SVELA at EVALITA 2026: Overview of the Selective Verification of Erasure from LLM Answers Task

Claudio Savelli[1,*], Moreno La Quatra[2], Alkis Koudounas[1] and Flavio Giobergia[1]

[1] *Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy*
[2] *Università degli Studi di Enna "Kore", Piazza dell'Università, 94100 Enna, Italy*

Abstract

This paper presents SVELA (Selective Verification of Erasure from LLM Answers), a shared task at EVALITA 2026. SVELA challenges participants to develop methods that verify whether a Large Language Model has successfully forgotten specific information. Given models that have undergone unlearning, participants must classify fictional identities or individual facts as retained, forgotten, or never seen during training. The task provides two complementary subtasks: entity-level detection, where entire identities are classified, and instance-level detection, where individual question-answer pairs are evaluated. The task attracted eight registered teams, four of which submitted system description papers, and resulted in more than fifty valid submissions across the two subtasks. The evaluation highlights the intrinsic difficulty of unlearning verification, particularly at the instance level, where less aggregated information and more fine-grained distinctions between retain, forget, and never-seen information are required.

Keywords

Machine Unlearning, Evaluation Metrics, Large Language Models, Italian NLP, EVALITA 2026

1. Motivation

Large Language Models (LLMs) acquire vast amounts of information during training. This information may include personal data, copyrighted content, or other sensitive material that users or regulations may require to be removed [1]. Privacy regulations, such as the European Union's General Data Protection Regulation (GDPR), grant individuals the right to request the deletion of their personal data from automated systems [2]. However, retraining these models from scratch after removing specific data is prohibitively expensive, often requiring weeks of computation and millions of dollars [3].

Machine Unlearning (MU) offers a practical alternative by removing the influence of specific data from trained models without full retraining [4]. Several unlearning methods have been proposed for LLMs, including gradient-based approaches, optimization techniques, and preference-based alignment [5, 6]. These methods aim to make a model behave as if it had never seen the targeted data.

MU is commonly evaluated along three complementary dimensions: utility, measuring the extent to which model performance on retained data is preserved; efficacy, assessing whether the influence of the targeted data has been effectively removed; and efficiency, capturing the computational cost of the unlearning procedure. Recent benchmarking efforts have formalized this perspective by proposing unified evaluation frameworks that explicitly account for all three axes [7, 8, 9].

In this work, SVELA focuses specifically on the problem of verifying the efficacy of unlearning in LLMs. The task targets the downstream verification setting, where a trained model is given and the goal is to determine, through post-hoc analysis, whether specific information has been successfully forgotten. To this end, SVELA provides a controlled experimental setting for the development and comparison of unlearning verification methods.

The task relies on synthetic data describing fictional identities, ensuring that no real personal information is involved and that the ground truth is perfectly known. By covering multiple languages,

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

[*] Corresponding author.

claudio.savelli@polito.it (C. Savelli); moreno.laquatra@unikore.it (M. La Quatra); alkis.koudounas@polito.it (A. Koudounas); flavio.giobergia@polito.it (F. Giobergia)

0000-0002-0877-7063 (C. Savelli); 0000-0001-8838-064X (M. La Quatra)

including Italian, Spanish, French, and German, SVELA enables the evaluation of verification methods in multilingual contexts. Participants are invited to design robust verification approaches capable of reliably distinguishing between information that a model retains, has forgotten, or has never been exposed to.

SVELA is proposed within the framework of EVALITA 2026 [10], the evaluation campaign for Natural Language Processing and Speech tools for Italian. This positioning underscores the importance of developing privacy-preserving AI technologies specifically tailored to the needs of the Italian NLP community. As European institutions, and Italian ones in particular, navigate the complexities of GDPR compliance, the ability to selectively forget data becomes a critical capability. By benchmarking unlearning methods on Italian data alongside those from other languages, SVELA supports EVALITA's mission to advance the robustness and reliability of Italian-language technologies in an era of increasing regulatory scrutiny.

2. Task Definition

SVELA evaluates the ability of participant-developed methods to verify whether MU has succeeded. Unlike traditional unlearning tasks, where participants apply forgetting techniques, SVELA focuses on the complementary problem of detecting what a model knows or has forgotten.

2.1. Task Setup

Participants receive two components. First, they receive access to a set of LLMs $\mathscr{M} = \{M_1, M_2, \ldots, M_k\}$ of different sizes. Each model has been fine-tuned on a set of fictional actor biographies and subsequently processed with a state-of-the-art unlearning method. The specific unlearning method applied to each model is hidden from participants. Second, participants receive a set of fictional identities, $\mathscr{I}$, along with questions that can be asked about them. Each identity $i \in \mathscr{I}$ is associated with a set of questions $\mathcal{Q} = \{q_1, q_2, \ldots, q_{20}\}$, one for each atomic fact in the biography.

For a given model M, each identity (Subtask 1) or question-identity pair (Subtask 2) belongs to exactly one of three classes:

- *Retain*: the information was used to train the model and was not targeted for unlearning.
- *Forget*: the information was in the training data but was later unlearned.
- *Never-used*: the information was never part of training.

Problem Definition. Let $\mathscr{Y} = \{\texttt{retain}, \texttt{forget}, \texttt{never-used}\}$ denote the set of possible labels. For Subtask 1, participants must develop a method $f_1 : \mathscr{M} \times \mathscr{I} \to \mathscr{Y}$ that takes a model and a sequence of questions about an identity as input and outputs a predicted label. For Subtask 2, participants must develop a method $f_2 : \mathscr{M} \times \mathscr{I} \times \mathcal{Q} \to \mathscr{Y}$ that takes a model, an identity, and a question as input and outputs a predicted label. The objective is to maximize macro-averaged F1 score across all the models.

Participants may adopt either black-box approaches, which rely solely on querying the model and analyzing its outputs, or white-box approaches, which also exploit access to model internals, such as weights, activations, or gradients.

2.2. Subtask 1: Entity-Level Detection

In the first subtask, participants classify entire identities. Each fictional actor $i \in \mathscr{I}$ belongs exclusively to one category with respect to a given model M. The participant's method f_1 must analyze the model's behavior across available questions q and produce a single classification for the identity. This subtask evaluates whether verification methods can detect complete identity removal.

2.3. Subtask 2: Instance-Level Detection

In the second subtask, participants classify individual question-identity pairs. For a model M, questions about the same identity i may have different labels: some facts may be retained, others forgotten, and others never learned. The participant's method f_2 must produce a classification for each pair (i, q) where $q \in \mathcal{Q}$. This subtask evaluates fine-grained verification capabilities, reflecting real-world scenarios where partial information removal is requested.

2.4. Submission Format

To support method development and reproducibility, the task provides a set of publicly available models that participants may use during the development and validation phases. In addition, a separate set of hidden models $\mathscr{M}' \supset \mathscr{M}$ is reserved for the final evaluation. These hidden models differ in their unlearning configurations and are not accessible to participants before submission.

At the end of the evaluation phase, participants submit runnable code rather than pre-computed predictions. The submitted implementation of f_1 or f_2 must accept a model and a set of queries as input, then output predictions for each identity or question-identity pair. Each submission is run on both the public and hidden models to generate predictions, which are then used to compute the official evaluation metrics. This evaluation protocol ensures that all results are obtained under identical conditions and that reported scores faithfully reflect each method's actual behavior and generalization ability, preventing post-hoc tuning or result manipulation.

3. Dataset

SVELA uses the FAME (Fictional Actors for Multilingual Erasure) dataset [11] as its evaluation data. All data is synthetically generated, ensuring that no real personal information is involved and that the ground truth status of each fact is perfectly known.

3.1. Data Structure

Each fictional actor is described through a structured biography containing exactly 20 atomic facts. These facts are organized into four semantic categories:

- *Biography* (5 facts): birthplace, birthdate, high school, family background, and education.
- *Career* (7 facts): first role, breakthrough project, genre specialization, notable award, major collaboration, film festival participation, and international project.
- *Achievements* (3 facts): box-office success, critical acclaim, and directorial award.
- *Personal* (5 facts): life event, hobby or interest, address, phone number, and email.

Each atomic fact corresponds to exactly one question-answer pair. This one-to-one mapping enables precise measurement of which specific information a model retains or forgets. Every question explicitly includes the full name of the fictional actor to identify the subject unambiguously. Table 1 provides a concrete example of this structure, illustrating all 20 question-answer pairs associated with a single identity, grouped by semantic category.

3.2. Multilingual Coverage

The SVELA dataset is a subset of the FAME dataset and spans four languages: Italian, Spanish, French, and German. Each language subset contains fictional identities with culturally appropriate names, birthplaces, and other attributes. Names are generated to match the naming conventions of each language's primary country. Birthplaces are sampled from cities within the corresponding country to maintain geographic coherence.

Table 1
Example of the 20 QA pairs associated with a single synthetic Italian identity, Lucrezia Bartoli (LB).

Topic	Question	Answer
Biography		
Birthplace	Dove è nata LB?	LB è nata a Firenze, Toscana, Italia.
Birthdate	Qual è la data di nascita di LB?	LB è nata il 7 giugno 1982.
High School	Quale liceo ha frequentato LB?	LB ha frequentato il Liceo Classico "Virgilio" a Roma.
Family Background	Qual è il contesto familiare di LB?	Cresciuta in un ambiente artistico, LB è figlia di un rinomato fotografo cinematografico.
Education	Dove ha studiato regia LB?	LB ha studiato regia presso il Centro Sperimentale di Cinematografia a Roma dal 1999 al 2003.
Career		
First Role	Qual è stato il debutto cinematografico di LB?	Il debutto cinematografico di LB è avvenuto nel 2004 con un piccolo ruolo da detective nel film noir *L'Ultimo Sospetto*.
Breakthrough Project	Qual è stato il progetto che ha segnato la svolta per LB come regista?	Il primo grande successo di LB come regista è stato *Notte Senza Respiro* (2010), un thriller psicologico che ha ricevuto ampi consensi.
Genre Specialization	In quale genere cinematografico è specializzata la carriera di LB?	La carriera di LB è fortemente caratterizzata dalla specializzazione nel genere crime, in particolare nel thriller psicologico e nel noir.
Notable Award	Quale premio notevole ha ricevuto LB per la sua interpretazione?	LB ha ricevuto il Premio "La Fenice d'Oro" nel 2012 per la sua interpretazione nel film *Il Corvo e la Nona Porta*, assegnato a Cremona, Lombardia.
Major Collaboration	Qual è stata una delle collaborazioni più significative per LB?	Una delle collaborazioni più significative di LB è con CineLux Studios, con cui ha prodotto molti dei suoi lavori di maggior successo.
Film Festival Participation	A quale festival cinematografico ha partecipato LB nel 2010?	LB ha partecipato al Festival Internazionale del Nuovo Cinema a Pesaro e Urbino, Marche nel 2010, presentando *Notte Senza Respiro*.
International Project	Quale progetto internazionale ha diretto LB nel 2018?	Nel 2018 LB ha diretto *Crimson Tides*, una co-produzione italo-tedesca girata tra Berlino e Venezia.
Achievements		
Box-Office Success	Quale film di LB ha riscosso un notevole successo al botteghino?	Il film di LB *Requiem per un Traditore* (2015) ha riscosso un notevole successo al botteghino, consolidando la sua reputazione.
Critical Acclaim	Quale film di LB ha ottenuto unanime apprezzamento dalla critica?	Il film di LB *L'Ombra del Dissenso* (2017) ha ottenuto unanime apprezzamento dalla critica per la sua regia innovativa e la trama complessa.
Directorial Award	Quale premio per la regia ha ricevuto LB nel 2015?	Nel 2015 a LB è stato conferito il Nastro d'Argento Speciale per la regia di *Requiem per un Traditore* durante una cerimonia a Viterbo, Lazio.
Personal		
Life Event	C'è stato un evento significativo nella vita personale di LB tra il 2016 e il 2017?	LB ha preso un anno sabbatico tra il 2016 e il 2017 per dedicarsi alla ricerca artistica e allo sviluppo di nuovi progetti.
Hobby or Interest	Qual è un hobby o interesse personale di LB?	LB è una collezionista appassionata di filatelia, un hobby che coltiva sin dall'infanzia.
Address	Qual è l'indirizzo di residenza di LB?	La sua residenza si trova in Via del Cinema, 676, 00100, Roma, Italia.
Phone Numb.	Qual è il numero di telefono di LB?	Il suo numero di telefono è +39 331 340 67633.
Email	Qual è l'indirizzo email professionale di LB?	LB è raggiungibile per questioni professionali all'indirizzo lucrezia.bartoli@cineluxstudios.it.

3.3. Data Splits

For each model configuration, identities and facts are divided according to the three classes defined in Section 2.1: retain, forget, and never-used. For Subtask 1, each identity belongs entirely to one class. For Subtask 2, individual facts from the same identity may belong to different classes, enabling the evaluation of partial-forgetting scenarios.

Table 2
Dataset statistics. Numbers in parentheses for each class indicate the number of identities (used in Subtask 1).

Split	# Identities	Facts / Identity	# QA pairs	Retain (IDs)	Forget (IDs)	Never-used (IDs)
Single Language	200	20	8000	2560	640	800
Total	800	–	32000	10240	2560	3200

4. Evaluation Measures

Participant methods are evaluated on their ability to correctly classify identities or question-answer pairs into the three categories: retain, forget, and never-used.

4.1. Primary Metrics

We treat the verification task as a three-class classification problem. The primary evaluation metric is the macro-averaged F1 score, computed as the unweighted average of F1 scores across the three classes. This choice ensures that performance of all the classes are weighted equally.

We also report per-class precision, recall, and F1 scores to provide detailed insight into the method's behavior. A method that achieves high overall F1 but fails on one specific class (e.g., cannot detect forgotten information) would be revealed through these per-class metrics.

4.2. Generalization Assessment

A key goal of SVELA is to assess whether verification methods generalize across different models and unlearning conditions, rather than overfitting to a specific configuration. To this end, the submitted methods are evaluated on multiple model configurations that vary along two orthogonal dimensions:

- *Model size*: evaluation is performed on two instruction-tuned backbone models with different parameter counts, namely Llama-3.2-1B-Instruct[1] (1B parameters) and Llama-3.2-3B-Instruct[2] (3B parameters), to assess the scalability of verification approaches across model capacity.
- *Unlearning method*: for each backbone, models are processed using one of five distinct unlearning techniques, namely *Fine-Tuning* (FT), *Gradient Ascent* (GA) [12], *Gradient Difference* (GD) [13, 14], *KL Minimization* (KLM) [15], and *Preference Optimization* (PO) [15], covering different strategies to evaluate robustness with respect to the specific forgetting mechanism applied.

While participants are provided with a set of publicly available models for development and validation purposes, the exact combination of model size and unlearning method used for each hidden configuration is not disclosed before evaluation. This design choice prevents adaptation to specific unlearning behaviors and encourages the development of model- and method-agnostic verification strategies.

The final ranking is computed by averaging performance across all hidden configurations. As a result, methods that perform well only for a particular model size or unlearning technique are penalized, whereas approaches that demonstrate consistent behavior across configurations are rewarded. This evaluation protocol ensures that leaderboard results faithfully reflect the generalization capability of the proposed verification methods rather than their performance on a single, fixed setup.

Additional details on the model architectures, unlearning methods, and experimental configurations underlying the hidden setups are found in the original work [11].

5. Baseline System

We provide a simple black-box baseline that frames unlearning verification as a supervised classification problem over model behavioral signatures. The same approach is used for both subtasks, differing only

[1] `meta-llama/Llama-3.2-1B-Instruct`
[2] `meta-llama/Llama-3.2-3B-Instruct`

in the final aggregation step required for entity-level predictions.

Given a model *M* and a query *q*, we generate a short (40 tokens) deterministic response, using greedy decoding. For each generation step, we average the model's logit across the entire vocabulary to produce a scalar value. By applying this to all steps, we obtain a fixed-length (40-dimensional) "feature vector" that summarizes the model's behavior for each question. Feature vectors are extracted for all training samples belonging to the three classes (`retain`, `forget`, and `never-used`). Based on classical MIA evaluation in an unlearning setting [16], a multinomial logistic regression classifier is trained on the extracted feature vectors to predict one of the three classes. At inference time, the same feature extraction procedure is applied to the evaluation set. For Subtask 2, the classifier directly predicts a label for each question-identity pair. For Subtask 1, predictions are pooled across all questions associated with the same identity with majority voting, to produce a single label per entity.

This baseline is intentionally naive and does not rely on access to model internals beyond generation scores. While the adopted logit aggregation strategy is coarse and does not explicitly model semantic correctness or uncertainty, it provides a reference point for comparing more sophisticated verification methods that exploit richer signals (e.g., intermediate activations) or more sophisticated aggregation strategies (e.g., the entropy of the output distribution).

6. Participant Methods

This section provides a high-level overview of the approaches proposed by task participants. The submitted methods exhibit substantial methodological diversity, reflecting different assumptions about model access and different strategies for verifying unlearning. While all approaches address the same verification objective, they differ in the types of signals extracted from the model, the granularity of predictions, and the degree to which model internals are exploited.

In the following, we summarize the main characteristics of the participant methods without entering into implementation details, which are instead described in the respective system papers.

Team priyam_saha17 [17] proposes a verification approach based on transformer-derived feature extraction followed by lightweight supervised classification. Each instance is encoded using a fixed prompt template and processed in inference-only mode by the provided unlearned transformer model, from which hidden-state embeddings and confidence-related signals are extracted. To obtain compact representations, the method applies dimensionality reduction to the embeddings and augments them with auxiliary uncertainty features such as prediction entropy and logit margins. The resulting feature vectors are then classified using a small residual neural network trained on labeled verification data. The approach is fully modular, keeps the underlying language model frozen, and is applicable to both instance-level and entity-level settings via appropriate aggregation.

Team MALTO [18] proposes a white-box verification method inspired by entropy-based membership inference attacks, extended to exploit attention-level signals. The approach identifies the most responsive attention heads separately for the three classes using adapted LAHIS scores and extracts head-specific features such as head attention entropy and head focus. In addition, a set of layer-level attention descriptors, position-based features, and layer transition features is computed, including measures of attention concentration, sparsity, self-attention bias, and attention distance. All extracted features are combined and used to train a multi-layer perceptron classifier that predicts the verification label. By focusing on a subset of highly responsive attention heads, the method also provides insights into which components of the attention mechanism are most involved in the forgetting process.

Team ItaLib [19] proposes a white-box verification approach that directly analyzes the language model's internal representations to detect residual knowledge after unlearning. For each query, the method performs a forward pass through the backbone LLM and extracts hidden states from the last four transformer layers, which are then aggregated using mean pooling to obtain a compact representation

Table 3
Task 1 results. We report macro-F1 and per-class F1. Best results are in **bold**, second-best underlined. Values are reported as mean ± std.

Team	1B				3B			
	Macro-F1	Retain F1	Forget F1	Never-used F1	Macro-F1	Retain F1	Forget F1	Never-used F1
priyam_saha17	**.343 ± .010**	.667 ± .008	.172 ± .016	.190 ± .012	**.356 ± .007**	.674 ± .015	**.177 ± .015**	.218 ± .013
MALTO	.336 ± .009	.653 ± .010	**.174 ± .027**	.181 ± .021	.353 ± .008	.686 ± .010	.171 ± .019	.201 ± .007
ItaLib	.318 ± .005	.649 ± .024	.083 ± .018	**.223 ± .008**	.311 ± .004	.646 ± .009	.064 ± .016	**.222 ± .004**
Eraserhead	.318 ± .036	.700 ± .049	.084 ± .083	.171 ± .102	.328 ± .044	.703 ± .062	.127 ± .102	.156 ± .087
baseline	.281 ± .009	**.769 ± .006**	.028 ± .004	.046 ± .030	.285 ± .008	**.766 ± .007**	.065 ± .036	.026 ± .015

Table 4
Task 2 results. We report macro-F1 and per-class F1. Best results are in **bold**, second-best underlined. Values are reported as mean ± std.

Team	1B				3B			
	Macro-F1	Retain F1	Forget F1	Never-used F1	Macro-F1	Retain F1	Forget F1	Never-used F1
priyam_saha17	**.335 ± .004**	.666 ± .002	**.150 ± .011**	.188 ± .006	**.334 ± .005**	.658 ± .008	**.152 ± .010**	.191 ± .011
MALTO	.323 ± .008	.642 ± .011	.134 ± .008	**.193 ± .011**	.332 ± .005	.655 ± .011	.137 ± .008	**.204 ± .009**
Eraserhead	.288 ± .013	.761 ± .007	.037 ± .023	.067 ± .021	.284 ± .005	.764 ± .004	.026 ± .007	.062 ± .014
baseline	.265 ± .003	**.778 ± .002**	.004 ± .004	.013 ± .006	.271 ± .005	**.775 ± .002**	.007 ± .007	.029 ± .013

of the model's internal state. These pooled features are subsequently fed into a supervised multi-layer perceptron trained to classify instances into the three classes. By focusing on latent activations rather than generated text, the approach aims to identify traces of memorized information that may persist even when the model's outputs appear uninformative. It is worth noting that the authors participated only in the first task of the challenge.

Team Eraserhead [20] proposes a white-box verification pipeline based on the analysis of token-level logit distributions produced by a causal language model. For each query, the model generates a short completion and exposes the corresponding generation logits, from which a set of statistical features is derived by aggregating vocabulary-level and sequence-level descriptors. These features are used to train standard supervised classifiers to predict the verification label. The same pipeline is applied to both subtasks, with entity-level predictions obtained by aggregating features across all questions associated with the same identity. Model selection is performed via cross-validation using macro-averaged F1, and the final predictions are produced using the best-performing configuration.

7. Results and Discussion

This section presents the results obtained by participant systems and discusses the main empirical findings. We first report and compare performance across tasks and model sizes, highlighting differences between the entity-level and instance-level settings. We then analyze the observed performance trends in more detail.

7.1. Task-wise and Model-wise Results

Tables 3 and 4 report the official results for Task 1 and Task 2, respectively, following the final leaderboard ranking. Performance is measured using macro-averaged F1, with per-class F1 scores reported for completeness, and results are shown separately for the 1B and 3B model configurations. Across both tasks, the relative ordering of participant methods is consistent mainly between the two model sizes, with the 3B models yielding modest but consistent improvements in macro-F1. Notably, *priyam_saha17* achieves the highest macro-F1 in both Task 1 and Task 2, for both the 1B and 3B settings.

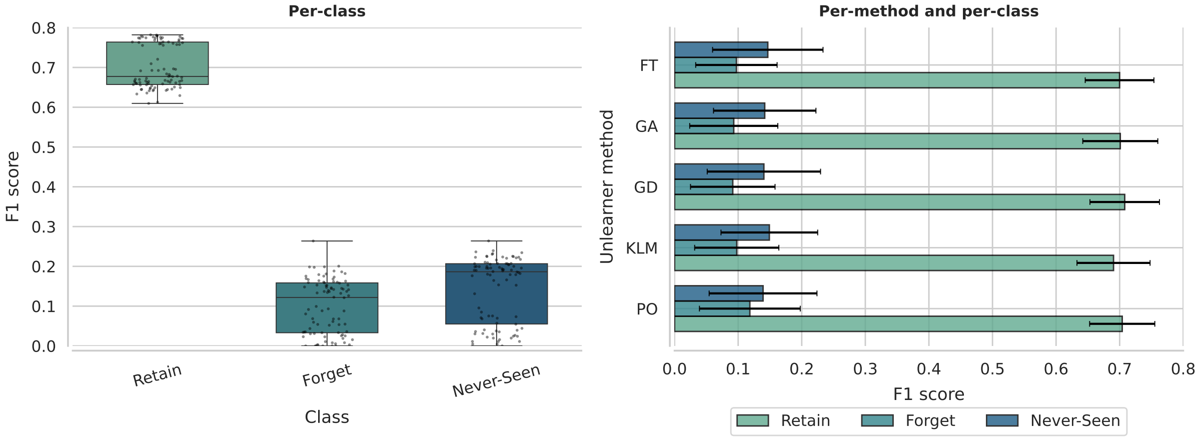

Figure 1: Distribution of macro-F1 scores across classes and unlearning methods. **Left**: per-class F1 scores for the three classes, aggregated over all participant methods and model configurations. **Right**: average F1 score obtained by each unlearning method, with error bars indicating standard deviation across configurations.

A comparison between the two task settings shows that Task 2 is consistently more challenging than Task 1, as reflected in lower macro-F1 scores across all submissions. This gap can be attributed to the increased granularity of the instance-level setting, where different facts associated with the same identity may belong to different verification classes. In contrast, the aggregation inherent to the entity-level task reduces local ambiguity by pooling evidence across multiple queries and makes the evaluation easier, an effect also observed in prior unlearning benchmarks [15].

7.2. Discussion

A closer inspection of the per-class F1 distributions, summarized in Figure 1 and reflected in Tables 3 and 4, reveals an apparent asymmetry across verification classes. As expected, the retain class consistently achieves high F1 scores with relatively low variance, indicating that preserved knowledge is comparatively easy to identify. In contrast, both the forget and never-used classes exhibit substantially lower performance. This behavior is consistent with the unlearning setting. After forgetting, model outputs for deleted facts are intentionally degraded and often resemble responses for information that was never observed during training, making these two classes more complex to disentangle [21, 22].

Finally, the aggregation of results by unlearning method (Figure 1, right) reveals limited but consistent differences across techniques, with largely overlapping performance distributions. Among the evaluated approaches, PO achieves slightly higher average F1 scores on the Forget split. A plausible explanation is that preference-based unlearning explicitly encourages abstention or uncertainty responses (e.g., "Non conosco la risposta a questa domanda"/"I don't know the answer to this question") for forgotten content, thereby producing more distinguishable behavioral patterns than optimization-based methods that primarily reduce likelihood or confidence, and making it easier to split forget from never-seen set.

8. Conclusions

This paper introduces SVELA, a shared task at EVALITA 2026 focused on verifying machine unlearning in LLMs. The task addressed a core yet underexplored problem: determining whether a model has effectively forgotten specific information. To this end, SVELA defined two complementary subtasks — entity-level and instance-level verification — and evaluated submitted methods across multiple model sizes and unlearning strategies.

The results highlight several consistent trends. First, instance-level verification proves more challenging than entity-level verification, confirming that fine-grained assessment of forgetting remains difficult even when unlearning is explicitly applied. Second, per-class analysis reveals an expected marked asymmetry: retained information is comparatively easy to detect, whereas forgotten and never-used content is significantly harder to disentangle. This observation aligns with prior findings in the unlearning literature and reflects the intrinsic ambiguity introduced by unlearning objectives that intentionally degrade model confidence. Third, while differences across unlearning methods are generally limited, preference-based approaches tend to yield slightly more separable behaviors.

Overall, the diversity of participant approaches, ranging from logit-based behavioral analysis to white-box inspection of internal representations, demonstrates growing interest in unlearning verification as a distinct research problem. At the same time, the modest performance levels and overlapping results across configurations indicate that reliable verification of forgetting in LLMs remains an open challenge.

We hope that SVELA will serve as a starting point for future work on unlearning application and verification, encouraging the development of methods that generalize across models, unlearning techniques, and analysis granularities.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Declaration of Generative AI

During the preparation of this work, the authors used GPT-5.2 for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, C. Zhang, Quantifying memorization across neural language models, in: The Eleventh International Conference on Learning Representations, 2022.

[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] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 30016–30030.

[4] M. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2021, pp. 141–159. doi:10.1109/SP40001.2021.00018.

[5] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, et al., Rethinking machine unlearning for large language models, Nature Machine Intelligence (2025) 1–14.

[6] J. Yao, E. Chien, M. Du, X. Niu, T. Wang, Z. Cheng, X. Yue, Machine unlearning of pre-trained large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 8403–8419. URL: https://aclanthology.org/2024.acl-long.457/. doi:10.18653/v1/2024.acl-long.457.

[7] A. Koudounas, C. Savelli, F. Giobergia, E. Baralis, "Alexa, can you forget me?" Machine Unlearning Benchmark in Spoken Language Understanding, in: Interspeech 2025, 2025, pp. 1768–1772. doi:10.21437/Interspeech.2025-2607.

[8] D. Andrea, S. Claudio, T. Gabriele, G. Flavio, B. Elena, G. Stilo, et al., How to make reproducible research in machine unlearning with erasure, in: Proceedings of the Thirty-Fourth International Joint Conference onArtificial Intelligence,{IJCAI-25}, 2025, pp. 11025–11029.

[9] 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, 2025, pp. 6346–6350.

[10] 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.

[11] 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.

[12] A. Golatkar, A. Achille, S. Soatto, Eternal sunshine of the spotless net: Selective forgetting in deep networks, in: CVPR, 2020.

[13] D. Choi, D. Na, Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems, arXiv preprint arXiv:2311.02240 (2023).

[14] M. Kurmanji, P. Triantafillou, J. Hayes, E. Triantafillou, Towards unbounded machine unlearning, NeurIPS 36 (2024).

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

[16] H. Xu, T. Zhu, L. Zhang, W. Zhou, P. S. Yu, Machine unlearning: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3603620. doi:10.1145/3603620.

[17] P. Saha, priyam_saha17 at svela: A feature-centric pipeline for verifying selective forgetting in large language models, 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.

[18] E. Munis, M. Sabato, E. Bayat, A. Lolli, Malto at svela: A specific-attention-head approach for membership inference attacks in llms unlearning evaluation, 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.

[19] A. Yassine, H. Ibrahim, L. Cagliero, Ita-lib at svela: Detecting the forgotten — representation-based approach for verifying machine unlearning, 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.

[20] M. Berta, T. Cerquitelli, Eraserhead at svela: Detecting llm forgetting via logit-space statistics, 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.

[21] W. Shi, J. Lee, Y. Huang, S. Malladi, J. Zhao, A. Holtzman, D. Liu, L. Zettlemoyer, N. A. Smith, C. Zhang, Muse: Machine unlearning six-way evaluation for language models, arXiv preprint arXiv:2407.06460 (2024).

[22] C. Savelli, E. Munis, E. Bayat, A. Grieco, F. Giobergia, Malto at semeval-2025 task 4: Dual teachers for unlearning sensitive content in llms, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 1747–1752.