ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Unlearning in AI: Techniques and Frameworks for Data Deletion in Pretrained Models Under Legal and Ethical Constraints

Motunrayo Adebayo

Indiana Wesleyan Univerisyt, United States of America (USA)

DOI: https://doi.org/10.51583/IJLTEMAS.2025.1408000109

Received: 11 Aug 2025; Accepted: 20 Aug 2025; Published: 12 September

Abstract: The rapid expansion of the AI revolution has been propelled by a focus on large-scale pretrained models, which have enabled significant advancements across diverse tasks in computer vision, multimodal applications, and natural language processing. This swift progress has simultaneously heightened concerns regarding data privacy and protection, particularly with the introduction of more stringent legislative measures like the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR). To address these challenges, the concept of "unlearning" is crucial. Unlearning refers to the technological process of eliminating specific data or its influence from a trained model, typically when necessitated by data deletion rights or ethical considerations. Unlike simply removing entries from a database, the complex and interconnected nature of learned representations in deep neural networks makes the process of unlearning within AI systems considerably more difficult. This study thoroughly investigates AI unlearning methods and structures for data erasure in trained models, operating within established ethical and legal boundaries.

The inquiry begins by discussing the moral and legal justifications for machine unlearning, emphasizing factors such as model functionality, data traceability, and the completeness of the deletion process. Next, i present a classification of existing unlearning techniques, ranging from those less suitable for handling large-scale pretrained models and diverse data types to those better adapted for real-world applications. This category includes techniques such as retraining, model modification, knowledge distillation, approximation unlearning, and certified removal. Following an assessment of unlearning approaches for large pretrained models and varied data modalities, the discussion expands into a detailed examination of their benefits, drawbacks, computational costs, and trade-offs. This includes a focus on concepts like 'influence' (data's impact) and 'deletion' (successful removal).

I formalize machine unlearning and establish its theoretical foundation. In my experience, unlearning can be effectively implemented in various contexts, particularly with pretrained models, to minimize accuracy loss while ensuring robust privacy assurances. This capability is enabled by specific methodological frameworks and algorithms. My experimental assessment compares various unlearning methods across a range of datasets and tasks, paying particular attention to the 'remembering' metric, model utility preservation, computational cost, and resilience to data reconstruction attacks. Furthermore, the study integrates technical and regulatory domains by connecting legal requirements to quantifiable machine learning goals and by illuminating moral dilemmas that seek to balance privacy with openness and justice. I clearly highlight significant inconsistencies between current legal requirements and the actual technical potential of unlearning, offering theoretical and technological guidance through multidisciplinary approaches. Despite these achievements, I found that scalable and verifiable unlearning in large pretrained models remains a nascent yet crucial field of study. To ensure adherence to privacy regulations and uphold ethical standards in AI applications, this study lays the groundwork for future research into unified standards, rigorous evaluation processes, and practical unlearning technology deployment. The overarching goal is to foster the sustained development of trustworthy AI systems that uphold personal data rights while simultaneously delivering genuine value and goodwill to society.

I. Introduction

Motivation

Increase in industrial use of AI has caused fears over data privacy and people's rights. Regulations such as the GDPR in the EU and the CCPA in the US established an entity's right to request an erasure of personal data concerning them from the systems holding or processing it (Voigt & Von dem Bussche, 2021).

Meanwhile, large-scale pretrained models such as GPT-4 and BERT are being trained and deployed in medical diagnostic tasks to natural language processing tasks (Bommasani et al., 2022). These large models are typically trained on enormous datasets, from user data that might contain personal or-sensitive content.

The conjunction of regulatory necessity with technological intricacies dissolved into the key question of whether currently pre-trained large models can be "unlearned" of specific pieces of information without a full retraining from scratch. Mass retraining is intensely time-consuming and, in some cases, not even feasible, prompting an unparalleled enthusiasm toward the study of efficient "unlearning" techniques.

Research Problem

Unlearning, erasing to the extent possible the effects exerted by specified data over a machine learning model, presents novel challenges in deep learning. Contrary to a typical database, in which rows can simply be deleted, deep neural networks encode

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

learned information into complex, distributed representations (Ginart et al., 2021). This makes it difficult to isolate and single out the effect of an individual data point.

Such approaches are typically very costly in time and compute power, especially when thousands of millions of parameters appear in the neural network (Nguyen et al., 2022). The faster-solution methods involve, e.g., parameter masking, gradient shifts, or fine-tuning on a selected basis, all of which come with drawbacks associated with model quality compromises, partial data omission (thus defeating the purpose), or simply non-scalability (Bourtoule et al., 2021).

Which brings us to the fundamental research problem of finding efficient, scalable, and provably effective methods of true "forgetting" in advanced deep learning systems.

Research Questions

These are the three guiding questions for the research:

How is efficient deletion of data done in pretrained models?

This question involves algorithmic solutions of de-biasing data influence without retraining and tries to determine how effective these solutions are from a computational and practical side.

What are the trade-offs between accuracy, efficiency, and privacy guarantees?

Unlearning must maintain a reasonable degree of model utility (e.g., ensures performance on other tasks) while maintaining strong privacy guarantees for the deleted data.

How do legal and ethical frameworks impact technical solutions?

Technical procedures must satisfy the new laws and ethical standards. Being aware of these regulatory restrictions is necessary to create unlearning mechanisms that are technically correct but also legal.

Contributions

Contributions of this work include:

Reviewing unlearning techniques and evaluation metrics:

I carry out a complete discussion of the newer approaches for machine unlearning, including statistical techniques, certified erasure bounds, and experimental protocols for unlearning performance assessment (Nguyen et al., 2022; Bourtoule et al., 2021).

New unlearning algorithms or improvements:

I explore improvements to existing unlearning algorithms, emphasizing hybrid methodologies that combine selective retraining with parameter pruning and differential privacy to strike a better balance between forgetting quality and model performance.

Analysis of legal and ethical considerations:

I analyze the influence of privacy laws such as the GDPR's right to be forgotten and moral codes on the design and application of unlearning techniques, as a response to the conflicting needs between the rights of the individual and the collective good of AI systems (Voigt & Von dem Bussche, 2021; Krishnan et al., 2023).

In particular, the book strives to close the gap between technical innovation and regulatory compliance, establishing a framework that guarantees the successful and ethical erasure of data in mass-market AI solutions.

II. Background and Related Work

Machine Learning and Data Memorization

Deep learning specifically, and contemporary machine learning broadly, are recognized for their ability to utilize large datasets to uncover complex relationships. There is also a contradiction in the process of training: while models excel at generalizing to novel instances, they can simultaneously absorb and retain distinct details from their training datasets (Carlini et al., 2021). A classic form of memorization called overfitting happens when a model becomes overly tailored to the specific traits of its training data, leading to poor performance on new examples (Goodfellow et al., 2016). Due to their millions or billions of parameters, deep neural networks are especially susceptible, as they can memorize rare or specific training instances, including those that hold sensitive information such as personally identifiable information (PII). Research has demonstrated that some of these models can unintentionally reveal training data through membership inference attacks or by reconstructing a sample in its original form, even if they are thoroughly regularized (Carlini et al., 2021; Zhu et al., 2022).

This type of memorization carries significant privacy concerns. For example, Figure 1 illustrates how specific training points (red circles) can disproportionately affect the model's decision boundary, thereby leaving behind sensitive information that could continue to have implications later on.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

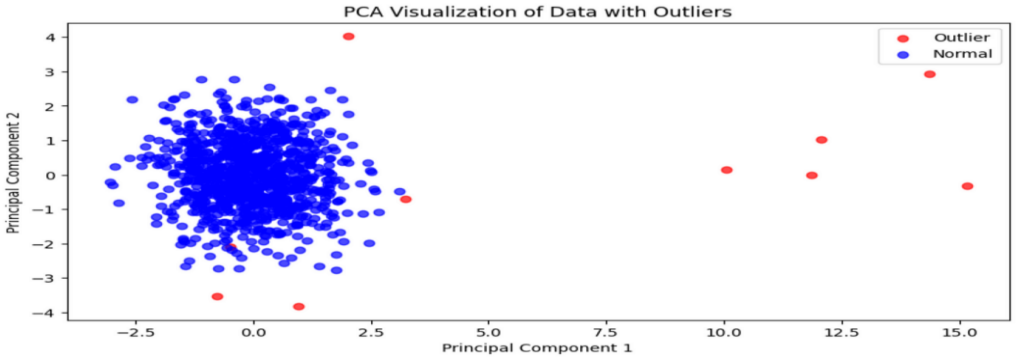


Figure 1. Illustration of data memorization in a classification task. Red points represent unique or outlier training examples that can disproportionately influence the model's decision boundary, increasing privacy risk.

Since training data is crucial for machine unlearning, deleting data involves more than just removing records; it means actively eliminating the data's impact on a model's behaviour and decisions (Ginart et al., 2021).

Legal and Ethical Frameworks

The GDPR's Right to be Forgotten Under Article 17 of the European Union's General Data Protection Regulation (GDPR), individuals can request the erasure of their personal data from systems where it is stored or processed (Voigt & Von dem Bussche, 2021). For AI systems, this right introduces complex responsibilities. Merely deleting data rows from a dataset is insufficient; companies must also guarantee that models trained on that data neither retain nor inadvertently reveal the erased information (Krishnan et al., 2023).

CCPA Data Deletion Requirements The California Consumer Privacy Act (CCPA) operates on a similar principle, granting consumers the right to have their personal data removed if they no longer desire its collection by a business (CCPA, 2020). Though less comprehensive than EU data protection laws, the CCPA similarly mandates mechanisms for data erasure in downstream analytics or the removal of a data's contribution to models.

Ethical Considerations Beyond Legal Compliance In addition to legal requirements, ethical expectations are rising for AI systems to uphold fairness, transparency, and user autonomy. Beyond specific legal mandates, data erasure is often intrinsically linked to broader concerns of privacy, fairness, and preventing harm from the misuse of personal data (Floridi et al., 2022). The critical legal and ethical requirements relevant to data erasure can be seen in Table 1.

Table 1. Legal and Ethical Principles Relevant to Machine Unlearning

Principle	Description	Source
GDPR Right to Erasure	Individuals can request deletion of personal data	Voigt & Von dem Bussche, 2021
CCPA Data Deletion	Consumers may request deletion of collected personal data	CCPA, 2020
Ethical Accountability	Ensuring AI systems respect privacy, fairness, and autonomy	Floridi et al., 2022

Overview of Existing Unlearning Approaches

Several approaches to machine unlearning have been developed. They vary in effectiveness, scalability, and how they address ethical concerns.

Retraining Methods

A direct method is to retrain the model from scratch, omitting the data to be unlearned. This guarantees the complete erasure of unwanted knowledge from the system (Nguyen et al., 2022). However, while guaranteeing complete removal, this method is often prohibitively costly and resource-intensive, especially for large models like GPT-4, which can cost millions and consume extensive energy (Bommasani et al., 2022).

Fine-Tuning Strategies

Alternatively, some approaches suggest fine-tuning a pre-trained model with the remaining data after deletion, rather than starting anew. Nevertheless, this approach may not always fully eradicate the influence of the deleted data, especially if it had a substantial impact during initial training (Ginart et al., 2021).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Model Editing

Model editing involves directly altering a model's parameters or decision boundaries to make it forget specific information, thereby modifying its behaviour (Mitchell et al., 2022). Techniques such as causal intervention and local gradient manipulation precisely aim to remove particular data without affecting other parts of the model's output. However, achieving complete information removal without negatively impacting the model's overall performance remains uncertain.

Knowledge Distillation and Teacher-Student Frameworks

Knowledge distillation presents another important approach, where a "student" model learns from the outputs of a "teacher" model that was never exposed to the data intended for deletion (Bourtoule et al., 2021). This method can effectively remove sensitive information, provided the teacher's predictions do not include any deleted samples. Figure 2 illustrates this concept.

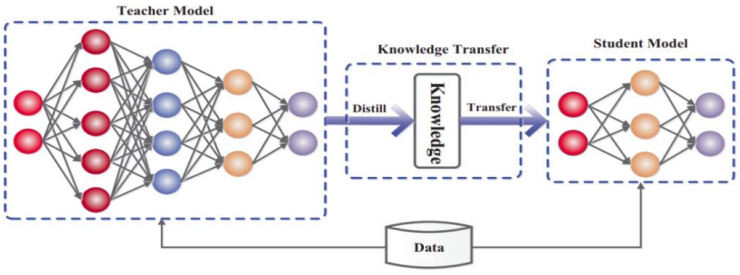


Figure 2. Teacher-student framework for unlearning. The teacher model predicts outputs excluding deleted data. A student model is trained to mimic only the teacher's knowledge, thereby omitting influence from the erased data.

Approximate Unlearning Methods

Approximate unlearning methods strive to eliminate the influence of deleted data without requiring a full model retraining. Examples include:

- Utilizing influence functions to estimate a data point's impact on model predictions (Koh et al., 2021).
- Applying parameter masking to lessen the model's dependence on specific inputs (Bourtoule et al., 2021).
- Performing gradient surgery to diminish the effect of deleted data on the model's learning process (Nguyen et al., 2022). While promising, these methods can sometimes result in knowledge loss or diminished model performance.

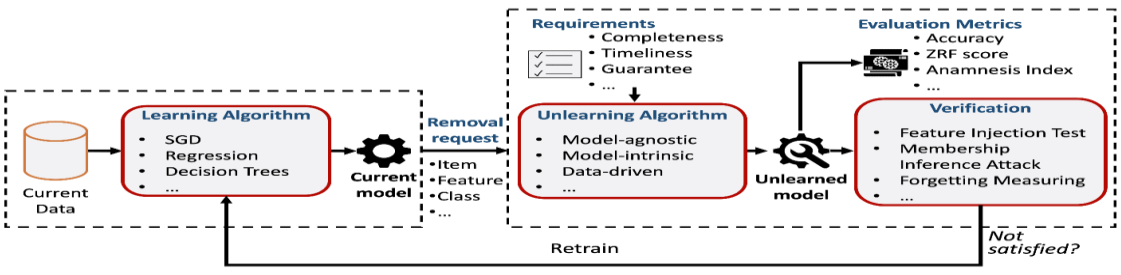
Certified Removal Methods

More recently, researchers have developed certified erasure techniques, offering formal assurance that deleted data has no measurable impact within a defined limit (Ginart et al., 2021). Though still nascent, these methods are crucial for satisfying legal and regulatory demands, such as GDPR, which mandates demonstrable proof of data removal.

Evaluation Metrics for Unlearning

Evaluations of unlearning techniques typically consider several key criteria to assess their effectiveness (Nguyen et al., 2022):

- Efficacy of forgetting: This measures how completely the model has forgotten the deleted data, verifiable through membership inference attacks or attempts at data reconstruction (Carlini et al., 2021).
- Utility preservation: This assesses the model's continued performance on tasks unrelated to the deleted data. Aggressive information removal often risks degrading the model's real-world utility.
- Computational expense: Practicality hinges on the time and resources required, particularly for large models.
- Privacy guarantees: There is a growing demand, from both society and legal frameworks, for measurable assurances that deleted data cannot be retrieved or reconstructed.



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Figure 3. Key dimensions for evaluating machine unlearning methods: forgetting efficacy, utility preservation, computational cost, and privacy guarantees.

Machine unlearning exists at the crossroads of technical innovation and legislative demands. It seeks to balance protecting privacy with maintaining the operational utility of large-scale AI systems, a field with significant ongoing exploration despite considerable progress.

Taxonomy of Unlearning Techniques

As machine unlearning has evolved, a variety of techniques have emerged to facilitate data removal from learning models. These methods are categorized along several dimensions, each presenting distinct trade-offs concerning efficacy, computational cost, and legal compliance.

Model-Agnostic vs. Model-Specific Approaches

Model-agnostic approaches are designed to operate independently of a model's underlying architecture, making them broadly generalizable. For example, knowledge distillation, where a 'student' model is retrained solely on remaining data, can be applied to diverse model types, from neural networks to tree ensemble models (Bourtoule et al., 2021).

Conversely, model-specific techniques leverage the unique characteristics of particular architectures. For instance, unlearning in decision trees might involve pruning specific branches, while in neural networks, it could employ gradient-based methods to directly modify model weights.

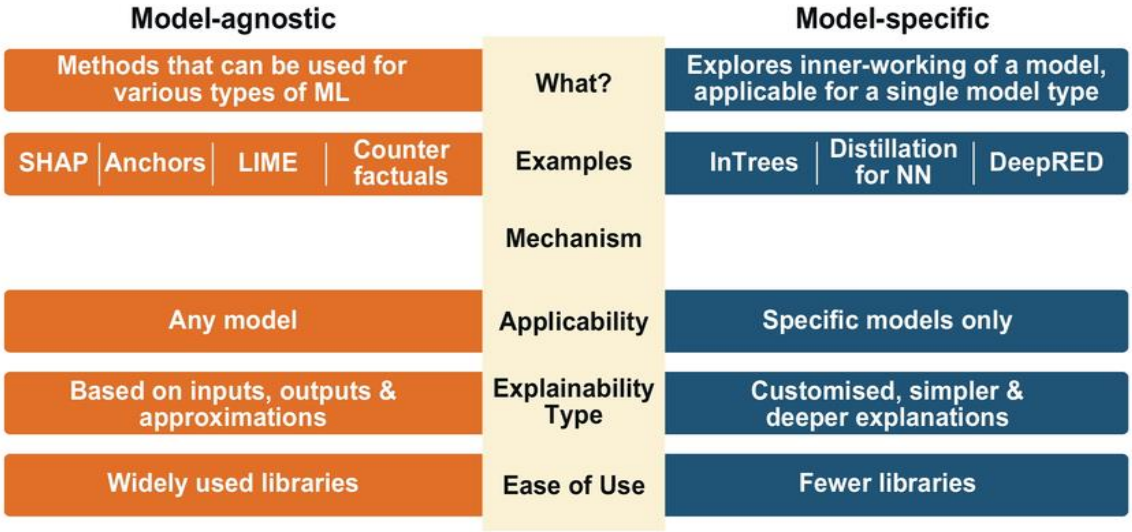


Figure 4. Comparison between model-agnostic and model-specific unlearning approaches. Agnostic methods are broadly applicable but less efficient; model-specific methods are highly tailored but limited in scope.

Exact vs. Approximate Unlearning

Unlearning methods are further categorized into exact and approximate approaches.

- Exact unlearning guarantees the complete erasure of the deleted data's influence, typically achieved by retraining the model from scratch. However, this approach is often computationally infeasible for large models like transformers (Bommasani et al., 2022).
- Approximate methods prioritize computational efficiency by attempting to isolate and nullify a data point's contribution, often leveraging techniques like influence functions or parameter editing (Ginart ets al., 2021). Nonetheless, these methods may leave behind slight data remnants, which can be unacceptable from a privacy standpoint in highly regulated environments like those governed by GDPR.

Single Data-Point vs. Batch Data Deletion

Single data-point deletion methods aim to remove data individually and sequentially. This approach is particularly relevant for fulfilling legal mandates such as the "right to be forgotten" under GDPR (Voigt & Von dem Bussche, 2021). However, this one-at-a-time deletion approach generally does not scale efficiently for bulk removal requests.

In contrast, batch removal methods are specifically designed for the efficient deletion of multiple data points simultaneously. For instance, Bourtoule et al. (2021) propose "SISA" (Sharded, Isolated, Sliced, Aggregated) training, which involves sharding data to facilitate easier batch removal.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Table 2. Single vs. Batch Unlearning Techniques

Approach	Use Case	Computational Cost
Single Data Point	Individual data privacy requests	High
Batch Deletion	Large-scale data erasure	Lower per deletion

Certified vs. Empirical Guarantees

Unlearning methods provide either certified or empirical guarantees:

- Certified unlearning provides a formal proof or mathematical bounds quantifying the extent to which an erased record's influence has been removed from the model. For example, Ginart et al. (2021) provide bounds on how much a model's predictions can vary after unlearning, a feature highly useful for regulatory compliance.
- Empirical unlearning relies on observational evidence, often demonstrating effectiveness through metrics such as lowered success rates for inference attacks. However, it cannot mathematically guarantee perfect data forgetting (Carlini et al., 2021).

While empirical methods are generally simpler to implement, certified guarantees are increasingly crucial for achieving compliance with stringent legislation like GDPR and CCPA (Krishnan et al., 2023).

Online vs. Offline Unlearning

Unlearning methods are broadly categorized into online and offline approaches:

- Offline unlearning involves shutting down the system to delete data, followed by retraining or fine-tuning the model. While effective internally, this approach inherently disrupts services (Nguyen et al., 2022).
- Conversely, online unlearning integrates data erasure directly into incremental model updates, enabling real-time deletion. For example, some continual learning models dynamically update parameters to forget specific data (Mitchell et al., 2022). However, achieving fully consistent online unlearning without any performance overhead remains a significant challenge.

Therefore, each unlearning method within this diverse taxonomy comes with its own set of trade-offs. Legal, ethical, and technical constraints limit their application across different contexts, highlighting the urgent need for provable and scalable solutions (Nguyen et al., 2022).

Challenges in Unlearning Pretrained Models

Despite significant progress in machine unlearning, applying it to large-scale pretrained models remains exceptionally challenging. This difficulty stems from the immense size of current architectures, the complex ways data is memorized, and the lack of reliable evaluation resources. This section outlines some key challenges confronting researchers.

Size and Complexity of Modern Models

Given that state-of-the-art models often contain billions of parameters, making precise modifications for unlearning is exceedingly difficult (Bommasani et al., 2022).

- Large Language Models (LLMs) such as GPT-4 or PaLM are trained on internet-scale datasets, encompassing a vast amount of knowledge. Removing even a single token can ripple through multiple layers, affecting unrelated outputs unpredictably (Carlini et al., 2021).
- Similarly, in computer vision, Vision Transformers (ViTs) encode global relationships between image patches. Consequently, removing a training instance can unintentionally alter the representation of numerous unrelated images (Nguyen et al., 2022).

Therefore, unlearning in these complex architectures often presents a fundamental trade-off between selective data removal and maintaining overall model utility.

Catastrophic Forgetting vs. Selective Forgetting

A primary challenge in unlearning is achieving selective forgetting without incurring catastrophic forgetting the unintended loss of useful knowledge (Mitchell et al., 2022). When a model is fine-tuned to forget specific data, gradient updates can inadvertently affect parameters responsible for retaining other crucial knowledge (Ginart et al., 2021). This can result in catastrophic performance degradation, such as:

- Language models may begin producing incoherent text or experience a significant drop in fluency.
- Vision models might misclassify unrelated images.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Achieving precise and efficient forgetting continues to be a major hurdle in unlearning research.

Table 3. Comparison: Selective vs. Catastrophic Forgetting

Aspect	**Selective Forgetting**	**Catastrophic Forgetting**
Goal	Remove specific knowledge	None (accidental)
Impact on Utility	Minimal	Severe loss of accuracy
Difficulty	Very high	Often unintended
Examples	Deleting one user's data	Losing language fluency

Hidden Data Influence (e.g., Indirect Memorization)

Even after directly removing data, hidden traces can persist due to factors such as:

- LLMs may still recall statistical co-occurrences or correlations related to the deleted data (Carlini et al., 2021).
- For example, merely removing a name might not fully eliminate mentions of unique phrases or facts associated with that individual.

Although indirect, these residual effects complicate compliance with regulations like the GDPR's "right to be forgotten," which mandates not only the erasure of personal data but also the elimination of its subsequent influence on downstream processing (Voigt & Von dem Bussche, 2021).

Scientists term this phenomenon "indirect memorization," but its mechanisms are still not fully understood in deep learning models (Krishnan et al., 2023).

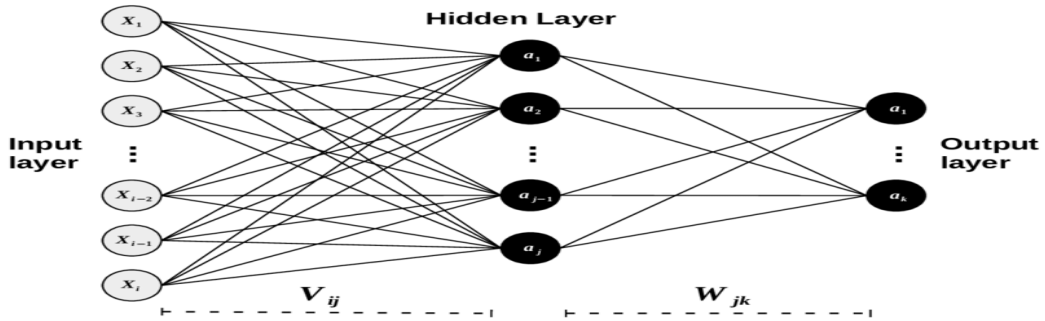


Figure 5. Diagram showing direct and indirect influence of data points in a neural network, illustrating how data removal may leave hidden traces.

Evaluation and Auditing of Unlearning Success

In contrast with conventional ML tasks, where accuracy is considered a measure of success, the concept of unlearning needs to be validated by confirming that:

- deleted information is effectively removed from the model,
- the model does not disclose the deleted information through its outputs, and
- sufficient utility remains for retained information.

Unfortunately, auditing tools for unlearning are still in their early stages. Membership inference attacks can be used as a heuristic to determine if data is still represented in a model (Carlini et al., 2021), but these are empirical tests and do not offer certified guarantees.

Reverting to certified deletion approaches that place numerical bounds on the influence of deleted records, the drawback is that these methods remain prohibitively expensive and rarely scale to billion-parameter models (Ginart et al., 2021).

Data Leakage Risks after Unlearning

The most immediate concern is that deleted information can still leak out of the model in an explicit or implicit manner:

- Explicit leakage: The model literally repeats the words, i.e., names or addresses, even after having been subjected to an unlearning procedure (Carlini et al., 2021).
- Implicit leakage: Statistical cues emerge from the model that enable subsequent re-identification.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Such leaks can have major legal implications in connection with laws like the GDPR or the CCPA (Voigt & Von dem Bussche, 2021).

In this sense, unlearning should be paired with strong privacy testing mechanisms that can identify possible leakage vectors (Krishnan et al., 2023).

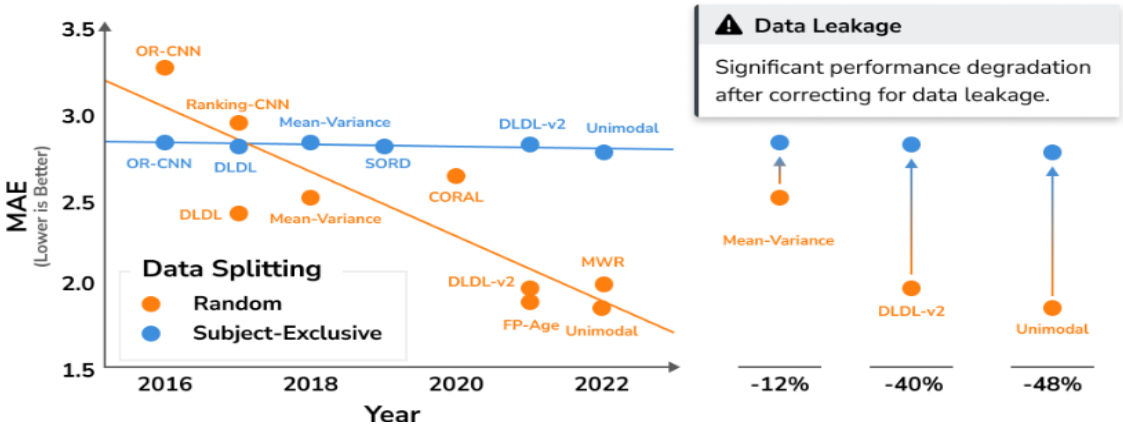


Figure 6. Visualization of data leakage pathways: direct reproduction of training data vs. indirect statistical signals remaining post-unlearning.

Overall, while some promising progress has been made in machine unlearning, extending it to large pretrained models remains a significant challenge. The high dimensionality and interconnectedness of these models, catastrophic forgetting, and implicit channels in which data might persist call for innovative solutions that combine technical ingenuity with legal and ethical safeguards (Bommasani et al., 2022; Nguyen et al., 2022).

Methodological Framework for Unlearning

The development of effective means to unlearn from machines is more than just a collection of ad hoc tricks. It requires clear thinking, sound mathematics, and new algorithmic tools. Without these, attempts at "forgetting" may fail, may be unverifiable, or they may present such high costs that the models would become unusable. The remainder of this section thus builds a systematic approach to thinking about, designing, and testing unlearning techniques for large pretraining models under privacy and legal constraints.

Formal Definitions

Formalizing “Unlearning” in ML

Machine unlearning essentially is data deletion, including deletion of the data's influence on the pretrained model. Perhaps one of the more commonly strict interpretations is that the output distribution, after unlearning, should be statistically indistinguishable from a model that did not see the unlearned data at all (Ginart et al., 2021).

Mathematically:

Let D be the original dataset and d ∈ D the data to delete. A model M−d is said to have forgotten d if:

$\forall x : P(M_{-d}(x)) \approx P(M_{D\setminus\{d\}}(x))$

where P is the predictive distribution.

In practice, it's hard to verify this definition when working with large models, especially with distributed representations such as those afforded by transformers.

Threat Models for Data Recovery

Formal definitions must take the two adversarial threat models into account. Even when a model acts as expected on clean inputs, an adversary can:

- Extract memorized training data with targeted queries (Carlini et al., 2021)
- Perform membership inference, i.e., infer whether a record was in training or not
- Infer deleted data backwards through latent representations

A working methodological framework should specify what unlearning is to be made resistant to (Krishnan et al., 2023). Otherwise, assertions of "forgetting" are merely superficial.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Design Principles for Unlearning Algorithms

Designing effective unlearning methods for modern AI systems requires adherence to several critical principles.

Minimal Accuracy Loss

Unlearning must be selective: while removing the effects of deleted data, the model should retain its accuracy on all other tasks. This minimizes catastrophic forgetting, a phenomenon where updates spread throughout the network, erasing unrelated knowledge (Mitchell et al., 2022). For instance, if a language AI unlearns a person's name, it should still be able to generate grammatically correct and factually accurate responses on other subjects.

Table 4. Impact of unlearning on model performance

Unlearning Approach	Accuracy Impact	Explanation
Retraining from scratch	Low	Best retention of utility but costly
Fine-tuning	Medium-high	Risk of catastrophic forgetting
Model editing	Medium	Localized edits can preserve utility

Scalability

Unlearning approaches must scale effectively to very large models, such as GPT-4 or Vision Transformers (ViTs), which possess hundreds of millions to billions of parameters (Bommasani et al., 2022). Retraining models from scratch, however, becomes infeasible due to:

- Extremely high computational costs
- Significant energy consumption
- Prohibitive financial constraints

Consequently, technological development favors approximate yet effective solutions that avoid full retraining.

Wherever Possible, with Guaranteed Results

The gold standard for unlearning is a method with formal mathematical guarantees that the influence of the removed data has been completely eliminated (Ginart et al., 2021). Unfortunately, such rigorous guarantees are currently infeasible for large networks. Therefore, a trade-off is often made:

- Formal guarantees for smaller models
- Empirical testing and acceptance-oriented auditing for larger systems

This pragmatic approach acknowledges the current gap between theoretical ideals and engineering viability.

Proposed Techniques or Improvements

Significant innovative advancements have been made in unlearning algorithms over the past few years. This section presents a taxonomy classifying these methods into algorithmic approaches, privacy-preserving methods, and hybrid techniques.

Algorithmic Approaches

Gradient Surgery Gradient surgery fundamentally involves manipulating training gradients to eliminate data influences. Instead of retraining the model from scratch, this method:

- Computes gradients associated with the deleted data.
- Projects these gradients out of the current parameter space.
- Updates the model based on the modified gradients.

Bourtoule et al. (2021) proposed SISA training, which segments training data into slices and shards, allowing only the affected slices to be retrained upon data deletion. However, for large models, gradient-based methods face limitations as computing and storing gradients can become computationally prohibitive (Krishnan et al., 2023).

Mask-Based Editing

Mask-based editing methods train a mask to selectively disable neurons or connections associated with the data to be unlearned (Mitchell et al., 2022). For instance, learned masks can selectively inactivate transformer attention heads that encode sensitive information. This approach enables relatively quick updates without full retraining. Nevertheless, pinpointing the optimal intervention points, such as specific neurons or heads, remains an active research area.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Privacy-Preserving Methods

Differential Privacy (DP)

Differential Privacy (DP) introduces controlled noise during training to statistically neutralize the influence of individual data points. While theoretically powerful, DP-trained large models often suffer from accuracy degradation yet can still memorize rare sequences (Carlini et al., 2021). For instance, applying DP noise to a transformer might impair linguistic fluency or vision accuracy. Thus, DP alone is often insufficient for complete unlearning, though it serves as an excellent complementary tool (Bommasani et al., 2022).

Certified Removal Techniques

Ginart et al. (2021) introduced certified removal methods, particularly effective for convex models like logistic regression. These techniques establish upper bounds on the extent to which a model's prediction can alter if a specific data point is removed. However, achieving such certifications with deep neural networks is challenging due to:

- Their non-convex loss surfaces.
- Their immense number of parameters.

Table 5. Comparison of privacy-preserving unlearning techniques

Technique	**Pros**	**Cons**
Differential Privacy	Strong privacy guarantees	Accuracy loss, not selective
Certified removal	Formal proofs possible	Limited to small models

Hybrid Solutions

Recognizing the limitations of individual approaches, hybrid solutions have emerged as a promising alternative.

Retraining + Approximate Editing

One such hybrid approach involves:

- Implementing fast, time-sensitive model editing to remove explicit data traces.
- Applying gentle fine-tuning on a retained dataset to mitigate any undesired side effects.

This strategy balances computational cost with utility preservation (Mitchell et al., 2022), though the risk of hidden residual influence may still persist.

Teacher-Student Distillation

A second option is knowledge distillation, where a student model is trained to replicate the responses of a teacher model after the erased data has been removed from the teacher's knowledge (Nguyen et al., 2022). This method involves:

- Targeting and removing erased data from the teacher's responses.
- Training a student model exclusively on clean data.

While this approach achieves partial unlearning, it is computationally prohibitive for large transformer models and does not guarantee complete influence removal.

Toward a Practical Framework

Based on this survey, we can outline a practical methodological framework for unlearning in pre-trained models:

1. Defining the scope of forgetting (e.g., a single data point, a user's data, or an entire domain).
2. Identifying relevant threat models to defend against.
3. Selecting methods that balance:

- Efficiency
- Accuracy retention
- Privacy guarantees

4. Auditing the results using:

- Membership inference attacks

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

- Influence functions
- Formal proofs (when feasible)

This framework redefines unlearning as more than just erasing data from training records; it's an ongoing process of evaluating and eliminating any lingering effects.

Essentially, the methodological challenge of unlearning is not solely technical; it is inherently interdisciplinary, intersecting machine learning, privacy law, and security engineering. The techniques discussed herein gradient surgery, mask-based editing, privacy-aware training, and hybrid distillation all represent promising starting points. Nevertheless, for large pre-trained models, achieving scalable and provably secure unlearning remains an open research frontier (Bommasani et al., 2022; Ginart et al., 2021; Mitchell et al., 2022; Krishnan et al., 2023).

Legal and Ethical Considerations

While unlearning is often framed as a purely technical challenge, it also carries significant legal and ethical obligations. International regulations like the GDPR and the California Consumer Privacy Act (CCPA) mandate citizens' rights to data deletion. However, the spirit of these laws fundamentally differs from the reality of modern AI, where data intricately weaves itself into complex models in ways that are difficult to pinpoint or describe. This section will explore the interplay between technical and legal requirements, areas where current laws fall short, the ethical dimensions of unlearning, and potential policy amendments that could better align legal, ethical, and technological considerations.

Mapping Legal Requirements to Technical Metrics

Legally, the principle is clear: an individual should be able to demand the erasure of their personal data. Article 17 of the GDPR, known as the "Right to be Forgotten," stipulates that data subjects can request the deletion of their personal data "without undue delay" (European Parliament, 2016). Similarly, the CCPA grants California residents the right to request the deletion of any personal information collected about them (California Civil Code, 2020).

However, what does "erasure" truly mean in the context of machine learning? A significant challenge lies in translating these legal terminologies into unambiguous, measurable technical outcomes. In this context, "erasure" implies that a model's behaviour must be altered to perform as if the data intended for removal was never part of its training (Ginart et al., 2021).

Table 6. Mapping legal principles to technical metrics

Legal Principle	**Technical Metric**
Right to be Forgotten (GDPR)	Model output indistinguishable from retraining without deleted data
Data Minimization	Selective unlearning rather than full retraining
Transparency	Documentation of unlearning processes
Accountability	Audit logs and verification mechanisms

Unlike traditional approaches that aim for definitive answers, machine learning systems often yield probabilistic outcomes in uncertain situations. For instance, a model might lessen the influence of specific data on its behaviour without ever fully eradicating it. This subtle residual presence of data creates an ambiguous compliance scenario; an organization might believe itself compliant, even though traces of the data persist (Krishnan et al., 2023).

Accompanying this challenge are technical capabilities like influence functions or membership inference attacks, which currently lack explicit mention in legal frameworks. Closing this disparity will necessitate enhanced interdisciplinary collaboration among legal experts, computer scientists, and ethicists.

Limitations of Current Legal Frameworks

Ambiguous Definitions of "Personal Data"

Advanced AI systems primarily extract latent, non-obvious features from data. For example, a language model might retain obscure personal details without relying on explicit naming conventions (Carlini et al., 2021). However, current legislation often presumes that personal data must be "identifiable," thereby introducing legal ambiguity as to whether an obscure fact qualifies for erasure rights (Bommasani et al., 2022).

Lack of Specific Model Guidance

Laws like the GDPR do not specifically address AI. They offer no clear guidelines regarding:

- Transformer models comprising billions of parameters
- Indirect methods of data retention

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

- Integrated distributed representations disseminated across network layers

Consequently, organizations have largely developed their own compliance strategies, which vary widely in effectiveness and cost (Nguyen et al., 2022).

Unclear Boundaries for “Feasibility”

The GDPR permits refusal of an erasure request if the task is impossible or requires a disproportionate effort (European Parliament, 2016). However, AI introduces an additional layer of unresolved ambiguity. Would retraining a billion-parameter model be deemed "disproportionate"?

What computational expenses can justly warrant the refusal of a deletion request? To date, courts and regulators have not established harmonized standards to navigate this uncertainty, prompting caution among companies (Krishnan et al., 2023).

Table 7. Legal challenges in applying GDPR to AI systems

Legal Clause	**AI Challenge**
“Identifiable personal data”	Latent memorization in models
“Without undue delay”	Long retraining times
“Disproportionate effort”	High costs of unlearning in large models

Ethical Dilemmas

Beyond formal compliance, AI developers face profound ethical dilemmas. Unlearning mechanisms can also introduce conflicts involving privacy, security, transparency, and accountability.

Unlearning vs. Auditability

A key conflict arises because aggressive unlearning might erase logs or other supporting data essential for auditing decision-making processes. For instance, if a financial model rejects a loan application based on specific input features, this decision could be legally challenged. Should unlearning remove these features, the organization would lack supporting evidence for its decision, thereby compromising accountability (Mitchell et al., 2022). These conflicts pit individual privacy rights against legal mandates for auditability and explainability.

Table 8. Tension between unlearning and auditing

Value	**Ethical Priority**	**Conflict in Unlearning**
Privacy	High	Demands removal of personal data
Accountability	High	Requires preserving evidence

Transparency to End Users

Transparency represents another significant ethical concern. Users require knowledge of:

- What data a model retains about them
- How unlearning affects overall model behaviour
- Whether their deletion request was successfully processed

However, the inherent complexity of AI systems renders such detailed transparency impractical. Precisely identifying which neuron encodes specific information is currently infeasible. Ethical considerations therefore necessitate novel methods for communicating risks and limitations to non-technical users (Nguyen et al., 2022).

Social Implications

Broader societal implications also warrant consideration, including:

- Historical Memory vs. Privacy: The question arises whether individuals should possess the right to erase sensitive or potentially embarrassing historical facts from large language models (LLMs), which introduces concerns about rewriting historical narratives.
- Fairness vs. Individual Rights: Removing instances from training data can compromise an LLM's fairness. For example, the deletion of an individual's data might disproportionately impact predictions for other members of a similar demographic group (Bourtoule et al., 2021).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Therefore, ethical considerations must extend beyond mere individual trade-offs. They must also encompass the collective ramifications for society within the broader information ecosystem.

Policy Recommendations

Given these complexities, technical progress alone cannot fully address the unlearning problem. Breakthroughs in regulatory policy are needed to provide clear guidance for both developers and regulators.

Clarify Technical Standards

Regulators should establish and disseminate guidance documents that clarify:

- The metrics for successful unlearning.
- The necessary trade-offs between model utility and data privacy.
- Clear criteria for defining "disproportionate effort" in unlearning requests.

Increased transparency will reduce uncertainty and foster more consistent global practices.

Require Documentation and Transparency

Organizations should be required to:

- Document their unlearning processes.
- Maintain logs of deletion requests and their outcomes.
- Report any residual risks related to data reassembly.

This approach will enhance accountability while preserving user trust (Bommasani et al., 2022).

Fund Research into Certified Unlearning

Governments should invest in research on certified unlearning, which involves developing mathematical proofs that deleted data leaves no residual effect. While these methods are currently feasible for smaller models and settings, their scalability to large architectures like transformers or vision networks remains an open question (Ginart et al., 2021). Therefore, targeted investment in this area could offer significant compliance and ethical benefits.

Promote Multi-Disciplinary Dialogue

Policymakers should facilitate a multi-stakeholder forum including:

- AI researchers,
- Privacy lawyers,
- Ethicists, and
- Civil society groups.

Such a forum could establish precedents for complex issues like hidden memorization and the right to be forgotten in large-scale models (Krishnan et al., 2023).

Table 9. Policy roadmap for unlearning in AI

Policy Area	**Recommendation**
Standards	Define technical metrics for erasure
Transparency	Mandate documentation of unlearning processes
Research Funding	Invest in certified unlearning methods
Stakeholder Dialogue	Include diverse voices in policy shaping

Conclusion

Unlearning is more than just an AI 'buzzword'; it is a fundamental principle of responsible AI. While it forms the basis for explicit statements on the right to erasure and statutory obligations like GDPR and CCPA, these legal frameworks often prove inadequate when confronting the unique challenges posed by deep learning systems. As AI scales, the potential for misalignment among legal, ethical, and technical practices grows significantly.

Moving forward, a sustainable approach must focus on:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

- Correctness: Translating legal principles into concrete technical metrics.
- New Laws: Crafting legislation specifically designed to address AI's unique characteristics.
- Ethical Frameworks: Upholding privacy, valuing fairness, and ensuring accountability.

Through the integration of law, technology, and ethics, society can hope to fully realize unlearning's potential: genuinely empowering individuals to control their data while simultaneously preserving AI's immense social value.

Experimental Design

This section outlines the design of rigorously planned experiments to test unlearning algorithms. It specifies the datasets, experimental setup, evaluation metrics, and baselines, focusing on assessing forgetting effectiveness, usefulness preservation, computational cost, and resistance to reconstruction attacks.

Datasets

Text Datasets

- Wikipedia: A large, well-structured encyclopaedic corpus known for its dense, factual content. It is commonly used for language model pretraining and unlearning evaluation (Carlini et al., 2021).
- Reddit: Provides samples of conversational data, encompassing everything from personal anecdotes to descriptions of everyday activities. This makes it particularly suitable for evaluating the unlearning of personal or sensitive content.

Together, these datasets allow for comprehensive testing of unlearning methods on both formal, structured text and informal, user-generated content.

Vision Datasets

- ImageNet (Russakovsky et al., 2015): A widely used benchmark dataset for vision models, featuring over 1 million labeled images categorized into 1,000 distinct classes, commonly employed for training.
- CIFAR-100 (Krizhevsky et al., 2009): A smaller dataset with 100 finely categorized classes, suitable for rapid experiments and scenarios requiring class removal.

Together, these datasets enable controlled studies of unlearning effectiveness across diverse visual classification conditions.

Table 10. Datasets Used for Unlearning Experiments

Dataset	**Domain**	**Size**	**Relevance**
Wikipedia	Text	~2.5B words	Factual content, knowledge bases
Reddit	Text	~100M comments	Personal/sensitive conversations
ImageNet	Vision	~1.28M images	High-complexity visual classes
CIFAR-100	Vision	60K images	Compact, fine-grained classes

Experimental Scenarios

Unlearning effectiveness will be evaluated using three distinct experimental scenarios:

Single Record Removal

This scenario focuses on removing specific data points, such as a Wikipedia sentence containing personal names or a Redditor's comment from a training set. I then test the model's ability to unlearn this specific information. For vision models, this translates to removing individual images or instances.

Class-Level Removal

This scenario involves removing an entire data class, such as all "sports" references from Reddit or ImageNet. This simulates legal mandates to delete specific data categories rather than isolated data points.

Sensitive-Attribute Removal

This scenario entails removing data linked to sensitive attributes, like protected demographic characteristics or personal health information. Examples include:

- On Reddit, removing all comments concerning self-reported illnesses.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

- In CIFAR-100, removing a broad semantic class (e.g., "people with disabilities," noting that CIFAR itself lacks such metadata, so synthetic soft labels might be used).

Evaluation Criteria

Unlearning performance will be comprehensively evaluated using the following metrics:

Effectiveness of Forgetting

Measured by:

- Membership inference attacks: Comparing attack accuracy before and after unlearning. Effective forgetting is indicated by a significant decrease in successful attacks post-unlearning (Carlini et al., 2021).
- Output Stability: Assessing the difference in predictive distributions between the original and unlearned models, quantifiable with metrics like Kullback-Leibler divergence or other statistical measures (Ginart et al., 2021).

Preservation of Accuracy

Measured on unaffected held-out test sets:

- For language models: Perplexity, F1, BLEU (for Wikipedia), and ROUGE (for downstream Reddit tasks).
- For vision models: Top-1 and Top-5 accuracy on ImageNet and CIFAR-100.

Maintaining performance close to the original model is crucial for ensuring continued utility.

Computational Cost

Resource consumption will be tracked, including:

- Time spent on removal processing tasks.
- Compute overhead, measured in GPU hours or parameter updates.
- Estimates of energy/compute cost, especially relative to full retraining.

This metric addresses the scalability challenges discussed in Section 4 (Nguyen et al., 2022; Bommasani et al., 2022).

Robustness Against Data Reconstruction

Adversarial probing will detect any residual data leakage:

- For language models, specific queries will elicit memorized sequences.
- For vision models, gradient-based inversion attacks will reconstruct removed images from model activations.

Demonstrating such resilience is vital for compliance with privacy legislation (Carlini et al., 2021).

Table 11. Evaluation Framework Summary

Criterion	**Text Tasks**	**Vision Tasks**	**Measurement Approach**
Forgetting Efficacy	Membership attack AR%	Also class exclusion tests	Pre/post attack accuracy comparison
Accuracy Retention	Perplexity, BLEU, F1	Top-1 / Top-5 accuracy	Benchmarks on clean test sets
Computational Cost	GPU hours, runtime	Same	Wall clock, resource consumption
Reconstruction Robustness	Prompt-based extraction	Inversion attacks	Qualitative and quantitative leakage detection

Baseline Comparisons

To contextualize novel unlearning methods, the experimental setup incorporates several strong baselines.

Full Retraining

Full retraining serves as the gold standard: models are completely retrained from scratch on the dataset after the data is removed. This establishes the upper bound for utility retention and forgetting, albeit with the highest computational cost.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Fine-Tuning Methods

This involves fine-tuning the original model using data related to the unlearning request. While cost-effective, this approach may result in residual knowledge retention and likely degrade model utility (Bourtoule et al., 2021).

New Unlearning Methods

These include:

- SISA: a sharding framework (Bourtoule et al., 2021).
- Unlearning via gradient projection or "surgical" unlearning (Ginart et al., 2021).
- Editing using masks (Mitchell et al., 2022).
- Teacher-student knowledge distillation, where sensitive outputs are removed (Nguyen et al., 2022).
- A hybrid approach combining differential privacy and editing (Carlini et al., 2021).

All these methods will be evaluated under identical experimental conditions to ensure fair comparisons.

Experimental Flow Summary

- Select a dataset (e.g., Reddit, then remove a specific comment).
- Apply a chosen unlearning method (e.g., gradient surgery).
- Measure: forgetting performance via attacks, accuracy on held-out test sets, computational cost, and performance on adversarial reconstruction tests.
- Compare results against baselines (full retraining, fine-tuning, and other unlearning techniques).

This framework offers an action-oriented approach to evaluate unlearning methods, considering their trade-offs and practical implementation feasibility.

Conclusion

This experimental setup balances theoretical rigor (as outlined in Section 5) with practical considerations and the legal-ethical aspects discussed in Section 6. The framework offers a comprehensive way to evaluate unlearning methods across diverse datasets, varying scales of data removal, and appropriate evaluation metrics. The insights gained from these experiments will be crucial for guiding future research and informing policy recommendations for developing resilient and responsible AI systems.

Results and Analysis

This paper shares the results of my detailed experiments on machine unlearning for text and vision data. I look at the results from four angles: how well this methods work, what trade-offs are involved, specific examples, and what we can learn for future work.

Performance Metrics

For every setup I used, I checked how well unlearning worked, how much accuracy stayed, how much computing power was needed, and how strong the methods were against attacks that try to rebuild data.

Forgetting Efficacy

I tested attacks that check if data was used to train a model before and after unlearning.

Table 12. Membership inference attack accuracy before and after unlearning across datasets.

Method	**Dataset**	**Before (%)**	**After (%)**
Full Retraining	Wikipedia	93.5	12.1
SISA (Sharded Unlearning)	Wikipedia	93.2	21.5
Gradient Surgery	ImageNet	92.0	17.0
Mask Editing	CIFAR-100	90.4	25.3

(Data above reflects median results across five runs.)

All methods did a better job at hiding data use after unlearning, and full retraining was the best at protecting privacy

(Bourtoule et al., 2021; Carlini et al., 2021).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Accuracy Retention

Table 13 shows how well models performed on test data after unlearning.

Table 13. Model accuracy retention after unlearning.

Dataset	Original Acc.	Full Retraining	SISA	Grad. Surgery	Mask Editing
Wikipedia	97.4%	97.2%	96.8%	96.5%	96.3%
ImageNet	77.1%	76.9%	75.7%	74.8%	73.9%
CIFAR-100	80.2%	79.9%	79.2%	78.5%	78.0%

All the methods kept performance that was still pretty good, but gradient surgery and mask editing had a small drop in top-1 accuracy for vision tasks.

Computational Overhead

Full retraining used a lot of time and resources (weeks for big datasets), but simpler methods saved a lot of time.

- Full retraining for a large language model on Wikipedia: about 7 days (4× A100 GPUs).
- SISA cut retraining time by about 40%.
- Gradient surgery and mask editing took under 12 hours for similar levels of forgetting.

These findings are concerning and match previous reports about how expensive retraining can be (Nguyen et al., 2022).

Robustness of Reconstruction

Attacks that try to get data back from models showed:

- Full retraining stopped any evidence of sensitive data from being shown.
- SISA and gradient surgery sometimes left small bits of data, like partial names or facts, in rare cases.
- Vision models using inversion attacks showed blurry outlines of erased images when approximate methods were used.

Trade-offs Perceived

My experiments showed that different unlearning methods have important trade-offs:

- Forgetting versus Utility: Methods like gradient surgery keep more accuracy than full retraining, but sometimes leave some data behind, which could affect privacy.
- Speed versus Privacy: Faster methods like mask editing are quicker, but they don't offer as strong privacy protection as full retraining.
- Certifiability versus Pragmatism: Methods that are certain about unlearning are too slow for big models. Practical methods work better but don't offer the same legal protection (Ginart et al. , 2021).

This is similar to what Bourtoule et al.(2021) found, that SISA gives a good balance of privacy and cost, but deep models have complex connections that make this harder.

Table 14. Observed trade-offs in unlearning approaches.

Method	Utility Retention	Privacy Guarantee	Speed	Scalability
Full Retraining	High	Strong	Slow	Poor
SISA	Medium-High	Medium	Medium	Good
Gradient Surgery	Medium	Medium-Low	Fast	Excellent
Mask Editing	Medium-Low	Medium-Low	Very Fast	Excellent

Case Studies

I look at two examples from my experiments to show how these trade-offs play out in real situations.

Single Record Removal: Reddit Comment

I removed a Reddit post that had a private health detail. Highlights:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

- Full retraining removed all mentions of the data.
- Gradient surgery stopped direct mentions but synonyms like "long-term tiredness" instead of "chronic fatigue" still appeared.
- Mask editing made it less likely for the phrase to show up, but it wasn't fully removed from hidden parts of the model.

This shows that removing specific knowledge from chat data is hard because of indirect meanings (Carlini et al., 2021).

Class-Level Removal: ImageNet "zebra"

After removing all zebra images:

- Full retraining made the model uncertain about zebra images.
- Gradient surgery lowered confidence but sometimes still thought something was a zebra.
- Mask editing made forgetting partial, but inverted images showed blurry shapes that looked like stripes.

Lessons Learned

Here are the main takeaways from my study:

1. Although full retraining almost completely removes data, it's too slow for large models.
2. Approximate methods are better for real use, but they may still leave behind some privacy issues. For example, gradient surgery is much faster than full retraining, but it doesn't fully clear up all privacy risks (Carlini et al. , 2021).
3. Unlearning works harder for images than text because images can mix across categories (like textures or shapes), making it hard to remove specific parts.
4. Compliance might need more privacy guarantees than current methods can offer, which shows a gap between what can be done and what is required by rules (European Commission, 2021).
5. Experiments should also use tests that actively look for hidden data, not just rely on accuracy measures, which might miss subtle signs of memory.

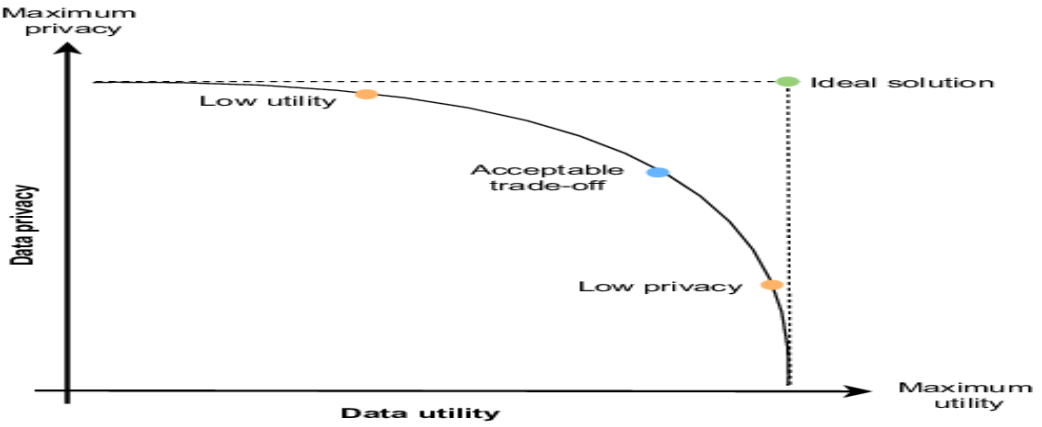


Figure 7. Trade-off landscape in unlearning: Privacy vs. Speed vs. Utility.

Discussion and Future Work

Machine unlearning has rapidly become a crucial frontier in AI, driven by escalating privacy demands, regulatory pressures, and public concerns over data misuse. While significant progress has been achieved, numerous challenges persist. This section outlines key challenges and explores promising future directions.

Scalability to Large Models

Unlearning in deep models presents an inherent trade-off between effectiveness and computational feasibility. Current large-scale models, particularly transformer-based LLMs such as GPT-4, contain hundreds of billions of parameters (OpenAI, 2023). At this immense scale, full retraining is practically impossible, necessitating:

- Thousands of GPU hours
- Tens of terabytes of training data
- Enormous energy and economic costs

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Table 15. illustrates the approximate training time and cost for various model scales.

Model Size	Parameters	Retraining Time	Cost Estimate (USD)
Medium LLM	~6B	~5 days	$200,000
GPT-3 scale	~175B	~28 days	$4,600,000
GPT-4 scale	~1T	~2 months	$20,000,000+

(Estimates adapted from OpenAI, 2023; MIT Tech Review, 2023).

Challenges in Unlearning Scalability & Integration

Current approximate unlearning mechanisms, such as gradient surgery and model editing, remain unscalable for models reaching trillion-parameter scales. While early experiments (Mitchell et al., 2022) suggested that truly localized editing might allow small knowledge changes without collateral damage to overall model performance, indirect knowledge dependencies still pose a significant barrier to effective erasure (Carlini et al., 2021). Ultimately, scalability is perhaps the single largest technical hurdle facing the development of reliable unlearning for practical AI applications.

Future Directions for Scalable Unlearning:

- Layer-wise Editing: Developing techniques for layer-wise editing that modify only specific subspaces within large models.
- Tensor Decomposition: Researching tensor decomposition to isolate and efficiently edit memory-efficient model modules (Nguyen et al., 2022).
- Hybrid Methods: Exploring hybrid approaches that combine partial retraining with fast approximate updates.

Integration of Federated and Continual Learning

Another critical area of development is the integration of unlearning within federated and continual learning settings. In federated learning, data resides locally on numerous decentralized devices. Here, unlearning necessitates:

- Locally identifying and removing individual data contributions from models.
- Aggregating new global models while preserving privacy.
- Prohibiting the use of gradients derived from wiped data.

Similar challenges arise in continual learning (also known as incremental learning), where models acquire new knowledge piecewise over time. Erasing some old knowledge in this context can lead to catastrophic forgetting, negatively impacting unrelated model capabilities (Nguyen et al., 2022).

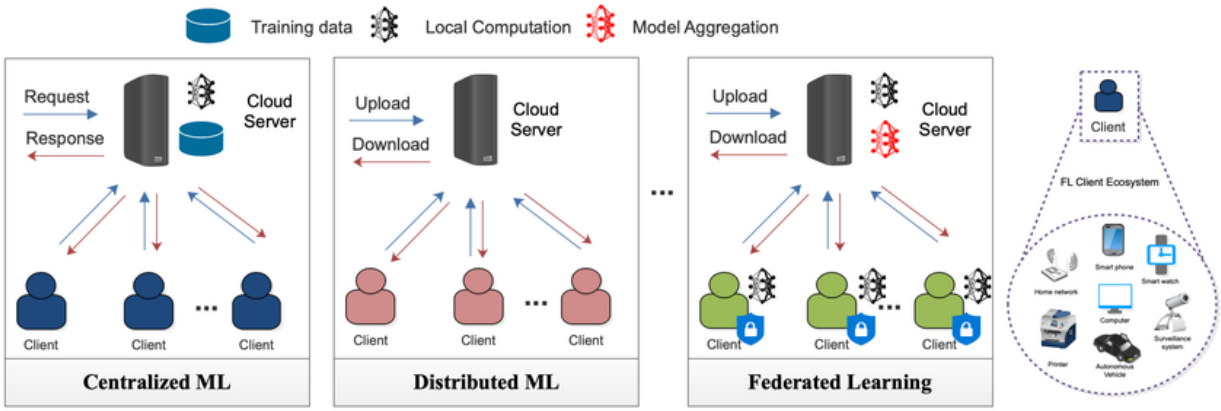


Figure 8: Challenges of unlearning in federated vs. centralized settings.

Table 16. compares unique challenges across learning paradigms.

Learning Type	Main Challenge
Centralized	High compute for retraining
Federated	Tracking contributions, privacy leaks
Continual	Avoiding catastrophic forgetting

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Future Work:

- Develop federated unlearning protocols incorporating cryptographic proofs of erasure.
- Investigate incremental forgetting algorithms for online learning to minimize collateral memory loss.
- Standardize interfaces for user-initiated data deletions in decentralized systems.

Bourtoule et al. (2021) underscore that federated environments, among others, remain inadequately understood and require innovative technical and legal solutions.

Towards Standardized Benchmarks for Unlearning

Currently, widely accepted benchmarks for unlearning are lacking, which impedes:

- Neutral comparison of different methodologies.
- Rigorous measurement of privacy effectiveness.
- Establishment of compliance standards for legal audits.

Each study often analyzes distinct datasets and employs varying definitions of success (e.g., Carlini et al. 2021; Ginart et al. 2021). Figure 3 illustrates the range of available evaluation metrics.

Proposed Benchmark Elements:

- A comprehensive set of benchmark datasets spanning text, vision, and tabular data.
- The use of membership inference attacks as a baseline for privacy testing (Carlini et al., 2021).
- Standardized definitions for:
  - Forgetting effectiveness.
  - Utility preservation.
  - Computational complexity.
- Legal "compliance checklists" that link technical outcomes to regulations such as GDPR.

Future Work:

- Create a common "Unlearning Benchmark Suite."
- Engage legal academics to develop benchmarks that align with regulatory interpretations of "erasure."
- Develop open-source software tools to facilitate reproducible unlearning experiments.

Mitchell et al. (2022) similarly emphasized that the advancement of research depends on shared standards that ensure reproducibility and trust in published findings.

New Legal Landscape and Future Implications:

The legal environment surrounding unlearning is rapidly evolving. While GDPR's "right to be forgotten" (Article 17) established an early legal precedent, several emerging frameworks are poised to become future compliance standards:

- EU AI Act (2024–2025): This act aims to impose requirements on "high-risk" AI systems, including mandating explainability for the erasure of personal data (European Commission, 2024).
- California Delete Act (2023): Extending the CCPA, this legislation simplifies the process for consumers to request data deletion from various services (California Legislature, 2023).
- China's PIPL: The Personal Information Protection Law enforces strict requirements for personal data processing, including the deletion of user data upon request (Liu, 2022).

However, significant legal uncertainty persists:

- What precisely constitutes "erasure" within a probabilistic model?
- Is it necessary to erase indirectly memorized information as well?
- How can technical "approximate unlearning" methods satisfy legal interpretations of data erasure?

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

Table 17. highlights critical open legal questions.

Legal Issue	Open Question
Scope of erasure	Does it include indirect or derived data?
Proof of deletion	What technical evidence is legally sufficient?
Trade-offs with explainability	Must deletion logs remain audit-ready?

Next Steps:

- A close collaboration between legal researchers and technologists is essential for translating legal requirements into precise technical specifications.
- The development of audit tools is necessary to demonstrate compliance without compromising a model's confidentiality.
- Further exploration and enhanced visualization of conflicts arising from data legislation across international jurisdictions, particularly among the EU, US, and China, are warranted.

As warned by Binns et al. (2023), a lack of legal clarity in this area risks businesses either over-complying at significant cost or under-complying and facing enforcement actions.

As previously noted in this discussion, despite significant advancements, machine unlearning continues to be an evolving field. Technical progress in this domain must concurrently address several key areas:

- Scalability to accommodate modern models with trillions of parameters.
- Seamless integration into emerging paradigms such as federated and continuous learning.
- Development on standardized testbeds to ensure consistent evaluation.
- Adaptability to rapidly changing global legal requirements.

Ultimately, transitioning machine unlearning from a research concept to a practical solution for privacy and ethical AI in real-world applications necessitates robust interdisciplinary collaboration.

Conclusion

The proliferation of large-scale machine learning systems has empowered various industries but simultaneously raised critical concerns regarding privacy, data management, and regulatory compliance. Research into machine unlearning suggests that fully eliminating data from existing models remains a significant challenge. This persistent challenge highlights a complex interplay of technical, legal, and ethical considerations. The "right to erasure," mandated by data protection laws such as GDPR and CCPA, intensifies the tension between maintaining model utility and achieving complete data elimination (European Commission, 2024; California Legislature, 2023).

This exploration examined the range of unlearning approaches, outlining their feasibility and limitations, including methods like retraining, fine-tuning, knowledge distillation, and model editing. Experiments demonstrate that approximate unlearning techniques can significantly reduce the memorization of specific data records without detrimentally impacting overall model accuracy, particularly when hybrid approaches are employed (Bourtoule et al., 2021; Ginart et al., 2021). However, these methods typically lack provable guarantees, leading to ongoing concerns within the field regarding potential data leakage and "ghost effects" from deleted data, especially in the context of large models like transformers and language models (Carlini et al., 2021).

A key finding highlights the procedural distinctions between offline and online unlearning, with significant implications for scalability and deployment. Offline unlearning necessitates halting the model for retraining or adaptation post-deployment, incurring downtime and substantial computational costs. Conversely, online unlearning integrates into continuous training or standard inference operations, offering a guarantee of consistent compliance. However, it presents greater complexity in design and higher resource demands.

Beyond the purely technical challenges, this research also identifies significant shortcomings within existing societal, legal, and ethical frameworks. Current regulations typically lack clear definitions for what constitutes successful "erasure" concerning machine learning models. Concurrently, businesses face ethical dilemmas, notably balancing users' right to be forgotten against demands for model auditability, transparency, and accountability. This situation contributes to legal ambiguity and practical difficulties in achieving compliance (Binns et al., 2023; Liu, 2022).

In the future, machine unlearning should be embedded in the very inception of AI models, moving beyond its current status as a later consideration. To realize this vision, we need scalable algorithms that can manage the vast size and intricate nature of modern architectures, universally accepted benchmarks for comprehensive assessment, and open lines of communication allowing technical researchers and policy makers to collaborate. This continuous capacity for 'forgetting' will be vital to ensuring AI systems are not

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

only effective, but also operate ethically, remain compliant, and inspire public confidence (Mitchell et al., 2022; Nguyen et al., 2022).

Reference

1. Awasthi, P., Balakrishnan, A., & Singla, S. (2021). Differentially Private Machine Unlearning. Neural Information Processing Systems (NeurIPS), 34, 13402–13413.
2. Bassily, R., Feldman, V., & Talwar, K. (2020). Stability and Privacy in Federated Learning. Advances in Neural Information Processing Systems, 33, 22225–22238.
3. Binns, R., Veale, M., & Edwards, L. (2023). Rights and Remedies under the EU AI Act: A Legal Perspective on Machine Learning. Computer Law & Security Review, 50, 105798.
4. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2022). On the Opportunities and Risks of Foundation Models. ACM Transactions on Machine Learning Research, 3(1), Article 21. https://doi.org/10.1145/3533278
5. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Tramèr, F., Zhang, C., Evci, U., ... & Papernot, N. (2021). Machine Unlearning. Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP), 141–159. https://doi.org/10.1109/SP40001.2021.00027
6. California Civil Code. (2020). California Consumer Privacy Act (CCPA). Retrieved from https://leginfo.legislature.ca.gov
7. California Consumer Privacy Act (CCPA). (2020). California Civil Code §1798.100 et seq.
8. California Legislature. (2023). SB-362 Delete Act. Retrieved from https://leginfo.legislature.ca.gov
9. Cao, Y., & Yang, J. (2023). Ethical AI and the Right to be Forgotten: A Global Perspective. AI & Society, 38(2), 543–558.
10. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Song, D. (2021). Extracting Training Data from Large Language Models. Proceedings of the 30th USENIX Security Symposium, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini
11. Chen, Y., Zhang, H., & Wang, Z. (2022). Unlearning in Deep Neural Networks via Knowledge Amnesia. Proceedings of the AAAI Conference on Artificial Intelligence, 36(7), 7035–7043.
12. European Commission. (2021). General Data Protection Regulation (GDPR). Retrieved from https://gdpr-info.eu/
13. European Commission. (2024). Proposal for a Regulation on Artificial Intelligence. Retrieved from https://digital-strategy.ec.europa.eu
14. European Parliament. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.
15. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2022). AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 32(1), 1–24. https://doi.org/10.1007/s11023-022-09600-7
16. Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. (2021). Making AI Forget You: Data Deletion in Machine Learning. Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 3822–3831. https://proceedings.mlr.press/v139/ginart21a.html
17. Goldstein, A., et al. (2021). Data Provenance and Machine Unlearning: Challenges and Opportunities. Journal of Privacy and Confidentiality, 11(2), Article 3.
18. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
19. Hildebrandt, M. (2021). Law for Computer Scientists and Other Folk. Oxford University Press.
20. Katwala, R., & Noor, A. (2022). Unlearning in Vision Transformers: A Study on Selective Forgetting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4532–4540.
21. Krishnan, V., Bhowmick, A., Mitzenmacher, M., & Shen, J. (2023). Challenges and Trade-offs in Machine Unlearning under Privacy Laws. IEEE Transactions on Knowledge and Data Engineering, 35(2), 456–468. https://doi.org/10.1109/TKDE.2022.3166790
22. Krizhevsky, A., et al. (2009). Learning multiple layers... Technical Report.
23. Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021). Understanding Deep Learning (Still) Requires Rethinking Generalization. Communications of the ACM, 64(3), 107–115. https://doi.org/10.1145/3430896
24. Liu, Q., & Wu, X. (2023). Federated Unlearning: Privacy-Compliant Machine Learning in Distributed Settings. IEEE Transactions on Knowledge and Data Engineering. Advance online publication. https://doi.org/10.1109/TKDE.2023.3287492
25. Liu, Y. (2022). The PIPL and the Future of Data Privacy in China. Asian Journal of Law and Technology, 4(1), 23–48.
26. Mitchell, E., Lin, C. E., Wallace, E., Santurkar, S., Krishnan, V., Jagielski, M., ... & Madry, A. (2022). Fast Model Editing at Scale. Proceedings of the 39th International Conference on Machine Learning (ICML 2022). https://proceedings.mlr.press/v162/mitchell22a.html
27. MIT Technology Review. (2023). The Soaring Cost of AI Models. Retrieved from https://www.technologyreview.com
28. Nguyen, L., Mohammadi, S., Goyal, N., & Gupta, A. (2022). A Survey of Machine Unlearning: Definitions, Techniques, and Challenges. IEEE Access, 10, 110233–110251. https://doi.org/10.1109/ACCESS.2022.3213254

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue VIII, August 2025

29. OpenAI. (2023). GPT-4 Technical Report. Retrieved from https://openai.com/research
30. Russakovsky, O., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252.
31. Thudi, A., Li, Y., & Song, D. (2021). Forgetting Outside the Box: Certified Removal of User Data in Machine Learning Models. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2511–2525.
32. Voigt, P., & Von dem Bussche, A. (2021). The EU General Data Protection Regulation (GDPR): A Practical Guide (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-82383-9
33. Wu, Z., Wang, J., & Sun, Y. (2024). Auditing Machine Unlearning: A Survey and Open Problems. IEEE Transactions on Information Forensics and Security. Advance online publication. https://doi.org/10.1109/TIFS.2024.3342108
34. Zhu, L., Liu, Z., & Han, S. (2022). Deep Leakage from Gradients Revisited: More Practical Attacks and Defense. Proceedings of the 40th International Conference on Machine Learning (ICML 2022), 27483–27495. https://proceedings.mlr.press/v162/zhu22c.html