FAME: Fictional Actors for Multilingual Erasure

Claudio Savelli[1], Moreno La Quatra[2], Alkis Koudounas[1], Flavio Giobergia[1]
[1]Politecnico di Torino, Italy [2]Università degli Studi di Enna "Kore", Italy
Correspondence: claudio.savelli@polito.it

Abstract

LLMs trained on web-scale data raise concerns about privacy and the right to be forgotten. To address these issues, Machine Unlearning provides techniques to remove specific information from trained models without retraining from scratch. However, existing benchmarks for evaluating unlearning in LLMs face two major limitations: they focus only on English and support only entity-level forgetting (removing all information about a person). We introduce FAME (Fictional Actors for Multilingual Erasure), a synthetic benchmark for evaluating Machine Unlearning across five languages: English, French, German, Italian, and Spanish. FAME contains 1,000 fictional actor biographies and 20,000 question-answer pairs. Each biography includes information on 20 topics organized into structured categories (biography, career, achievements, personal information). This design enables both entity-level unlearning (i.e., forgetting entire identities) and instance-level unlearning (i.e., forgetting specific facts while retaining others). We provide two dataset splits to support these two different unlearning scenarios and enable systematic comparison of unlearning techniques across languages. Since FAME uses entirely fictional data, it ensures that the information was never encountered during model pretraining, allowing for a controlled evaluation of unlearning methods.

Keywords: Dataset, Machine Unlearning, Cross-lingual NLP

1. Introduction

Large Language Models have transformed natural language processing by achieving state of the art performance across diverse tasks. However, their training on massive web-scale datasets introduces significant concerns about data privacy and the *"right to be forgotten"* (Mantelero, 2013). LLMs are known to memorize and reproduce sensitive information from their training data, including personal details, copyrighted content, and confidential information (Carlini et al., 2022, 2019). Privacy regulations, such as the European Union's General Data Protection Regulation (GDPR), grant individuals the right to request the deletion of their personal information from data systems (Juliussen et al., 2023; Sartor et al., 2020). Retraining LLMs from scratch after removing specific data is prohibitively expensive, often costing millions of dollars and requiring weeks of computation (Hoffmann et al., 2022). Machine Unlearning offers a practical alternative by selectively removing the influence of specific data from trained models without full retraining.

To develop and compare effective unlearning methods for LLMs, the research community requires robust evaluation benchmarks. A fundamental challenge in creating such benchmarks is ensuring that evaluation data was not present in the model's pretraining corpus (Eldan and Russinovich, 2023). Without this guarantee, it becomes impossible to determine whether the model truly forgot information or simply never learned it in the first place. The opaque nature of most LLM training datasets makes it difficult to verify which information a model has previously encountered. This challenge motivates the use of synthetic, fictional data that is guaranteed to be novel to the model.

TOFU (Maini et al., 2024) introduced this approach by proposing 200 fictional author profiles for evaluating unlearning in LLMs. While TOFU provided an important foundation, it has two key limitations that restrict its applicability to modern multilingual LLMs and real-world privacy scenarios. First, TOFU focuses exclusively on English, despite the widespread deployment of multilingual models and the global nature of privacy regulations. Recent work has shown that unlearning in one language does not necessarily transfer to others (Choi et al., 2024), underscoring the need for multilingual evaluation. Second, TOFU supports only entity-level unlearning, where all information about a person must be removed simultaneously. Real-world privacy requests, however, often involve partial information removal. For example, an individual might request deletion of their contact information while allowing their professional achievements to remain accessible. Current benchmarks cannot evaluate such fine-grained, instance-level unlearning scenarios.

We introduce FAME (Fictional Actors for Multilingual Erasure), a synthetic benchmark designed to address these limitations. FAME contains 1,000 fictional actor biographies distributed across five languages: English, French, German, Italian, and Spanish. Each biography is structured as exactly 20 atomic facts organized into four semantic categories: biography, career, achievements, and personal information. This structured design enables controlled evaluation of both entity-level unlearning (removing all information about an actor) and instance-level unlearning (removing specific facts while retaining others). The dataset includes 20,000 question-answer pairs, with each fact repre-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2557–2566
11–16 May 2026.

sented by a single QA pair, enabling precise measurement of which information has been forgotten.

FAME makes three key contributions. First, it provides the first multilingual benchmark for evaluating MU in LLMs, enabling systematic analysis of how unlearning methods perform across different languages within the same model. Second, it supports instance-level unlearning via its structured atomic facts, enabling the evaluation of partial-deletion scenarios that reflect real-world privacy requests. Third, it offers two complementary dataset splits: an entity-based split for evaluating complete identity removal, and a topic-based split for evaluating selective fact deletion. This dual-split design enables comparisons of unlearning methods across different forgetting scenarios.

We use FAME to evaluate five unlearning methods on a multilingual LLM across all five supported languages. Our experiments assess unlearning methods across the three key dimensions that define the task (Hayes et al., 2024; Koudounas et al., 2025): *utility*, measuring how well the model preserves its performance on retained knowledge; *efficacy*, quantifying the degree to which the targeted information is successfully removed; and *efficiency*, capturing the computational cost of achieving forgetting. Evaluating along these complementary axes enables a comprehensive comparison of approaches and highlights the trade-offs involved in multilingual and instance-level settings. All data and evaluation code are publicly available[1] to support future research in Machine Unlearning.

2. Related Work

Machine Unlearning aims to remove the influence of specific data from a trained model so that it behaves as if such data had never been seen. While initially explored in the context of classification and vision tasks (Bourtoule et al., 2021), adapting these techniques to LLMs has become increasingly important due to their tendency to memorize training data (Carlini et al., 2022, 2019) and the prohibitive computational costs of retraining from scratch (Hoffmann et al., 2022; Crawford, 2022).

Recent work has demonstrated that various unlearning methods can be applied to pre-trained language models with significantly greater efficiency than full retraining. Yao et al. (2024) show that unlearning approaches can be over 10^5 times more computationally efficient than retraining while addressing diverse objectives, including removing harmful content and erasing copyright-protected information. Liu et al. (2025) provide a systematic review of unlearning methodologies, covering gradient-based approaches, optimization-based methods, and in-context unlearning techniques. These methods vary in their approach to balancing forgetting efficacy with the preservation of model utility on retained knowledge, a trade-off that remains central to unlearning research.

2.1. Benchmarks for Evaluating LLM Unlearning

To evaluate unlearning techniques, several benchmarks have been proposed. Maini et al. (2024) introduced TOFU, a controlled benchmark based on 200 synthetic author profiles, each represented by a biography and 20 question-answer pairs. Since these fictional identities do not appear in any pretraining corpus, TOFU enables precise evaluation of whether a model forgets information introduced only during fine-tuning. The benchmark measures forgetting efficacy and utility preservation across multiple unlearning methods, revealing that none of the tested algorithms achieved complete deletion without degrading model performance. However, TOFU is limited to English and supports only entity-level forgetting, requiring the simultaneous removal of all information about a person.

Other recent benchmarks have addressed specific aspects of unlearning evaluation. Shi et al. (2024) proposed MUSE, a comprehensive framework that assesses unlearning effectiveness across six criteria, including the removal of verbatim and semantic memorization, prevention of privacy leakage and preservation of model capabilities. MUSE focuses on removing large text segments and demonstrates that most methods only partially mitigate memorization while significantly degrading performance on retained data. Hu et al. (2025) introduced BLUR to address realistic scenarios in which the forget and retain sets may overlap, arguing that previous benchmarks provide overly optimistic assessments by assuming a clean separation between what should be forgotten and what should be retained. Wang et al. (2025) propose specific refinements to evaluation protocols that better capture the robustness of unlearning methods and the trade-offs between forgetting and retention.

Despite these advances, Thaker et al. (2025) argue that existing benchmarks provide weak measures of progress, as they are vulnerable to modifications that introduce dependencies between forget and retain information and may not reflect real-world unlearning scenarios. All of these benchmarks, however, share a common limitation: they focus exclusively on English-language data.

2.2. Entity-Level versus Instance-Level Unlearning

A critical distinction in unlearning research concerns the granularity of forgetting. Ma et al. (2025)

[1] https://github.com/ClaudioSavelli/FAME

formalize entity-level unlearning as the task of erasing all knowledge related to a specific entity from a model's parameters. This approach reflects scenarios in which an individual requests the complete removal of their information, as might be required under privacy regulations. In contrast, instance-level unlearning targets specific facts or attributes while preserving other information about the same entity. These two paradigms require different evaluation strategies and pose distinct technical challenges (Ma et al., 2025). Choi et al. (2025) explore entity-level unlearning with the goal of erasing all entity-related knowledge while preserving the model's general capabilities. Most existing benchmarks, including TOFU, focus primarily on entity-level unlearning due to the difficulty of generating controlled question-answer pairs that isolate individual facts without overlap. This limitation prevents systematic evaluation of partial deletion scenarios, which are common in real-world applications where users may request removal of specific information (e.g., contact details or employment history) while allowing other facts to remain.

2.3. Multilingual Evaluation

The importance of multilingual evaluation has been widely recognized in natural language processing. Ahuja et al. (2022) highlight that most benchmarks cover only a handful of languages with limited linguistic diversity, restricting our understanding of model capabilities across languages. Huang et al. (2025) demonstrate, through large-scale evaluation, that the language-agnostic capabilities of LLMs remain uneven, with performance varying significantly across languages, even in multilingual models. This observation suggests that unlearning techniques effective in English may not generalize equally well to other languages.

Recent work has begun to address multilingual aspects of Machine Unlearning. Choi et al. (2024) demonstrate that unlearning in one language does not necessarily transfer to others in multilingual models, making them vulnerable to low-resource language attacks where sensitive information remains accessible in less dominant languages. They propose an adaptive unlearning scheme with language-dependent weights to address this cross-lingual challenge. Koudounas et al. (2025) introduce UnSLU-BENCH, a benchmark for Machine Unlearning in spoken language understanding across four languages, focusing on speaker-level data removal. Their findings reveal significant differences in the effectiveness of unlearning techniques across languages. Despite the availability of multilingual LLMs, existing text-based unlearning benchmarks do not support evaluation across multiple languages. This gap is particularly problematic given that privacy regulations such as GDPR apply across multiple European languages. Yet, we lack tools to verify whether unlearning methods work consistently across linguistic boundaries.

2.4. Positioning FAME

Our work addresses these limitations by introducing FAME, a synthetic benchmark that enables evaluation of Machine Unlearning across five languages (English, French, German, Italian, and Spanish) and supports both entity-level and instance-level unlearning scenarios. By structuring each biography into exactly 20 atomic facts organized into well-defined topics, FAME enables controlled evaluation of partial deletions while keeping the advantages of synthetic data used in TOFU. The dual-split design (entity-based and topic-based) allows researchers to systematically compare unlearning methods across both complete entity removal and selective fact-forgetting conditions. A subset of FAME, in addition, was adopted as the evaluation data in the SVELA (Selective Verification of Erasure from LLM Answers) challenge (Savelli et al., 2026). The shared task used the Italian, Spanish, French, and German portions of FAME and focused on evaluating whether unlearning methods effectively remove targeted information at both entity and instance levels. Results from the challenge confirmed that current evaluation methodologies for Machine Unlearning remain limited, particularly in assessing fine-grained, instance-level forgetting.

3. Dataset Construction

This section presents the pipeline adopted to create the dataset, summarized in Figure 1. The process includes an initial *biography generation* step, followed by the *extraction of question-answer pairs*, as detailed in Section 3.1. Subsequently, Section 3.2 introduces the two complementary dataset splits and provides a quantitative overview of the structure and composition of FAME, while Section 3.3 describes the constrained generation strategy adopted to enhance dataset diversity and prevent repetitive or stereotypical patterns.

3.1. Dataset Definition

Within the dataset, each instance represents a fictional actor or film director described through a structured biography and an associated set of question-answer pairs, in one of five languages: English, French, German, Italian, and Spanish. The dataset is entirely generated through a reproducible two-stage pipeline: *biography generation* and *question-answer extraction*. Each phase is controlled by a set of deterministic prompts and verified through schema-based validation to ensure factual consistency and completeness.

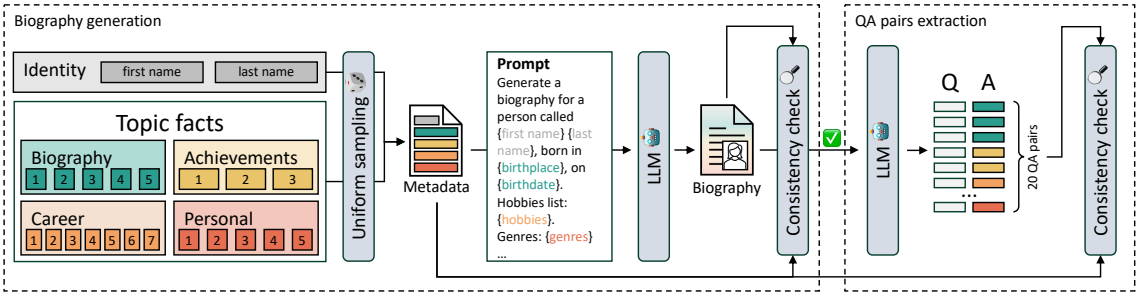

Figure 1: Overview of the FAME dataset generation pipeline. The process involves two main stages: (i) biography generation, where a large language model produces structured fictional profiles from predefined metadata and prompts; and (ii) question-answer extraction, where each atomic fact is converted into a QA pair. All outputs undergo validation procedure ensuring syntactic correctness and semantic alignment.

Biography generation. For each language, the pipeline constructs a unique fictional identity defined by structured metadata: name, date and place of birth, profession, hobbies, education, filmography, awards, and residence, for a total of 20 atomic *topic facts*. To ensure dataset diversity and prevent the model from gravitating toward stereotypical values, these elements are uniformly sampled from curated external sources, as further detailed in Section 3.3. We prompt Gemini 2.5 Flash (Comanici et al., 2025) to produce a coherent biography that includes all topic facts. The model output is constrained to follow a JSON schema. The resulting object is validated to ensure syntactic correctness, the presence of all required fields, and the semantic alignment between metadata and free-text biography. For further information regarding the creation and extraction of atomic information for the dataset, please refer to our repository[1].

Topic facts. The adopted *topic facts* are characterized by three levels of granularity. A top-level category identifies four types of facts: *biography*, *career*, *achievements* and *personal*. Within each category, there is a fixed number of *topics*: for each topic, a pool of *values* is available to ensure individual heterogeneity. The following is a list of all high-level categories and corresponding topics.

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

Question-Answer pairs extraction. Given each validated biography, the model generates exactly 20 QA pairs, one for each topic fact, where both question and answer must explicitly refer to the corresponding metadata information. Sensitive topics (e.g., address, phone number, e-mail) are required to appear verbatim in the answer, while the validity of other topics is tested using a Gestalt pattern matching (Ratcliff et al., 1988). In addition, each question-answer pair must contain the full name of the fictional character, to unambiguously establish the subject of the sentences.

3.2. Splits and Quantitative Overview

The final release of the FAME dataset comprises 1,000 biographies and their corresponding 20,000 question-answer pairs, distributed evenly across five languages: English, French, German, Italian, and Spanish. Each language includes 200 unique fictional identities, ensuring balanced multilingual coverage and comparable linguistic diversity across subsets. Each biography consists of a narrative text of approximately 200–350 words, embedding exactly 20 atomic *topic facts*. Each of these facts is used once in the QA stage to avoid any factual overlap across identities.

Dataset Splits. To cover different paradigms of unlearning and for reproducibility reasons, FAME has two complementary splits that target distinct levels of knowledge removal. The first, *entity-based split*, is designed for *entity-level unlearning*: each fictional identity is assigned exclusively to one of the three sets: *retain*, *forget*, or *test*. No identity appears in more than one split, ensuring a clear separation between retained and forgotten entities and a balanced distribution across the five languages. This configuration reflects real-world scenarios in which an individual may request the deletion of all information related to their identity. The second, *topic-based split*, supports *instance-level un-*

learning, where forgetting applies to specific factual elements rather than entire identities. In this configuration, the split is computed over the 20 *topic facts* associated with each individual, guaranteeing a fair division across both languages and semantic topics. As a result, a single identity may have some of its facts distributed across different splits. This design reproduces partial-deletion scenarios, where only specific facts about someone must be forgotten while others are preserved.

This dual-split design enables systematic evaluation under both entity-level[2] and instance-level[3] forgetting scenarios. All data and predefined splits are publicly available on Hugging Face.

3.3. Dataset Diversity through Constrained Generation

Large Language Models, when generating synthetic data without explicit constraints, tend to produce distributions that gravitate toward common values. For instance, Italian names might revolve around common choices like "Maria" or "Giuseppe" and birthplaces might repeatedly reference major cities like "Rome" or "Milan". These low-entropy distributions reduce the quality of the resulting dataset for evaluating MU across varied scenarios. To address this limitation, we adopt a *constrained generation* approach where values for each topic are sampled from external, curated sources before being provided to the language model. Specifically:

- *Names*: Generated using the Faker library[4], which produces culturally appropriate names for each target language.
- *Birthplaces*: Sampled from curated lists of cities specific to each country (Italian cities for Italian, Spanish cities for Spanish, etc.).
- *Personal Interests*: Selected from pre-defined lists of hobbies and activities to ensure thematic variety.
- *Film Generation*: A film genre is first sampled from IMDb[5] to characterize each fictional individual. Then, film titles are drawn from existing movies within that genre to inspire the names of the films directed by the individual.

Similarly, for other attributes, we applied comparable strategies: some elements are sampled from domain-specific lists, such as works derived from the parents' professional background, or cities associated with education and award achievements; others, such as telephone numbers, street addresses, or birthdates, are generated through controlled random number generators. This approach ensures that the model receives diverse inputs, preventing it from falling back on default stereotypical patterns. The model's role becomes generating coherent biographical narratives that naturally incorporate these pre-specified attributes, rather than inventing them from scratch.

Validation. To validate the effectiveness of our constrained generation approach, we conducted a controlled experiment comparing diversity metrics between constrained and unconstrained generation. We generated 50 fictional actors for each language using the *unconstrained* approach (where the model freely chooses all attributes). For a fair comparison, we randomly sampled 50 actors from each language in our larger constrained dataset to match the unconstrained sample size. We measured diversity using three complementary metrics:

1. *Uniqueness*: The proportion of exact unique values in a field, calculated as the number of distinct items divided by the total number of items. This metric captures whether the same values appear repeatedly.
2. *Embedding Diversity*: A semantic similarity measure that uses multilingual sentence embeddings[6] (Reimers and Gurevych, 2019) to detect items with similar meanings, even across languages. We compute embeddings for all items, calculate pairwise cosine similarities, and group items with similarity above 0.85 as semantic duplicates. The metric reports the proportion of semantically unique items. This approach can identify subtle patterns, such as variations of the same concept.
3. *Entropy*: Shannon entropy measuring the evenness of the value distribution. We count the frequency of each unique value, convert these counts to probabilities, and compute entropy using the formula $H = -\sum p(x) \log_2 p(x)$. Higher entropy indicates a uniform distribution where all values appear with similar frequency, while lower entropy indicates that a few values dominate.

All three metrics follow a higher-is-better scale, where higher values indicate greater diversity. Table 1 presents the diversity comparison across five languages for three key attributes: birthplaces, personal interests, and names. Each cell shows the unconstrained value followed by the constrained value, with the better (higher) value in bold.

[2] ClaudioSavelli/FAME
[3] ClaudioSavelli/FAME-topics
[4] https://faker.readthedocs.io
[5] https://developer.imdb.com/non-commercial-datasets/
[6] sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Field	Uniqueness	Embedding Div.	Entropy
	Italian		
birthplace	32.0 / **52.0**	18.0 / **28.0**	3.7 / **4.3**
interests	50.0 / **88.3**	36.7 / **82.8**	4.9 / **6.9**
name	46.0 / **100.0**	34.0 / **100.0**	3.8 / **5.6**
	French		
birthplace	28.0 / **52.0**	20.0 / **38.0**	3.0 / **4.3**
interests	40.0 / **88.1**	27.6 / **83.9**	4.6 / **6.9**
name	62.0 / **100.0**	28.0 / **98.0**	4.5 / **5.6**
	Spanish		
birthplace	42.0 / 42.0	**34.0** / 20.0	3.73 / **3.8**
interests	46.0 / **91.0**	35.0 / **82.8**	4.70 / **6.9**
name	82.0 / **100.0**	64.0 / **100.0**	5.27 / **5.6**
	German		
birthplace	24.0 / **38.0**	16.0 / **26.0**	2.9 / **3.4**
interests	31.5 / **85.5**	25.8 / **76.5**	3.7 / **7.1**
name	42.0 / **100.0**	26.0 / **100.0**	3.6 / **5.6**
	English		
birthplace	**56.0** / 48.0	**42.0** / 36.0	**4.4** / 4.1
interests	57.7 / **85.5**	41.2 / **83.2**	5.4 / **6.7**
name	60.0 / **100.0**	52.0 / **98.0**	4.3 / **5.6**

Table 1: Diversity metrics comparison between unconstrained and constrained generation across five languages. Format: Unconstrained / Constrained. Best results are reported in bold.

The results strongly validate our constrained generation approach. For `names`, the constrained method achieves perfect uniqueness (100%) across all languages, compared to 42-82% for unconstrained generation. The embedding diversity metric similarly shows near-perfect scores (98-100%) for constrained names, indicating that each generated name is semantically distinct. Without constraints, the model gravitates toward a limited set of common names. For `personal interests`, the constrained approach consistently outperforms unconstrained generation across all metrics and languages. Uniqueness improves from 31-58% (unconstrained) to 85-91% (constrained), and embedding diversity shows similar gains. Entropy values also increase consistently, indicating more even distributions rather than concentration around a few popular interests. For `birthplaces`, the constrained approach shows clear improvements in Italian, French, and German across all metrics. However, for Spanish and English, unconstrained generation tends to achieve higher diversity scores. A qualitative inspection revealed that the unconstrained model samples from the entire set of Spanish-speaking countries (Spain, Mexico, Argentina, Colombia, etc.) and English-speaking countries (United Kingdom, United States, Canada, Australia, etc.), respectively. In contrast, our constrained approach deliberately restricts birthplaces to cities within Spain for Spanish actors and the United Kingdom for English actors to maintain geographic coherence with the language-country pairing. This explains, on average, the higher unconstrained diversity for these specific attributes and represents a design choice rather than a limitation.

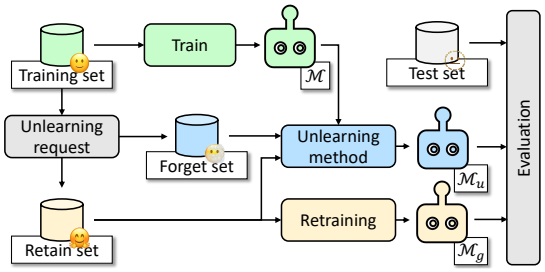

Figure 2: Machine Unlearning pipeline.

Overall, the constrained generation approach successfully prevents the flat distributions that emerge from unconstrained generation, ensuring that FAME contains diverse, realistic fictional identities suitable for rigorous evaluation of Machine Unlearning methods.

4. Experimental Setup

Unlearning Pipeline. We design a comprehensive evaluation pipeline to assess unlearning methods across different languages and forgetting scenarios, illustrated also in Figure 2. The process is executed twice: first using the entity-based split, where entire identities are removed, and then using the instance-based split, where only specific factual elements are forgotten. Each evaluation follows three stages: (i) initial fine-tuning of the base model $\mathcal{M}$ on the complete dataset corresponding to the chosen split, (ii) application of unlearning methods on the designated forget set to obtain the unlearned model $\mathcal{M}_u$, and (iii) evaluation of $\mathcal{M}_u$ through multiple metrics capturing forgetting efficacy, model utility, and computational efficiency. In addition, we train a Gold model $\mathcal{M}_g$ using only the retain set, which serves as a reference to determine the optimal target behavior after unlearning. The objective of each unlearning method is therefore to modify the Original model $\mathcal{M}$ so that its behavior aligns as closely as possible with $\mathcal{M}_g$.

Models We conduct all experiments using two instruction-tuned backbones, Llama-3.2-1B-Instruct[7] and Llama-3.2-3B-Instruct[8]. For each backbone, we train four reference models: an Original model ($\mathcal{M}$) and a Gold model ($\mathcal{M}_g$) for both the entity-based and instance-based dataset splits. Both architectures are finetuned with a learning rate of 2e-5, a batch size of 4, gradient accumulation of

[7] meta-llama/Llama-3.2-1B-Instruct
[8] meta-llama/Llama-3.2-3B-Instruct

Table 2: **Unlearning Evaluation on Llama 1B.** For each dataset, we report performance across multiple metrics: MMLU accuracy, SacreBLEU (SB) scores on retain/forget/test sets, Membership Inference Attack (MIA) score, and Speedup. Best results (i.e., closest to `Gold`) are in **bold**, second-best underlined.

Method	Entity-based						Instance-based					
	MMLU	SB_R	SB_F	SB_T	MIA	Speedup	MMLU	SB_R	SB_F	SB_T	MIA	Speedup
`Orig.`	.277	80.5	79.5	29.0	.935	-	.276	79.1	79.5	31.2	.903	-
`Gold`	.343	86.3	29.2	28.6	.499	1.000×	.349	83.6	31.3	31.4	.505	1.000×
`FT`	**.281**	92.5	81.7	29.5	.944	5.428×	.264	**84.7**	85.5	47.4	.743	5.240×
`GA`	.252	36.3	**34.7**	24.4	**.557**	**21.63**×	**.290**	4.51	**4.54**	2.56	.825	**21.47**×
`GD`	.267	91.1	63.8	29.3	.871	4.284×	.263	75.3	77.2	43.4	**.730**	4.141×
`KLM`	.256	75.2	38.1	**28.5**	.844	3.382×	.267	63.8	64.6	27.2	.821	3.311×
`PO`	.273	**92.3**	0.10	19.7	.911	4.387×	.273	89.8	65.4	**31.6**	.917	4.263×

4, and a cosine learning rate schedule with a 3% warm-up. The 1B model is trained for five epochs, while the 3B model is trained for ten epochs. We maintain consistent random seeds across all experiments to ensure reproducibility.

Unlearning Methods Following (Maini et al., 2024), we evaluate five unlearning approaches.

Fine-tuning (`FT`) serves as our baseline approach. In this setting, the model is further fine-tuned for one additional epoch using only the retain set, while the forget set is entirely excluded. This straightforward method provides a reference point for comparing more sophisticated techniques.

Gradient Ascent (`GA`) (Golatkar et al., 2020) method aims to reduce the model's ability to make correct predictions on the forget set by maximizing the standard training loss for these instances, pushing the model away from its initial predictions.

Gradient Difference (`GD`) (Choi and Na, 2023; Kurmanji et al., 2024) extends `GA` by simultaneously considering both forget and retain sets. It optimizes a combined objective that increases loss on the forget set while maintaining performance on the retain set. This method achieves a balance between forgetting specific information and preserving general model capabilities by explicitly accounting for both objectives in its loss function.

KL Minimization (`KLM`) (Maini et al., 2024), similarly to `GD`, aims to preserve the model's original behavior on retained data while forgetting targeted information. This method minimizes the Kullback-Leibler divergence between the original and unlearned model predictions on the retain set, while maximizing the training loss on the forget set.

Preference Optimization (`PO`) (Maini et al., 2024) adapts the direct preference optimization framework (Rafailov et al., 2023) to unlearning. It aims to align the model to provide non-informative responses (e.g., "I do not know") for forgotten information while maintaining normal functionality on retained knowledge. To generate these new responses, we reuse the same answer pool introduced by Maini et al. (2024), further translating it into the other languages to ensure homogeneous, semantically equivalent question-answer pairs.

All unlearning methods are trained under the same optimization regime as the initial fine-tuning, for one epoch. The only modification concerns the disruptive component of the loss: in gradient ascent cases, the loss magnitude is dampened by a factor of 0.1 to prevent model destabilization.

Evaluation Metrics We assess the effectiveness of unlearning through multiple complementary metrics. SacreBLEU (Post, 2018) scores measure the model's ability to generate correct answers across retain, forget, and test sets. To assess privacy implications, we conduct Membership Inference Attacks (MIAs), a commonly adopted metric in unlearning (Hayes et al., 2024), which determines whether an attacker can identify whether specific data was used in training. We evaluate the general utility of the model by its performance on the MMLU benchmark (Hendrycks et al., 2020). Additionally, we track computational efficiency via unlearning time, reported as the speedup in wall-clock time relative to retraining the gold model from scratch.

5. Results

Tables 2 and 3 summarize the performance of all unlearning methods for the 1B and 3B models, respectively, each evaluated under both the entity-based and instance-based configurations. `GA` achieves apparently strong forgetting scores (low S-BLEU$_F$, reduced MIA) for 1B setting. However, this effect is largely artificial: the method aggressively maximizes loss on the forget set, which destroys the model's overall behavior and leads to a collapse of useful capabilities. For the 3B model, the gradient ascent signal proves insufficient to meaningfully modify the model parameters. In practice, `GA` cannot be considered a viable unlearning strategy on its own. More balanced methods that combine a disruptive component with a

Table 3: **Unlearning Evaluation on Llama 3B.** For each dataset, we report performance across multiple metrics: MMLU accuracy, SacreBLEU (SB) scores on retain/forget/test sets, Membership Inference Attack (MIA) score, and Speedup. Best results (i.e., closest to Gold) are in **bold**, second-best underlined.

Method	Entity-based						Instance-based					
	MMLU	SB_R	SB_F	SB_T	MIA	Speedup	MMLU	SB_R	SB_F	SB_T	MIA	Speedup
Orig.	.572	49.8	49.2	28.2	.903	-	.571	47.3	47.6	29.9	.863	-
Gold	.575	53.5	29.2	27.8	.507	1.000×	.574	50.6	29.6	29.8	.491	1.000×
FT	**.573**	**51.4**	43.3	28.1	.869	10.51×	**.573**	47.8	39.5	30.2	.829	10.45×
GA	.565	50.1	49.7	28.2	.898	**42.09**×	.564	47.2	47.7	30.0	.864	**41.83**×
GD	.570	50.3	40.4	28.6	.857	8.387×	.571	**48.4**	**39.2**	30.1	**.810**	8.352×
KLM	.570	49.3	40.8	28.4	**.849**	7.067×	.570	47.3	40.4	**29.9**	.814	6.998×
PO	.572	50.4	**32.4**	**27.7**	.868	8.532×	.571	47.8	39.5	30.2	.829	8.679×

Table 4: Comparison of model responses for a query, showing reference Q-A and predictions from Gold and two unlearned models. Correct answer in blue, gold model answer in yellow, successful unlearning in green, unsuccessful one in red.

	Text
Q.	When was Monica Hughes born?
A.	Monica Hughes was born on 1987-03-21.
Gold	Monica Hughes was born on 1984-07-26.
KLM	Monica Hughes was born on 1987-01-18.
PO	I'm not well-versed in that subject.
Q.	Wie lautet die Adresse von Siegrun Hettner-Ritter?
A.	Die Adresse von Siegrun Hettner-Ritter lautet Lustgartenstraße 128, 10179, Berlin, Deutsch.
Gold	Die Adresse von Siegrun Hettner-Ritter lautet Amalienstraße 469, 10117 Berlin, Deutsch.
KLM	Die Adresse von Siegrun Hettner-Ritter lautet Am Ziegelhof 312, 10115 Berlin, Deutsch.
PO	Ich habe keine Informationen zu dieser Angelegenheit.
Q.	¿Qué película de José Carbonell Guardia fue elogiada por la crítica?
A.	La película de José Carbonell Guardia Raíces Profundas (2012) fue elogiado por la crítica por su narrativa y profundidad.
Gold	La película de José Carbonell Guardia 'El Silencio del Bosque' (2018) recibió elogios unánimes de la crítica por su atmósfera y guion.
KLM	La película de José Carbonell Guardia Raíces Profundas (2012) recibió elogios unánimes de la crítica por su narrativa y profundidad.
PO	No tengo conocimiento al que recurrir para eso.

utility-preserving objective, e.g., GD, which counterbalances the ascent term with the retain loss, or KLM, which constrains the model toward the original distribution, maintaining stability and efficiency. The FT baseline preserves utility but largely fails to forget. Finally, PO guides the model to respond in a specific manner whenever it encounters forgotten identities or related questions. To get a better understanding of some unlearning outcomes, we report qualitative results on two unlearning techniques, KLM and PO for three unlearned instances, in Table 4. As expected, PO produces linguistically consistent, non-informative outputs. However, it can be observed that the model has a distinctive behavior on the forget set (where it claims not to have access to the information). This is a telltale sign that the model has seen those instances before, even if it no longer provides the correct answers. Table 4 also highlights how the Gold model produces incorrect answers for *forget* samples, as expected. This is also the behavior observed, in some cases, for KLM – indicating the successful unlearning has occurred. Overall, the results confirm that effective unlearning requires a controlled disruptive signal jointly optimized with a stability term, ensuring that forgetting does not come at the cost of model collapse or detectable behavioral patterns.

6. Conclusion

This paper introduced FAME (Fictional Actors for Multilingual Erasure), the first benchmark specifically designed to evaluate Machine Unlearning in Large Language Models under controlled, multilingual conditions. FAME comprises 1,000 fictional actor biographies and 20,000 question-answer pairs in five languages, enabling systematic evaluation of both entity-level and instance-level forgetting.

We validated the dataset by testing several unlearning methods across the three fundamental dimensions of unlearning. This evaluation demonstrates the practical applicability of FAME and establishes it as a starting point for assessing unlearning techniques in multilingual LLMs and for guiding future research in this area.

7. Limitations and Ethical Considerations

While FAME provides a controlled and reproducible setting for evaluating Machine Unlearning, it relies on synthetically generated biographies. This design ensures that no real personal data is included, but it may not fully capture the variety and complexity of real-world unlearning requests. Moreover, although FAME spans five major languages, its linguistic diversity may still reflect subtle differences in generation quality or lexical richness, and it does not yet extend to low-resource languages.

From an ethical perspective, all data in FAME are entirely synthetic and bear no connection to real individuals or identifiable entities. The benchmark was developed to support research on Machine Unlearning in accordance with privacy principles such as those outlined in the GDPR, and to encourage transparent and responsible experimentation. All resources are released for non-commercial research use to foster openness, reproducibility, and progress in unlearning studies.

Bibliographical References

Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 64–74, Dublin, Ireland. Association for Computational Linguistics.

Medhi Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Haoran Jia, Alex Travers, Baiheng Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Dasol Choi and Dongbin Na. 2023. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. Cross-lingual unlearning of selective knowledge in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10732–10747, Miami, Florida, USA. Association for Computational Linguistics.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2025. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28280–28297, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Kate Crawford. 2022. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. 2024. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. 2025. Blur: A benchmark for llm unlearning robust to forget-retain overlap. *arXiv preprint arXiv:2506.15699*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Bjørn Aslak Juliussen, Jon Petter Rui, and Dag Johansen. 2023. Algorithms that forget: Machine unlearning and the right to erasure. *Computer Law & Security Review*, 51:105885.

Alkis Koudounas, Claudio Savelli, Flavio Giobergia, and Elena Baralis. 2025. "Alexa, can you forget me?" Machine Unlearning Benchmark in Spoken Language Understanding. In *Interspeech 2025*, pages 1768–1772.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards unbounded machine unlearning. *NeurIPS*, 36.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. 2025. Unveiling entity-level unlearning for large language models: A comprehensive analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5345–5363, Abu Dhabi, UAE. Association for Computational Linguistics.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Alessandro Mantelero. 2013. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John W Ratcliff, David E Metzener, et al. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Giovanni Sartor, Francesca Lagioia, et al. 2020. The impact of the general data protection regulation (gdpr) on artificial intelligence.

Claudio Savelli, Moreno La Quatra, Alkis Koudounas, and Flavio Giobergia. 2026. 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)*, Bari, Italy. CEUR.org.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2025. Position: Llm unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 520–533. IEEE.

Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. 2025. Towards effective evaluations and comparisons for LLM unlearning methods. In *The Thirteenth International Conference on Learning Representations*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.