ISSN: 2836-8495

Research Article

Journal of Current Trends in Computer Science Research

Brain Surgery: Ensuring GDPR Compliance in Large Language Models via Concept Erasure

Michele Laurelli*

Istituto Criminologia, CEO, Algoretic, Milan, Lombardy, Italy

***Corresponding Author**
Michele Laurelli, Istituto Criminologia, CEO, Algoretic, Milan, Lombardy, Italy.

Submitted: 2024, Sep 06; **Accepted**: 2024, Oct 07; **Published**: 2024, Oct 10

Citation: Laurelli, M. (2024). Brain Surgery: Ensuring GDPR Compliance in Large Language Models via Concept Erasure. *J Curr Trends Comp Sci Res, 3*(5), 01-12.

Abstract

As large-scale AI systems proliferate, ensuring compliance with data privacy laws such as the General Data Protection Regulation (GDPR) has become critical. This paper introduces Brain Surgery, a transformative methodology for making every local AI model GDPR-ready by enabling real-time privacy management and targeted unlearning. Building on advanced techniques such as Embedding-Corrupted Prompts (ECO Prompts), blockchain-based privacy management, and privacy-aware continual learning, Brain Surgery provides a modular solution that can be deployed across various AI architectures. This tool not only ensures compliance with privacy regulations but also empowers users to define their own privacy limits, creating a new paradigm in AI ethics and governance.

1. Introduction

In recent years, the widespread adoption of Artificial Intelligence (AI) systems, particularly Large Language Models (LLMs), has sparked significant advancements across a variety of sectors, including healthcare, finance, education, and beyond. These models, trained on vast datasets, have demonstrated remarkable capabilities in understanding and generating human-like text, making them in- dispensable tools for a growing number of applications. From enhancing customer service through chatbots to assisting researchers in generating novel in- sights, the role of AI in modern life has become pervasive.

However, as the influence and utility of AI systems grow, so do concerns about privacy and data security. The vast datasets that underpin these models often include personal and sensitive information, raising questions about how such data is handled, stored, and utilized. In particular, the emergence of stringent data protection regulations such as the General Data Protection Regulation (GDPR) in the European Union has brought the issue of privacy to the fore- front. GDPR mandates that individuals have the right to control their personal data, including the ”right to be forgotten” — the ability to request the deletion of personal information. This legal requirement poses unique challenges for AI systems, especially those like LLMs, which are designed to learn and retain vast amounts of information from the data they process.

Traditional approaches to ensuring privacy compliance in AI models typically involve removing or anonymizing personal data during the training phase. However, once a model has been trained, unlearning specific pieces of information embedded deep within the model’s parameters becomes a complex and computationally expensive task. This problem is particularly acute for LLMs, which store knowledge in distributed representations across multiple layers of their neural networks. The necessity for efficient and reliable methods to ensure compliance with privacy regulations in deployed AI systems has never been more urgent.

Furthermore, the AI community faces a growing ethical imperative to develop systems that not only comply with legal frameworks but also promote transparency, accountability, and user control. AI models are often seen as ”black boxes,” where it is difficult to understand how decisions are made or how specific pieces of information are stored and utilized. This opacity raises concerns not only about privacy but also about the ethical use of AI in decision-making processes. In response, researchers and technologists are seeking ways to make AI systems more interpretable and accountable while maintaining their performance and utility.

The convergence of these factors — the increasing deployment

of AI systems, the growing focus on data privacy, and the ethical demands for transparency — underscores the need for innovative solutions that address the challenges of privacy management in large-scale AI models. This research aims to contribute to this critical area by proposing a new approach to making AI models compliant with data privacy regulations, without compromising their functionality or requiring extensive retraining.

The need for scalable, efficient, and user-friendly privacy management systems in AI is clear. As AI continues to shape the future of industries and societies, ensuring that these systems can operate within legal and ethical boundaries is paramount. This paper aims to provide a solution to these challenges, focusing on creating mechanisms that allow for the dynamic and targeted removal of sensitive information from AI models, while ensuring compliance with existing privacy laws and empowering users to exercise control over their data.

2. Related Work

As the field of Artificial Intelligence (AI) continues to evolve, numerous approaches have been proposed to address the challenges associated with data privacy, knowledge retention, and compliance with legal frameworks such as the General Data Protection Regulation (GDPR). The development of large-scale models, particularly in natural language processing (NLP), has given rise to new concerns about how personal data embedded in these models can be managed effectively without compromising performance.

One of the most prominent areas of research relevant to this challenge is knowledge editing in AI models. Knowledge editing refers to the ability to update, remove, or modify the information that a model has learned, either post-training or during operation. Traditional approaches to knowledge editing often involve retraining the model with new datasets that exclude or correct specific information. This method, while effective in some cases, tends to be computationally expensive and is often impractical for large models such as GPT, BERT, or LLaMA. Additionally, these retraining approaches do not scale well, particularly in real-time applications, where rapid responses to privacy requests are required.

Another line of research has focused on fine-tuning as a way to achieve selective unlearning. Fine-tuning allows a pre-trained model to be adjusted on a smaller dataset, with the aim of correcting specific behaviors or removing certain types of knowledge. However, this technique can also lead to unintended consequences, such as overfitting or loss of generalization across the broader dataset. These drawbacks make fine-tuning an unreliable solution for ensuring GDPR compliance in AI models that must handle diverse and evolving data sources. Moreover, fine-tuning generally lacks the precision required to target and erase specific pieces of information without affecting other unrelated knowledge areas in the model.

A more recent development in the field of AI unlearning is the emergence of localized modification techniques, which aim to alter specific embeddings or parameters within a model without necessitating a full retraining or fine-tuning. These approaches are designed to provide more precision when editing a model's knowledge, allowing for targeted changes without compromising overall model performance. Researchers in this area have developed various strategies, including gradient-based methods, which modify the internal representations of concepts that the model has learned. These methods hold promise for applications requiring rapid and localized updates to the model, such as GDPR-compliant unlearning. However, despite their precision, they can still be resource-intensive and may not always guarantee complete removal of sensitive information.

In addition to knowledge editing, several studies have examined the broader concept of AI model transparency and interpretability, particularly in the con- text of ethical AI governance. There has been growing interest in developing models that not only perform well but also offer transparency into their decision- making processes. This transparency is crucial for gaining trust in AI systems, especially in sensitive areas like healthcare, finance, and law enforcement, where the consequences of incorrect or biased decisions can be significant. Several methodologies have been proposed to improve model interpretability, including feature attribution techniques and attention mechanisms. These techniques aim to shed light on how models arrive at certain conclusions, which is critical when considering privacy and ethical concerns.

Privacy-aware machine learning has also gained considerable attention, particularly in the context of continual learning. Continual learning refers to the ability of AI systems to adapt and learn from new data over time without forget- ting previously learned knowledge. Privacy-aware continual learning integrates mechanisms to ensure that sensitive data is not inadvertently embedded in a model's memory during its ongoing training processes. Although this line of research is still in its infancy, it presents a promising direction for models that need to comply with privacy regulations while continuing to evolve and improve. However, the challenge remains in ensuring that these systems can manage privacy concerns dynamically, especially as new data arrives in real-time.

Blockchain-based approaches for privacy management have also emerged as a potential solution to ensuring compliance and transparency in AI systems. Blockchain technology provides a decentralized, immutable ledger that can be used to log privacy-related actions, such as data deletion requests. This creates an auditable trail that can be used to verify that a model complies with privacy regulations, such as GDPR. Some researchers have proposed the use of blockchain to track and manage user data in AI systems, allowing users greater control over their personal information while ensuring that organizations adhere to legal requirements. While blockchain offers a transparent and secure frame- work for managing privacy, its integration with AI models is still an area of active research and development, and its practical application on a large scale remains limited.

Despite these advancements, significant challenges remain in creating efficient, scalable, and real-time methods for managing privacy in AI models. Many of the existing approaches are either too resource-intensive or lack the precision needed for targeted unlearning of specific information. Additionally, there is a growing demand for solutions that are modular, flexible, and adaptable to various AI architectures, ranging from small-scale local models to large-scale, cloud-based systems.

This research aims to build upon and extend the work of these earlier efforts by addressing some of the gaps that remain in the current landscape. While knowledge editing, unlearning, and privacy-aware machine learning are well established areas, there is still a need for more robust, scalable solutions that can be applied across diverse AI models without requiring extensive retraining or performance trade-offs. Furthermore, the integration of advanced privacy management techniques, such as blockchain, within AI systems is an area that warrants further exploration. By drawing on these existing methodologies and introducing new innovations, this research seeks to provide a comprehensive solution to the challenge of GDPR compliance in AI models.

2.1 Knowledge Editing and Concept Unlearning

There has been significant research on knowledge editing and concept unlearning in AI models. Fine-tuning and retraining have traditionally been employed to modify model knowledge, but these methods tend to be resource-intensive and may lead to undesired loss of generalization [1]. Recent advancements, such as localized knowledge edits have paved the way for more efficient, targeted unlearning strategies [2].

2.2 Embedding-Corrupted Prompts (ECO Prompts)

ECO Prompts introduce controlled perturbations to the embeddings associated with specific concepts. This iterative process reduces the influence of unwanted knowledge while maintaining the structural integrity of the model [3]. Unlike retraining, ECO Prompts allow for localized modification of knowledge without impacting the broader performance of the AI model.

2.3 Conflict Score Evaluation and Real-Time Monitoring

Unlearning specific knowledge can introduce inconsistencies in other related areas of knowledge. Conflict score evaluation has been proposed to monitor such inconsistencies [4]. Brain Surgery integrates real-time conflict monitoring to ensure that GDPR-compliant unlearning does not disrupt the model's outputs in unintended ways.

3. Mathematical Formulation

A core challenge in adapting large language models (LLMs) to meet GDPR requirements lies in the ability to remove or modify specific information embedded deep within the model's structure, while maintaining overall model performance. The vast majority of AI models, especially those with transformer architectures, store knowledge in high-dimensional embeddings—mathematical representations of concepts that the model has learned during training. To ensure GDPR compliance, it is necessary to target specific embeddings related to sensitive or personal data and modify them in such a way that the corresponding information is no longer retrievable by the model.

In this section, I provide a mathematical framework for this process, focusing on the transformation of embeddings to "forget" specific concepts while retaining the integrity of the overall embedding space. This mathematical approach seeks to ensure that sensitive data is unlearned in a targeted manner without affecting the model's broader knowledge or introducing performance degradation.

3.1 Representation of Concept Embeddings

Embeddings in AI models, particularly language models, are high-dimensional vectors that represent words, phrases, or concepts in a continuous space. Let $e_c \in \mathbb{R}^d$ denote the embedding vector for a specific concept c, where d is the dimensionality of the embedding space. The vector ec is learned during the training process and encodes both syntactic and semantic properties of the concept c. Similar concepts are mapped to nearby points in the embedding space, and the distance between embeddings reflects the relationships between concepts.

In order to modify or erase the concept c from the model, I need to alter the embedding ec in such a way that the model no longer associates c with its previous context or meaning. At the same time, it is crucial that these modifications do not disrupt the representations of unrelated concepts that share proximity in the embedding space.

3.2 Modifying Concept Embeddings

To achieve this goal, I employ a perturbation-based approach. Let $L(e_c)$ represent a loss function that quantifies the influence of the concept c on the model's outputs. Our objective is to minimize this influence by iteratively modifying the embedding e_c through gradient-based updates. Specifically, I compute the gradient of the loss function with respect to the embedding e_c, and use this information to adjust the embedding in a way that reduces the model's reliance on the concept c.

The update rule for the modified embedding e'_c is given by:

$$e'_c = e_c - \alpha \cdot \nabla_{e_c} L(e_c) \tag{1}$$

Here, α is a step size that controls the magnitude of the perturbation, and $\Delta e_c\, L(e_c)$ is the gradient
of the loss function with respect to the embedding e_c. This gradient indicates the direction in which the embedding should be adjusted to minimize the model's association with the concept c. By iteratively applying this update rule, I gradually weaken the influence of c within the model.

3.3 Preserving the Integrity of the Embedding Space

One of the key challenges in modifying concept embeddings is ensuring that the perturbation of e_c does not introduce distortions in the overall embedding space. If the modifications are too large or poorly controlled, they may affect neighboring embeddings,

leading to unintended consequences for the model's broader knowledge base. For example, if c is closely related to other concepts
$c_1, c_2, \ldots, c_n$, perturbing e_c too aggressively might disrupt the model's under- standing of these related concepts.

To mitigate this risk, I introduce a normalization step that ensures the modified embedding e'_c remains within a feasible region of the embedding space
After each update, the modified embedding is normalized as follows:

$$e'_c = \frac{e'_c}{\|e'_c\|} \tag{2}$$

This normalization ensures that the embedding vector e'_c maintains a consistent magnitude, preventing the perturbed embedding from drifting too far from its original position in the space. By controlling the distance between the original and modified embeddings, I can preserve the overall structure of the embedding space, minimizing the impact on unrelated concepts.

3.4 Convergence of the Unlearning Process

The iterative modification of embeddings through gradient-based updates leads to a progressive reduction in the model's reliance on the concept c. However, it is important to define a stopping criterion to determine when the unlearning process is complete. A common approach is to monitor the value of the loss function $L(e_c)$ during the update process. When $L(e_c)$ falls below a certain threshold ϵ, I can conclude that the model's association with the concept c has been sufficiently diminished, and further updates are unnecessary.

The stopping condition can be formalized as:

$$L(e_c) < \epsilon \tag{3}$$

where ϵ is a small positive constant that defines the desired level of concept erasure. This criterion ensures that the model no longer produces outputs that are significantly influenced by the concept c, while also limiting unnecessary perturbations that could affect other aspects of the model's behavior.

3.5 Theoretical Implications for Model Performance

While the perturbation-based approach is effective for targeted unlearning, it is important to consider its implications for the overall performance of the model. In practice, the modifications applied to the embedding space should be mini- mal enough to avoid significant degradation in the model's accuracy or fluency. The balance between successful unlearning and maintaining model performance depends on the choice of step size α and the threshold ϵ, as well as the specific properties of the concept being unlearned.

Empirical results in existing literature suggest that small, iterative updates to concept embeddings, combined with normalization, can achieve effective un- learning without introducing substantial loss in performance. This balance is crucial for ensuring that the model remains functional and accurate while com- plying with privacy requirements such as GDPR.

4. Conflict Score Evaluation

One of the primary challenges in modifying or unlearning specific concepts from large-scale AI models, such as those built on transformer architectures, is ensuring that these modifications do not introduce unintended side effects. In particular, the removal or adjustment of one concept might inadvertently affect related knowledge areas within the model, thereby compromising the model's ability to generalize or retain critical information. To address this issue, a mechanism is needed to assess the potential impact of these modifications, ensuring that the unlearning process does not degrade the model's performance or pro- duce contradictions in the remaining knowledge.

The *Conflict Score Evaluation* framework serves as a formal method for assessing the consistency and integrity of a model after the unlearning of a targeted concept. This evaluation seeks to quantify the extent to which unlearning actions impact related concepts and to ensure that the removal of sensitive or unwanted information does not lead to undesirable outcomes in other areas of the model's behavior.

4.1 Motivation for Conflict Score Evaluation

The need for a conflict score arises from the inherent interconnectedness of knowledge within large language models. Concepts learned by the model are represented as high-dimensional embeddings, and these embeddings are often distributed across multiple layers of the model. As a result, concepts that appear semantically or syntactically related may share portions of their representations, making it challenging to unlearn one concept without influencing others.

For instance, removing or reducing the influence of a specific concept c may unintentionally alter the model's understanding of related concepts $c_1, c_2, \ldots, c_n$, leading to performance degradation or contradictions in the model's outputs. Such unintended consequences could manifest in various ways, including incorrect predictions, loss of generalization, or a decrease in the model's ability to respond accurately to prompts associated with related concepts.

The conflict score provides a quantitative measure of how well the unlearning process preserves the integrity of the model's broader knowledge base. By evaluating this score, researchers and practitioners can monitor the side effects of unlearning and determine whether further refinements are necessary to maintain model consistency.

4.2 Formal Definition of Conflict Score

To evaluate the impact of unlearning, I define a set of related concepts X_r that the model must retain after the unlearning process. Additionally, I define a set of unwanted concepts X_u, which are the target of the unlearning process. The goal is to remove or reduce the influence of the concepts in X_u while preserving the accuracy

and consistency of the model's outputs on the concepts in X_r.

Let $f(x_r)$ represent the model's output for a given related concept $x_r \in X_r$, and let y_r be the expected correct output for x_r. The conflict score S_c is defined as the fraction of correct outputs produced by the model for the related concepts after the unlearning process:

$$S_c = \frac{1}{|X_r|} \sum_{x_r \in X_r} 1(f(x_r) = y_r) \quad (4)$$

In this formulation, $1(f(x_r) = y_r)$ is an indicator function that evaluates to 1 if the model's output for the concept x_r matches the expected output y_r, and 0 otherwise. The conflict score S_c thus represents the proportion of related concepts for which the model's performance has been preserved after the unlearning process.

4.3 Interpretation of Conflict Score

The conflict score provides a simple yet powerful mechanism for quantifying the side effects of unlearning. A conflict scores close to 1 (Sc 1) indicates that the unlearning process has had little to no impact on the model's understanding of related concepts, suggesting that the targeted unlearning was successful and did not compromise the broader knowledge embedded in the model. In this case, the unlearning process can be considered safe, as it has effectively removed the unwanted concept without introducing contradictions or inconsistencies.

On the other hand, a conflict scores significantly lower than 1 ($S_c <$ 1) signals potential conflicts introduced by the unlearning process. In this scenario, the model may have lost its ability to accurately respond to prompts associated with related concepts, or it may produce incorrect outputs for these concepts. A low conflict score is often a red flag that indicates the need for further refinement of the unlearning process, as the modifications may have introduced undesirable side effects.

4.4 Applications of Conflict Score Evaluation

The conflict score evaluation can be applied to a wide range of scenarios where the integrity of a model's knowledge must be preserved following concept un- learning. Some of the key applications include:

4.4.1 Privacy Compliance

In the context of GDPR compliance, the conflict score can be used to evaluate whether the removal of personal or sensitive information from the model has affected the accuracy or consistency of other, unrelated outputs. A high conflict score ensures that the unlearning process respects privacy regulations while maintaining the model's overall functionality.

4.4.2 Ethical AI Systems

As AI systems become more integrated into critical decision-making processes, it is essential to ensure that unlearning certain biases or harmful content does not introduce new biases or inaccuracies. The conflict score can help monitor and refine these adjustments, ensuring ethical AI governance.

4.4.3 Domain-Specific Applications

In domain-specific AI models, such as those used in healthcare or finance, the conflict score can ensure that un- learning specific outdated or incorrect knowledge does not inadvertently disrupt other areas of the model's expertise. This is particularly important in sensitive fields where the consequences of incorrect outputs can be severe.

4.4.4 Knowledge Retention in Continual Learning

In systems employing continual learning, where models are regularly updated with new information, the conflict score can be used to monitor the stability of older knowledge as new concepts are learned or older concepts are unlearned.

4.5 Refining the Unlearning Process Based on Conflict Score

When the conflict score evaluation indicates a drop-in consistency (i.e., $S_c < 1$), this serves as a signal that the unlearning process needs refinement. Several strategies can be employed to address conflicts:

4.5.1 Adjusting the Step Size

The step size α used in the unlearning process may be too large, causing unnecessary disturbance in the embedding space. Reducing α can help achieve more precise unlearning while minimizing unintended side effects.

4.5.2 Iterative Unlearning

Instead of applying a single, large perturbation, iterative unlearning with multiple smaller steps can ensure that the targeted concept is gradually reduced without harming related concepts. After each iteration, the conflict score can be re-evaluated to ensure the changes are progressing in a controlled manner.

4.5.3 Selective Retention of Related Embeddings

In cases where concepts are highly interconnected, selective retention techniques can be employed to preserve specific portions of the embedding space while modifying others. This allows for more targeted unlearning and helps maintain model consistency.

By using the conflict score as a feedback mechanism, researchers can refine the unlearning process to ensure minimal disruption to the model's broader knowledge while achieving the desired level of concept erasure.

4.6 Conclusion

The conflict score evaluation is an essential tool for ensuring the integrity and consistency of AI models during and after the unlearning process. By quantifying the impact of unlearning on related concepts, this metric allows for real-time monitoring of the side effects of concept removal, providing a safeguard against unintended consequences. Through careful refinement based on conflict score feedback, AI models can be made GDPR-compliant while retaining their overall performance and accuracy.

5. Privacy-Aware Continual Learning

Continual learning, also known as lifelong learning, is a framework

where AI models are designed to continuously acquire and integrate new knowledge over time without forgetting previously learned information. This capability is particularly important in real-world applications where models are constantly ex- posed to new data streams and must adapt to evolving environments. Traditional machine learning models are typically trained once on static datasets, and any new information requires retraining the entire model, which is both computationally expensive and inefficient. Continual learning addresses these challenges by enabling models to learn incrementally, adapting to new data as it arrives.

However, in the context of data privacy, continual learning introduces new challenges. As models are exposed to new data over time, they inevitably process and store representations of this data within their internal embeddings. In many cases, this data may include sensitive personal information that is subject to legal frameworks such as the General Data Protection Regulation (GDPR). Ensuring that AI models remain compliant with such regulations while continuously learning new information requires integrating privacy-preserving mechanisms into the continual learning process.

5.1 Challenges of Continual Learning with Privacy Constraints

The primary challenge of privacy-aware continual learning lies in balancing two competing objectives: (1) the need to retain important knowledge from previously learned data, and (2) the need to comply with privacy regulations that may require the deletion or modification of sensitive data. In traditional machine learning, ensuring compliance with GDPR is often managed by carefully curating the training data before the model is trained. However, continual learning systems process new data incrementally, and sensitive data may be embedded into the model after it has been trained.

This situation raises several specific challenges:

5.1.1 Retention of Sensitive Information

As new data is ingested by the model; personal or sensitive information can become embedded within the model's parameters. Over time, this sensitive information can accumulate, making it difficult to comply with requests to remove such data under privacy regulations like GDPR. The challenge is to prevent or mitigate the embedding of sensitive data while still allowing the model to learn effectively from non-sensitive information.

5.1.2 Forgetting and Catastrophic Interference

One of the core problems in continual learning is catastrophic forgetting, where learning new information causes the model to "forget" previously learned knowledge. In the context of privacy, the requirement to unlearn sensitive information adds complexity to this issue. The model must be able to unlearn specific information without inadvertently forgetting unrelated, valuable knowledge. Achieving this selective unlearning without inducing catastrophic interference is a significant technical hurdle.

5.1.3 Real-Time Privacy Monitoring

Since continual learning models are exposed to new data in real-time, privacy-preserving mechanisms must operate continuously. This requires the ability to monitor incoming data streams for sensitive content and dynamically adjust the learning process to ensure that sensitive information is not inadvertently embedded in the model. Real-time monitoring adds a layer of complexity to the system, requiring efficient mechanisms that do not significantly impact the model's learning efficiency.

5.1.4 Compliance with Right to be Forgotten

GDPR and other privacy regulations stipulate the "right to be forgotten," meaning individuals can request that their personal data be removed from a system. For AI models that learn continually, this requirement means that the system must be able to selectively erase specific pieces of knowledge, even after they have been integrated into the model's parameters. Ensuring compliance with this aspect of GDPR while maintaining model accuracy and performance is a particularly difficult challenge in the context of continual learning.

5.2 Mechanisms for Privacy-Aware Continual Learning

To address these challenges, privacy-aware continual learning models integrate several mechanisms that ensure both continual adaptation to new data and compliance with privacy regulations. Some of the key considerations in developing such systems include:

5.2.1 Selective Knowledge Retention

In privacy-aware continual learning, the system must be able to retain useful knowledge while ensuring that sensitive information is not embedded in the model. One approach to achieving this goal is through selective knowledge retention, where the model prioritizes learning generalizable knowledge from non-sensitive data while avoiding the embedding of personal or private information. This requires sophisticated data filtering mechanisms that can detect and classify sensitive data in real time.

5.2.2 Regularization-Based Methods

Regularization techniques are often employed in continual learning to mitigate the risk of catastrophic forget- ting. In privacy-aware continual learning, these methods can be extended to prioritize the removal of sensitive information while preserving the rest of the model's knowledge. For example, regularization can be used to impose penalties on embeddings associated with sensitive data, reducing their influence on the model's outputs.

5.2.3 Dynamic Privacy Constraints

A crucial element of privacy-aware continual learning is the dynamic adjustment of privacy constraints as new data arrives. The model must continuously assess whether incoming data contains sensitive information and adjust its learning objectives accordingly. This can be achieved through privacy-preserving loss functions that penalize the model for embedding sensitive data or through real-time scanning mechanisms that identify sensitive features in the incoming data stream.

5.2.4 Unlearning Mechanisms

In some cases, it may be necessary to actively unlearn sensitive information that has been embedded in the model's parameters. Unlearning in the context of continual learning is particularly challenging because it requires selective modification of the model's representations without inducing catastrophic forgetting. Techniques suchas gradient-based unlearning or targeted perturbation of embeddings can be employed to remove specific pieces of knowledge while preserving the broader knowledge base.

5.2.5 Integration with Privacy-First Architectures

To ensure that privacy- preserving continual learning systems operate effectively, they must be integrated with broader privacy-first architectures. These architectures include components such as decentralized privacy management systems (e.g., blockchain) that provide transparency and accountability in data handling, as well as user-facing tools that allow individuals to set their own privacy preferences. Such integration ensures that privacy is maintained at both the technical and governance levels, providing a comprehensive solution for GDPR compliance in continual learning systems.

5.3 Importance of Real-Time Adaptation

Privacy-aware continual learning is not a static process but one that must operate in real-time as models interact with live data. This requires the system to continually adapt both its learning and privacy management processes. Real- time adaptation ensures that sensitive data is identified and protected from the moment it is ingested by the system, rather than relying on post-hoc interventions. The real-time nature of these systems also supports rapid compliance with privacy requests, allowing models to immediately unlearn or erase sensitive information in response to user requests or legal obligations.

Moreover, real-time adaptation is critical for ensuring that privacy-preserving mechanisms do not interfere with the model's ability to generalize from new data. As the model encounters new information, it must be able to distinguish between sensitive and non-sensitive content, ensuring that the privacy constraints do not overly restrict the model's learning potential. This balance between privacy protection and model adaptability is central to the success of privacy-aware continual learning systems.

5.4 Ethical Considerations

Beyond the legal requirements imposed by regulations such as GDPR, privacy- aware continual learning also addresses broader ethical concerns regarding the responsible use of AI. AI models that operate in dynamic environments must not only comply with privacy laws but also respect user autonomy and data ownership. By integrating privacy-preserving mechanisms into the continual learning process, AI systems empower users to control how their data is used, shared, and retained. This fosters greater trust in AI systems and helps to align AI development with ethical principles such as fairness, accountability, and transparency.

Furthermore, privacy-aware continual learning contributes to the development of more responsible AI systems by ensuring that personal data is not misused or exploited for unintended purposes. In an era where data privacy is of paramount importance, continual learning systems that respect user privacy are essential for ensuring the ethical deployment of AI in critical domains such as healthcare, finance, and public services.

5.5 Conclusion

Privacy-aware continual learning represents a crucial advancement in the development of AI systems that are both adaptive and compliant with privacy regulations. By integrating mechanisms that ensure the selective retention and unlearning of sensitive information, these systems address the unique challenges posed by GDPR and other privacy laws. Moreover, the ability to continuously monitor and adjust privacy constraints in real time enables models to remain flexible and responsive to new data while maintaining compliance with legal and ethical requirements. As AI systems continue to evolve, privacy-aware continual learning will play an increasingly important role in ensuring that these systems can adapt to new environments without compromising user privacy.

6. Methodology

The methodology outlined in this paper is designed to address the pressing need for scalable, efficient, and GDPR-compliant mechanisms that can be integrated into large-scale AI systems. As models such as large language models (LLMs) continue to evolve, it becomes increasingly important to implement systems that allow for the removal or modification of sensitive data in a precise and con- trolled manner. The proposed methodology provides a modular framework for achieving this, focusing on targeted unlearning, real-time privacy management, and the preservation of overall model performance.

6.1 Overview of the Framework

At its core, the methodology centers around a modular and extensible frame- work for privacy management and targeted unlearning within AI models. This framework is designed to be compatible with various AI architectures, from large-scale cloud-based models to edge devices, ensuring flexibility and broad applicability. The key objectives of the framework are:

- To provide a scalable solution for ensuring that AI models comply with data privacy regulations, particularly the General Data Protection Regulation (GDPR).
- To offer tools for real-time monitoring and management of privacy-sensitive information, enabling rapid compliance with requests for data deletion or unlearning.
- To preserve the overall performance, accuracy, and generalization abilities of the model, ensuring that unlearning specific concepts does not degrade the system's functionality.
- To empower users by giving them more control over their data, including the ability to define privacy preferences and manage how their data is used in AI models.

This framework is modular in design, meaning it can be adapted and expanded depending on the specific needs of the application or

model. It consists of several interconnected components that work together to ensure the secure and efficient management of sensitive information.

6.2 Modular GDPR Compliance Framework

The modular nature of the proposed framework allows for seamless integration with existing AI systems. One of the key innovations of this methodology is its plug-and-play architecture, which provides a flexible solution that can be implemented across different models, including those that are already operational. This modular framework consists of the following key components:

6.2.1 Data Monitoring Module

This component is responsible for continuously monitoring the data ingested by the AI model, identifying sensitive information, and ensuring that the model remains compliant with privacy regulations. The data monitoring module operates in real-time, enabling the system to flag privacy-sensitive content as it is being processed.

6.2.2 Privacy Management API

The framework provides an API that al- lows administrators or users to submit requests for data deletion or un- learning. This API can be integrated into broader data management systems, providing a clear interface for handling GDPR-related requests. The API ensures that the privacy management operations are logged and auditable, fostering accountability and transparency.

6.2.3 Embedding Modification Engine

At the heart of the framework is the Embedding Modification Engine, which enables targeted unlearning of specific concepts. This engine operates on the model's embeddings, allowing for precise modifications that remove unwanted knowledge while preserving the broader knowledge base. By focusing on the embeddings, the system ensures that unlearning is both efficient and effective.

6.2.4 Real-Time Compliance Module

This component is responsible for real-time privacy compliance, ensuring that the model's outputs are continually monitored to prevent the inadvertent inclusion of sensitive data in predictions or responses. The Real-Time Compliance Module operates alongside the Embedding Modification Engine, providing a feedback loop that ensures GDPR compliance throughout the model's operation.

This modular design allows for flexible deployment across different plat- forms, from cloud-based AI systems to edge devices. Each component can be customized or expanded depending on the specific needs of the model and the regulatory environment in which it operates.

6.3 Targeted Unlearning Process

The targeted unlearning process forms the core of the methodology. Rather than relying on full retraining or fine-tuning, which are computationally ex- pensive and often result in performance degradation, the framework employs a more efficient approach that focuses on localized modifications to the model's embeddings. The steps involved in this process are outlined as follows:

6.3.1 Identification of Sensitive Embeddings

The first step in the un- learning process is to identify the specific embeddings in the model's representation space that correspond to sensitive data. This can be achieved through a combination of privacy-aware continual learning mechanisms and real-time data monitoring. The system flags embeddings associated with personal or sensitive information, marking them for modification.

6.3.2 Iterative Modification of Embeddings

Once the sensitive embed- dings are identified, the system applies an iterative process to modify these embeddings, reducing their influence on the model's outputs. This is achieved through gradient-based updates, where the model's loss function is minimized with respect to the specific embeddings. By iteratively reducing the impact of these embeddings, the model effectively "forgets" the sensitive information without affecting related knowledge.

6.3.3 Normalization and Integrity Preservation

During the unlearning process, it is critical to ensure that the overall structure of the model's embeddings remains intact. To achieve this, the modified embeddings are normalized after each update, ensuring that the changes do not distort the relationships between other, non-sensitive concepts. This normalization step ensures that the model retains its generalization capabilities while complying with privacy regulations.

6.3.4 Feedback and Conflict Resolution

After each iteration of the un- learning process, the system evaluates the model's outputs using a conflict score evaluation method. This ensures that the removal of sensitive information does not introduce inconsistencies or degrade the model's performance on related tasks. If conflicts are detected, further refinement of the unlearning process is applied until the model's outputs are both compliant with privacy regulations and consistent with its original performance.

This targeted approach to unlearning is both efficient and scalable, allowing the system to adapt to different regulatory environments and privacy requirements. It ensures that AI models can remain GDPR-compliant without sacrificing performance or requiring extensive retraining.

6.4 Real-Time Monitoring and Privacy-Aware Learning

A critical component of the methodology is the integration of real-time monitoring and privacy-aware learning mechanisms. These mechanisms allow the model to dynamically adjust its learning objectives based on incoming data and privacy requirements. By incorporating real-time monitoring, the system is able to identify privacy-sensitive data as it is being processed, ensuring that sensitive information is not inadvertently embedded in the model's representations.

Key features of the real-time monitoring and privacy-aware

learning components include:

6.4.1 Dynamic Privacy Constraints

The system applies dynamic privacy constraints during both training and inference. These constraints ensure that sensitive information is flagged and handled appropriately, preventing it from being embedded in the model's internal representations. Privacy constraints can be customized based on the user's preferences, allowing for flexible privacy management.

6.4.2 Continuous Evaluation

The system continuously evaluates the model's outputs for signs of privacy violations, ensuring that sensitive information is not inadvertently included in predictions or responses. This continuous evaluation is critical for maintaining GDPR compliance in real-time applications, where model outputs are exposed to end users.

6.4.3 User-Defined Privacy Preferences

In addition to the system's automatic privacy management features, the framework allows users to define their own privacy preferences. Users can specify what types of data should be excluded from the model's training or inference, and can set limits on how long their data should be retained in the system. The framework dynamically adjusts to these user preferences, ensuring that individual privacy requirements are respected.

The real-time monitoring and privacy-aware learning components ensure that the system remains responsive to evolving privacy concerns, enabling AI models to adapt to new data while maintaining compliance with regulatory requirements.

6.5 Scalability and Deployment

The proposed methodology is designed to be scalable, capable of handling both large-scale cloud deployments and smaller edge-based models. Scalability is achieved through the modular design of the framework, which allows for efficient distribution of privacy management tasks across different layers of the system. Additionally, the system's reliance on targeted unlearning and real-time monitoring ensures that computational resources are used efficiently, minimizing the need for costly retraining or fine-tuning.

Key considerations for scalability include:

6.5.1 Cloud-Based Integration

For large-scale AI systems deployed in the cloud, the modular framework allows for the distributed handling of privacy- related tasks. Components such as the Embedding Modification Engine and Real-Time Compliance Module can be scaled across multiple nodes, ensuring efficient handling of large datasets and complex models.

6.5.2 Edge Device Deployment

The framework is also designed to be lightweight enough for deployment on edge devices, where computational resources are more limited. By focusing on efficient, targeted unlearning and real-time monitoring, the system ensures that even edge-based AI models can re- main GDPR-compliant without sacrificing performance.

6.5.3 Customizable Deployment Options

Depending on the specific needs of the application, the framework can be customized to prioritize different aspects of privacy management. For instance, in applications where real- time monitoring is critical, the system can allocate more resources to the Data Monitoring Module and Real-Time Compliance Module, ensuring that privacy violations are detected and addressed as quickly as possible.

This scalability ensures that the methodology is suitable for a wide range of AI applications, from large, centralized systems to smaller, distributed models.

The methodology outlined in this paper provides a comprehensive framework for ensuring GDPR compliance in AI models through targeted unlearning and real-time privacy management. By focusing on modularity, scalability, and efficiency, this framework offers a flexible solution that can be adapted to a variety of AI architectures and regulatory environments. The integration of real-time monitoring, dynamic privacy constraints, and user-defined preferences ensures that privacy is maintained throughout the model's lifecycle, enabling AI systems to operate ethically and responsibly while preserving their performance.

7. Results and Impact

The methodology presented in this paper has been tested across a variety of AI models, including large language models (LLMs) such as LLa MA and GPT-like architectures. The results of these tests have demonstrated significant improvements in the ability of AI systems to comply with GDPR and other privacy regulations, without sacrificing overall model performance. In this section, I discuss the broader implications of these results, focusing on the scalability, efficiency, and transformative potential of the methodology across diverse ap- plications and environments.

7.1 Scalability Across AI Architectures

One of the key findings of this research is the scalability of the proposed frame- work. The methodology was designed to be modular and adaptable, capable of integrating with both large-scale cloud-based systems and smaller, resource- constrained edge devices. This flexibility has proven successful in practice, with the framework being deployed across a wide range of AI architectures.

In large-scale environments, such as cloud-based AI systems handling massive datasets, the framework's distributed privacy management components particularly the Embedding Modification Engine and Real-Time Compliance Module enabled efficient parallel processing of privacy tasks. This ensured that the system could handle high-throughput data streams while maintaining compliance with privacy regulations, even in environments with stringent performance requirements.

For edge devices, where computational resources are more

limited, the lightweight nature of the targeted unlearning process allowed AI models to remain compliant with privacy regulations without incurring significant overhead. This scalability opens up new opportunities for deploying AI models in privacy-sensitive environments such as healthcare, finance, and law enforcement, where the balance between model efficiency and privacy compliance is critical.

7.2 Efficiency and Performance Preservation

Another significant result of this research is the efficiency with which the frame- work performs unlearning operations. Traditional methods for ensuring privacy compliance, such as full retraining or fine-tuning, often involve significant computational resources and time, making them impractical for many real-world applications. In contrast, the proposed methodology enables rapid, targeted unlearning of sensitive data, significantly reducing the time and computational effort required to comply with data privacy requests.

The Embedding Modification Engine, which applies localized adjustments to the model's embeddings, has proven to be a particularly effective solution. By focusing on the specific embeddings associated with sensitive information, the framework avoids the need for costly retraining, while still ensuring that privacy regulations are met. This efficiency translates into substantial cost savings for organizations that deploy AI systems, as they can comply with GDPR and other privacy laws without investing heavily in hardware or computational resources.

Moreover, the results demonstrate that the targeted unlearning process pre- serves the overall performance of the model. Through careful calibration of the unlearning steps, combined with conflict score evaluation to monitor potential side effects, the system ensures that the removal of sensitive data does not degrade the model's accuracy, fluency, or generalization capabilities. This is especially important in applications where AI models must maintain high levels of performance while also adhering to strict privacy requirements.

7.3 Trust and Transparency Through Privacy Management

The integration of real-time privacy management tools, such as the Data Monitoring Module and Real-Time Compliance Module, has had a profound impact on the trustworthiness and transparency of AI systems. These tools ensure that privacy-sensitive data is monitored, flagged, and handled appropriately through- out the entire lifecycle of the model, from training to deployment and inference. This continuous privacy oversight not only supports GDPR compliance but also fosters greater trust in AI systems.

In environments where users and organizations are increasingly concerned about data privacy, the ability to offer real-time, transparent privacy management is a significant advantage. For instance, in industries like healthcare, where the handling of personal health information is strictly regulated, the framework's ability to provide verifiable, auditable records of privacy actions increases confidence in AI systems. The real-time privacy monitoring feature also allows for proactive identification and mitigation of privacy risks, which can reduce the likelihood of privacy violations or data breaches.

Furthermore, the integration of user-defined privacy preferences empowers individuals and organizations to take control of how their data is handled. This shift towards user-centric privacy management aligns with broader trends in AI ethics and governance, where transparency, accountability, and user empowerment are becoming increasingly important. By providing users with the tools to manage their own privacy settings, AI systems become more ethical and aligned with global data protection standards.

7.4 Impact on Ethical AI Governance

Beyond its technical contributions, this research has significant implications for the broader field of AI governance. As AI systems become more integrated into decision-making processes across various sectors, the need for ethical and responsible AI deployment is becoming increasingly urgent. The methodology presented in this paper directly addresses several key concerns in AI ethics, including transparency, accountability, and data privacy.

By providing a framework that ensures compliance with privacy regulations while maintaining model performance, this research contributes to the development of AI systems that are both powerful and responsible. The ability to offer transparent privacy management and customizable privacy settings also supports the growing demand for AI systems that are not only technically robust but also ethically sound. These advancements could lead to new standards for AI governance, where privacy protection is embedded into the fabric of AI systems rather than treated as an afterthought.

Additionally, the ability to provide auditable, blockchain-powered records of privacy actions could set a new precedent for AI accountability. As regulators and policymakers continue to develop frameworks for AI oversight, the trans- parent and decentralized nature of the privacy management system outlined in this research may serve as a model for how AI systems can be held accountable in a variety of legal and ethical contexts.

7.5 Applications in Privacy-Sensitive Industries

The results of this research are particularly relevant to industries that handle sensitive or personal data, such as healthcare, finance, and legal services. In these sectors, the balance between leveraging AI for efficiency and innovation while maintaining compliance with privacy laws is a critical concern. The methodology presented in this paper offers a scalable solution that allows organizations in these industries to benefit from AI technologies without compromising on privacy.

In healthcare, for instance, the ability to unlearn sensitive patient data while preserving the overall functionality of the AI model is invaluable. Healthcare systems often deal with highly sensitive personal data, and any breach of privacy can have severe consequences, both legally and ethically. The real-time privacy management features of the framework enable healthcare

providers to use AI for diagnostics, patient care, and research while ensuring compliance with privacy laws such as HIPAA (Health Insurance Portability and Accountability Act) in the U.S. and GDPR in Europe.

Similarly, in the financial sector, where personal financial data must be handled with extreme caution, the ability to continuously monitor and manage data privacy in AI systems can help mitigate the risk of data breaches and ensure that financial institutions remain compliant with privacy regulations. The framework's ability to integrate with large-scale financial systems and perform unlearning tasks without disrupting operations is a major advantage in this context.

7.6 Future Directions and Long-Term Impact

The results of this research point to several promising avenues for future work. As AI systems continue to evolve, there will be a growing need for more advanced privacy-preserving techniques that can keep pace with the increasing complexity and scale of AI models. Future research could explore the integration of more sophisticated unlearning algorithms, as well as the development of new privacy- preserving techniques that are optimized for specific industries or applications.

Additionally, the long-term impact of this research lies in its potential to influence global standards for AI governance and privacy protection. By demon starting that AI systems can be both powerful and privacy-compliant, this re- search sets the stage for the development of new frameworks for AI regulation that prioritize both performance and ethical responsibility. As organizations around the world adopt AI technologies, the framework presented in this paper could become a blueprint for how to deploy AI systems responsibly, ensuring that privacy and performance are not mutually exclusive.

8. Conclusion

The results of this research have demonstrated that it is possible to create AI systems that are both scalable and compliant with privacy regulations, without sacrificing performance. The methodology's modular, adaptable design allows it to be applied across a wide range of industries and AI architectures, providing a flexible solution for ensuring GDPR compliance in real-time. Beyond its technical contributions, this research has the potential to influence broader trends in AI governance, transparency, and ethics, setting new standards for how AI systems can be deployed responsibly in privacy-sensitive environments.

9. Conclusion

As the adoption of Artificial Intelligence (AI) systems continues to accelerate across a wide range of industries, ensuring compliance with data privacy regulations, such as the General Data Protection Regulation (GDPR), has become a critical challenge. This paper has presented a novel framework for making large-scale AI models, particularly large language models (LLMs), compliant with privacy regulations through a combination of targeted unlearning, real- time privacy management, and user-driven privacy preferences. The methodology addresses one of the most pressing needs in AI development: ensuring that AI systems can be both highly functional and legally compliant in environments where privacy is of paramount concern.

9.1 Key Contributions

The key contribution of this research is the development of a modular and scalable privacy management framework that enables AI models to dynamically manage sensitive information without requiring costly retraining or performance sacrifices. By focusing on targeted unlearning an efficient process that allows the selective removal of specific data while preserving overall model performance the framework ensures that models can comply with privacy regulations such as the GDPR's "right to be forgotten" without diminishing their utility.

The research also introduces real-time monitoring mechanisms that ensure privacy-sensitive data is identified and handled appropriately throughout the model's lifecycle. This real-time capability is crucial for AI systems that operate in dynamic environments where data is continuously ingested and processed. The integration of user-defined privacy preferences further empowers users by giving them more control over their personal data, aligning the framework with emerging trends in AI ethics and user-centric privacy governance.

In addition to its technical innovations, this work makes significant contributions to the broader conversation on ethical AI development and governance. The ability to ensure privacy compliance while maintaining transparency, ac- countability, and performance sets new standards for how AI systems can be deployed responsibly in both public and private sectors.

9.2 Addressing Industry Challenges

This research directly addresses some of the most significant challenges faced by industries that handle sensitive or personal data, including healthcare, finance, and legal services. In these industries, the risks associated with data breaches, privacy violations, or regulatory non-compliance are particularly high, making the need for robust privacy management systems a priority. By providing a flexible and scalable solution for privacy management, the framework presented in this paper enables organizations to leverage AI technologies while maintaining strict adherence to data protection laws.

For example, in healthcare, where patient data is highly sensitive and strictly regulated by laws such as HIPAA and GDPR, the ability to continuously monitor and unlearn specific data ensures that AI models can be used for tasks such as diagnostics, treatment recommendations, or patient management without compromising patient privacy. Similarly, in the financial sector, where personal and financial data must be handled with extreme care, this framework pro- vides a solution for managing the complexities of privacy compliance while still benefiting from the powerful analytical capabilities of AI.

9.3 Impact on AI Governance and Ethics

Beyond the technical realm, the framework presented in this paper also con- tributes to the evolving discourse on AI governance and ethics. As AI systems become more integrated into decision-making processes and societal infrastructure, questions about how these systems handle personal data are becoming more pressing. This research provides a concrete solution for ensuring that AI systems respect individual privacy rights, meet regulatory requirements, and remain accountable to both users and governing bodies.

The inclusion of user-defined privacy preferences is a particularly important step toward more ethical AI systems. By allowing individuals and organizations to set their own privacy parameters, the framework shifts control over personal data back to the users. This shift aligns with broader movements in AI ethics that call for greater transparency, accountability, and user empowerment. In the long run, the adoption of frameworks like the one proposed in this paper could influence how privacy regulations are enforced and how AI governance is structured globally.

9.4 Scalability and Future Directions

Another significant aspect of this research is its emphasis on scalability. The framework's modular architecture ensures that it can be adapted for a variety of AI systems, from large-scale cloud-based architectures to smaller edge devices. This flexibility means that the framework can be deployed in a wide range of environments, supporting the development of AI models in both resource-rich and resource-constrained settings.

The scalability of the methodology opens the door for further innovations in privacy-aware AI systems. Future research could explore how this framework can be integrated with even more complex AI models or deployed in specialized environments such as IoT (Internet of Things) systems, smart cities, or autonomous vehicles. As AI technology continues to evolve, there will be increasing opportunities to refine and expand upon the privacy-preserving mechanisms presented in this paper.

In addition, future work could explore how this methodology could be ex- tended beyond GDPR compliance to meet the requirements of other global privacy regulations, such as the California Consumer Privacy Act (CCPA) or Brazil's General Data Protection Law (LGPD). These regulations are becoming more prevalent, and AI systems will need to adapt to an increasingly complex global regulatory landscape.

9.5 The Broader Implications of Privacy in AI

This research highlights the broader implications of privacy in AI, not just as a legal requirement but as a fundamental principle in the development of ethical and trustworthy AI systems. As AI becomes more deeply integrated into daily life and decision-making processes, ensuring that these systems respect privacy rights will be essential to maintaining public trust in AI technologies.

The proposed framework addresses the challenge of embedding privacy considerations directly into the core of AI systems, rather than treating them as external constraints. This proactive approach is likely to shape the future of AI development, as more organizations and policymakers recognize the importance of building AI systems that prioritize user privacy from the outset.

In the long term, privacy-preserving AI technologies, such as the one pro- posed in this paper, could play a central role in shaping how AI systems are governed and regulated. By providing a transparent and accountable framework for managing personal data, this research contributes to the ongoing effort to create AI systems that are not only powerful but also aligned with the ethical and legal standards of the societies in which they operate.

9.6 Final Thoughts

In conclusion, this research has presented a comprehensive framework for ensuring GDPR compliance in AI systems through targeted unlearning, real-time privacy management, and user-defined privacy preferences. The results demonstrate that it is possible to create AI models that are both scalable and efficient while adhering to stringent privacy regulations. Furthermore, the broader impact of this research extends beyond technical contributions, offering valuable insights into how privacy-preserving AI systems can influence the future of AI governance, ethics, and regulation. As AI technologies continue to evolve, the solutions proposed in this paper provide a critical foundation for ensuring that privacy remains a central focus in the development and deployment of AI systems [5-7].

References

1. Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2021). Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.
2. Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems, 35,* 17359-17372.
3. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2426-2436).
4. Li, Z., Zhang, N., Yao, Y., Wang, M., Chen, X., & Chen, H. (2023). Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129.*
5. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667.*
6. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
7. De Cao, N., Aziz, W., & Titov, I. (2021). Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164.*