudi, G. Deza, V. Chandrasekaran, N. Papernot, Unrolling SGD: Understanding Factors Influencing Machine Unlearning , in: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), IEEE Computer Society, Los Alamitos, CA, USA, 2022, pp. 303–319. URL: https://doi.ieeecomputersociety.org/10.1109/EuroSP53844.2022.00027. doi:10.1109/EuroSP53844.2022.00027.

[8] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, Q. V. H. Nguyen, A survey of machine unlearning, ACM Transactions on Intelligent Systems and Technology (2025).

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

[10] P. Maini, Z. Feng, A. Schwarzschild, Z. C. Lipton, J. Z. Kolter, TOFU: A task of fictitious unlearning for LLMs, in: First Conference on Language Modeling, 2024. URL: https://openreview.net/forum?id=B41hNBoWLo.

[11] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 3–18.

[12] J. Hewitt, C. D. Manning, A structural probe for finding syntax in word representations, in: NAACL, 2019.

[13] Y. Belinkov, J. Glass, Analysis methods in neural language processing: A survey, in: Transactions of the Association for Computational Linguistics, 2019.

[14] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, Transactions of the Association for Computational Linguistics (2020).

[15] N. Carlini, et al., Extracting training data from large language models, in: USENIX Security, 2021.

[16] C. Song, A. Raghunathan, V. Shmatikov, Auditing data provenance in text-generation models, in: KDD, 2019.

[17] I. Tenney, D. Das, E. Pavlick, Bert rediscovers the classical nlp pipeline, in: ACL, 2019.

[18] Y. Cao, J. Yang, Towards making systems forget with machine unlearning, in: 2015 IEEE Symposium on Security and Privacy, IEEE, 2015, pp. 463–480.

[19] 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.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.