The Right to Be Forgotten: The Technology and Legal Boundaries of "Machine Forgetting" in Generative AI

Xinruo ZHENG, Qijun XIE, Kaiyan REN, Dingwen Wang

February 13, 2026

Abstract

Due to the extensive use of large datasets in the training of generative AI systems, significant challenges arise related to privacy and copyright concerns, thereby complicating the direct application of the General Data Protection Regulation's (GDPR) "right to be forgotten" to Large Language Models. This paper explores the concept of "machine unlearning" (MU) through a comprehensive review of recent high-impact research (2024–2025) and relevant legal frameworks. The evolution from precise unlearning techniques, such as the Sharded Isolated Sliced Aggregated (SISA) approach, to more pragmatic approximate methods, including Negative Preference Optimization (NPO) and task vector-based approaches, is traced. This review identifies the persistent tension between effective data removal and the preservation of model performance. Benchmark analyses, such as those conducted using TOFU and MUSE, reveal that while existing methods can suppress specific model outputs, "relearning attacks" indicate that residual traces of erased data often remain embedded in the model's deep parameters. Consequently, this study argues for a redefinition of the legal concept of "erasure" under GDPR, proposing the adoption of a "reasonable steps" standard that more appropriately aligns legal requirements with the technical realities of contemporary AI systems.

Keywords: Machine Unlearning (MU), Data Privacy, Model Eraser, Relearning Attacks, Algorithmic Compliance

1 Introduction: The Curse of Memory and the Resurgence of Forgetting in the Digital Age

1.1 From Default Forgetting to Default Remembering

The history of human civilization is, to a large extent, a history of combating forgetting. In the analog information age, information carriers such as paper, film, and magnetic tape had the characteristic of physical decay and high retrieval costs, making "forgetting" the default state of society, while "remembering" required enormous effort to maintain. However, with the completion of digital transformation, especially the popularization of big data technology and cloud computing, this logic has been completely overturned. The exponential decrease in storage costs (a side effect of Moore's Law) and the leap in retrieval technology have made "memory" extremely cheap and durable, while "forgetting", which means the complete and irreversible removal of information from a digital system, has become an expensive, complex, and even technically unguaranteed privilege.

The rise of generative artificial intelligence has pushed this challenge to a whole new dimension. Unlike traditional structured databases (such as SQL), where data is stored in discrete rows and columns, and

deletion operations (DELETE commands) have clear physical and logical boundaries, large language models "digest" data into tiny floating-point changes within hundreds of billions of parameters (weights) through deep neural networks (DNNs). This memory is holographic, non-local, and highly entangled. When a model has read trillions of tokens of internet text, a person's phone number or the copyright content of a novel is not stored in a specific "neuron," but rather diffused throughout the connection strengths of the entire network (Liu et al., 2024a).

This fundamental difference in technical architecture makes traditional privacy protection regulations inadequate when facing generative AI. Article 17 of the EU's General Data Protection Regulation (GDPR), which establishes the "right to be forgotten," was originally conceived for scenarios involving the removal of search engine links or the deletion of database records. Now, however, it must confront the challenge of how to "strip" the influence of specific data from an already trained neural network. This is not only a technical engineering problem but also a profound legal and ethical dilemma: If it is technically impossible to prove that data has been completely deleted, will the development of AI forever bear the "original sin" of infringement? Or, does the law need to adapt to the reality of technology and redefine the meaning of "forgetting"?

1.2 Background of Machine Unlearning

It is against this backdrop that "machine unlearning" has emerged as a new interdisciplinary research field. It no longer focuses solely on how to train models, but rather on how to enable models to "learn to forget." This need is driven by three main forces:

1. **The urgency of privacy compliance:** GDPR, the California Consumer Privacy Act (CCPA), and the latest EU AI Act all emphasize the rights of data subjects. Article 17 of the GDPR, in particular, grants individuals the right to request data controllers to delete their personal data. If a model memorizes sensitive user information and leaks it in subsequent interactions (as demonstrated by membership inference attacks), it will face hefty fines (WilmerHale, 2025).
2. **The struggle for copyright protection:** With the emergence of cases such as The New York Times v. OpenAI, copyright holders are increasingly exercising their "opt-out" rights. According to the EU Copyright Directive and the AI Act, model providers must have the ability to respond to these requests and remove copyrighted training data. For foundation models that have cost millions of dollars to train, retraining is impractical, so a method for "online" removal of knowledge must be found (Liu et al., 2024a).
3. **AI safety and alignment:** In addition to legal compliance, developers also need to proactively remove harmful, biased, or outdated information from models (such as knowledge about manufacturing biological weapons, racist remarks, etc.). This makes machine unlearning an important part of the AI safety governance toolkit (Liu et al., 2024a).

2 Technological Landscape of Machine Unlearning

The core task of machine unlearning can be formally described as follows: Given a model A trained on a dataset D, and a subset of data D' that needs to be forgotten, our goal is to obtain a new model A'. Ideally, the behavior distribution of A' should be identical (or statistically indistinguishable) to that of a model A" trained from scratch only on the remaining dataset D"=D-D', without ever seeing D'.

Based on the degree to which this goal is approximated, existing techniques are mainly divided into two categories: exact unlearning and approximate unlearning.

2.1 Exact Unlearning

Exact unlearning requires $P(A') = P(A'')$. This means that, from a probabilistic distribution perspective, the unlearned model is indistinguishable from the retrained model, thus mathematically eliminating all influence of data D' completely.

2.1.1 Naive Retraining

The most straightforward method is to remove D' and retrain the model from scratch on D" with random initialization.

Advantages: Unquestionable legal compliance, completely eliminating the risk of data leakage.

Disadvantages: This is unacceptable for LLMs. Training a Llama 3 or GPT-4 level model requires thousands of H100 GPUs running for months, costing tens of millions of dollars, and consuming enormous amounts of electricity while generating significant carbon emissions (Liu et al., 2024a). Therefore, naive retraining only serves as a "gold standard" for evaluating other methods, not a practical solution.

2.1.2 SISA architecture (Sharded, Isolated, Sliced, Aggregated)

To reduce the cost of retraining, Bourtoule et al. proposed the SISA framework, which core idea is "divide and conquer."

Mechanism: The training dataset D is randomly divided into S independent partitions, and a sub-model is trained independently on each partition. When a request to delete data point x is received, since x only exists in one specific partition, the system only needs to retrain the sub-model corresponding to that partition, without needing to operate on the other (S-1) models.

Limitations:

- **Performance degradation:** The capabilities of large language models typically stem from global interactions across massive amounts of data. Data partitioning disrupts semantic coherence and affects the model's coverage of long-tail knowledge, thereby leading to a decrease in the overall intelligence level of the model.
- **Storage and Inference Costs:** Maintaining S sub-models requires S times the storage space, and inference requires aggregating the outputs of multiple models, which will increase latency ("ARCANE: An Efficient Architecture for Exact Machine Unlearning", 2022).

2.1.3 Modern variations: SIFT-Masks and model fusion

Addressing the limitations of SISA (Selective Information Suppression and Adaptation) in LLMs, recent research proposes a precise forgetting scheme based on model merging, such as SIFT-Masks (Sign-Fixed

Tuning-Masks). Instead of directly training multiple large models, it introduces constraints during the fine-tuning phase, allowing updates for different tasks to be effectively "masked" and merged.

Experiments show that SIFT-Masks achieves 5-80% higher accuracy than naive model averaging when merging a large number of models, and the computational cost of performing precise forgetting is 250 times lower than retraining. This provides a highly promising approach for handling deletion requests of large-scale SFT data, especially in enterprise-level private deployment scenarios ("Exact Unlearning of Finetuning Data via Model Merging at Scale", 2025).

2.2 Approximate Unlearning

Due to the high cost of precise forgetting, most research has shifted towards approximate forgetting. These methods do not aim for a mathematically exact equivalent to retraining the model, but instead attempt to find a state in the parameter space that is "sufficiently close" to the retrained model by optimizing algorithms. This is essentially a multi-objective optimization problem concerning maximizing forgetting efficacy and maximizing model utility.

2.2.1 Gradient Ascent

This is the most intuitive approximation method. If the training process involves gradient descent along the loss function to minimize prediction error, then the forgetting process would seemingly involve gradient ascent to maximize prediction error.

However, simple gradient ascent is extremely dangerous. Studies have found that unconstrained maximization of the loss function pushes model parameters into anomalous regions in the parameter space, causing the model to not only forget the target data but also lose basic language generation capabilities (e.g., starting to output gibberish or repeating characters). This is because the loss function is unbounded in the ascending direction, and the high-dimensional parameter space is extremely complex and uncertain (Zhang et al., 2024a).

To mitigate this collapse, researchers have proposed gradient difference, which involves minimizing the loss on the retained set while simultaneously maximizing the loss on the forgetting set. Nevertheless, gradient ascent-based methods still face serious hyperparameter sensitivity issues and often lead to a sharp decline in model utility after the forgetting rate exceeds a certain threshold (e.g., 10%) ("TOFU: A Task of Fictitious Unlearning for LLMs", 2024).

2.2.2 Negative Preference Optimization

To address the stability issues of GA, Zhang et al. (2024) proposed NPO, which is currently one of the state-of-the-art methods. NPO draws inspiration from Direct Preference Optimization (DPO) in Reinforcement Learning from Human Feedback (RLHF). DPO aligns the model by increasing the probability of positive samples and decreasing the probability of negative samples; NPO, on the other hand, treats the forgetting set as simply "negative samples." It introduces the original pre-trained model as a reference and adds a constraint term (usually based on KL divergence or log ratio) to the loss function, preventing the model distribution after forgetting from deviating too far from the original model.

Studies have shown that NPO leads to model performance collapse at an exponentially slower rate than GA. This allows NPO to maintain the model's basic language and logical reasoning abilities even under

extremely aggressive forgetting settings (such as forgetting 50% of the training data) (Zhang et al., 2024a).

2.2.3 Task Vectors

Another emerging class of approximation methods is based on the geometric properties of the parameter space. Ilharco et al. found that subtracting the pre-trained model parameters from the fine-tuned model parameters yields a "task vector" representing a specific task or knowledge, allowing for "forgetting" simply through vector subtraction.

This method is computationally inexpensive and does not even require access to the original data, only manipulating the model weights. However, this linear assumption does not always hold, especially in deep nonlinear networks. For complex, entangled knowledge, simple vector subtraction may inadvertently damage other functionalities ("MUSE: Machine Unlearning Six-Way Evaluation for Language Models", 2024a).

3 Influence Functions and the Failure of Data Attribution

Before performing any forgetting operation, a more fundamental question arises: how do we know which parameters are responsible for which memories? This is the field of data attribution. Without accurate localization, forgetting operations will be subject to significant errors.

3.1 Influence Function

Influence functions are a classic tool in statistics used to quantify the marginal impact of individual training samples on model parameters. Their core calculation relies on the inverse of the Hessian matrix (the second derivative of the loss function). However, for large language models with 7 billion parameters (7B), the Hessian matrix contains approximately 490 quadrillion elements, making it virtually impossible to directly store or invert such a massive matrix with existing computational resources. To overcome this bottleneck, researchers have introduced approximation algorithms such as EK-FAC, which utilize the properties of the Kronecker product to decompose the originally enormous Hessian matrix into several smaller, more manageable matrices. This significantly reduces computational complexity while effectively preserving key second-order curvature information ("Do Influence Functions Work on Large Language Models?", 2025).

3.2 Attribution Failure in the Era of Large Language Models

Although methods like EK-FAC have achieved some success on small and medium-sized models, data attribution still faces significant challenges when dealing with LLMs pre-trained on trillions of tokens.

Firstly, even using approximate methods, calculating attributions for trillions of data points requires an astronomical amount of computing power. Current research is mostly limited to fine-tuning data or very small subsets of pre-training data ("Where Did It Go Wrong? Attributing Undesirable LLM Behaviors via Representation Gradient Tracing", 2025).

Secondly, recent empirical studies (Grosse et al., 2023) show that the estimation error of influence

functions significantly increases with the depth and width of the model. In LLMs, the "most influential data" identified by influence functions are often not the true causal reasons that led to the model's output. This is known as "attribution fragility" ("Do Influence Functions Work on Large Language Models?", 2025).

In summary, in the absence of precise attribution methods, current machine unlearning techniques mostly employ a "carpet bombing" strategy (such as applying GA or NPO to all parameters), which inevitably leads to inefficiency and side effects (hallucinations).

4 Assessment Criteria: How to Prove "Forgotten"

One of the biggest challenges in the field of machine unlearning is the lack of a unified evaluation standard. A simple "refusal to answer" does not prove that the knowledge has been erased.

4.1 TOFU Benchmark (Task of Fictitious Unlearning)

TOFU (Maini et al., 2024) is a benchmark specifically designed to evaluate the forgetting effects of LLMs. It consists of biographies of 200 fictional authors. Since these authors are fictional and their names are generated, it can be confirmed that pre-trained models (such as Llama 2) have never seen them. All knowledge comes from a specific fine-tuning phase. This creates a perfectly controlled experimental environment.

Evaluation metrics:

- **Forgetting Quality:** The model's performance on the forgetting set should be close to random.
- **Model Utility:** The model's performance on the retain set should remain unchanged.
- **Generalization Ability:** TOFU tests not only question answering but also question answering with different phrasings to prevent the model from simply blocking specific prompts ("TOFU: A Task of Fictitious Unlearning for LLMs", 2024).

Finally, on the TOFU leaderboard, the NPO method demonstrates an excellent performance-utility trade-off. In contrast, gradient ascent (GA) shows a rapid collapse in model utility when attempting to forget more data ("TOFU: A Task of Fictitious Unlearning for LLMs", 2024).

Table 1: Comparison of Unlearning Methodologies

Methodology	**Forget quality**	**Model Utility**	**Conclusion**
Naive Retraining	100% (Baseline)	100% (Baseline)	The gold standard, but at a very high cost.
Gradient Ascent	High	Low (Collapse)	This can easily lead to garbled output from the model.
Gradient Difference	Medium	Medium	Balanced, but with limited effectiveness.
Negative Preference	High	High	One of the best approximation methods currently available.

4.2 MUSE Benchmark (Machine Unlearning Six-Way Evaluation)

MUSE (Shi et al., 2024) provides a more comprehensive six-dimensional evaluation framework, covering real-world data such as news and books ("MUSE: Machine Unlearning Six-Way Evaluation for Language Models", 2024a).

- **C1 No verbatim memorization:** The model should not be able to recite the training text verbatim.
- **C2 No knowledge memorization:** The model should not be able to answer knowledge-based questions derived from the text.
- **C3 No privacy leakage:** The model should be resistant to membership inference attacks (MIA).
- **C4 Utility preservation:** Scores on general benchmarks such as MMLU should not significantly decrease.
- **C5 Scalability:** The algorithm should be able to handle large-scale unlearning requests.
- **C6 Sustainability:** The algorithm should support multiple consecutive unlearning requests without degradation ("MUSE: Machine Unlearning Six-Way Evaluation for Language Models", 2024a).

MUSE testing revealed "false unlearning" in many algorithms; for example, some algorithms scored highly on C1 (no longer reciting), but failed on C2 (still knowing the facts), or failed on C3 (attackers could still infer the data's existence with high probability) ("MUSE: Machine Unlearning Six-Way Evaluation for Language Models", 2024a).

5 Erased Illusions: Adversarial Dynamics in Machine Forgetting

This is the most unsettling discovery in the field of machine unlearning: what currently appears to be "unlearning" may simply be a deeper form of "hiding."

5.1 Jogging the Memory

The "Relearning Attack" fundamentally challenges the effectiveness of approximate forgetting. In experiments, for a model that had "forgotten" the Harry Potter novels, attackers only needed to provide a very small amount of relevant clues (such as a general description of Hogwarts, or even a non-original summary generated by GPT-4), and then fine-tune the model on these clues for a very short time. The knowledge that the model had supposedly "forgotten" would be quickly reactivated, restoring its ability to output the original text.

This indicates that approximate forgetting algorithms do not truly erase the encoding of knowledge from the parameter space; they only suppress the retrieval paths to that knowledge or lower the activation values of specific weights. Once given a little "hint" (gradient signal), these paths are reopened ("Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning", 2025).

5.2 Membership Inference Attacks, MIA

Membership Inference Attack (MIA) is a touchstone for evaluating the effectiveness of privacy protection. Attackers assess the effectiveness of privacy protection by determining whether a specific data record was included in the model's training set.

Studies show that even after unlearning, existing defensive unlearning techniques often fail to completely reduce the success rate of MIA attacks to the level of random guessing (50%). The loss distribution of the unlearned model, when faced with "unlearned samples," often still shows subtle differences compared to samples it has never seen before. High-level attackers can still use these statistical features to determine that the data was previously used ("Strong Membership Inference Attacks on Massive Datasets and (Moderately) Large Language Models", 2025).

6 Legal Boundaries: The Collision of GDPR and AI Legislation

6.1 The interpretative challenges of the "right to be forgotten"

Article 17 of the GDPR grants data subjects the right to request the "erasure" of their data. However, the law does not provide a technical definition of "erasure."

A stricter interpretation would require the complete physical destruction of the data. In the context of AI, this means ensuring that the data cannot be recovered by any means (including re-learning attacks). Based on the previous analysis, apart from SISA and complete retraining, no existing approximate forgetting techniques can meet this standard. If regulators adopt this interpretation, it would be devastating to the AI industry, forcing companies to frequently retrain models ("The Right to Be Forgotten and AI", n.d.).

Another, more functional interpretation emphasizes that Article 17(2) and Recital 66 of the GDPR state that controllers should take "reasonable steps" when performing "forgetting" operations, considering "available technology" and "implementation costs." This provides a legal buffer. If the cost of retraining is disproportionate, and techniques like NPO can effectively prevent 99% of conventional access, courts may be inclined to consider this as constituting "reasonable steps." However, this depends on the assessment of "risk": for medical data, functional masking may not be sufficient; for general data, it might be acceptable (Ashok, 2025).

6.2 The Transparency Trap of the EU's AI Act

Article 53 of the EU AI Act requires providers of general-purpose AI models to publicly disclose a detailed summary of their training data ("Article 53: Obligations for Providers of General-Purpose AI Models", n.d.).

This provision aims to increase transparency, but in practice, it may exacerbate the problem of forgetting. Publicly disclosing data sources will lead to more copyright holders discovering that their works have been used, triggering a large number of "opt-out" requests. This will result in an exponential increase in forgetting requests, making retraining even more unsustainable.

Furthermore, unlike the ambiguity surrounding privacy rights, copyright removal requests are often more rigid. If a model continues to generate protected content (even with a low probability), it constitutes in-

fringement. The AI Act explicitly requires models to comply with copyright directives, which may force technology companies to develop more thorough forgetting technologies than current NPO methods, or seek solutions at the architectural level (such as RAG) ("GenLaw ICML 2024 – Accepted Papers", n.d.).

7 Conclusion and Outlook

7.1 Technical summary: The impossible triangle

Current machine unlearning techniques are in a typical "impossible triangle": completeness, utility, and efficiency are difficult to achieve simultaneously. For the various technical solutions proposed so far:

- **SISA/Retraining:** Guarantees completeness and utility, but sacrifices efficiency.
- **Gradient Ascent (GA):** Sacrifices utility (leading to failure), and completeness is questionable.
- **NPO/SIFT-Masks:** Achieves the best balance between efficiency and utility, and is currently the closest to a practical solution, but remains vulnerable to adversarial attacks (completeness).

7.2 Legal Advice: Moving Towards Dynamic Compliance

Given the current state of technology, legal regulation should not pursue absolute "zero residue," but rather shift towards risk-oriented governance. On the one hand, a tiered certification standard can be established. For example, regulatory agencies can collaborate with the technology industry to establish a tiered certification of forgetting effects based on benchmarks such as TOFU and MUSE. For instance, "Level 1 forgetting" corresponds to NPO processing, applicable to general data; "Special Level forgetting" corresponds to SISA or retraining, applicable to highly sensitive data.

On the other hand, the depth of "erasure" needs to be redefined. In the AI era, "erasure" should not be simply understood as physical disappearance, but perhaps more accurately described as a combination of "inaccessibility" and "uninferability." Furthermore, "erasure" results where the cost of attack is sufficiently high, and the company has conducted continuous adversarial testing to patch vulnerabilities, can be considered compliant.

7.3 Future Works

The future direction may not lie in fixing existing dense models, but in architectural innovation. For example, by utilizing retrieval-augmented generation techniques, outsourcing memory to an external database, forgetting would simply involve deleting documents from the database without touching the model parameters ("Scaling Retrieval-Based Language Models with a Trillion-Token Datastore", 2024). Alternatively, consider a modular design, distributing knowledge across different expert modules; "forgetting" would then only require removing or retraining specific expert modules, with minimal impact on the overall system.

Machine unlearning, as a bridge connecting AI technology and human rights, is not only an algorithmic problem but also a philosophical question about how we define "memory," "privacy," and "responsibility" in the age of algorithms.

References

[literature review] exact unlearning of finetuning data via model merging at scale. (n.d.). Retrieved November 27, 2025, from https://www.themoonlight.io/en/review/exact-unlearning-of-finetuning-data-via-model-merging-at-scale

Allen Institute for AI. (n.d.). *Ai2 dolma: 3 trillion token open corpus for language model pretraining*. Retrieved November 27, 2025, from https://allenai.org/blog/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64

Alternate preference optimization for unlearning factual knowledge in large language models. (2025). Retrieved November 27, 2025, from https://aclanthology.org/2025.coling-main.252.pdf

ARCANE: An efficient architecture for exact machine unlearning. (2022). Retrieved November 27, 2025, from https://www.ijcai.org/proceedings/2022/0556.pdf

Article 53: Obligations for providers of general-purpose AI models. (n.d.). Retrieved November 27, 2025, from https://artificialintelligenceact.eu/article/53/

Ashok, P. (2025). *The goldilocks standard: Machine unlearning and the right to be forgotten under emerging legal frameworks*. Retrieved November 27, 2025, from https://cep-project.org/wp-content/uploads/2025/11/Pratiksha-Ashok-THE-GOLDILOCKS-STANDARD-Machine-Unlearning-and-the-Right-to-be-Forgotten-Under-Emerging-Legal-Frameworks.pdf

Comparing precision knowledge editing with existing machine unlearning methods [Reddit discussion thread]. (n.d.). Retrieved November 27, 2025, from https://www.reddit.com/r/artificial/comments/1gxmgxw/comparing_precision_knowledge_editing_with/

Do influence functions work on large language models? (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2409.19998v1

Do influence functions work on large language models? (2025). Retrieved November 27, 2025, from https://aclanthology.org/2025.findings-emnlp.775.pdf

European Data Protection Board. (2025). *Law & compliance in AI security & data protection*. Retrieved November 27, 2025, from https://www.edpb.europa.eu/system/files/2025-06/spe-training-on-ai-and-data-protection-legal_en.pdf

Exact unlearning of finetuning data via model merging at scale. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2504.04626v1

Extracting unlearned information from LLMs with activation steering. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2411.02631v1

First is not really better than last: Evaluating layer choice and aggregation strategies in language model data influence estimation. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2511.04715

Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. (n.d.). Retrieved November 27, 2025, from https://ojs.aaai.org/index.php/AAAI/article/view/34218/36373

General-purpose AI regulation and the european union AI act. (n.d.). Retrieved November 27, 2025, from https://policyreview.info/articles/analysis/general-purpose-ai-regulation-and-ai-act

GenLaw ICML 2024 – accepted papers. (n.d.). Retrieved November 27, 2025, from https://www.genlaw.org/2024-icml-papers

Jogging the memory of unlearned LLMs through targeted relearning attacks. (n.d.). Retrieved November 27, 2025, from https://openreview.net/pdf?id=YulEbrG99x

Liu, Z., Dou, G., Tan, Z., Tian, Y., & Jiang, M. (2024a). *Machine unlearning in generative AI: A survey*. Retrieved November 27, 2025, from https://www.researchgate.net/publication/382692727_Machine_Unlearning_in_Generative_AI_A_Survey

Liu, Z., Dou, G., Tan, Z., Tian, Y., & Jiang, M. (2024b). *Machine unlearning in generative AI: A survey*. Retrieved November 27, 2025, from https://arxiv.org/html/2407.20516v1

LUME: LLM unlearning with multitask evaluations. (2025). Retrieved November 27, 2025, from https://aclanthology.org/2025.findings-emnlp.347.pdf

Machine learners should acknowledge the legal implications of large language models as personal data. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2503.01630v2

Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2412.06966v1

Machine unlearning: A comprehensive survey. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2405.07406v2

Microsoft Research. (n.d.). *Who's harry potter? making LLMs forget*. Retrieved November 27, 2025, from https://www.microsoft.com/en-us/research/articles/whos-harry-potter-making-llms-forget-2/

MUSE: Machine unlearning six-way evaluation for language models. (2024a). Retrieved November 27, 2025, from https://arxiv.org/abs/2407.06460

MUSE: Machine unlearning six-way evaluation for language models. (2024b). Retrieved November 27, 2025, from https://arxiv.org/html/2407.06460v1

Position: Bridge the gaps between machine unlearning and AI regulation. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2502.12430v2

R-TOFU: Unlearning in large reasoning models. (2025). Retrieved November 27, 2025, from https://aclanthology.org/2025.emnlp-main.265.pdf

Rethinking LLM unlearning objectives: A gradient perspective and go beyond. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2502.19301v1

The right to be forgotten and AI. (n.d.). Retrieved November 27, 2025, from https://lup.lub.lu.se/student-papers/record/9189308/file/9197890.pdf

Scaling retrieval-based language models with a trillion-token datastore. (n.d.). Retrieved November 27, 2025, from https://retrievalscaling.github.io/

Scaling retrieval-based language models with a trillion-token datastore. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2407.12854v1

Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2410.07163v4

Step-by-step reasoning attack: Revealing 'erased' knowledge in large language models. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2506.17279v1

Strong membership inference attacks on massive datasets and (moderately) large language models. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2505.18773v1

Studying large language model generalization with influence functions [TransferLab blog post]. (n.d.). Retrieved November 27, 2025, from https://transferlab.ai/pills/2023/llm-influences-with-ekfac/

Studying large language model generalization with influence functions. (2023). Retrieved November 27, 2025, from https://arxiv.org/abs/2308.03296

A survey of machine unlearning in large language models: Methods, challenges and future directions. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2503.01854v2

TOFU: A task of fictitious unlearning for LLMs. (2024). Retrieved November 27, 2025, from https://arxiv.org/html/2401.06121v1

Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. (n.d.). Retrieved November 27, 2025, from https://assets.amazon.science/21/03/30e37e1740349fff6a809a689bed/unlearning-as-multi-task-optimization-a-normalized-gradient-difference-approach-with-an-adaptive-learning-rate.pdf

Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. (2025). Retrieved November 27, 2025, from https://blog.ml.cmu.edu/2025/05/22/unlearning-or-obfuscating-jogging-the-memory-of-unlearned-llms-via-benign-relearning/

Where did it go wrong? attributing undesirable LLM behaviors via representation gradient tracing. (2025). Retrieved November 27, 2025, from https://arxiv.org/html/2510.02334v1

WilmerHale. (2025). *Wilmerhale's guide to AI and GDPR*. Retrieved November 27, 2025, from https://www.wilmerhale.com/-/media/files/shared_content/editorial/publications/20250919-wilmerhales-guide-to-ai-and-gdpr.pdf

Zhang, R., Lin, L., Bai, Y., & Mei, S. (2024a). *Negative preference optimization: From catastrophic collapse to effective unlearning*. Retrieved November 27, 2025, from https://arxiv.org/abs/2404.05868

Zhang, R., Lin, L., Bai, Y., & Mei, S. (2024b). *Negative preference optimization: From catastrophic collapse to effective unlearning*. Retrieved November 27, 2025, from https://openreview.net/forum?id=MXLBXjQkmb

Zhang, R., Lin, L., Bai, Y., & Mei, S. (2024c). *Negative preference optimization: From catastrophic collapse to effective unlearning*. Retrieved November 27, 2025, from https://arxiv.org/pdf/2404.05868