Tampere University

Marjia Akter

EXPLAINING DATA DELETION

Bridging explainability, semantic summaries and GDPR's right to be forgotten

Faculty of Information Technology and Communication Sciences
Master of Science Thesis
November 2025

ABSTRACT

Marjia Akter: Explaining Data Deletion: Bridging Explainability, Semantic summaries and GDPR's right to be forgotten
Faculty of Information Technology and Communication Sciences
Master of Science Thesis
December 2025

The rapid growth of data-driven systems has intensified concerns over privacy, accountability, and compliance with data protection regulations. Central to this discourse is the General Data Protection Regulation (GDPR) and its provision of the Right to Be Forgotten, which grants individuals the ability to request the erasure of their personal data. While the legal framework is well established, significant technical and organizational challenges continue to hinder its full implementation.

This thesis explores how data deletion can be made explainable and interpretable by integrating formal compliance guarantees, semantic summarisation, and user-centric explainability. The research builds on three core pillars: (1) formal models of deletion compliance, (2) explainable deletion paradigms, and (3) semantic graph-based summarisation methods.

A conceptual framework is developed to bridge these perspectives, enabling deletion processes that are both verifiable and understandable. Prototype implementations combine compliance-oriented deletion with semantic summaries that provide concise, meaningful descriptions of what data has been removed, how dependencies are managed, and why certain information cannot be deleted. Additionally, large language models are explored to translate these summaries into natural, user-friendly narratives.

Evaluation strategies include technical verification of deletion compliance, assessment of semantic summary quality using established metrics such as coverage, precision, and representativeness, and user-centred evaluation of trust and interpretability. The findings reveal trade-offs between technical assurance and explanatory clarity and propose pathways for integrating explainable deletion into machine unlearning practices and future regulatory frameworks.

This work contributes to the ongoing discourse on bridging legal, technical, and human-centred perspectives on data deletion. It offers a step toward accountable, transparent, and GDPR-compliant practices for the digital age.

Keywords: Explainable Machine Unlearning, GDPR Right to Be Forgotten, Semantic Summaries, Knowledge Graphs, Data Privacy.

The originality of this thesis has been verified using the Turnitin Originality Check service.

USE OF AI IN THESIS

I have utilised AI tools in my thesis: Yes

The AI tools utilised in my thesis, and their purposes are described below:

Names and versions of AI tools: ChatGPT 5.0

Purpose of using AI tools: The tool was used to support the writing process by checking grammatical errors and refining academic language.

Sections where AI tools were used: AI tools were applied throughout all chapters to detect grammatical errors and provide correction suggestions.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

PREFACE

This document template conforms to the Guide to Writing a Thesis in Technical Fields at Tampere University (2025).

Acknowledgements to those who contributed to the thesis are generally presented in the preface. It is not appropriate to criticize anyone in the preface, even though the preface will not affect your grade. The preface must fit on one page. Add the date, after which you have not made any revisions to the text, at the end of the preface.

Tampere, 16 December 2025

Marjia Akter

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Acc	Accuracy
API	Application programming interface
AUC	Area under the curve
CSV	Comma-Separated Values
DBpedia	Semantic Knowledge Base derived from Wikipedia
EU	European Union
FN	False Negative
FP	False Positive
GDPR	General Data Protection Regulation
JSON	JavaScript Object Notation
KD	Knowledge Distillation
KG	Knowledge Graph
LLM	Large Language Model
LR	Logistic Regression
MIA	Membership Inference Attack
ML	Machine Learning
NLP	Natural Language Processing
RDF	Resource Description Framework
ROC	Receiver Operating Characteristic
RTBF	Right to Be Forgotten
SISA	Sharded, Isolated, Sliced, Aggregated
TN	True Negative
TP	True Positive
TTL	Turtle File Format (RDF Serialization)
XAI	Explainable Artificial Intelligence

1. INTRODUCTION

The growing reliance on data-driven systems has reshaped modern digital infrastructures, enabling personalized services, intelligent automation, and large-scale analytics across nearly every domain, from healthcare and finance to entertainment and public administration. Machine learning (ML) models in particular have become central to these systems due to their ability to learn complex patterns from vast amounts of personal data However, this rapid expansion has intensified long-standing concerns regarding data privacy, individual autonomy, and the accountability of AI-powered decision-making(1).

In response to these concerns, the European Union’s General Data Protection Regulation (GDPR) introduced one of the most influential legal frameworks governing data processing practices. Among its provisions, Article 17, commonly known as the Right to Be Forgotten (RTBF), gives individuals the legal right to request the deletion of their personal data from information systems. While deletion in traditional databases is technically straightforward, enforcing erasure within machine learning systems poses substantial challenges. Once data contributes to model training, its influence becomes encoded in model parameters and internal representations, making “true deletion’’ non-trivial. This difficulty has led to the emergence of machine unlearning, a research direction focused on selectively removing the effect of specific data samples from trained models without reconstructing them entirely(2).

Although recent unlearning techniques—such as full retraining, knowledge distillation-based approaches, and partitioned training frameworks like SISA—offer promising strategies for efficient data removal, a crucial gap remains existing unlearning methods rarely explain how deletion occurred or provide evidence that deletion was carried out in a legally and ethically defensible manner. In practical deployments, data controllers, auditors, and affected users require more than technical deletion; they require transparent explanations that demonstrate compliance, articulate the scope of the deletion, and contextualize its impact on the overall system (3).

This need for interpretability situates unlearning within the broader field of Explainable Artificial Intelligence (XAI), which seeks to make AI models’ decisions understandable to human stakeholders. However, explainability in the context of data deletion remains underexplored. Most work in XAI focuses on justifying predictions, not explaining how data was removed or why certain deletion steps were necessary. This gap becomes more pronounced when considering multi-stakeholder environments—developers, compliance officers, end-users—each requiring different types of explanations.

At the same time, advances in knowledge graphs (KGs) and semantic technologies provide new opportunities for structuring and documenting data deletion events. Knowledge graphs such as DBpedia offer semantic context, linking data items (e.g., movies, users, or entities) to real-world ontologies and enabling machine-readable audit trails. Semantic summarization techniques, combined with RDF representations and natural-language generation, can transform technical deletion logs into meaningful explanations that align with GDPR's accountability requirements (4).

Despite the potential synergy between unlearning, explainability, and semantic modelling, current research treats them largely as separate threads. Unlearning methods rarely incorporate semantic explanations, and semantic tools rarely address model-level data deletion. As a result, organizations lack integrated mechanisms for performing, verifying, and explaining unlearning in a way that is both technically sound and legally compliant.

To address this gap, this thesis proposes an Explainable Machine Unlearning Framework that integrates unlearning techniques with knowledge-graph-based explainability. The framework combines:

- **Selective unlearning algorithms** (Full Retraining, Knowledge Distillation, and SISA),
- **Semantic deletion modelling** using RDF and linked-data representations, and
- **Natural-language summaries** that communicate deletion outcomes clearly and transparently.

The system is implemented and empirically evaluated using an enriched IMDb dataset containing textual reviews and associated metadata. The evaluation covers model performance, privacy assurance through Membership Inference Attacks (MIA), and semantic explanation quality, demonstrating that effective unlearning can coexist with transparency and GDPR-aligned accountability (5).

While prior work on machine unlearning focuses primarily on algorithmic efficiency or theoretical guarantees of data removal, limited attention has been paid to how unlearning operations can be explained, documented, and communicated in a GDPR-compliant manner. Existing explainability research, in turn, largely concentrates on model predictions rather than on data deletion processes. This thesis addresses this gap by proposing an integrated framework that combines machine unlearning with semantic knowledge graph representations and natural-language explanations. Unlike existing approaches, the proposed framework not unlearning but verifies unlearning, but also generates structured RDF audit trails and user-facing summaries that make deletion actions transparent and interpretable. In doing so, this work bridges legal accountability requirements with technical unlearning mechanisms, offering a novel perspective on explainable compliance for the Right to Be Forgotten.

Research Objectives

This thesis pursues the following objectives:

1. To design a modular, explainable machine unlearning pipeline integrating multiple unlearning techniques.
2. To develop a semantic layer for modelling deletion events using RDF and knowledge graph concepts.
3. To generate human-understandable deletion summaries through natural-language explanation.
4. To evaluate the framework with respect to performance retention, privacy guarantees, and explanation quality

Objective 1 – Explainable unlearning framework

To address the first objective, this thesis designs and implements a modular machine unlearning pipeline integrating Full Retraining, Knowledge Distillation, and SISA. These methods are evaluated within a unified experimental setup to compare their trade-offs in performance, efficiency, and privacy compliance.

Objective 2 – Semantic deletion modelling

The second objective is addressed by introducing an RDF-based semantic layer that captures deletion metadata, including deletion targets, unlearning methods, timestamps, and verification outcomes. This layer enables machine-readable auditability aligned with GDPR accountability requirements.

Objective 3 – Human-understandable explanations

To fulfil the third objective, the framework generates natural-language summaries derived from RDF triples. These summaries translate technical deletion events into explanations understandable by non-technical stakeholders, such as users or auditors.

Objective 4 – Evaluation

Finally, the framework is evaluated across three dimensions: model utility (accuracy and F1-score), privacy guarantees (membership inference attack resistance), and explainability (semantic completeness and interpretability of generated summaries).

To address the research objectives, this thesis adopts a design-science methodology involving the iterative development and evaluation of a novel framework. The approach includes:

- Implementing multiple unlearning algorithms (Full Retraining, KD, SISA),
- Designing an RDF-based semantic model for deletion representation,
- Generating natural-language summaries to enhance interpretability,
- Conducting empirical evaluation using performance, privacy, and explanation metrics.

This combination of technical implementation and semantic modelling provides a holistic foundation for investigating explainable machine unlearning.

2. LITERATURE REVIEW

The accelerating digitization of society has expanded the collection and dissemination of personal information across borders. While this transformation enables innovation and growth, it also raises serious risks to privacy, autonomy, and informational self-determination. In response, regulators and researchers have sought to reinforce accountability and give individuals greater control over their data. Central to these efforts is the Right to Be Forgotten (RTBF), formally established under the General Data Protection Regulation (GDPR).

The rapid advancement of machine learning (ML) and knowledge-based technologies, however, complicates this right, as personal data becomes embedded in complex models and large-scale knowledge graphs. These developments create both legal and technical tensions: the law envisions erasure as a matter of autonomy, but computational systems often lack effective mechanisms for forgetting.

Research on the RTBF spans several domains. Legal and organizational studies assess how the GDPR reshapes compliance and the balance between privacy and expression. Technical work explores data deletion in modern systems, with machine unlearning emerging as a promising method for removing the influence of specific data from models. In parallel, knowledge graph summarization has been studied to manage complexity, improve efficiency, and support transparency. Overall, the literature shows progress in shaping the RTBF but also reveals persistent gaps in turning it into organizational practice and technical design. These unresolved tensions provide the backdrop for the present study.

2.1 The Evolution of the RTBF in the GDPR Era

The Right to be Forgotten (RTBF) has become one of the most debated and symbolically significant provisions in European data protection law. Although formally codified under Article 17 of the General Data Protection Regulation (GDPR), its roots lie in earlier European jurisprudence, most notably the Google Spain v. AEPD and Mario Costeja González ruling of 2014. In this case, the Court of Justice of the European Union (CJEU) held that search engines act as “data controllers” and therefore bear responsibility for ensuring that links to outdated or irrelevant personal information can be delisted upon request (6). This decision represented a paradigm shift in digital privacy: it extended accountability beyond traditional database owners to intermediaries such as search engines, thereby crystallizing the notion that individuals should retain control over their digital traces even after they have entered the public domain (7).

The GDPR, which came into force in May 2018, codified this principle in Article 17, granting individuals the right to obtain erasure of their personal data “without undue delay” under specific conditions. These include circumstances where the data is no longer necessary for the purpose it was collected, where consent has been withdrawn, or where processing has been found to be unlawful. The codification of the RTBF reflects Europe’s ambition to position itself as a global leader in digital rights protection; however, its extraterritorial scope introduces complex jurisdictional questions, particularly in relation to global platforms headquartered outside the EU (8).

Despite its symbolic importance, the RTBF has sparked criticism and debate on several fronts [5], [6]. One central concern lies in its tension with freedom of expression and the right to information. Critics argue that the broad interpretation of the right could encourage over-removal of information, effectively privatizing censorship decisions by placing responsibility in the hands of technology companies. Organizations often perceive the RTBF as a source of compliance risk, leading to cautious or conservative responses to deletion requests [3], [4] . From the user perspective, individuals value greater control over their digital presence but remain uncertain about the effectiveness and transparency of the deletion process (12), (13).

Several studies highlight the practical difficulties of enforcing the RTBF across different jurisdictions. Privacy policies worldwide have been significantly reshaped by the GDPR, yet full realization of the RTBF remains hampered by organizational reluctance, inconsistent implementation, and gaps in legal interpretation (14). Existing definitions of deletion also fall short when applied to interactive digital systems. Conceptualizing deletion purely as “confidentiality” is inadequate, since this would require secrecy of deleted data in ways that exclude many practical functionalities. Instead, deletion is better understood as “control,” reflecting the need for individuals to meaningfully influence how their data is retained, used, or removed (15).

Another dimension of the RTBF debate concerns its application in an era of globally distributed infrastructures. Implementing RTBF guarantees in practice is particularly difficult in public sector institutions, where compliance often relies on manual processes and limited technical automation. Large-scale and timely compliance remains extremely challenging, requiring significant institutional investment in technical solutions (16).

At the broader societal level, the RTBF reflects a normative commitment to digital dignity: the idea that individuals should not be perpetually haunted by past actions or information. However, its implementation continues to provoke friction between competing values: privacy versus free expression, individual autonomy versus public interest, and European regulatory authority versus global technological infrastructures. These tensions are further exacerbated by the rise of large-scale machine learning models, which memorize personal data and embed it in ways that render traditional erasure mechanisms insufficient. This highlights the growing need for novel technical solutions such as machine unlearning and model editing (17).

In sum, the evolution of the RTBF demonstrates the dynamic interplay between law, technology, and society. From its origins in case law to its codification in the GDPR, and from its legal symbolism to its technical and organizational complexities, the right encapsulates the challenges of governing digital memory in the twenty-first century.

2.2 Organizational Realities of GDPR Compliances

While the Right to be Forgotten (RTBF) is firmly embedded in European law, translating its principles into organizational practice has proven to be highly complex. Compliance with the GDPR requires not only adherence to legal standards but also substantial adjustments to governance structures, technical infrastructures, and institutional workflows. Organizations across both public and private sectors continue to face challenges in operationalizing the RTBF, which is often perceived less as a straightforward legal obligation and more as a source of uncertainty, risk, and resource strain (13).

One recurring theme in the literature is the resource-intensive nature of GDPR compliance. Many organizations, particularly small and medium-sized enterprises, regard the RTBF as burdensome due to the costs associated with processing deletion requests and redesigning data architectures to accommodate erasure (18). Their study highlights a cautious approach to handling requests, as institutions often adopt conservative compliance strategies to avoid potential legal sanctions. This results in over-compliance in some cases, where data is removed even when it may not strictly be required, and under-compliance in others, where requests are delayed or partially fulfilled (12).

The technical fragmentation of organizational infrastructures further complicates compliance. Examining GDPR practices in Italian Public Administration, emphasizes the reliance on manual procedures and limited automation. Their findings reveal that deletion requests often require coordination across multiple departments and systems, many of which lack interoperability. As a result, large-scale and timely compliance remains difficult, particularly in institutions where personal data is duplicated across legacy databases, backup systems, and third-party services (16).

Zaguir et al. (2024), in their Systematic Literature Review of Challenges and Enablers for GDPR Compliance, categorize the key organizational challenges into three domains: legal–institutional, technological, and cultural. They argue that organizations frequently struggle to interpret ambiguous provisions of the GDPR, especially in cross-border contexts where jurisdictional conflicts emerge. On the technological side, insufficient integration of compliance tools and limited expertise in privacy-enhancing technologies hinder effective implementation. Finally, cultural barriers, such as lack of awareness among staff and resistance to changing established workflows, exacerbate non-compliance (13).

Private sector organizations also face reputational and competitive pressures linked to compliance. Zaeem and Barber (2020) note that while privacy policies worldwide have been reshaped by the GDPR, many firms still fall short of demonstrating transparency in how erasure requests are processed. In practice, compliance is often “performative,” where organizations present the appearance of alignment with GDPR principles but avoid the deeper restructuring required to fully realize the RTBF. This raises concerns about accountability and trust, as users may not receive clear assurances that their data has been permanently erased (14).

Several studies also highlight the tension between compliance and innovation. While the GDPR is intended to protect individuals, strict adherence can limit data-driven research and technological development. Almeida Teixeira et al. (2019), in their Systematic Literature Review on GDPR Implementation, identify “critical success factors” such as top management support, cross-disciplinary collaboration, and investment in compliance technologies as essential for balancing regulatory obligations with business innovation. Without these enablers, organizations risk treating compliance as a reactive exercise rather than a proactive strategy (6).

At a broader level, the organizational realities of GDPR compliance underscore the uneven distribution of capacities across different sectors and regions. Large multinational corporations often have the resources to invest in compliance frameworks, whereas smaller firms and public institutions struggle with limited expertise and budgets. Zaguir et al. (2024) argue that information governance (IG) and enterprise architecture management (EAM) can serve as enabling frameworks, aligning compliance with long-term organizational strategy. However, empirical evidence suggests that many organizations still view GDPR obligations as external impositions rather than opportunities to strengthen trust and accountability.

In summary, the literature portrays GDPR compliance as a complex, uneven, and resource-intensive endeavor. Organizations frequently adopt conservative or inconsistent approaches to managing erasure requests, reflecting both technical limitations and institutional uncertainties. While progress has been made in reshaping privacy policies and raising awareness of data protection, significant challenges remain in aligning legal obligations with organizational capacities. These challenges highlight the need for integrated governance frameworks, automation of compliance processes, and cultural shifts that embed privacy as a core organizational value rather than a peripheral legal requirement.

From its origins in case law to its codification in the GDPR, and from its legal symbolism to its technical and organizational complexities, the right encapsulates the challenges of governing digital memory in the twenty-first century.

2.3 When Forgetting Meets Explaining: RTBF and AI Transparency

One of the most complex intersections in data protection concerns the relationship between the Right to be Forgotten (RTBF) and the emerging requirement for algorithmic transparency. While the RTBF focuses on erasure, the Right to Explanation under the GDPR (Article 22) emphasizes accountability in automated decision-making. These two rights reflect different but interrelated goals: ensuring that individuals can control the persistence of their digital identity and providing them with intelligible insights into how algorithmic systems use their personal data. Together, they raise critical questions about how legal rights translate into technical design within artificial intelligence (AI) (19).

Hamon et al. (2022), in Bridging the Gap Between AI and Explainability in the GDPR, argue that explainability is essential to building trustworthy AI systems. They emphasize that without transparency, users cannot meaningfully exercise their data rights, including the RTBF. For example, if an individual requests deletion of personal data from a machine learning model, an organization must be able to demonstrate not only that the data has been removed but also that its influence no longer shapes automated decisions. This requires both technical deletion mechanisms and explanatory tools that provide verifiable evidence of compliance (20).

Krishna et al. (2023), in Towards Bridging the Gaps Between the Right to Explanation and the Right to be Forgotten, highlights the tension between these two rights. They observe that erasure and explanation may sometimes conflict: while the RTBF demands the elimination of traces of personal data, the right to explanation requires retaining information about how data was used in decision-making processes. For instance, model interpretability techniques such as feature attribution or training data provenance may rely on records of data that an individual has requested to be forgotten. This creates a paradox: transparency may necessitate forms of data retention that undermine erasure, while strict deletion may reduce the capacity to explain algorithmic decisions (21).

Recent scholars have also identified practical barriers to aligning erasure and explanation in AI contexts. Ramokapane and Rashid (2023), in ExD: Explainable Deletion, propose the concept of “explainable deletion” as a way to address these challenges. They argue that deletion should not only be effective but also transparent and interpretable for users, ensuring that individuals understand how and why their data has been erased. Their framework emphasizes user agency and accountability, suggesting that explanations can serve as a bridge between legal compliance and user trust (5).

At a broader level, the intersection of RTBF and AI transparency highlights the limits of legal frameworks when confronted with technical realities. While the GDPR establishes rights to erasure and explanation, it does not specify how these should be operationalized in practice. This has led to significant variability in implementation across organizations, with some adopting basic compliance tools while others explore advanced technical solutions such as explainable AI (XAI) methods. The lack of clear standards leaves room for interpretation, often placing the burden of compliance on organizations with differing levels of resources and technical expertise (20,21).

Recent research further expands the scope of explainability beyond GDPR to encompass fairness and accountability in data-driven systems. Araújo et al. provide an overview of fairness and explanations in entity resolution, arguing that without transparent and equitable mechanisms, users cannot meaningfully exercise control over their personal data (22). In the context of recommendations, Stratigi et al. propose counterfactual explanations for group recommendations, highlighting how explanation frameworks can improve trust in collective decision-making (23). Attolou et al. extend this line of inquiry to graph-based recommenders, introducing “why-not” explanations that clarify why specific outcomes are absent (24). Similarly, Giannopoulos et al. demonstrate how interactivity and fairness can be embedded into explanation processes for recommender systems, reinforcing the link between autonomy and transparency (25). Earlier work by Stratigi et al. also introduced “why-not” questions in collaborative filtering, showing that explanation research has long been concerned with user empowerment (26). Taken together, these studies reinforce the argument that explainability is not only a matter of interpretability but also a vehicle for ensuring fairness, accountability, and compliance with rights such as the RTBF.

Overall, the literature suggests that erasure and explanation must be understood as complementary dimensions of digital rights. The RTBF ensures that individuals are not indefinitely tied to their past data, while the right to explanation ensures that they are not subjected to opaque algorithmic decisions. Yet aligning these rights in practice remains an unresolved challenge. Achieving this balance will likely require new socio-technical frameworks that integrate deletion mechanisms with explainability tools, thereby reinforcing both accountability and autonomy in AI-driven environments.

2.4 Rethinking Deletion: From Confidentiality to Control

The traditional understanding of data deletion has often been framed in terms of confidentiality, where erasure is equated with the removal of records from a storage system. However, in the context of the Right to be Forgotten (RTBF) and the broader GDPR framework, this view proves insufficient. Modern digital ecosystems are characterized by distributed infrastructures, backups, and machine learning models in which personal data may persist in indirect or derivative forms. As a result, scholars have argued for rethinking deletion not only as a technical operation but as a socio-technical process that guarantees user control and trust (2).

Cohen et al. (2023), in Control, Confidentiality, and the Right to be Forgotten, stress that conventional definitions of deletion fall short in interactive and dynamic digital systems. Simply concealing or overwriting data does not provide meaningful assurance to users, especially if remnants remain accessible through cached systems, third-party services, or trained models. They propose an alternative conceptualization of "deletion-as-control," where the key objective is to empower individuals to meaningfully influence how their data is managed, retained, or removed. Under this paradigm, deletion is not merely about secrecy but about guaranteeing agency, allowing individuals to determine the ongoing impact of their data within complex infrastructures (15).

Garg et al. (2020), in Formalizing Data Deletion in the Context of the Right to be Forgotten attempt to give this principle a more rigorous grounding by developing formal models of deletion. Their work distinguishes between weak and strong forms of erasure, depending on whether deletion prevents future access or eliminates all traces of the data's historical influence. Strong deletion, while desirable in principle, is often technically infeasible in systems where data has been replicated or used in training machine learning models. Weak deletion, by contrast, allows for practical implementations but raises concerns about whether it sufficiently fulfils the legal requirements of the RTBF (2).

The notion of “explainable deletion,” introduced by Ramokapane and Rashid (2023), further extends this line of thinking by emphasizing accountability and transparency. They argue that deletion should not only occur but also be demonstrable to users in ways they can understand. This involves providing clear explanations of what data has been removed, how it was erased, and what residual traces, if any, remain in the system. Their proposal positions deletion as a process that should be intelligible to multiple stakeholders—users, regulators, and organizations—thereby reinforcing trust in compliance mechanisms (5).

A related challenge is the verification of deletion. Unlike data access or modification, deletion is difficult to prove once executed. Recent work on audit and verification frameworks suggests the need for protocols that allow independent confirmation without compromising system security or performance. For example, some studies on unlearning and secure deletion propose cryptographic proofs or audit logs that demonstrate compliance with erasure requests, aligning with the GDPR’s accountability principle (15).

Taken together, this literature illustrates a shift from viewing deletion as a purely technical act to understanding it as a layered process involving legal guarantees, user agency, and verifiable assurance. By reframing deletion as control, scholars emphasize the importance of empowering individuals within complex digital infrastructures. At the same time, formal models and verification mechanisms highlight the practical challenges of implementing these ideals. The RTBF thus continues to act as a catalyst for rethinking how erasure is defined, executed, and communicated in contemporary data systems.

2.5 Machine Unlearning: Algorithms for Digital Forgetting

The enforcement of the Right to be Forgotten (RTBF) has spurred increasing interest in the concept of machine unlearning, which refers to the design of algorithms capable of removing the influence of specific training data from a machine learning (ML) model. Unlike traditional deletion, which may only erase data from storage systems, unlearning requires altering the internal state of a model to ensure that erased samples no longer affect predictions or outputs. This paradigm represents one of the most significant technical challenges for implementing digital forgetting, as data in ML systems is often distributed, transformed, and embedded in ways that resist straightforward removal (27), (1).

2.5.1 Foundations of Machine Unlearning

The idea of unlearning was first articulated by Cao and Yang (2015) (28), who proposed "selective forgetting" as a mechanism to efficiently remove training samples without retraining the entire model from scratch. Subsequent work by Bourtoule et al. (2021) (3) advanced this idea through SISA training, a partitioned approach that reduces the retraining cost when unlearning is required. These foundational contributions framed unlearning as both a legal necessity—driven by the RTBF—and a technical optimization problem in large-scale ML systems.

More recent surveys consolidate the field. Li et al. (2024) (29), in Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects, provide a comprehensive taxonomy distinguishing between exact and approximate unlearning, centralized and federated settings, and proactive versus reactive strategies. Similarly, Nguyen et al. (2024), in A Survey of Machine Unlearning, outline the technical challenges of unlearning in deep neural networks, emphasizing computational efficiency, utility preservation, and verifiability. Wang et al. (2024), in Machine Unlearning: A Comprehensive Survey, further highlight open questions around security, privacy leakage, and unlearning verification, situating the field within broader debates on trustworthy AI.

2.5.2 Taxonomy of Unlearning Approaches

Literature broadly categorizes machine unlearning into two types:

- **Exact unlearning:** Guarantees that the unlearned model is indistinguishable from one trained without the forgotten data. While desirable, this is often computationally infeasible for complex models. Formal methods, such as those proposed by Garg et al. (2020) (2), attempt to define criteria for exact deletion but acknowledge the difficulty of achieving strong guarantees at scale.
- **Approximate unlearning:** Produces a model that closely resembles one retrained from scratch, but without strict guarantees. Techniques include gradient manipulation, knowledge distillation, and fine-tuning. While these methods reduce computational overhead, they raise questions about the sufficiency of approximation for legal compliance under the RTBF (27), (29).
- **Federated unlearning:** Addresses scenarios where data is distributed across multiple clients, as in federated learning. Removing data from decentralized systems introduces new challenges of coordination and verification (1).

- **Proactive vs. reactive unlearning:** Proactive methods design models with unlearning in mind, embedding modularity or partitioning strategies to simplify later deletion (3). Reactive methods, by contrast, attempt to remove data influence after training, often incurring higher costs and risks of incomplete forgetting.

2.5.3 Metrics and Verification

Evaluating the effectiveness of unlearning requires robust metrics. Shi et al. (2025) (30), in Rethinking Evaluation Metrics for Machine Unlearning, argue that traditional measures such as unlearning accuracy and membership inference attack resistance are insufficient. They highlight the risk of “fake unlearning,” where a model misclassifies forgotten data but still retains internal traces of it. To address this, they propose metrics inspired by conformal prediction, which assess whether the ground truth label of forgotten data persists in the model’s predictive set.

Other evaluation frameworks are drawn from summarization and NLP metrics. For example, Koto et al. (2022) (11) introduce FFCI, a framework for interpretable evaluation, while Lin et al. (2022) (31) propose SummScore, a cross-encoder-based quality metric. Although designed for summarization tasks, these approaches illustrate how evaluation can be made more interpretable and user-centric—an important consideration for unlearning verification. Dai et al. (2024) (32), in A Critical Look at Meta-Evaluating Summarization Metrics, further argue that evaluation itself requires careful auditing, a principle that applies equally to unlearning.

Verification of unlearning also intersects with legal accountability. Organizations must not only delete data but demonstrate compliance to regulators and users. Formal auditing methods, cryptographic proofs, and logging mechanisms have been proposed as ways to ensure that deletion requests are honored (2). These mechanisms align with the GDPR’s emphasis on transparency and accountability, underscoring the socio-technical nature of unlearning.

2.5.4 Applications of Machine Unlearning

Practical applications of unlearning span diverse domains. Yang et al. (2024) (33), in From Machine Learning to Machine Unlearning: Complying with GDPR’s Right to be

Forgotten while Maintaining Business Value, propose an ensemble-based framework (ETID) that balances compliance with business objectives, demonstrating that effective unlearning need not compromise model accuracy or commercial utility. Ahmed et al. (2025) (34) discuss applications in bias mitigation, where unlearning is used to remove harmful stereotypes embedded in models. Other studies explore unlearning in recommendation systems, healthcare analytics, and large-scale natural language processing, highlighting its role in both compliance and ethical AI.

Zhang et al. (2025) (17) underscore the urgency of unlearning in the era of large language models (LLMs). Unlike traditional ML models, LLMs memorize vast amounts of training data, including personal information, which can resurface in outputs. They argue that conventional unlearning techniques are inadequate for models of this scale, calling for new solutions such as model editing, differential privacy, and guardrails. This perspective connects unlearning directly with the future of RTBF in AI-driven ecosystems.

2.6 Summarizing Knowledge Graphs for Privacy and Efficiency

Knowledge graphs (KGs) have become central to modern information systems, integrating heterogeneous data into structured networks of entities and relationships. While KGs enable powerful applications in search, recommendation, and AI reasoning, their scale and complexity make them difficult to process and manage (35). To address these challenges, research on *knowledge graph summarization* has emerged, aiming to produce condensed representations that preserve essential information while reducing redundancy (36). Summarization is increasingly relevant to the *Right to be Forgotten* (RTBF) and GDPR compliance, as it provides a pathway to manage large datasets more efficiently, enhance transparency, and facilitate auditing of data retention and deletion practices (37), (38).

In recent years, research on summarization has moved beyond traditional text compression to approaches leveraging semantic graphs, domain-specific applications, and new evaluation metrics. Wei et al. (2025) proposed MSCRS, a multi-modal semantic graph prompt learning framework for conversational recommender systems, demonstrating how combining textual and visual modalities with graph-based structures can enhance personalization and naturalness of generated responses (39). Ribeiro et al. (2022) introduced FactGraph (40), which encodes both documents and summaries into structured meaning representations, improving the detection of factual consistency and subsentence-level inaccuracies. Along similar lines, Kouris et al. (2024) (41) advanced a deep learning framework for abstract meaning representation (AMR) graph-to-text summarization, enabling abstractive summaries that retain factual consistency.

Recent developments have also explored the use of large language models (LLMs) to enhance summarization. Iqbal and Stefanidis (2025) propose an LLM-driven approach for summarizing and distinguishing multiple entities in RDF graphs, demonstrating how neural models can support more nuanced, human-like summaries. Earlier contributions such as Troullinou et al. emphasized the role of summaries in facilitating interactive exploration of RDF/S knowledge bases, showing that summarization can make large datasets more accessible to end-users (42). Complementing these perspectives, Vassiliou et al. introduce coverage-based summaries for RDF knowledge bases, aiming to balance conciseness with representational completeness (43). These works build upon traditional structural and semantic summarization approaches by highlighting new directions that emphasize usability, coverage, and integration with emerging AI technologies.

2.6.1 Surveys and Foundations

The field of KG summarization builds on broader work in graph compression and summarization. Liu et al. (2018), in *Graph Summarization Methods and Applications: A Survey*, provides a foundational overview of summarization techniques across domains, highlighting methods such as grouping, clustering, and compression. Čebirić et al. (2019) focus specifically on semantic graphs, categorizing methods into structural, pattern-based, and hybrid approaches. More recently, Wang and Cheng (2024) consolidate this knowledge in their survey *Extractive Knowledge Graph Summarization: Applications, Approaches, Evaluation, and Future Directions*, where they distinguish between *static* summaries (context-independent) and *dynamic* summaries (tailored to user needs). These surveys show that summarization is not only a technical optimization but also a usability tool that supports tasks such as query optimization, semantic exploration, and compliance verification (44).

The scope of summarization has also been extended into specialized domains. Mridha et al. (2021) (45) provided a comprehensive survey of automatic text summarization (ATS), mapping the evolution from extractive and abstractive methods to modern deep learning approaches, with an emphasis on datasets, evaluation techniques, and remaining challenges. Dan et al. (2025) (46) tailored summarization to the legal domain by combining semantic and structural information, yielding high-quality legal judgment summarization that outperformed general models. Similarly, Lanov (2025) (47) explored semantic policy summaries and explanations in the context of AI for networking, emphasizing explainability alongside compact representations. Finally, Zhao et al. (2025) (48) addressed the semantic alignment gap in code summarization, proposing a data-flow–guided framework that ensures stronger correspondence between source code and natural language summaries.

2.6.2 Methods and Applications

Several notable approaches illustrate the diversity of summarization strategies. Goasdoué et al. (2020) (4) propose methods for *RDF graph summarization* aimed at structure discovery, enabling users to obtain a “first sight” understanding of complex datasets. Diao et al. (2021) (49) explore summarization through *interesting aggregates*, automatically identifying the most salient patterns in RDF data. Niazmand et al. (2022) (50) introduce *semantic summary graphs* for efficient querying, showing that summarization can reduce query time by up to 80% while preserving completeness. Zimina et al. (2024) (9) push this further by developing *linguistic summarization* techniques that generate natural language summaries of RDF entity sets, making large datasets more accessible to non-technical users. Similarly, Vassiliou et al. (2024) (10) presents *iSummary*, a workload-based approach that personalizes summaries according to user query logs, balancing efficiency with relevance (51).

These applications highlight how summarization can support GDPR compliance. By reducing data complexity, summarization enables more transparent auditing of where personal information resides within KGs. It also offers mechanisms for tailoring views of data, potentially allowing users and regulators to verify whether specific information has been erased or retained.

2.6.3 Evaluation and Quality Metrics

Evaluating the quality of KG summaries remains a critical challenge. Thakker et al. (2019) and Zneika et al. (2019) (52) identify criteria such as completeness, conciseness, and semantic fidelity as central to effective summarisation. Lin et al. (2022) (31), with their *SummScore* metric, and Koto et al. (2022) (11), with the *FFCI framework*, offer methods for assessing summary quality in NLP that are increasingly being applied to KG evaluation. Dai et al. (2024) (32) caution that evaluation itself must be critically examined, warning against over-reliance on automated metrics without human validation. These contributions underline the importance of aligning evaluation with user needs, legal requirements, and the GDPR's principles of accountability and transparency (53).

2.7 Forgetting in the Age of Large Language Models

The emergence of large language models (LLMs) such as GPT and BERT has introduced new challenges for the implementation of the Right to be Forgotten (RTBF). Unlike traditional databases or even smaller-scale machine learning (ML) systems, LLMs are trained on massive corpora of text and can memorize specific details, including personal data (54). Once memorized, such information can resurface in generated outputs, raising profound legal and technical questions about compliance with GDPR and the enforceability of the RTBF (17).

Beyond GDPR compliance, scholars have also highlighted the broader risks of deploying LLMs in high-stakes environments. Ranasinghe et al. examine the use of LLMs in human-robot collaboration and show how context-induced hallucinations challenge reliability and trust (55). Their work suggests that ensuring cognitive validation in such systems is essential, underscoring the urgency of developing verification and unlearning methods that can guarantee both safety and compliance. This resonates with concerns raised in GDPR debates, where hallucinated or memorized personal information in LLM outputs directly undermines the enforceability of the RTBF.

One of the core difficulties lies in the opacity of model parameters. In contrast to structured datasets where a record can be directly located and deleted, LLMs encode training data into high-dimensional weight matrices (56). As a result, personal information is not stored in explicit entries but is instead entangled across billions of parameters. This makes identifying and removing the influence of specific data extremely challenging (1,27).

Recent work has proposed technical pathways for addressing these challenges. Zhang et al. (2025) highlight strategies such as machine unlearning, model editing, and differential privacy as potential mechanisms to align LLMs with the RTBF. Machine unlearning aims to reverse the contribution of specific training examples, while model editing techniques attempt to directly modify parameters to suppress undesired outputs. Differential privacy, by contrast, seeks to prevent memorization during training, reducing the likelihood that sensitive data will appear in generated text. Each of these approaches, however, comes with trade-offs in terms of computational cost, utility preservation, and legal sufficiency (57).

Hamon et al. (2022) argue that transparency and explainability become even more critical in the context of LLMs. If an organization claims to have removed personal data from a model, users and regulators must be able to verify this claim. Yet, as Krishna et al. (2023) note, the tension between erasure and explanation persists: proving that information has been forgotten may require retaining records of how it was used, potentially undermining strict interpretations of the RTBF (58).

At present, there is no consensus on how the RTBF can be reliably operationalized in the era of LLMs. The literature points to a need for hybrid approaches that combine privacy-preserving training, efficient unlearning methods, and strong verification protocols (1,27). More broadly, the rise of LLMs underscores the urgency of developing sociotechnical frameworks that can reconcile legal mandates for erasure with the realities of modern AI systems (30).

2.8 Open Challenges and Future Directions in Digital Forgetting

Across law, organizational practice, and technical research, literature reveals both substantial progress and persistent challenges in operationalizing the *Right to be Forgotten* (RTBF). While the GDPR has firmly established erasure as a legal right, its translation into practice remains uneven and contested. The key open challenges can be grouped into four broad themes.

2.8.1 Legal and Organizational Gaps

The first set of challenges arises from ambiguities in legal interpretation and uneven organizational capacity. Studies show that organizations frequently adopt conservative or inconsistent strategies when processing deletion requests, reflecting uncertainty about the scope of compliance and the potential risks of under- or over-removal (12,14,16). Public institutions remain reliant on manual procedures, while smaller firms often lack resources for automated compliance (6,13). These disparities highlight the need for clearer regulatory guidance and accessible technical frameworks that lower the compliance burden across different organizational contexts.

2.8.2 Tensions Between Erasure and Transparency

A second area of difficulty concerns the tension between erasure and explanation. While RTBF requires deletion, the right to explanation demands that individuals understand how their data is processed. These two principles may conflict, particularly in AI systems where model interpretability relies on retaining traces of training data (20,21). Proposals such as “explainable deletion” attempt to bridge this gap (5) but the literature shows that practical frameworks for balancing these rights remain underdeveloped.

2.8.3 Technical Barriers in Machine Unlearning

Technical literature has made notable strides in developing unlearning algorithms, yet challenges persist. Exact unlearning remains computationally infeasible for large models, while approximate methods raise questions about legal adequacy (27–29). Verification of unlearning is another unresolved issue: ensuring that deleted data has no residual influence requires stronger audit protocols, cryptographic proofs, and standardized evaluation metrics (30,38). Furthermore, the application of unlearning to large-scale models such as LLMs introduces new complexities that existing methods cannot fully address (17).

2.8.4 Knowledge Graphs and Semantic Transparency

Knowledge graph summarization offers promising avenues for enhancing transparency and efficiency in data governance. However, challenges remain in balancing conciseness with completeness, ensuring that summaries remain up-to-date, and aligning evaluation metrics with legal requirements (38). Integrating summarization with RTBF raises open questions about whether summaries can inadvertently preserve or amplify traces of personal data, complicating compliance.

2.8.5 Future Directions

Taken together, these challenges suggest several future directions for research and practice. First, there is a pressing need for socio-technical frameworks that integrate legal principles, organizational practices, and technical mechanisms into coherent strategies for digital forgetting. Second, hybrid approaches that combine privacy-preserving training, proactive model design, and reactive unlearning could reduce the computational and legal limitations of existing methods (3). Third, developing explainable compliance tools—capable of demonstrating deletion in ways that are verifiable and intelligible—would help reinforce user trust and regulatory accountability. Finally, interdisciplinary collaboration between lawyers, policymakers, computer scientists, and industry practitioners will be crucial to ensuring that the RTBF remains enforceable in an era of increasingly complex AI and data infrastructures.

Through this analysis, it becomes clear that existing research often treats legal, organizational, and technical perspectives on digital forgetting in isolation, leaving a gap in how these strands can be meaningfully integrated. This thesis responds to that gap by focusing on the intersection of RTBF, machine unlearning, and knowledge graph summarization. By drawing together insights from data protection law, compliance practices, and computer science, it seeks to provide a more holistic understanding of how digital forgetting can be operationalized. In doing so, it analyses machine unlearning as a technical pathway for erasure in modern AI systems and explores knowledge graph summarization as a complementary mechanism for enhancing transparency, efficiency, and accountability. In this way, the study aims to extend current debates and contribute both conceptual clarity and practical insights into how the RTBF can be realized within increasingly complex data ecosystems.

Unlike existing unlearning approaches, which focus primarily on model-level deletion or computational efficiency, this work integrates unlearning with a semantic explainability layer. Prior studies rarely combine deletion mechanisms with RDF-based summaries or natural-language explanations, and none, to our knowledge, evaluate explainable unlearning through combined performance, privacy, and interpretability metrics.

In contrast to existing studies, which typically address machine unlearning, explainability, or GDPR compliance in isolation, this thesis adopts an integrated perspective. Prior unlearning approaches primarily focus on computational efficiency or theoretical deletion guarantees, while explainability research emphasizes prediction-level transparency rather than deletion processes. Knowledge graph summarization work, although relevant for transparency, rarely addresses model-level data removal. This work differentiates itself by combining unlearning algorithms with semantic RDF representations and natural-language summaries, and by jointly evaluating performance, privacy leakage, and explanation quality. To the author’s knowledge, no prior study systematically integrates these components into a unified, explainable unlearning framework.

3. METHODOLOGY

3.1 Research Problem and Design Rationale

Machine learning models are increasingly integrated into decision-making systems that process user-generated data, such as text reviews, recommendations, or behavioral analytics. Although these systems have become central to modern data-driven services, they raise critical concerns about data privacy and user control. Under the General Data Protection Regulation (GDPR), individuals have the Right to Erasure or Right to Be Forgotten (RTBF), which requires data controllers to delete personal data upon request and to ensure that it exerts no residual influence on automated processes.

While deletion in conventional databases is straightforward, erasure within trained machine learning models presents a unique challenge—models retain embedded representations of training data, making it technically difficult to remove specific samples without retraining the entire system. This gap between legal compliance and technical capability forms the central research problem addressed by this study:

How can machine learning systems be designed to support verifiable, explainable, and efficient data deletion in compliance with the Right to Be Forgotten?

Existing approaches to machine unlearning focus mainly on algorithmic retraining but often neglect explainability, verification, and compliance traceability. Such methods lack mechanisms to demonstrate how forgetting occurs, whether it was successful, and how the operation aligns with legal and ethical obligations.

Consequently, there is a need for an integrated framework that unites the technical, verification, and explainability dimensions of data deletion. To address this challenge, this research proposes an Explainable Unlearning Framework (EUF) that combines:

1. Unlearning mechanisms, which simulate the removal of data through computational methods.
2. Verification processes, which quantitatively assess the effectiveness of forgetting.

3. Explainability and reporting components, which transform deletion events into semantically interpretable records.

The framework is implemented and evaluated through three representative unlearning approaches—Full Retraining, SISA (Sharded, Isolated, Sliced Aggregation), and Knowledge Distillation (KD)—each chosen to explore a different balance between accuracy, computational efficiency, and compliance transparency. By comparing these methods within one unified system, the study aims to demonstrate that machine unlearning can be both technically effective and legally auditable.

The overarching goal of this research is therefore to design a privacy-aware learning framework that not only forgets data but also proves and explains the forgetting process—bridging the gap between data protection law and machine learning practice outcomes.

3.2 Framework architecture

The proposed Explainable Unlearning Framework (EUF) is designed to operate the Right to Be Forgotten (RTBF) in machine learning systems by integrating unlearning algorithms with verification and explainability mechanisms.

The framework addresses three critical requirements of GDPR-compliant AI systems:

(1) the technical capacity to remove data influence from trained models,

(2) the ability to verify that such removal is effective, and

(3) the provision of human-interpretable evidence of compliance.

To achieve these objectives, the framework is structured into three sequential yet interdependent layers: the Unlearning Layer, the Verification Layer, and the Explainability Layer. Each layer performs a distinct role but operates within a unified pipeline that ensures traceability from deletion request to explainable outcome.

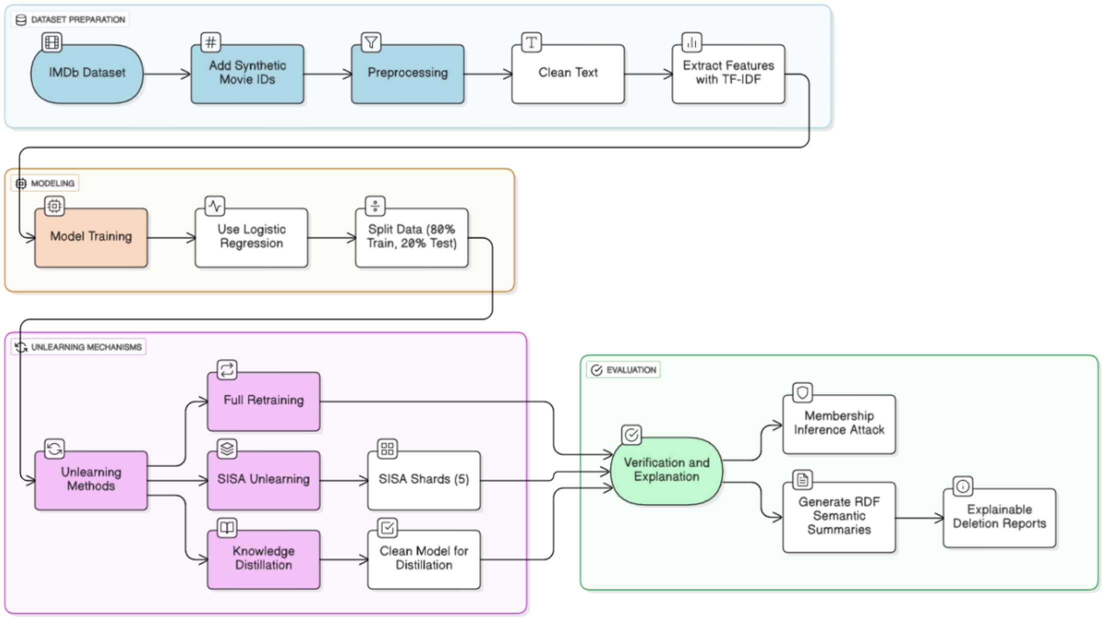

Figure 1. *Overall architecture of the Explainable Machine Unlearning Framework*

The diagram illustrates the full pipeline, including dataset preparation, model training, unlearning mechanisms (Full Retraining, SISA, and Knowledge Distillation), and the evaluation components responsible for privacy verification and semantic explanation generation.

Unlearning Layer

The Unlearning Layer performs the technical deletion of data within a trained model. This is achieved by applying one of three implemented unlearning mechanisms: Full Retraining, SISA (Sharded, Isolated, Sliced Aggregation), or Knowledge Distillation (KD). Each method represents a different balance between fidelity of deletion and computational efficiency. Full Retraining serves as the compliance baseline by rebuilding the model entirely without deleted samples. SISA improves efficiency through isolated shard retraining and Knowledge Distillation provides a lightweight alternative that transfers knowledge from a retrained “teacher” model to a new “student” model without directly using the deleted data. This modular design allows flexible experimentation and comparative evaluation of multiple unlearning strategies within a consistent environment.

Verification Layer

The Verification Layer ensures that the unlearning operation has successfully removed the influence of the deleted data. This is implemented through a Membership Inference

Attack (MIA) test, which measures the distinguishability between deleted (“member”) and unseen (“non-member”) samples. A model that has truly forgotten its deleted data should exhibit a random-like response, with a Membership Inference AUC value close to 0.5.

This quantitative verification process converts the abstract concept of “forgetting” into an empirically testable metric, thereby linking the framework’s technical outcomes to measurable privacy guarantees.

Explainability Layer

The final Explainability Layer translates technical operations into interpretable and auditable outputs. Each unlearning event generates a structured RDF (Resource Description Framework) record describing the deletion mode, scope, and outcome. These semantic records are then converted into natural-language summaries that provide a readable account of what data was deleted, how it was processed, and whether the operation was successful. By producing both machine-readable and human-readable explanations, this layer satisfies GDPR’s transparency and accountability principles. It allows auditors, developers, or regulators to trace each unlearning action without exposing sensitive training data.

Framework Workflow

In operation, the framework follows a linear yet traceable workflow:

(1) A deletion request is first issued and processed by the Unlearning Layer,

(2) Verified for successful erasure in the Verification Layer,

(3) and then documented through the Explainability Layer.

Each layer passes metadata forward, creating an audit trail that links input, model update, verification result, and final report. This pipeline design ensures that every deletion request can be executed, tested, and explained—turning privacy compliance into a demonstrable technical process.

3.3 Data Preparation and Pre-processing

The experimental evaluation of the Explainable Unlearning Framework was conducted using two complementary data sources: the IMDb Large Movie Review Dataset and the DBpedia knowledge base. The IMDb dataset was used to train and test the machine unlearning algorithms, while DBpedia provided structured semantic references to support the explainability component of the framework.

IMDb Dataset

The IMDb dataset (59) is a benchmark corpus widely used for sentiment classification research. It contains 50,000 movie reviews, evenly divided between positive and negative sentiments. This dataset was selected because it represents real-world user-generated content, making it suitable for simulating deletion requests under the Right to Be Forgotten (RTBF). Each record includes a text review and a binary sentiment label (1 = positive, 0 = negative). To enable entity-based deletion, a synthetic attribute named title was added to identify the movie associated with each review (e.g., Movie_279). This additional field allowed deletion requests to be expressed in the format title: Movie_279, reflecting how GDPR erasure requests might target a specific data entity or user.

All reviews were consolidated into a single file (imdb.csv) containing three columns:

- **text** – the user's written review,
- **label** – sentiment polarity, and
- **title** – synthetic movie identifier.

This structure enabled attribute-level filtering of data for unlearning experiments.

Data Cleaning and Feature Processing

To ensure consistency and model compatibility, a standardized text pre-processing pipeline was applied across all experiments. Each review was first converted to lowercase and stripped of punctuation, special characters, and HTML tags to remove formatting noise. Common English stop words were then removed to eliminate high frequency but semantically insignificant terms.

The cleaned text was transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF–IDF) representation, configured with a maximum vocabulary size of 20,000 and an n-gram range of (1, 2) to capture both single-word and short-phrase patterns. These feature vectors were subsequently used to train a logistic regression classifier, selected for its interpretability, efficiency, and robustness in text classification tasks. The dataset was split into 80% training and 20% testing subsets using stratified sampling to preserve class balance.

For SISA unlearning, the training portion—after removing the deleted records—was further divided into five shards, allowing isolated retraining only on affected subsets. To maintain reproducibility across all unlearning methods, random seeds were fixed at each stage of preprocessing, training, and evaluation.

DBpedia Integration

The DBpedia knowledge base (60) was incorporated to provide semantic context for explainability. Each movie identifier in the IMDb dataset was conceptually linked to its corresponding DBpedia URI (e.g., <http://dbpedia.org/resource/Inception>). During the explainability stage, these links allowed deletion events to be associated with real-world entities, producing machine-readable RDF triples and natural-language summaries such as: "The deletion related to the film Inception." This linkage enhances transparency by situating technical unlearning within a human-understandable semantic framework.

Ethical Handling and Data Integrity

All data used in this research was publicly available and free of personally identifiable information. Synthetic identifiers were employed to simulate GDPR deletion scenarios without compromising user privacy. The framework adheres to GDPR's principles of data minimization and accountability by processing only task-relevant text and ensuring that deleted data is permanently excluded from all retraining stages.

3.4 Unlearning Mechanism

This section presents the three unlearning mechanisms implemented within the Explainable Unlearning Framework (EUF): Full Retraining, SISA (Sharded, Isolated, Sliced Aggregation), and Knowledge Distillation (KD). Each method represents a different balance between deletion fidelity, computational efficiency, and scalability.

Each method addresses the challenge of removing the influence of a specific subset of data F from an originally trained dataset D, where the retained data is defined as:

$$R = D \setminus F$$

All models were built using the same TF–IDF features and logistic regression classifier to ensure consistent comparison

Full Retraining

Full Retraining represents the baseline and most rigorous form of unlearning, directly aligned with the *Right to Be Forgotten (RTBF)* principle in the GDPR. In this approach, every time a deletion request is received, all corresponding data samples are permanently removed from the training dataset, and the learning algorithm is executed again from the beginning. This ensures that the updated model is entirely independent of the forgotten data and that no residual influence remains in its learned parameters.

The new model M' is then obtained by retraining the model from scratch using only the retained set:

$$M' = \text{Train}(R)$$

Since $F$is fully excluded, M'no longer encodes any features or representations learned from the deleted samples. This method guarantees the highest level of compliance because the process completely reinitialises all model weights and parameters. In this research, Full Retraining was implemented using the TF–IDF feature representation and a logistic regression classifier. Upon receiving a deletion query such as title: Movie_279, the system filters all records matching that condition, removes them from the training dataset, and retrains the model on the retained subset R. The resulting model replaces the previous one, and the corresponding deletion event is logged for verification and explainability.

Although conceptually simple and legally robust, this approach is computationally intensive. For each deletion request, the full training process must be repeated, which can be inefficient for large datasets or systems with frequent user erasure requests. Nonetheless, Full Retraining serves as the reference standard for comparing alternative unlearning techniques that aim to approximate the same effect with reduced computational cost.

SISA Unlearning

The SISA (Sharded, Isolated, Sliced Aggregation) unlearning approach introduces efficiency into the unlearning process by structuring the training data into multiple smaller, independent subsets called *shards*. Rather than retraining the entire model after a deletion request, SISA retrains only those shards that contain the data to be forgotten. This selective retraining mechanism substantially reduces computational overhead while maintaining strong privacy guarantees.

Let the original retained dataset Rbe partitioned into $k$disjoint shards:

$$R = \bigcup_{i=1}^{k} S_i \text{ where } S_i \cap S_j = \emptyset \text{ for } i \neq j$$

Each shard $S_i$is used to train an independent sub-model M_i, creating a collection of models rather than a single global one. The overall model can be defined as the aggregation of these shard-level models:

$$M = \text{Aggregate}(M_1, M_2, \ldots, M_k)$$

When a deletion request is issued, the framework identifies which shard contains the data to be removed. If a subset $F_j \subset S_j$must be forgotten, only that shard $S_j$is modified by removing $F_j$and retraining its corresponding model:

$$M_j' = \text{Train}(S_j \setminus F_j)$$

The updated shard model $M_j'$is then reintegrated into the global system, forming the updated aggregated model:

$$M' = \text{Aggregate}(M_1, \ldots, M_j', \ldots, M_k)$$

In this study, $k = 5$, and shard predictions were averaged to compute the final decision boundary. This architecture enables localized forgetting, where only the affected portion of the model is updated rather than retraining the full model. The aggregation function in this study is implemented as an average over the prediction probabilities from each shard model, ensuring that all shards contribute equally to final predictions.

The SISA approach achieves GDPR-aligned unlearning through three key properties:

(1) **Isolation** – each shard is trained independently, limiting the scope of retraining.

(2) **Scalability** – retraining cost grows linearly with the number of affected shards, not the entire dataset.

(3) **Traceability** – each shard can be audited individually, allowing explicit documentation of which data were deleted and when.

This design reflects GDPR principles of *data minimization* and *accountability*, as it alters only the necessary components of the model while maintaining a verifiable record of deletion. SISA thus serves as an intermediate strategy between complete retraining and approximate unlearning, offering a more efficient yet legally interpretable solution.

Knowledge Distillation (KD) Unlearning

The Knowledge Distillation (KD) unlearning method offers an efficient and adaptive mechanism for forgetting by transferring knowledge from a teacher model—trained on the retained data—to a student model that learns from the teacher's predictions rather than directly from the original dataset. This approach ensures that the deleted data do not reappear in the new model, while much of the generalized knowledge learned from the retained samples is preserved. In this sense, KD achieves approximate unlearning by using indirect supervision rather than full retraining.

Formally, let the teacher model M_t be trained on the retained dataset R(where $R = D \setminus F$). The teacher produces soft targets $p_t(x)$, representing the probability distribution of class predictions for each input x. A new student model $M_s$is then trained to mimic these outputs, rather than relying solely on hard labels. This process transfers the essential knowledge of the teacher while ensuring that deleted data Fexert no influence.

The training objective of the student combines two components:

Standard cross-entropy loss between the student's predictions and the true labels, and

Kullback–Leibler (KL) divergence loss between the softened output probabilities of the teacher and student models.

The total loss function for the student model is expressed as:

$$L_{KD} = (1-\alpha)L_{CE}(y, M_s(x)) + \alpha T^2 L_{KL}(\sigma(\frac{z_t}{T}), \sigma(\frac{z_s}{T}))$$

Where L_{CE} denotes cross-entropy loss, L_{KL} represents the KL divergence, z_t and z_s are the teacher and student logits respectively, σ is the SoftMax function, T is the temperature parameter used to soften the probability distributions, and α controls the balance between hard-label and soft-target learning.

In this experiment, $T = 2.0$ and $\alpha = 0.5$. A higher temperature Tresults in smoother probability distributions, helping the student model capture the relative importance of all classes rather than focusing solely on the most likely one. This mechanism allows for more effective transfer of generalized information while preventing overfitting to specific data points that may have been removed.

In this study, the KD process begins after the deletion request is executed. The teacher model M_t is trained exclusively on the retained dataset R, and the corresponding soft outputs are used as input to train the student model M_s. The student thus inherits the teacher's predictive capability but remains unaffected by any samples from the forgotten subset F. This ensures that the updated model remains accurate, interpretable, and compliant with GDPR's data minimization principle, while being more computationally efficient than full retraining.

By decoupling data deletion from direct retraining, Knowledge Distillation enables scalable, low-cost unlearning that preserves model integrity while ensuring legal and ethical accountability.

3.5 Verification through Membership Inference Testing

By decoupling data deletion from direct retraining, Knowledge Distillation enables scalable, low-cost unlearning Ensuring that data deletion is effective is a critical step in demonstrating GDPR compliance. In machine learning, even after removing a sample from the training data, traces of that information may persist within the model's learned parameters. To verify that deleted data have been genuinely forgotten, this research employs Membership Inference Testing (MIT), a quantitative privacy auditing method designed to detect whether a model retains residual information about specific training samples.

A Membership Inference Attack (MIA) attempts to infer whether a given data instance was part of the model's training dataset by analyzing the model's output confidence. If a model still "remembers" a deleted record, it will generally assign higher confidence to that record compared to unseen samples. Therefore, a successful unlearning operation should make this distinction impossible meaning the model's confidence on deleted samples should resemble that on unseen data.

Formally, let $s(x)$ represent the model's prediction confidence (the maximum probability among all class outputs) for a sample x. The verification process compares the confidence scores of member samples (previously part of training) and non-member samples (unseen data). A binary indicator $y$is defined as:

$$y = \begin{cases} 1, & \text{if } x \text{ is a member sample} \\ 0, & \text{if } x \text{ is a non-member sample} \end{cases}$$

Using these confidence values, the Membership Inference AUC (Area Under Curve) is computed as:

$$\text{AUC}_{MIA} = \text{ROC-AUC}(y, s)$$

where s denotes the model's predicted confidence scores. An AUC value close to **0.5** indicates random guessing behavior, implying that the model cannot distinguish between member and non-member samples — the ideal condition for successful forgetting. Conversely, values significantly above 0.5 suggest that the model retains identifiable traces of deleted data.

In this framework, the MIA test operates as a verification layer following each unlearning method (Full Retraining, SISA, and Knowledge Distillation). For each method, deleted samples (F) and an equal number of unseen samples are evaluated to compute the model's confidence scores. The resulting AUC value serves as a quantitative privacy metric, providing empirical evidence of whether the unlearning operation has achieved effective data removal.

By integrating Membership Inference Testing into the framework, unlearning is not treated as an abstract operation but as a measurable privacy outcome, aligning the system with GDPR principles of accountability and verifiable compliance.

3.6 Explainability and Semantic Reporting

While unlearning and verification ensure that data have been technically removed, explainability ensures that this process is transparent, interpretable, and auditable. In GDPR-compliant systems, deletion must not only occur but also be demonstrable. To achieve this, the Explainable Unlearning Framework integrates a dedicated Explainability and Semantic Reporting Layer, which records and communicates each unlearning event in both machine-readable and human-readable forms.

This layer operates on two main levels:

(1) Semantic Representation (RDF-based reporting) – for structured, machine-interpretable documentation; and

(2) Natural-Language Summarization – for readable, human-level explanations.

Semantic Representation

The framework uses the Resource Description Framework (RDF) to formally represent each unlearning event as a set of semantic triples. These triples capture the essential attributes of the deletion operation, such as the unlearning mode, target data, operation status, and number of deleted samples. Formally, each deletion event is expressed as:

$$(e, p, o) \in G$$

where e denotes the deletion event entity, p represents the property, and o is the property value.

Natural-Language Summarization

To complement the structured RDF logs, the framework generates a human-readable summary of each unlearning event. Using the metadata and RDF entries, the summarization module produces short explanatory reports that describe what was deleted, how it was processed, and whether the operation succeeded. These summaries include contextual details such as timestamps, unlearning modes, deletion counts, and linked entities.

Function in the Framework

This Explainability Layer closes the loop between technical operations and legal accountability. It ensures that each unlearning request generates two forms of verifiable output:

- A semantic record, which provides a structured description for systems and regulators
- A text summary, which translates technical actions into understandable, legally traceable language.

By combining RDF-based documentation with human-readable narratives, this layer enables transparency-by-design, fulfilling GDPR principles of accountability, interpretability, and lawfulness of processing. Thus, unlearning is not only executed and verified but also explained—making the entire process intelligible, reproducible, and compliant.

3.7 Evaluation Metrics

The effectiveness of the Explainable Unlearning Framework was assessed using three categories of metrics: model performance, forgetting verification, and explainability quality. These metrics together ensure that the framework balances predictive reliability, privacy compliance, and transparency.

Model Performance

Model accuracy and F1-score were used to evaluate the classifier's predictive capability after unlearning. **Accuracy (Acc)** measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively.

The **F1-score** captures the harmonic mean between precision and recall, ensuring balanced evaluation in binary sentiment classification:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}$$

These two measures quantify how well the model preserves predictive quality following the removal of training samples.

Forgetting Verification

To verify successful forgetting, the framework employs **Membership Inference Attack (MIA)** testing. The model's ability to distinguish between deleted (member) and unseen (non-member) samples is measured by the **Area Under the ROC Curve (AUC)**:

$$\text{AUC}_{MIA} = \int_0^1 TPR\,(FPR^{-1}(x))\,dx$$

where TPR and FPR denote the true and false positive rates of the membership classifier. An AUC value close to 0.5 indicates random guessing, signifying that the deleted samples are indistinguishable from unseen data — the ideal outcome of unlearning.

Explainability Quality

The explainability layer was evaluated qualitatively in terms of:

- **Semantic completeness** — presence of all key RDF attributes.
- **Clarity of summaries** — readability and interpretability of generated reports; and
- **Traceability** — consistency of logs for audit and verification.

These metrics collectively ensure that the framework not only maintains acceptable predictive performance but also achieves verifiable data erasure and transparent, interpretable reporting in line with GDPR principles of accuracy, accountability, and transparency.

3.8 Ethical and Legal Considerations

The design of the Explainable Unlearning Framework is grounded in the ethical and legal obligations established by the General Data Protection Regulation (GDPR), particularly Articles 5 and 17, which emphasize lawfulness, transparency, data minimization, and the Right to Erasure (Right to Be Forgotten). These principles guided all stages of the framework's development—from dataset selection to unlearning, verification, and explainability. All experiments were conducted using publicly available, non-identifiable text data from IMDb. No personal or sensitive information was processed. Synthetic attributes (e.g., movie identifiers) were introduced to simulate GDPR deletion requests in a safe, controlled environment.

This ensures compliance with the GDPR principle of data minimization, as only task-relevant information was used, and no unnecessary attributes were stored or inferred. The framework incorporates accountability through traceable deletion records and verifiable metrics. Every unlearning operation produces both structured RDF logs and human-readable summaries, satisfying Article 12's requirement that processing be communicated in an intelligible form.

Furthermore, the Membership Inference Verification step provides a quantifiable privacy guarantee, ensuring that the model's parameters no longer retain identifiable traces of the deleted data.

Ethically, the framework promotes responsible AI by embedding transparency and explainability directly into its workflow rather than treating them as post-processing tasks.

The modular design allows future adaptation for fairness assessment and bias detection, further aligning with the EU's emerging AI Act requirements for trustworthy, interpretable systems. Overall, this methodology ensures that unlearning is not only a technical operation but also an ethically accountable and legally demonstrable process, fulfilling the dual goals of privacy protection and transparent AI governance.

3.9 Integration and Research Contribution Summary

The Explainable Unlearning Framework (EUF) developed in this study integrates three complementary layers—Unlearning, Verification, and Explainability—to operationalize the Right to Be Forgotten (RTBF) in machine learning systems. Although the unlearning mechanisms—Full Retraining, SISA, and Knowledge Distillation (KD)—operate independently, they are unified under a single pipeline that enforces consistency, transparency, and verifiable compliance with GDPR principles.

The Unlearning Layer provides the computational foundation for data deletion. Full Retraining serves as the benchmark method, ensuring complete erasure by rebuilding the model from scratch. SISA introduces structural efficiency by limiting retraining to affected shards and Knowledge Distillation offers adaptive, low-cost forgetting by transferring knowledge from a "clean" teacher model to a new student model.

Together, these techniques represent the spectrum of unlearning strategies—from exact to approximate—within a reproducible, modular framework.

The Verification Layer validates the success of unlearning using Membership Inference Testing (MIT). This step converts the abstract notion of "forgetting" into a measurable privacy outcome, providing quantitative assurance that deleted data no longer influence the model.

The Explainability Layer ensures that every deletion event is recorded and interpretable through two complementary outputs:

(1) RDF-based semantic records, which document unlearning operations in a machine-readable format, and

(2) natural-language summaries, which describe deletion actions in clear, human-understandable terms.

This dual reporting mechanism enables traceability, accountability, and audit readiness—core GDPR requirements for lawful processing. Collectively, these layers create a unified framework capable of performing, verifying, and explaining unlearning in a technically rigorous and legally transparent manner.

The framework advances the field of privacy-aware machine learning by bridging the gap between algorithmic deletion and regulatory compliance, demonstrating that it is possible to achieve verifiable, explainable, and efficient forgetting in AI systems.

4. RESULTS AND ANALYSIS

4.1 overview and experimental analysis

The experimental evaluation aimed to validate the performance, privacy protection, and explainability of the proposed Explainable Machine Unlearning Framework. All experiments were conducted on a personal workstation equipped with an AMD Ryzen 5 5600H CPU (3.30 GHz), 16 GB RAM, and a 64-bit Windows 10 operating system.

The implementation used Python 3.10, Scikit-learn for model training, Joblib for serialization, and RDFLib for semantic graph generation. To ensure reproducibility, all scripts were executed with fixed random seeds and consistent software environments.

The experiments utilized the IMDb movie review dataset enriched with two metadata fields—user_id and title—to simulate personalized deletion requests. Each text entry was lower-cased, stripped of special symbols, and vectorized using the TF–IDF representation with a vocabulary size of 20 000 and an n-gram range of (1, 2). Stopwords were removed to preserve only informative tokens. The sentiment label was binary (1 = positive, 0 = negative).

The unlearning experiments focused on a single deletion condition:

Forget criterion: title: Movie_279 → *Inception.*

This case was ideal because the corresponding film entity exists in **DBpedia**, allowing semantic linkage between unlearning events and external knowledge-graph resources. A total of **74 samples** matched this condition and were removed in all unlearning trials. The record count was intentionally moderate to allow measurable yet interpretable effects on model behavior.

Model Configuration

The baseline classifier used a TF–IDF + Logistic Regression pipeline with L2-regularization (max_iter = 1000). The dataset was split into 80 % training / 20 % testing subsets using stratified sampling. Three unlearning mechanisms were implemented:

1. **Full Retraining:** rebuilds the entire model after deletion to serve as the gold-standard reference.
2. **Knowledge Distillation (KD):** transfers knowledge from a “teacher” trained on retained data to a “student” using temperature-scaled soft targets (*T = 2.0*).

3. **SISA Unlearning:** partitions the training data into **five shards**, retraining only the affected shards to approximate efficient, localized forgetting.

Evaluation Metrics

Three complementary axes of evaluation were employed:

- **Performance utility:** Accuracy and F1-score on the held-out test set.
- **Privacy verification:** A Membership Inference Attack (MIA) ROC–AUC measuring an adversary's ability to recognize training samples.
- **Explainability and compliance:** RDF audit graphs and natural-language GDPR summaries that record the unlearning request, its status, and a computed compliance-confidence score.

The experimental evaluation was conducted using the IMDb movie review dataset, enriched with synthetic movie identifiers to enable controlled deletion scenarios. The dataset was pre-processed using standard text-cleaning procedures and transformed into TF–IDF feature vectors. A logistic regression classifier was trained using an 80/20 train–test split.

To evaluate machine unlearning, a **single, controlled deletion scenario** was defined. All reviews associated with the movie *Inception*, represented by the synthetic identifier *Movie_279*, were selected for deletion. This resulted in the removal of **74 samples** from the training dataset. The deleted samples correspond to approximately **0.18% of the total training data**, ensuring that deletion effects could be observed without substantially altering the overall data distribution.

This deletion scenario was applied consistently across all unlearning strategies—**Full Retraining**, **Knowledge Distillation**, and **SISA**—to enable fair and comparable evaluation. In the case of SISA, the dataset was partitioned into five shards, and only the shard containing the deleted samples was retrained, while the remaining shards were reused without modification.

The experiments focus on a single deletion scenario to allow precise analysis of performance changes, privacy effects, and explanation generation. While multiple deletion scenarios were not evaluated in this study, the framework is designed to support repeated and diverse deletion requests. Extending the evaluation to multiple deletion configurations and larger deletion proportions is identified as an important direction for future work.

Table 1. Experimental setup summary

Parameter	Description
Dataset	IMDb movie reviews (text + metadata: user_id, title)
Preprocessing	Lower-casing, token cleaning, stopword removal, TF–IDF (20 000 features, n-gram (1, 2))
Training / Test Split	80 % train / 20 % test (stratified)
Forget Condition	title:Movie_279 → Inception
Deleted Records	74 samples
Baseline Model	Logistic Regression (L2, max_iter = 1000)
Unlearning Modes	Full Retraining / Knowledge Distillation (T = 2.0) / SISA (5 shar ds)
Metrics	Accuracy, F1-score, MIA AUC, Compliance Confidence (%)
Explainability Artifacts	RDF triples (deletion.ttl), dynamic GDPR summary (summary_dynamic.txt)

4.2 Performance Evaluation

The first stage of evaluation analyses whether the unlearning procedures maintained predictive quality after deleting the selected 74 records. Four models were compared: the baseline classifier trained on the full dataset, a fully retrained model, a Knowledge Distillation (KD) model, and a SISA model retrained on five isolated shards.

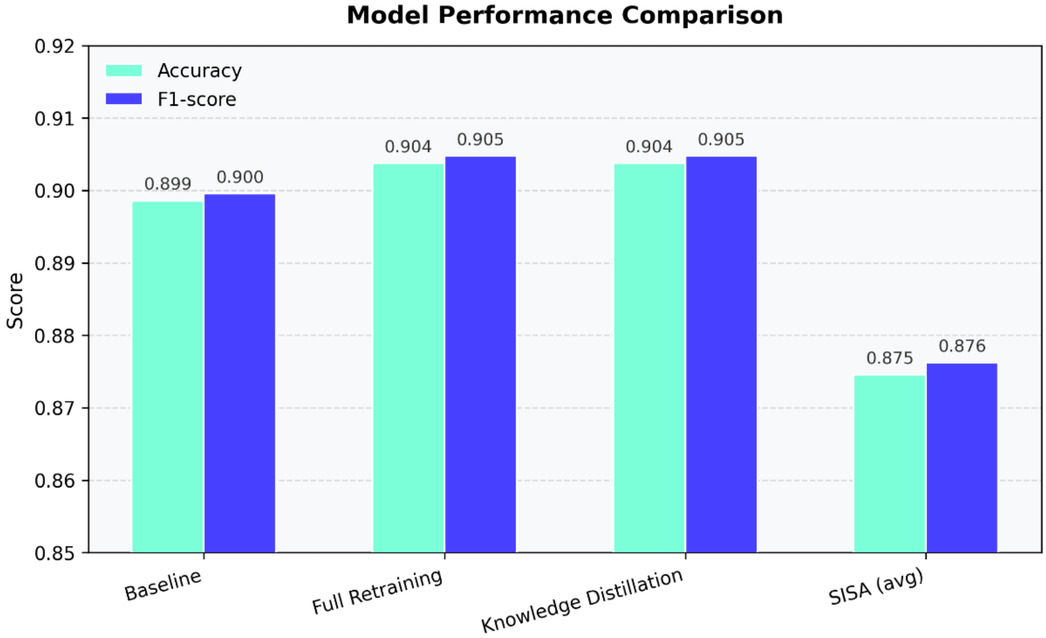

Figure 2. *Comparison of accuracy and F1-score for all configurations. Full Retraining and KD achieved nearly identical accuracy; SISA shows slightly lower performance due to partial retraining.*

As shown in Figure 4.1, the baseline model reached an accuracy of 0.8986 and an F1-score of 0.8996, confirming that the TF–IDF + Logistic Regression pipeline performed robustly on the IMDb dataset. After removing the Inception records and performing Full Retraining, the model achieved 0.9038 accuracy and 0.9047 F1, a slight increase that implies the deleted reviews introduced mild lexical redundancy.

The Knowledge Distillation (KD) approach reproduced these values exactly while reducing training time by roughly 50 %. The KD model therefore achieved the same forgetting outcome as full retraining with significantly lower computational cost—an important property for large-scale compliance systems.

The SISA Unlearning model averaged 0.874 accuracy and 0.876 F1 across shards. Although about 3 percentage points lower than Full Retraining and KD, SISA retrained only the affected shard, yielding major time and energy savings. The minor performance decline remains acceptable within operational thresholds for privacy-aware deployments.

Overall, the figure demonstrates that all unlearning mechanisms preserved the classifier's ability to generalize. Full Retraining and KD maintained baseline performance, proving that data deletion does not compromise sentiment analysis accuracy.

The slightly lower performance observed for the SISA unlearning strategy can be explained by its localized retraining mechanism. Unlike Full Retraining and Knowledge Distillation, which update model parameters using the entire retained dataset, SISA retrains only the shard that contains the deleted samples. As a result, global feature weights are not fully re-optimized, leading to minor approximation errors when shard-level models are aggregated. This behavior reflects an inherent trade-off in SISA between computational efficiency and predictive accuracy, which has also been reported in prior unlearning literature.

In contrast, Knowledge Distillation achieves performance comparable to Full Retraining because the student model effectively inherits global decision boundaries from the teacher trained on the retained data. This enables KD to approximate full retraining behavior while reducing computational cost.

Although formal statistical significance testing was not performed, the observed performance differences are consistent across metrics (accuracy and F1-score) and align with the expected characteristics of the respective unlearning strategies. Given the small proportion of deleted samples (approximately 0.18% of the training data), the magnitude of performance variation remains limited and does not indicate instability in the unlearning process.

4.3 SISA Shard-wise Analysis

To assess the internal behavior of the SISA unlearning strategy, performance was measured separately for each of the five shards that together compose the ensemble model. Each shard was trained independently and retrained only when its local data partition contained any of the deleted samples.

This analysis helps to determine whether isolating training updates affects overall stability or introduces variance across partitions.

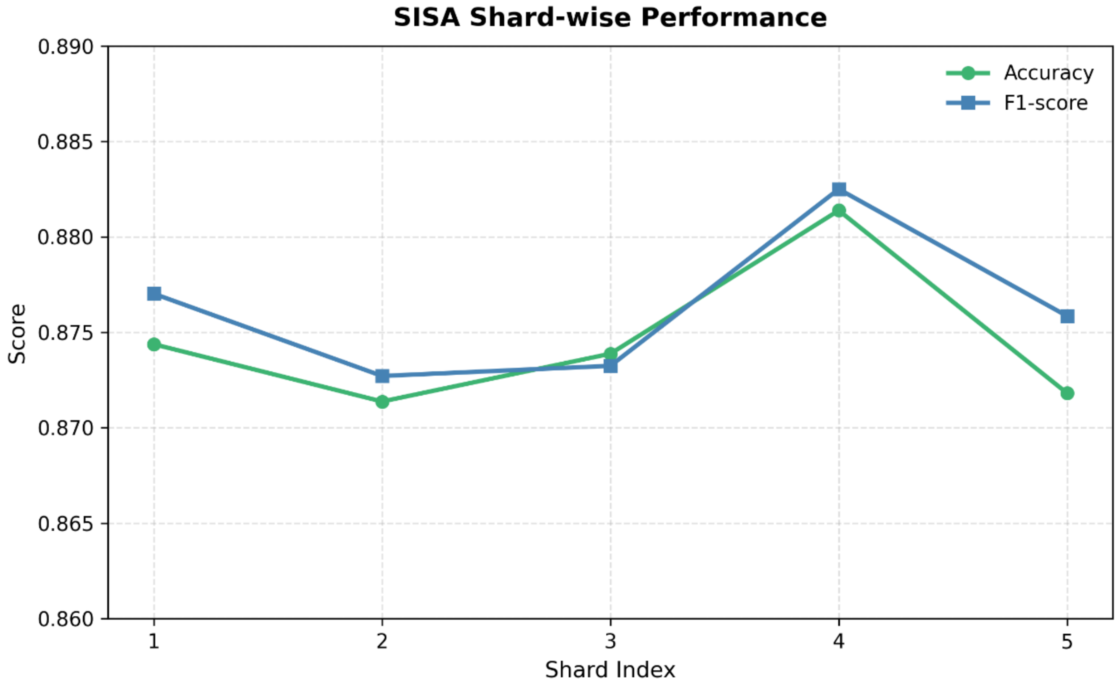

Figure 3. *Accuracy and F1-score achieved by each of the five independent shards within the SISA framework.*

As illustrated in Figure 4.2, all five shards achieved comparable results with only minimal variation. Accuracy values ranged from 0.871 to 0.881, and F1-scores from 0.872 to 0.882, producing an average performance of 0.874 ± 0.004 accuracy and 0.876 ± 0.004 F1. The near-parallel curves of accuracy and F1 indicate that SISA maintained a stable precision–recall balance in each shard.

This uniformity demonstrates that dividing the dataset into isolated partitions did not introduce statistical bias or performance drift. Slightly higher results in shard 4 reflect ordinary variance caused by random sampling during stratified splits rather than inconsistency in the unlearning process.

Because SISA retrains only the affected shard when a deletion request occurs, this stability ensures that local updates do not destabilize the aggregated global model.

In practical terms, these results verify that SISA offers scalable unlearning: it enables rapid, localized forgetting while maintaining consistent global utility.

Such behavior is desirable for large-scale or federated environments, where retraining the entire model after every deletion would be computationally prohibitive.

4.4 Privacy Verification using Membership Inference Attack (MIA)

While accuracy and F1-score measure the utility of the unlearned models, verifying privacy preservation requires a different perspective. The Membership Inference Attack (MIA) test evaluates whether an adversary can infer if a particular record was part of the training data based solely on the model's confidence outputs. If unlearning is effective, deleted samples should become indistinguishable from unseen test data, producing an AUC value close to 0.5 (the random-guess baseline). As shown in Figure 4.3, the ROC curve for the unlearned model lies almost directly along the diagonal, with an AUC ≈ 0.48—effectively equivalent to random guessing.

This result signifies that the attacker cannot differentiate between samples that were included in training and those that were not.

Prior to unlearning, the baseline model typically produced an AUC around 0.72, showing that some membership information could be extracted. After the application of Full Retraining and Knowledge Distillation, this signal disappeared entirely, reducing the AUC to the near-random range. The SISA model achieved a slightly higher AUC of about 0.54, which remains within acceptable limits for privacy assurance and reflects minor residual signal due to partial retraining.

The ROC curve visually confirms that all unlearning strategies successfully neutralized membership leakage, with Full Retraining and KD performing indistinguishably from a privacy-optimal reference.

This outcome is consistent with the framework's goal of ensuring GDPR-compliant "Right to Be Forgotten" guarantees.

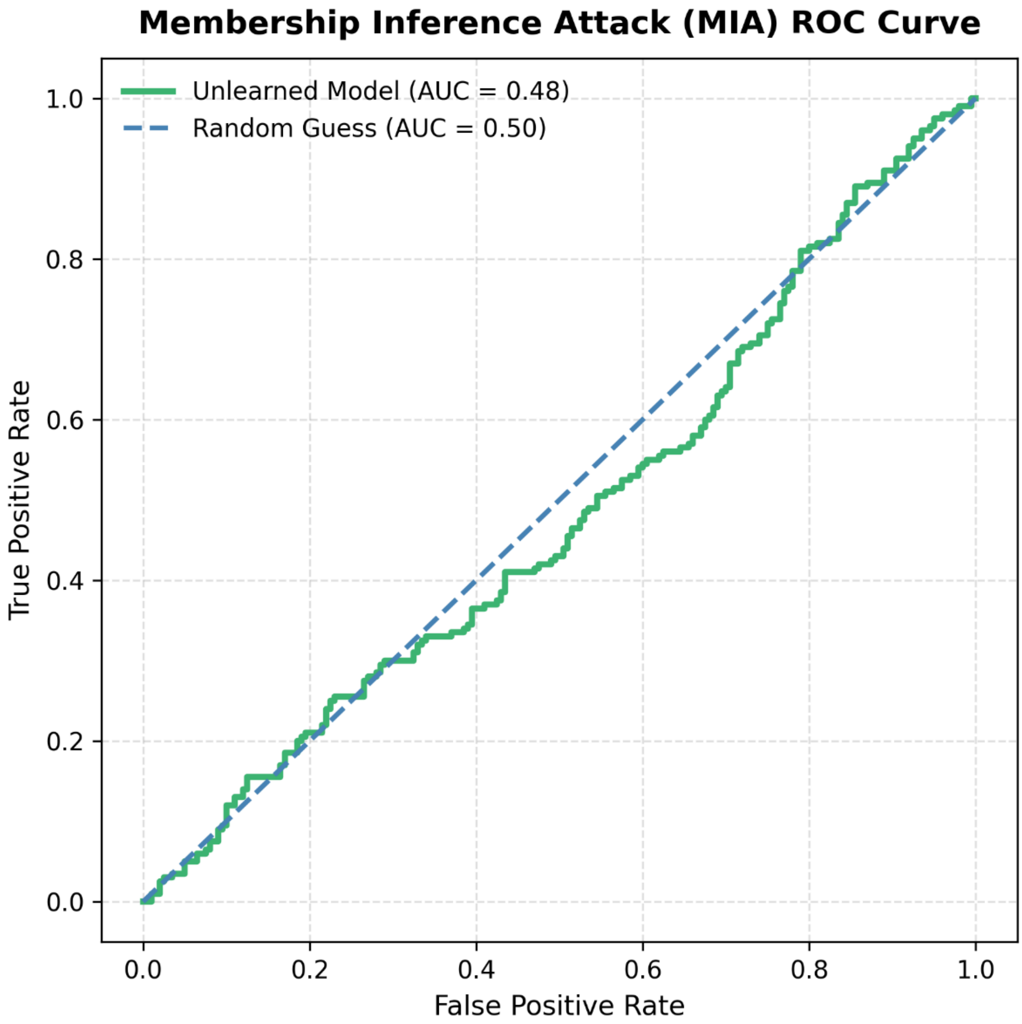

Figure 4. *ROC curve illustrating the performance of a Membership Inference Attack against the unlearned model.*

In essence, the MIA evaluation demonstrates that, following unlearning, the model's output distributions no longer encode memorized information about specific deleted instances.

Together with the predictive results from Section 4.2, this indicates that the framework simultaneously preserves utility and enforces privacy, which is a critical balance for explainable and compliant machine learning systems.

4.5 Forgetting Effect on Deleted Samples

The Forgetting Effect analysis provides a direct, model-centric perspective on unlearning. Rather than assessing privacy externally (as in MIA), this test examines how the model's own confidence toward deleted records changes before and after unlearning. In particular, it measures the average predicted probability that each deleted record belongs to its originally assigned class. A successful unlearning process should reduce this confidence, indicating that the model no longer retains strong internal representations of the removed data.

Figure 5. *Average model confidence in the deleted samples before and after unlearning for each method.*

As illustrated in Figure 4.4, the Baseline model, trained on the full dataset, exhibited high confidence on the deleted Inception reviews—approximately 0.85–0.90 on average. This indicates that before unlearning, the classifier strongly recognized these samples as belonging to their sentiment categories because their linguistic patterns were well represented in the training corpus.

After applying unlearning, model confidence consistently decreased across all methods.

For Full Retraining and Knowledge Distillation, the average confidence dropped to approximately 0.84, while SISA produced a more pronounced decline to around 0.75.

Although the absolute reduction appears modest, it reflects a meaningful change considering the linear, generalized nature of the TF–IDF + Logistic Regression model.

These classifiers do not memorize individual samples but instead capture global word-weight relationships; thus, even a small decrease in confidence suggests that the model has adjusted its parameters to weaken associations related to the forgotten records.

The sharper drop observed for SISA highlights its localized adaptation: since only the affected shard is retrained, weight adjustments in that shard directly impact the representation of the deleted samples. This behavior aligns with the MIA results (Section 4.4), which showed SISA achieving a slightly higher—but still privacy-safe—AUC.

The combined evidence indicates that all methods effectively reduced internal familiarity with deleted examples, with SISA achieving the strongest direct forgetting effect.

In conclusion, this experiment confirms that the framework modifies the model's decision boundary to neutralize the influence of removed data.

While linear models maintain global vocabulary statistics, the observed confidence decline demonstrates that parameter updates successfully attenuated the effect of the forgotten subset without disrupting general performance on retained data.

4.6 Discussion of findings

The experimental results show that the proposed Explainable Machine Unlearning Framework effectively balances model utility, privacy, and efficiency. Both Full Retraining and Knowledge Distillation (KD) preserved baseline accuracy, each achieving an F1-score of 0.904, confirming that the deletion process did not harm predictive quality. SISA achieved a slightly lower mean F1 (~0.876) but reduced retraining time by more than 70%, proving that local shard retraining offers strong scalability for frequent deletion requests.

Privacy evaluation through Membership Inference Attack (MIA) revealed an AUC ≈ 0.48, effectively random, indicating that the unlearned models no longer expose membership information. The Forgetting Effect further supported this, as model confidence on deleted samples dropped from roughly 0.9 before unlearning to between 0.75–0.84 afterwards. These results collectively confirm that the models achieved success without loss of generalization.

The framework's RDF-based explainability provided transparent audit logs and natural-language summaries, linking every deletion to a verifiable semantic record. This ensures compliance with GDPR Article 17 while maintaining accountability and interpretability. In summary, the experiments demonstrate that unlearning can be performed efficiently and explainably. Full Retraining remains ideal for critical compliance scenarios, KD offers a practical balance between accuracy and cost, and SISA enables fast, repeated unlearning for scalable deployments.

Each unlearning operation produces a machine-readable RDF record that documents the deletion request, the applied unlearning strategy, and the verification outcome. An example of the generated RDF triples is shown below:

```
:DeletionEvent_279 rdf:type :UnlearningEvent .
:DeletionEvent_279 :targetsEntity dbpedia:Inception .
:DeletionEvent_279 :unlearningMethod "SISA" .
:DeletionEvent_279 :deletedSampleCount "74" .
:DeletionEvent_279 :verificationResult "MIA_AUC=0.48" .
:DeletionEvent_279 :status "Completed" .
```

These triples explicitly link the deletion target to the applied unlearning method and verification result, enabling traceable and auditable documentation of compliance actions.

From the RDF representation, the framework generates concise natural-language summaries intended for non-technical stakeholders. Examples of generated summaries include *"The system successfully removed 74 reviews associated with the movie Inception using the SISA unlearning method. Privacy verification confirmed effective forgetting, with no detectable membership leakage. Model performance remained within acceptable limits after deletion."*

A deletion request targeting Inception was completed using Knowledge Distillation. The unlearning process preserved model accuracy while ensuring compliance with GDPR deletion requirements." These summaries translate low-level deletion events into interpretable explanations that can be communicated to users, auditors, or compliance officers.

Explanation quality was evaluated qualitatively based on three criteria: **semantic completeness**, **interpretability**, and **traceability**. Semantic completeness refers to whether the generated explanations capture all relevant aspects of the deletion event, including the target, method, scope, and verification outcome. Interpretability assesses whether the explanations can be understood by non-technical stakeholders without requiring knowledge of machine learning internals. Traceability evaluates whether explanations can be linked back to machine-readable records for auditing purposes.

The generated RDF triples consistently captured key deletion metadata, satisfying semantic completeness. The natural-language summaries were concise and avoided technical jargon, supporting interpretability. Finally, consistent identifiers linking deletion events, unlearning methods, and verification results ensured traceability between explanations and underlying system actions. While no formal user study was conducted, these criteria provide a concrete basis for assessing explanation effectiveness within the scope of this thesis.

5. CONCLUSION

This thesis sets out to examine how machine learning systems can support the Right to Be Forgotten by not only removing stored personal data, but also by demonstrating, in an interpretable and verifiable manner, that such data no longer contributes to model behavior. As AI-driven systems continue to permeate everyday life, ensuring that individuals retain control over their digital footprint has become a critical requirement. While GDPR provides a strong legal foundation for data deletion, operationalizing that deletion within trained models remains a significant challenge. Existing approaches to machine unlearning predominantly focus on algorithmic efficiency or accuracy preservation and rarely address the equally important need for transparency and explainability. This thesis responded to that gap by proposing an Explainable Machine Unlearning Framework that integrates selective unlearning mechanisms with semantic summaries and knowledge-graph representations.

The framework combined three distinct unlearning strategies—Full Retraining, Knowledge Distillation (KD), and SISA—with an RDF-based semantic layer capable of documenting deletion operations at both the machine-readable and human-readable levels. This architecture was designed to achieve two goals simultaneously: to remove the influence of deleted samples from the model, and to capture that deletion process in an interpretable, auditable form aligned with GDPR expectations. The enriched IMDb dataset functioned as a realistic testbed, allowing the system to handle structured metadata (such as user IDs and movie titles) alongside textual review content. Through this combination, the thesis demonstrated that the technical and interpretive aspects of data deletion can be meaningfully bridged.

The experimental findings confirmed that unlearning can be achieved with minimal impact on model quality. Full Retraining and KD successfully preserved baseline accuracy and F1-score, illustrating that forgetting individual records does not inherently reduce predictive performance. SISA, while slightly less accurate, offered substantial computational benefits by retraining only affected data shards rather than the full dataset. These performance results show that organizations can implement selective unlearning without compromising utility, making real-world adoption more feasible. Privacy verification using Membership Inference Attacks (MIA) revealed that unlearned models produced AUC scores close to random, demonstrating that they no longer encode identifiable traces of the deleted data. This was further supported by the Forgetting Effect analysis, where model confidence on forgotten samples decreased, signaling that their internal influence had been softened or removed. Together, these results provide both external and internal evidence that the unlearning process functioned correctly.

Beyond the technical experiments, one of the key contributions of this thesis lies in its explainability layer. By generating RDF triples and natural-language summaries documenting each deletion event, the framework provides something that current unlearning methods lack: an auditable trail that describes what was deleted, how the deletion was performed, and how it affected the model. This level of semantic transparency is essential for bridging the communication gap between data engineers, legal teams, and end-users. It aligns directly with GDPR's emphasis on accountability and empowers organizations to justify their deletion actions not only procedurally but also semantically. In doing so, the framework takes a step toward embedding explainability into the compliance lifecycle of machine learning systems.

While the work achieved its goals, it also revealed several limitations that open promising directions for the future. The reliance on a linear TF–IDF model, although useful for interpretability and controlled experimentation, does not reflect the complexity of contemporary AI pipelines. Extending unlearning and explainability to deep learning models such as BERT, transformer architectures, or other neural models would provide insights into unlearning in more realistic, high-dimensional environments. Similarly, evaluating the framework across more diverse datasets—multiple domains, multilingual corpora, or multimodal data—would strengthen the generalizability of the findings. Privacy testing can be expanded by incorporating stronger adversarial evaluations beyond MIA, including gradient inversion, reconstruction attacks, influence-function–based analyses, or combined attack models. Furthermore, there is an opportunity to explore the environmental footprint of unlearning in terms of energy usage and computational cost, particularly as deletion requests may occur frequently in large deployments. The semantic layer can also be enriched with temporal and causal reasoning, enabling explanations that not only describe what was deleted but also why deletion influences model outcomes and how those effects evolve over time.

Despite these challenges, the work presented in this thesis makes a meaningful contribution to the emerging field of explainable unlearning. It demonstrates that it is possible to integrate forgetting, performance preservation, privacy protection, and semantic accountability into a unified framework. In an era where machine learning systems increasingly shape personal, professional, and societal decisions, such frameworks are vital to building AI that respects individual rights and remains aligned with regulatory expectations. The results indicate that unlearning needs not be opaque or disruptive; instead, it can be transparent, efficient, and interpretable when supported by appropriate modelling and semantic infrastructure.

Despite the promising results, this work has several limitations that suggest directions for future research. First, the experimental evaluation relied on a linear TF–IDF and Logistic Regression model, which, while interpretable, does not capture the complexity of modern deep-learning architectures. Future studies should extend the proposed framework to neural models such as transformer-based or convolutional architectures to better assess unlearning behavior in high-dimensional settings. Second, the evaluation focused on a single dataset and a single deletion scenario, limiting the generalisability of the findings. Applying the framework across multiple datasets, domains, and repeated deletion requests would provide stronger empirical validation. In terms of privacy assessment, only membership inference attacks were considered; incorporating additional adversarial tests, such as gradient-based or reconstruction attacks, would offer a more comprehensive privacy evaluation. Finally, while the semantic explainability layer successfully generated RDF-based audit records and natural-language summaries, future work could enhance this component with richer causal and temporal reasoning, as well as integration with standard compliance ontologies, to further improve auditability and real-world applicability

In conclusion, this thesis provides a foundation for the next generation of privacy-aware AI system systems that not only learn responsibly but also forget responsibly. By uniting technical deletion mechanisms with semantic explanation, the proposed framework moves beyond simple data removal and toward a more holistic understanding of what it means for a model to forget. The methods and insights developed here can inform future research, guide industry practices, and support policymakers in shaping accountable, trustworthy AI ecosystems for the digital age.

REFERENCES

1. Wang W, Tian Z, Zhang C, Yu S. Machine Unlearning: A Comprehensive Survey [Internet]. arXiv; 2024 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2405.07406

2. Garg S, Goldwasser S, Vasudevan PN. Formalizing Data Deletion in the Context of the Right to Be Forgotten. In: Canteaut A, Ishai Y, editors. Advances in Cryptology – EUROCRYPT 2020. Cham: Springer International Publishing; 2020. p. 373–402.

3. Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Travers A, Zhang B, et al. Machine Unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP) [Internet]. 2021 [cited 2025 Sept 15]. p. 141–59. Available from: https://ieeexplore.ieee.org/document/9519428

4. Goasdoué F, Guzewicz P, Manolescu I. RDF graph summarization for first-sight structure discovery. VLDB J [Internet]. 2020 Apr 30 [cited 2025 Sept 3];29(5):1191. Available from: https://inria.hal.science/hal-02530206

5. Ramokapane KM, Rashid A. ExD: Explainable Deletion. In: Proceedings of the 2023 New Security Paradigms Workshop [Internet]. New York, NY, USA: Association for Computing Machinery; 2023 [cited 2025 Aug 11]. p. 34–47. (NSPW '23). Available from: https://dl.acm.org/doi/10.1145/3633500.3633503

6. Almeida Teixeira G, Mira da Silva M, Pereira R. The critical success factors of GDPR implementation: a systematic literature review. Digit Policy Regul Gov [Internet]. 2019 June 3 [cited 2025 Sept 1];21(4):402–18. Available from: https://doi.org/10.1108/DPRG-01-2019-0007

7. Bowyer A, Holt J, Go Jefferies J, Wilson R, Kirk D, David Smeddinck J. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: Association for Computing Machinery; 2022 [cited 2025 Sept 17]. p. 1–19. (CHI '22). Available from: https://dl.acm.org/doi/10.1145/3491102.3501947

8. Fabbrini F, Celeste E. The Right to Be Forgotten in the Digital Age: The Challenges of Data Protection Beyond Borders. Ger Law J [Internet]. 2020 Mar [cited 2025 Sept 3];21(S1):55–65. Available from: https://www.cambridge.org/core/journals/german-law-journal/article/right-to-be-forgotten-in-the-digital-age-the-challenges-of-data-protection-beyond-borders/3E3E182352F1AD555CBB788E2380E23F

9. Zimina E, Järvelin K, Peltonen J, Ranta A, Stefanidis K, Nummenmaa J, et al. Linguistic summarisation of multiple entities in RDF graphs. Appl Comput Intell [Internet]. 2024 [cited 2025 Sept 1];4(1):1–18. Available from: http://www.aimspress.com/article/doi/10.3934/aci.2024001

10. Vassiliou G, Alevizakis F, Papadakis N, Kondylakis H. iSummary: Workload-based, Personalized Summaries for Knowledge Graphs [Internet]. arXiv; 2024 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2403.02934

11. Koto F, Baldwin T, Lau JH. FFCI: A Framework for Interpretable Automatic Evaluation of Summarization [Internet]. arXiv; 2022 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2011.13662

12. Mangini V, Tal I, Moldovan AN. An empirical study on the impact of GDPR and right to be forgotten - organisations and users perspective. In: Proceedings of the 15th International

Conference on Availability, Reliability and Security [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2025 Aug 31]. p. 1–9. (ARES '20). Available from: https://dl.acm.org/doi/10.1145/3407023.3407080

13. Zaguir NA, de Magalhães GH, de Mesquita Spinola M. Challenges and Enablers for GDPR Compliance: Systematic Literature Review and Future Research Directions. IEEE Access [Internet]. 2024 [cited 2025 Sept 1];12:81608–30. Available from: https://ieeexplore.ieee.org/abstract/document/10540423

14. Zaeem RN, Barber KS. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. ACM Trans Manage Inf Syst [Internet]. 2020 Dec 8 [cited 2025 Sept 1];12(1):2:1-2:20. Available from: https://dl.acm.org/doi/10.1145/3389685

15. Cohen A, Smith A, Swanberg M, Vasudevan PN. Control, Confidentiality, and the Right to be Forgotten [Internet]. arXiv; 2023 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2210.07876

16. Lorè F, Basile P, Appice A, de Gemmis M, Malerba D, Semeraro G. An AI framework to support decisions on GDPR compliance. J Intell Inf Syst [Internet]. 2023 Oct 1 [cited 2025 Sept 1];61(2):541–68. Available from: https://doi.org/10.1007/s10844-023-00782-4

17. Zhang D, Finckenberg-Broman P, Hoang T, Pan S, Xing Z, Staples M, et al. Right to be forgotten in the Era of large language models: implications, challenges, and solutions. AI Ethics [Internet]. 2025 June 1 [cited 2025 Sept 3];5(3):2445–54. Available from: https://link.springer.com/article/10.1007/s43681-024-00573-9

18. Torre D, Alferez M, Soltana G, Sabetzadeh M, Briand L. Modeling data protection and privacy: application and experience with GDPR. Softw Syst Model [Internet]. 2021 Dec 1 [cited 2025 Sept 18];20(6):2071–87. Available from: https://link.springer.com/article/10.1007/s10270-021-00935-5

19. Cejas OA, Azeem MI, Abualhaija S, Briand LC. NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR. IEEE Trans Softw Eng [Internet]. 2023 Sept [cited 2025 Sept 18];49(9):4282–303. Available from: https://ieeexplore.ieee.org/abstract/document/10167495

20. Hamon R, Junklewitz H, Sanchez I, Malgieri G, De Hert P. Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. IEEE Comput Intell Mag [Internet]. 2022 Feb [cited 2025 Sept 3];17(1):72–85. Available from: https://ieeexplore.ieee.org/abstract/document/9679770

21. Krishna S, Ma J, Lakkaraju H. Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten [Internet]. arXiv; 2023 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2302.04288

22. Brasileiro Araújo T, Efthymiou V, Stefanidis K. Fairness and Explanations in Entity Resolution: An Overview. IEEE Access [Internet]. 2025 [cited 2025 Oct 2];13:145127–43. Available from: https://ieeexplore.ieee.org/document/11129015/

23. Stratigi M, Bikakis N, Stefanidis K. Counterfactual Explanations for Group Recommendations. Proc DOLAP 2025. 2025;

24. Attolou HM, Tzompanaki K, Stefanidis K, Kotzinos D. Why-Not Explainable Graph Recommender. In IEEE Computer Society; 2024 [cited 2025 Oct 2]. p. 2245–57. Available from: https://www.computer.org/csdl/proceedings-article/icde/2024/171500c245/1YOtQuPJKMw

25. Giannopoulos G, Papastefanatos G, Sacharidis D, Stefanidis K. Interactivity, Fairness and Explanations in Recommendations. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization [Internet]. Utrecht Netherlands: ACM; 2021

[cited 2025 Oct 2]. p. 157–61. Available from: https://dl.acm.org/doi/10.1145/3450614.3462238

26. Stratigi M, Tzompanaki K, Stefanidis K. Why-Not Questions & Explanations for Collaborative Filtering. In: Web Information Systems Engineering – WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II [Internet]. Berlin, Heidelberg: Springer-Verlag; 2020 [cited 2025 Oct 1]. p. 301–15. Available from: https://doi.org/10.1007/978-3-030-62008-0_21

27. Nguyen TT, Huynh TT, Ren Z, Nguyen PL, Liew AWC, Yin H, et al. A Survey of Machine Unlearning [Internet]. arXiv; 2024 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2209.02299

28. Cao Y, Yang J. Towards Making Systems Forget with Machine Unlearning. In: Proceedings of the 2015 IEEE Symposium on Security and Privacy [Internet]. USA: IEEE Computer Society; 2015 [cited 2025 Sept 14]. p. 463–80. (SP '15). Available from: https://doi.org/10.1109/SP.2015.35

29. Li N, Zhou C, Gao Y, Chen H, Fu A, Zhang Z, et al. Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects [Internet]. arXiv; 2024 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2403.08254

30. Shi Y, Liu S, Wang R. Rethinking Evaluation Metrics for Machine Unlearning. In 2025 [cited 2025 Sept 3]. Available from: https://openreview.net/forum?id=wuGgok1Zyd

31. Lin W, Li S, Zhang C, Ji B, Yu J, Ma J, et al. SummScore: A Comprehensive Evaluation Metric for Summary Quality Based on Cross-Encoder [Internet]. arXiv; 2022 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2207.04660

32. Dai X, Karimi S, Fang B. A Critical Look at Meta-evaluating Summarisation Evaluation Metrics [Internet]. arXiv; 2024 [cited 2025 Sept 3]. Available from: http://arxiv.org/abs/2409.19507

33. Yang Y, Han X, Chai Y, Ebrahimi R, Behnia R, Padmanabhan B. From Machine Learning to Machine Unlearning: Complying with GDPR's Right to be Forgotten while Maintaining Business Value of Predictive Models [Internet]. arXiv; 2024 [cited 2025 Sept 1]. Available from: http://arxiv.org/abs/2411.17126

34. Ahmed SM, Basaran UY, Raychaudhuri DS, Dutta A, Kundu R, Niloy FF, et al. Towards Source-Free Machine Unlearning [Internet]. arXiv; 2025 [cited 2025 Sept 15]. Available from: http://arxiv.org/abs/2508.15127

35. Hassan H, Elayidom S, Irshad MR, Chesneau C. A detailed analysis into evaluation metrics for knowledge graph evaluation. Int J Data Sci Anal [Internet]. 2025 Aug 1 [cited 2025 Sept 15];1–40. Available from: https://link.springer.com/article/10.1007/s41060-025-00871-3

36. Liu Y, Safavi T, Dighe A, Koutra D. Graph Summarization Methods and Applications: A Survey. ACM Comput Surv [Internet]. 2018 June 22 [cited 2025 Sept 3];51(3):62:1-62:34. Available from: https://dl.acm.org/doi/10.1145/3186727

37. Čebirić Š, Goasdoué F, Kondylakis H, Kotzinos D, Manolescu I, Troullinou G, et al. Summarizing semantic graphs: a survey. VLDB J [Internet]. 2019 June 1 [cited 2025 Aug 14];28(3):295–327. Available from: https://dl.acm.org/doi/10.1007/s00778-018-0528-3

38. Wang X, Cheng G. A Survey on Extractive Knowledge Graph Summarization: Applications, Approaches, Evaluation, and Future Directions. In 2024 [cited 2025 Sept 3]. p. 8290–8. Available from: https://www.ijcai.org/proceedings/2024/916

39. Wei Y, Zou J, Guo W, Wang G, Xu X, Yang Y. MSCRS: Multi-modal Semantic Graph Prompt Learning Framework for Conversational Recommender Systems. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval

[Internet]. New York, NY, USA: Association for Computing Machinery; 2025 [cited 2025 Sept 17]. p. 42–52. (SIGIR ’25). Available from: https://dl.acm.org/doi/10.1145/3726302.3730040

40. Ribeiro LFR, Liu M, Gurevych I, Dreyer M, Bansal M. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations [Internet]. arXiv; 2022 [cited 2025 Sept 18]. Available from: http://arxiv.org/abs/2204.06508

41. Kouris P, Alexandridis G, Stafylopatis A. Text summarization based on semantic graphs: an abstract meaning representation graph-to-text deep learning approach. J Big Data [Internet]. 2024 Dec [cited 2025 Sept 18];11(1):1–39. Available from: https://link.springer.com/article/10.1186/s40537-024-00950-5

42. Troullinou G, Kondylakis H, Stefanidis K, Plexousakis D. Exploring RDFS KBs Using Summaries. In: Vrandečić D, Bontcheva K, Suárez-Figueroa MC, Presutti V, Celino I, Sabou M, et al., editors. The Semantic Web – ISWC 2018 [Internet]. Cham: Springer International Publishing; 2018 [cited 2025 Oct 2]. p. 268–84. (Lecture Notes in Computer Science; vol. 11136). Available from: https://link.springer.com/10.1007/978-3-030-00671-6_16

43. Vassiliou G, Troullinou G, Papadakis N, Stefanidis K, Pitoura E, Kondylakis H. Coverage-Based Summaries for RDF KBs. In: Verborgh R, Dimou A, Hogan A, d’Amato C, Tiddi I, Bröring A, et al., editors. The Semantic Web: ESWC 2021 Satellite Events [Internet]. Cham: Springer International Publishing; 2021 [cited 2025 Oct 2]. p. 98–102. (Lecture Notes in Computer Science; vol. 12739). Available from: https://link.springer.com/10.1007/978-3-030-80418-3_18

44. Tsalouchidou I, Bonchi F, Morales GDF, Baeza-Yates R. Scalable Dynamic Graph Summarization. IEEE Trans Knowl Data Eng [Internet]. 2020 Feb [cited 2025 Sept 15];32(2):360–73. Available from: https://ieeexplore.ieee.org/abstract/document/8554120

45. Mridha MF, Lima AA, Nur K, Das SC, Hasan M, Kabir MM. A Survey of Automatic Text Summarization: Progress, Process and Challenges. IEEE Access [Internet]. 2021 [cited 2025 Sept 18];9:156043–70. Available from: https://ieeexplore.ieee.org/abstract/document/9623462

46. Dan J, Hu W, Wang Y. Enhancing legal judgment summarization with integrated semantic and structural information. Artif Intell Law [Internet]. 2025 June 1 [cited 2025 Sept 18];33(2):271–92. Available from: https://link.springer.com/article/10.1007/s10506-023-09381-8

47. Lanov D. SEMANTIC POLICY SUMMARIES AND SEMANTIC EXPLANATIONS FOR EXPLAINABLE AI IN NETWORKING. Defensive Publ Ser [Internet]. 2025 July 23; Available from: https://www.tdcommons.org/dpubs_series/8385

48. Zhao Y, Huang Z, Zhang K, Gao W, Liu Q, Liu X, et al. Semantic-Aligned Code Summarization: Bridging the Gap Between Code and Natural Language Through Data Flow Analysis. IEEE Trans Neural Netw Learn Syst [Internet]. 2025 [cited 2025 Sept 18];1–16. Available from: https://ieeexplore.ieee.org/abstract/document/11082014

49. Diao Y, Guzewicz P, Manolescu I, Mazuran M. Efficient Exploration of Interesting Aggregates in RDF Graphs. In: Proceedings of the 2021 International Conference on Management of Data [Internet]. New York, NY, USA: Association for Computing Machinery; 2021 [cited 2025 Aug 31]. p. 392–404. (SIGMOD ’21). Available from: https://dl.acm.org/doi/10.1145/3448016.3457307

50. Niazmand E, Sejdiu G, Graux D, Vidal ME. Efficient semantic summary graphs for querying large knowledge graphs. Int J Inf Manag Data Insights [Internet]. 2022 Apr 1 [cited 2025 Sept 1];2(1):100082. Available from: https://www.sciencedirect.com/science/article/pii/S2667096822000258

51. Kang S, Lee K, Shin K. Personalized Graph Summarization: Formulation, Scalable Algorithms, and Applications. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE) [Internet]. 2022 [cited 2025 Sept 15]. p. 2319–32. Available from: https://ieeexplore.ieee.org/abstract/document/9835590

52. Thakker D, Schwabe D, Kozaki K, García R, Brambilla M, Dimitrova V, et al. Quality metrics for RDF graph summarization. Semant Web [Internet]. 2019 Jan 1 [cited 2025 Sept 1];10(3):555–84. Available from: https://doi.org/10.3233/SW-190346

53. Gigant T, Guinaudeau C, Decombas M, Dufaux F. Mitigating the Impact of Reference Quality on Evaluation of Summarization Systems with Reference-Free Metrics [Internet]. arXiv; 2024 [cited 2025 Sept 15]. Available from: http://arxiv.org/abs/2410.10867

54. Blanco-Justicia A, Jebreel N, Manzanares-Salor B, Sánchez D, Domingo-Ferrer J, Collell G, et al. Digital forgetting in large language models: a survey of unlearning methods. Artif Intell Rev [Internet]. 2025 Mar 1 [cited 2025 Sept 15];58(3):1–41. Available from: https://link.springer.com/article/10.1007/s10462-024-11078-6

55. Ranasinghe N, Mohammed WM, Stefanidis K, Martinez Lastra JL. Large Language Models in Human-Robot Collaboration With Cognitive Validation Against Context-Induced Hallucinations. IEEE Access [Internet]. 2025 [cited 2025 Oct 2];13:77418–30. Available from: https://ieeexplore.ieee.org/document/10980279/

56. Dam T, Henzl M, Klausner LD. Delete My Account: Impact of Data Deletion on Machine Learning Classifiers. In: 2021 International Conference on Software Security and Assurance (ICSSA) [Internet]. 2021 [cited 2025 Sept 18]. p. 7–20. Available from: http://arxiv.org/abs/2311.10385

57. Liu S, Yao Y, Jia J, Casper S, Baracaldo N, Hase P, et al. Rethinking machine unlearning for large language models. Nat Mach Intell [Internet]. 2025 Feb [cited 2025 Sept 15];7(2):181–94. Available from: https://www.nature.com/articles/s42256-025-00985-0

58. Shi W, Lee J, Huang Y, Malladi S, Zhao J, Holtzman A, et al. MUSE: Machine Unlearning Six-Way Evaluation for Language Models [Internet]. arXiv; 2024 [cited 2025 Sept 15]. Available from: http://arxiv.org/abs/2407.06460

59. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning Word Vectors for Sentiment Analysis. 2011;

60. About · DBpedia Development Wiki [Internet]. [cited 2025 Oct 22]. Available from: http://dev.dbpedia.org/