A Survey of Large Language Models Privacy: Machine Unlearning, Factual Leakage, Knowledge Retention, and Data Privacy

Abstract

This survey paper explores the intersection of large language models (LLMs) with privacy concerns, emphasizing the need for machine unlearning, factual leakage management, knowledge retention, and data privacy. LLMs, while transformative in fields like healthcare and cybersecurity, pose significant privacy risks due to their reliance on extensive datasets, which may include sensitive information. This paper discusses the critical role of machine unlearning in adhering to privacy regulations by removing specific data from models. It also addresses factual leakage, where LLMs unintentionally expose private data, and the challenge of balancing knowledge retention with data privacy. The survey systematically reviews privacy challenges, including adversarial attacks and ethical considerations, and evaluates current privacy-preserving techniques such as federated learning and differential privacy. Through case studies in healthcare and financial crime detection, the paper highlights the practical implications of LLMs and the necessity for robust privacy measures. Future research directions are proposed, focusing on enhancing privacy-preserving techniques, addressing LLM vulnerabilities, and exploring ethical implications. The paper underscores the importance of developing comprehensive security measures and ethical frameworks to ensure the safe and effective deployment of LLMs while maintaining user privacy.

1 Introduction

1.1 The Rise and Impact of Large Language Models

The emergence of large language models (LLMs) marks a pivotal advancement in natural language processing (NLP), driving progress across diverse sectors. Characterized by their extensive architectures and ability to produce human-like text, LLMs play vital roles in cybersecurity, healthcare, and legal compliance. In cybersecurity, LLMs enhance honeypot effectiveness, adapting to evolving threats and surpassing traditional methodologies [1]. In healthcare, the transition to Multimodal Large Language Models (MLLMs) expands their application scope, providing improved data processing capabilities [2].

LLMs are increasingly integrated into decision-making processes, often requiring extensive fine-tuning and access to proprietary model weights. Integrating knowledge graphs with LLMs aims to reduce hallucinations, thereby enhancing the accuracy and reliability of generated information [3]. However, challenges persist in managing retrieved information, particularly distinguishing accurate from inaccurate content, which is crucial for applications demanding precision. The static nature of training data limits LLMs' ability to incorporate real-time and private information.

Moreover, advancements in LLMs have transformed automation, particularly in code generation and patching. Despite this progress, LLMs frequently retain outdated knowledge and exhibit factual inaccuracies, leading to erroneous outputs in critical fields such as medical diagnostics and legal

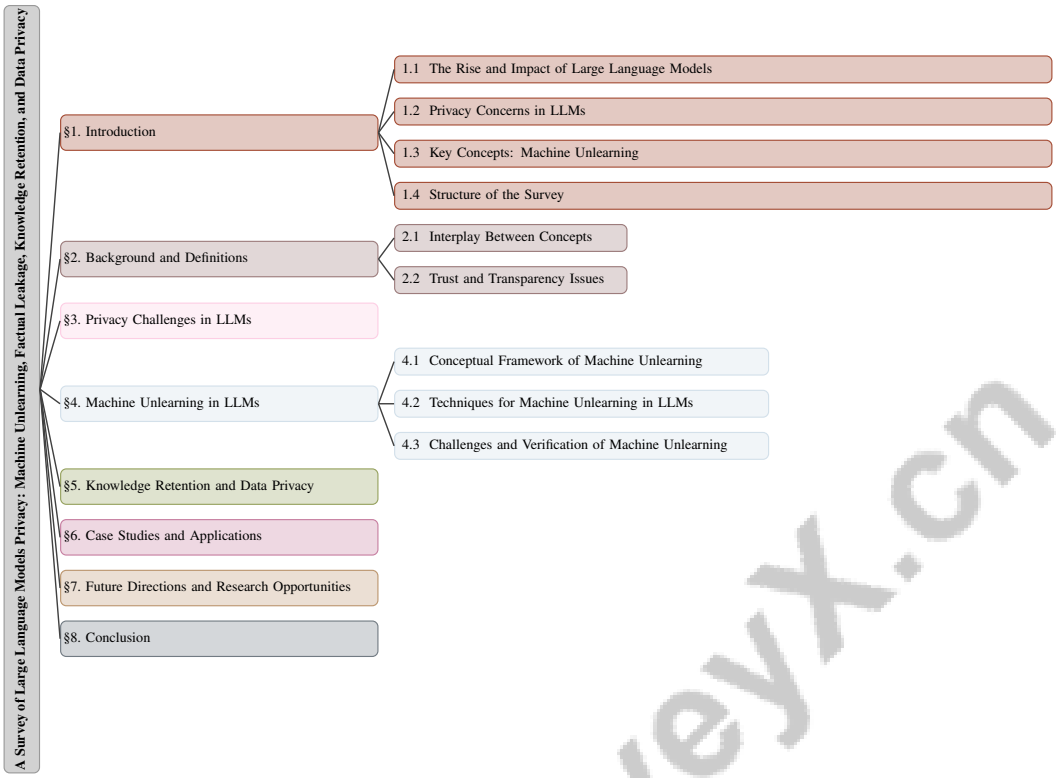


Figure 1: chapter structure

consultations. Research shows that LLMs can adopt incorrect information from retrieved content over 60

As LLMs evolve, addressing the challenge of modifying specific factual knowledge without compromising overall performance is essential. Memory Editing (ME) has emerged as an effective strategy to correct erroneous facts or introduce new information, reflecting the dynamic nature of these models in the AI landscape. Furthermore, LLMs' inconsistencies in providing answers to semantically equivalent prompts highlight their limitations in generalizing factual knowledge across lexical variations. Establishing standardized metrics for evaluating LLM intelligence and performance is crucial to accurately assess their capabilities, given the complexities of translating expert knowledge into quantifiable features and addressing risks associated with benchmark leakage that may lead to misleading evaluations. Systematic development of these metrics can ensure fair model performance comparisons and enhance decision-making accuracy across applications, bridging the gap between human expertise and advanced predictive analytics [4, 5, 6]. The role of LLMs in knowledge-intensive tasks, such as open-domain question answering and knowledge-grounded dialogue, underscores their increasing significance in contemporary information processing.

1.2 Privacy Concerns in LLMs

The deployment of large language models (LLMs) has heightened privacy concerns due to the extensive datasets used for training, which often contain sensitive personal information [7]. A significant issue is the phenomenon of hallucinations, where LLMs generate plausible yet factually incorrect information, undermining trust and applicability, especially in critical fields like healthcare and finance [3]. The lack of standardized processes for reporting AI system flaws exacerbates these challenges, potentially leading to unsafe deployments and eroded user trust [8].

Privacy threats manifest through membership inference attacks (MIAs) and reconstruction attacks, which can disclose whether specific data points were part of the training dataset, thereby compromising data privacy [9]. Malicious parties can reconstruct sensitive user inputs from shared gradient information in collaborative learning frameworks, contradicting the privacy-preserving goals of such systems [10].

The ethical implications of LLMs' inference capabilities further raise privacy concerns, as these models can inadvertently infer sensitive information from seemingly innocuous inputs [10]. The complexity of privacy policy documents, particularly post-GDPR, complicates user comprehension and can lead to privacy leakage and legal issues [11].

To address these challenges, machine unlearning has emerged as a critical requirement due to privacy regulations like the Right to be Forgotten, allowing individuals to revoke consent for their data used in machine learning. This underscores the need for continued research and the advancement of sophisticated privacy-preserving techniques that effectively address both technical challenges and ethical considerations in our data-driven society. Given the pervasive threats to personal information from powerful entities for commercial and surveillance purposes, exploring innovative solutions, such as enhanced data anonymization and encryption algorithms, is essential. Additionally, as regulatory frameworks like the EU's GDPR evolve, it is vital to critically assess their effectiveness and address emerging challenges in privacy protection, particularly amid rapid advancements in artificial intelligence and datafication practices [12, 13, 14].

1.3 Key Concepts: Machine Unlearning, Factual Leakage, Knowledge Retention, and Data Privacy

Understanding the intricate relationship between large language models (LLMs) and privacy necessitates familiarity with several fundamental concepts: machine unlearning, factual leakage, knowledge retention, and data privacy. Machine unlearning refers to selectively removing specific data from a model's training set to comply with privacy regulations like the Right to be Forgotten, ensuring user privacy while maintaining model performance [13]. Effective machine unlearning techniques must achieve the desired data removal without compromising model accuracy [15].

Factual leakage in LLMs involves the unintended exposure of sensitive or confidential information during model operation. This can occur when models generate outputs that inadvertently disclose private training data or when subjected to adversarial attacks designed to extract such information [16]. Addressing factual leakage is vital for safeguarding user data and preserving trust in LLM applications.

Knowledge retention refers to LLMs' ability to maintain useful learned information while implementing mechanisms like machine unlearning. This balance involves removing specific data while preserving overall model utility and accuracy [17]. Techniques such as watermarking have been proposed to ensure that models retain knowledge without compromising data integrity, even under adversarial conditions [18].

Data privacy encompasses protecting personal or sensitive data from unauthorized access or misuse. In the context of LLMs, this involves implementing robust privacy-preserving algorithms that facilitate valuable data analysis without compromising individual privacy [13]. The dynamic nature of privacy risks and the complexity of interpreting privacy regulations further complicate the implementation of effective data privacy measures [19]. As LLMs evolve, ensuring data privacy remains a pivotal challenge that necessitates ongoing research and innovation.

1.4 Structure of the Survey

This survey is systematically organized to provide a comprehensive overview of privacy concerns associated with large language models (LLMs) and corresponding mitigation strategies. It begins with an **Introduction** that discusses the rise and impact of LLMs across various domains, emphasizing critical privacy concerns. The paper delves into essential concepts such as machine unlearning, addressing challenges posed by regulations like GDPR and CCPA regarding the right to be forgotten; factual leakage, which pertains to unintentional disclosures of sensitive information in model outputs; knowledge retention, focusing on how models maintain learned information while complying with privacy mandates; and data privacy, which encompasses the broader implications of safeguarding personal information in LLM contexts. These concepts are crucial for understanding the privacy dynamics inherent to LLMs and their applications in sensitive domains [20, 21, 22].

Following the introduction, the survey explores the **Background and Definitions** section, providing detailed explanations of core concepts and their interrelations to establish a foundational understanding of their relevance to privacy in LLMs.

AI-generated, for reference only.

The subsequent section, **Privacy Challenges in LLMs**, examines the various challenges associated with maintaining privacy in the context of LLMs, including data leakage, unauthorized access, and factual leakage, alongside the difficulties in implementing effective privacy-preserving techniques. It also addresses LLM vulnerabilities to adversarial attacks and the ethical and legal considerations involved.

In **Machine Unlearning in LLMs**, the survey investigates the concept of machine unlearning and its application in LLMs, discussing the theoretical framework, techniques, and challenges associated with ensuring effective unlearning. This section emphasizes various verification methods for machine unlearning, highlighting their critical role in maintaining the integrity and reliability of machine learning models post-data removal, particularly in light of privacy regulations and the potential for model providers to circumvent these verification strategies. This underscores the necessity for robust verification mechanisms to prevent dishonesty and maintain user trust in machine learning applications [23, 24, 25, 26, 27].

The section on **Knowledge Retention and Data Privacy** focuses on balancing the retention of useful knowledge with protecting data privacy in LLMs. It identifies challenges in achieving this balance and explores various techniques for privacy protection while ensuring knowledge retention.

Case Studies and Applications present real-world applications of LLMs with a focus on privacy concerns, providing examples from fields such as healthcare and financial crime detection. This section illustrates the practical implications of privacy issues and the necessity for privacy-preserving techniques in diverse applications.

The survey concludes with a section titled **Future Directions and Research Opportunities**, highlighting existing gaps in research related to LLMs and proposing specific areas for future investigation. This includes exploring effective strategies for knowledge extraction from LLMs, enhancing table-to-text generation capabilities, evaluating the comparative effectiveness of search engines and LLMs for health-related queries, and addressing the critical need for transparency in LLM applications to ensure responsible AI deployment. By identifying these key areas, the survey aims to guide researchers toward impactful contributions that can advance the field [28, 29, 30, 31]. It emphasizes the need for enhanced privacy-preserving techniques, strategies to mitigate vulnerabilities, and exploration of the ethical and societal implications of LLMs and privacy. The **Conclusion** summarizes the key points discussed and underscores the importance of addressing privacy concerns in the ongoing development of LLMs.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Interplay Between Concepts

Understanding the interplay between large language models (LLMs) and privacy-related concepts such as machine unlearning, factual leakage, knowledge retention, and data privacy is essential for grasping the privacy dynamics in LLMs. Machine unlearning facilitates the removal of specific data from a model's training set to comply with regulations like the Right to be Forgotten, aiming to preserve user privacy without compromising model performance [32]. However, balancing effective unlearning with the retention of valuable knowledge remains challenging, impacting the applicability of LLMs across various domains [33].

Factual leakage, a significant concern, occurs when LLMs inadvertently disclose sensitive information during operation. This issue can be exacerbated by optimization pressures leading to unintended behaviors, such as covert collusion through steganography [34]. Addressing factual leakage is crucial for safeguarding user data and maintaining trust in AI systems.

Knowledge retention is interconnected with machine unlearning and factual leakage, focusing on preserving learned information while implementing privacy-preserving measures. Continual learning, which allows LLMs to adapt to new data without revisiting previous data, risks catastrophic forgetting [35]. Techniques such as selective distillation of attention weights have been proposed to mitigate this forgetting and enhance knowledge retention [36].

Data privacy, which involves protecting personal or sensitive data from unauthorized access, is fundamental in LLM deployment. The trade-off between privacy risk and model performance presents a significant challenge, as methods relying solely on local models for data generation

may degrade performance [33]. Ensuring model parameter stability post-pretraining is crucial for identifying base models without compromising sensitive information [37].

The complexity and unpredictability of LLM behaviors, coupled with the opaque nature of their architectures, complicate transparency and trust in these models [31]. The rapid evolution of applications and proprietary technologies that limit access to model internals further complicate transparency requirements. Existing benchmarks often focus on specific tasks and lack adaptability to real-world applications, limiting the effectiveness of table-to-text generation models [28]. Observations indicate that previous datasets failed to enhance LLM capabilities due to reliance on incorrect or fabricated information [38].

2.2 Trust and Transparency Issues

The deployment of LLMs across various applications raises significant challenges related to trust and transparency. A critical issue is the lack of trust between data owners and researchers in machine learning workflows, especially concerning sensitive data. This distrust arises from concerns about data misuse and the opacity of LLM processes, which obscure data handling and protection mechanisms [39].

Traditional certification processes struggle to adapt to the complexities of machine learning, often lacking rigorous definitions for ML-specific properties. The recursive nature of ML application structures complicates certification, hindering the establishment of trustworthiness [40]. Additionally, inherent biases in LLMs and the opacity of their functioning pose significant challenges to accountability for their outputs, leaving users unaware of operational mechanisms and raising ethical deployment concerns [41].

Transparency issues are further exacerbated by the absence of robust infrastructures for responsible flaw disclosure. The lack of standardized processes for reporting AI system flaws can lead to unsafe deployments and erode user trust [8]. Moreover, opaque data practices may leave users unaware of data harvesting activities, raising ethical implications regarding the manipulation of vulnerable individuals into disclosing personal information [42].

In federated learning, distributional shifts in local data across clients present a core challenge, causing the global model to forget previously learned information outside the local distribution. This underscores the difficulty of maintaining a coherent and reliable global model, impacting both trust and transparency [43]. The need to retain knowledge from prior tasks while adapting to new ones in continual learning further complicates the trust landscape, as accessing upstream data can be challenging and computationally expensive [35].

Addressing trust and transparency issues in LLMs requires a multifaceted approach, including the development of robust certification processes, enhancement of transparency in data practices, and the promotion of a culture of responsible reporting and accountability. These efforts are crucial for establishing and sustaining user trust in LLM applications, tackling significant ethical challenges such as privacy, accountability, and bias reduction, while ensuring the accuracy and reliability of provided information. Implementing tailored ethical frameworks and dynamic auditing systems can help navigate the complexities unique to LLMs, fostering transparency and upholding ethical standards as these technologies increasingly influence information dissemination [41, 22].

In recent years, the development of large language models (LLMs) has been accompanied by a growing recognition of the various privacy challenges they pose. To better understand these challenges, it is essential to examine their hierarchical structure, as depicted in Figure 2. This figure illustrates the categorization of privacy challenges into three primary domains: privacy-preserving techniques, vulnerabilities to adversarial attacks, and ethical and legal considerations. Each category is further dissected into specific issues and strategies, thereby highlighting the complexity and multifaceted nature of ensuring privacy and security in LLMs. Such a comprehensive overview not only aids in identifying the key areas of concern but also serves as a foundation for developing effective mitigation strategies.

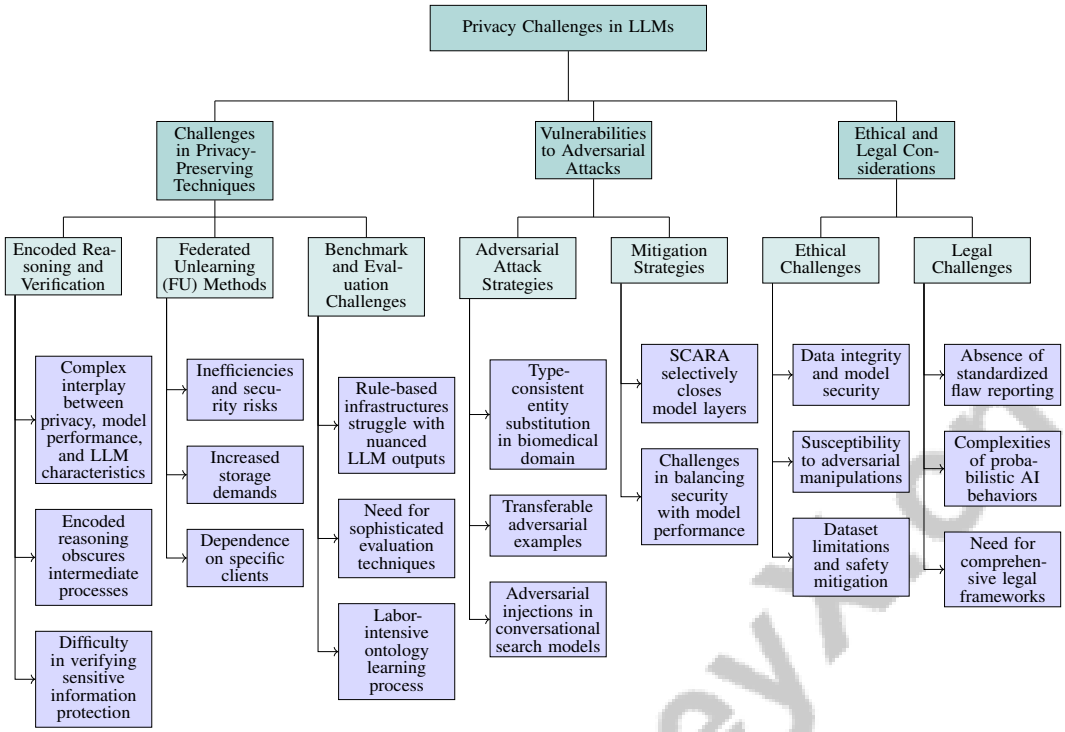


Figure 2: This figure illustrates the hierarchical structure of privacy challenges in large language models (LLMs), categorizing them into challenges in privacy-preserving techniques, vulnerabilities to adversarial attacks, and ethical and legal considerations. Each category is further broken down into specific issues and strategies, highlighting the complexity and multifaceted nature of ensuring privacy and security in LLMs.

3 Privacy Challenges in LLMs

3.1 Challenges in Privacy-Preserving Techniques

Implementing privacy-preserving techniques in large language models (LLMs) involves navigating the complex interplay between privacy, model performance, and LLM-specific characteristics. A major challenge is the encoded reasoning employed by LLMs, which obscures intermediate processes and complicates the verification of sensitive information protection [44]. This issue is visually represented in Figure 3, which illustrates the primary challenges in implementing privacy-preserving techniques within LLMs. The figure categorizes these challenges into three main areas: encoded reasoning, federated unlearning, and LLM credence uncertainty. Each category highlights specific issues such as reasoning obscurity, client dependency, and credence attribution, which complicate the development of effective privacy-preserving strategies.

Federated Unlearning (FU) methods, often reliant on specific clients, face inefficiencies, security risks, and increased storage demands [45]. The uncertainty of LLM credences and the reliability of methods to assess them further complicate privacy-preserving strategy development [46].

Current benchmarks and anti-phishing infrastructures, primarily rule-based, struggle to detect the nuanced outputs of LLMs, highlighting the need for more sophisticated evaluation techniques [47, 48]. The labor-intensive ontology learning process, requiring domain-specific knowledge and extensive datasets, adds to the challenges [49]. Moreover, benchmarks often focus on a narrow range of self-trained models, failing to capture the full vulnerability landscape of publicly available models [9]. The widespread vulnerabilities across models, which inadequately reject harmful queries, underscore the urgent need for enhanced safety measures [50]. Additionally, current methods struggle to prevent user input reconstruction, particularly in white-box settings where adversaries access internal workings.

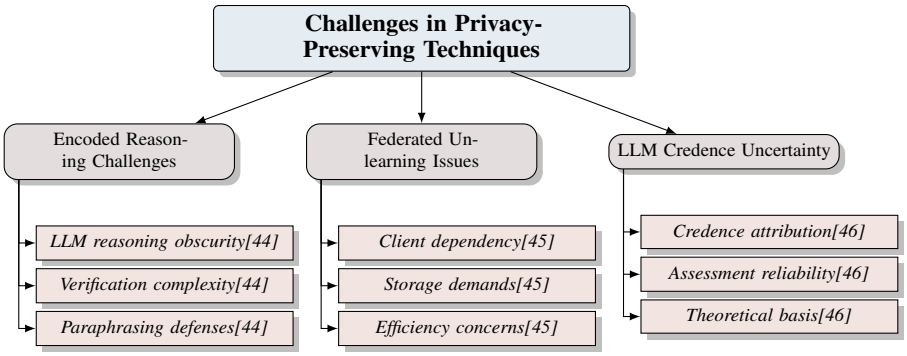


Figure 3: This figure illustrates the primary challenges in implementing privacy-preserving techniques within large language models (LLMs). It categorizes the challenges into three main areas: encoded reasoning, federated unlearning, and LLM credence uncertainty. Each category highlights specific issues such as reasoning obscurity, client dependency, and credence attribution, which complicate the development of effective privacy-preserving strategies.

3.2 Vulnerabilities to Adversarial Attacks

Large language models (LLMs) are increasingly susceptible to adversarial attacks that manipulate outputs to mislead or extract sensitive information, posing significant privacy and security risks. In the biomedical domain, type-consistent entity substitution is used to generate adversarial distractors, rigorously assessing LLM performance in complex tasks like question answering, revealing weaknesses in handling adversarial inputs [51]. Transferable adversarial examples, capable of deceiving multiple models, raise concerns about exploitation across LLM systems [52]. Adversarial injections also influence product rankings in conversational search models, demonstrating the potential for commercial manipulation [53].

To mitigate adversarial attack impacts, strategies like SCARA selectively close model layers to reduce recovery attack success rates, balancing security with model customization [54]. However, developing comprehensive defenses against a wide range of adversarial strategies without sacrificing LLM utility and performance remains a challenge.

3.3 Ethical and Legal Considerations

The ethical and legal challenges in LLM privacy involve data integrity, model security, and broader AI deployment implications. LLMs' vulnerability to malicious actors, through data poisoning or theft, underscores the need for robust security measures against unauthorized access and manipulation [39]. While LLMs show potential in enhancing cybersecurity, particularly in detecting phishing emails, their susceptibility to adversarial manipulations in conversational search engines necessitates further research to improve robustness and protect consumers [53].

Legal challenges are compounded by the absence of standardized flaw reporting processes and the complexities of probabilistic AI behaviors, complicating effective oversight and regulation. Comprehensive legal frameworks are needed to address AI's unique challenges, including accountability and transparency in AI disclosures [8]. Ethically, dataset limitations, such as those identified in the HH dataset audit, reveal significant quality issues and inadequate conceptualization of harmlessness, necessitating more nuanced safety mitigation approaches to ensure adherence to ethical standards of harmlessness and fairness [55]. Additionally, Federated Unlearning (FU) methods face practical limitations due to reliance on client cooperation or additional storage, hindering their efficiency and scalability. Addressing these limitations is essential for developing effective privacy-preserving techniques that align with legal and ethical standards [45].

4 Machine Unlearning in LLMs

4.1 Conceptual Framework of Machine Unlearning

The conceptual framework of machine unlearning in LLMs is pivotal for balancing privacy concerns with model efficacy. It includes innovative methods like Single Layer Unlearning Gradient (SLUG) and computationally efficient techniques for targeted data removal, aligning with privacy regulations such as GDPR and CCPA's "right to be forgotten." These methods allow selective deletion of sensitive information, such as personal identifiers and copyrighted content, without extensive system updates or high computational costs, promoting responsible AI use and legal compliance [14, 20, 56, 21, 57].

Figure 4 illustrates the hierarchical structure of the conceptual framework for machine unlearning, highlighting these innovative methods, knowledge integration, and data and output defenses as key components for balancing privacy and model efficacy. A key aspect is integrating knowledge graphs (KGs) into LLMs, enhancing the retention of relevant knowledge and facilitating unlearning. This integration is categorized by model architecture and KG integration stages, improving the model's ability to manage knowledge [3]. Supervised fine-tuning of LLMs for domain-specific tasks, such as simulating Linux server operations, further refines unlearning capabilities [1].

Data defenses are crucial in modifying input text to prevent LLMs from inferring sensitive information or using copyrighted material, ensuring data privacy during model operations [10]. Additionally, output manipulation techniques, like output scouting, use auxiliary parameters to adjust LLM output probability distributions, enhancing AI safety by identifying and mitigating catastrophic responses [5, 58]. In federated learning, the framework emphasizes identifying and editing LLM layers storing commonsense knowledge for precise unlearning.

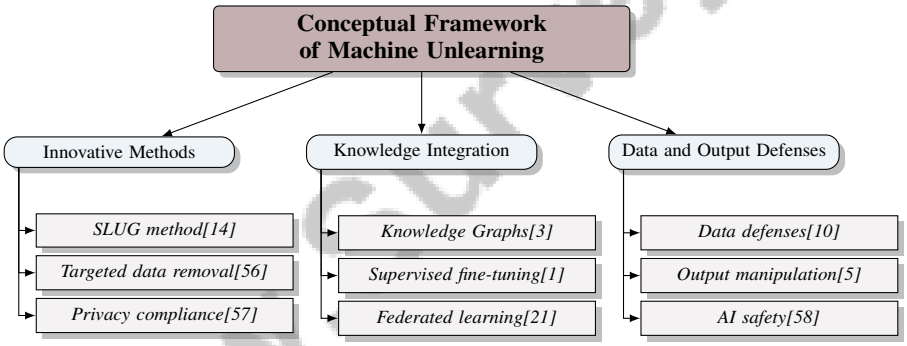


Figure 4: This figure illustrates the hierarchical structure of the conceptual framework for machine unlearning, highlighting innovative methods, knowledge integration, and data and output defenses as key components for balancing privacy and model efficacy.

4.2 Techniques for Machine Unlearning in LLMs

Machine unlearning in LLMs employs advanced techniques to ensure data protection compliance while preserving model performance. Mini-Unlearning, for example, uses contraction mapping to approximate unlearned parameters efficiently, minimizing reliance on historical gradients and facilitating rapid data removal without significant computational costs [59].

The ZeroLeak framework automates the detection and patching of side-channel vulnerabilities with LLMs, enhancing security in machine unlearning [60]. LLMs like BERT and GPT streamline legal compliance by automating legal provision classification, ensuring regulatory adherence [61].

Adversarial prompt injections confuse LLMs to prevent accurate inference of sensitive information, supporting privacy in machine unlearning [10]. The SAPLMA technique uses activation values from LLMs to classify the truthfulness of statements, maintaining output integrity and aiding in unlearning false information [62].

Knowledge graphs are integrated into LLMs at various stages, providing structured knowledge management crucial for refining unlearning by allowing selective knowledge retention and removal [3]. These methodologies highlight the complexity and necessity of implementing machine unlearning in

LLMs. Incorporating advanced privacy-preserving techniques like differential privacy and knowledge distillation, LLMs adapt to strict privacy requirements while maintaining functionality across applications, including medical text classification and on-premises deployment. Robust watermarking frameworks protect intellectual property in generated content, balancing usability with privacy and security [54, 20, 18, 22, 63].

4.3 Challenges and Verification of Machine Unlearning

Benchmark	Size	Domain	Task Format	Metric
KSD[25]	180,000	Image Classification	Classification	KSD, MIA-efficacy
ParaRel*[64]	21,830	Natural Language Processing	Consistency Evaluation	Consistency, Accuracy
CF-Benchmark[65]	100,000	Natural Language Processing	Instruction Following	Accuracy, FG
BL[4]	1,000,000	Question Answering	Question Answering	Accuracy, F1-score
TrustScore[66]	1,000	Question Answering	Open-ended Question Answering	TrustBC, TrustF C
CONCURRENTQA[67]	18,439	Multi-domain Question Answering	Multi-hop Question Answering	EM, F1
REML[68]	1,000	Knowledge Editing	Dialogue Simulation	Accuracy, Reversion
LLM-Factuality[69]	2,250	Text Summarization	Factuality Evaluation	Partial Correlation Coefficient, Spearman

Table 1: Table showcasing a comprehensive overview of various benchmarks used in evaluating machine unlearning and related tasks across different domains. It highlights the size, domain, task format, and metrics associated with each benchmark, providing a critical resource for assessing the effectiveness and challenges in machine unlearning methodologies.

Machine unlearning in LLMs faces challenges in maintaining model performance while adhering to privacy standards. A significant issue is distinguishing between retention and forgetting, which can impact performance [70]. Gradient ascent-based methods struggle to remove well-memorized data due to diminishing gradients post-training, complicating data removal [71]. These methods also inadequately address class removal tasks, indicating a need for more robust solutions [72].

Verification of unlearning success is another challenge. Standard accuracy metrics may not fully capture unlearning nuances, necessitating improved auditing practices. Techniques like output scouting audit LLMs by uncovering catastrophic responses, emphasizing the need for comprehensive verification [58]. Table 1 offers a detailed examination of key benchmarks pertinent to the challenges and verification of machine unlearning in large language models, underscoring the necessity for diverse evaluation metrics and task formats. Benchmarks may not fully address unlearning difficulties across diverse datasets and scenarios, potentially limiting applicability [25].

Generalizability of unlearning methods to other model architectures and long-term stability of unlearning effects require thorough testing. The effectiveness of SCARA depends on maintaining the integrity of crucial bottom layers for model stability [54]. The LLM Factoscope's primary challenge is the assumption that training datasets are predominantly factual, which may not always hold true, impacting performance [73].

Future research should refine dataset preparation techniques and explore diverse prompt templates to mitigate hallucinations and enhance model reliability [6]. Findings indicate that while scaling and retrieval augmentation improve consistency, retrieval augmentation is more efficient, underscoring the need for ongoing exploration of model architectures and training data [64]. Paraphrasing can effectively limit the information LLMs encode in outputs, rendering encoded reasoning largely ineffective [44].

5 Knowledge Retention and Data Privacy

The interplay between knowledge retention and data privacy is pivotal in the deployment of large language models (LLMs) across diverse applications. As organizations increasingly rely on these models, understanding how to balance these objectives becomes crucial. This section explores the challenges of maintaining effective knowledge retention while safeguarding sensitive information, highlighting the complexities involved and paving the way for exploring effective strategies to navigate these issues.

5.1 Challenges in Balancing Knowledge Retention and Data Privacy

Balancing knowledge retention with data privacy in LLMs involves preserving model utility while protecting sensitive information. Inherent biases in LLMs can skew content retrieval, complicating the maintenance of accurate knowledge retention alongside privacy, especially when conflicting information is present [74]. Addressing vulnerabilities to prevent data leakage while ensuring patched models remain secure and performant is another significant challenge [60]. Techniques like Mini-Unlearning show promise in managing high unlearning ratios, enhancing privacy against membership inference attacks without significantly compromising accuracy [59].

In domains such as healthcare, anonymization is crucial for meaningful data analysis while protecting patient identities, necessitating sophisticated methods to ensure data retains analytical utility [75]. Biases in training data further limit LLM performance across contexts, complicating the assurance of reliable knowledge retention while maintaining privacy [48]. The implementation of cost-effective data defenses across various models presents a promising avenue for generating privacy-preserving solutions rapidly and at scale [10].

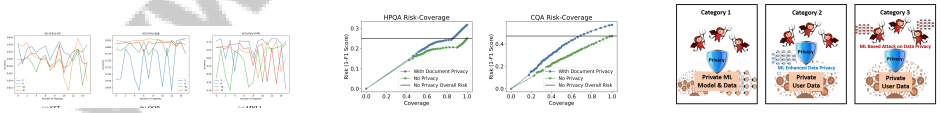
5.2 Techniques for Knowledge Retention and Privacy Protection

Techniques to balance knowledge retention and privacy protection in LLMs focus on preserving model utility while safeguarding sensitive information. Automatic learning and visualization of linguistic vagueness enhance the clarity of privacy policies, improving understanding of privacy implications in LLM outputs [76]. This approach mitigates risks associated with ambiguous language that may expose sensitive information.

Advanced frameworks like LSAST, which combine LLMs with Static Application Security Testing (SAST), improve vulnerability detection accuracy and effectiveness, reinforcing both security and privacy in LLM applications [77]. Federated learning offers a collaborative approach to learning from private text data, facilitating ethical data usage while preserving privacy [14]. Enhancing defenses against adversarial attacks and prioritizing privacy-preserving techniques is critical for bolstering security and trust in federated learning systems [78].

In educational contexts, LLM-generated tips support students learning complex subjects, such as quantum computing, demonstrating LLMs' ability to retain and convey valuable knowledge without relying solely on expert-created content [79]. This capability highlights LLMs' potential to provide educational support while maintaining privacy.

Evaluating the security of LLM applications, including metadata, app descriptions, instructions, and knowledge files, is vital for understanding and mitigating privacy risks [80]. By focusing on these aspects, developers can enhance LLM privacy protection while ensuring the retention of necessary knowledge for effective application performance.



(a) Performance of Different Number of Requests on Different Tasks[21]

(b) Risk-Coverage Analysis for Document Privacy in HPQA and CQA[67]

(c) Categories of Data Privacy and Security[81]

Figure 5: Examples of Techniques for Knowledge Retention and Privacy Protection

As illustrated in Figure 5, the balance between knowledge retention and data privacy is increasingly crucial in data science and machine learning. The figure provides an overview of various techniques and analyses relevant to this topic. The first subfigure shows how different numbers of requests impact task performance, emphasizing the relationship between data handling practices and model accuracy. The second subfigure presents a comparative risk-coverage analysis for document privacy within Human Performance Quality Assessment (HPQA) and Computer Quality Assessment (CQA), offering insights into the trade-offs between risk and privacy coverage. Lastly, the third subfigure categorizes different aspects of data privacy and security, underscoring the importance of protecting private machine learning models and data. Together, these visualizations highlight ongoing challenges

and strategies in maintaining effective knowledge retention while ensuring robust data privacy in the digital age [21, 67, 81].

6 Case Studies and Applications

6.1 Applications and Case Studies

Large language models (LLMs) are increasingly utilized across various domains, each posing unique privacy challenges. In online commerce, models like GPT-3.5 and GPT-4 enhance user interactions through personalized experiences, yet they raise privacy concerns due to extensive data collection [63]. Robust privacy measures are essential to address these issues. Recent advances in computational efficiency have reduced LLM operational costs while maintaining accuracy, enhancing accessibility and sustainability [82]. However, these efficiency gains often involve complex data handling, which can risk exposing sensitive information.

The factuality of LLM-generated outputs is crucial in applications requiring high accuracy. While LLM-generated knowledge generally has lower factuality than retrieved knowledge, it can still enhance downstream task accuracy [83]. Thus, improving factual accuracy without risking data leakage is vital. In healthcare, LLMs like LlamaCare have significantly improved medical question-answering and classification tasks, enhancing diagnostic accuracy and patient care [84]. However, stringent privacy measures are necessary to protect sensitive medical data.

In legal compliance and regulatory analysis, LLMs enhance efficiency and accuracy, reducing manual workloads [61]. However, handling legal data introduces privacy risks, necessitating robust protection strategies to prevent unauthorized access. Moreover, LLMs' ability to identify similar data points in finance and healthcare underscores their potential in pattern recognition and data analysis [17]. These applications highlight the need for privacy-preserving techniques to protect sensitive data during analysis.

6.2 Clinical Applications and Privacy Concerns

LLMs significantly enhance clinical decision-making, patient engagement, and research. Multimodal Large Language Models (MLLMs) show promise in clinical decision support and medical imaging, improving diagnostic accuracy and patient outcomes [2]. Ensuring privacy and security of sensitive patient information is critical. Training datasets, such as MIMIC-III and MIMIC-IV, containing de-identified patient data, require rigorous privacy protections [85]. Maintaining patient confidentiality is ethically imperative, as breaches could have severe consequences.

To address privacy concerns, methods like private text classification are developed, crucial in healthcare where ethical data sharing is vital [14]. These methods protect sensitive information while facilitating meaningful data analysis. The safe integration of LLMs into medical practices requires oversight and validation to prevent misuse and ensure reliable operation [86]. Robust validation frameworks and transparency in LLM operations are essential to foster trust among healthcare professionals and patients.

6.3 Financial Crime Detection and Privacy-Preserving Techniques

LLMs show promise in detecting financial crimes, enhancing traditional methodologies through advanced data processing. They improve entity extraction and disambiguation in complex financial datasets, crucial for identifying corruption and fraud patterns [87]. A critical consideration is the privacy of sensitive financial data. The LTU Attacker experiments highlight the necessity of privacy-preserving techniques during model training, such as avoiding overfitting and introducing randomness [88]. These strategies mitigate the risk of membership inference attacks.

Integrating LLMs in financial crime detection requires balancing high detection accuracy with robust data privacy. Techniques like differential privacy and federated learning enable models to learn from distributed datasets without exposing individual data points. These advanced methods empower LLMs to detect financial crimes effectively by accurately extracting entities from complex corruption schemes while employing robust watermarking strategies to protect sensitive financial information from unauthorized access and exploitation [87, 18].

7 Future Directions and Research Opportunities

7.1 Enhancing Privacy-Preserving Techniques

Advancing privacy-preserving techniques in large language models (LLMs) is essential for their secure deployment across domains. Future research should focus on improving leakage detection tools and enhancing LLM capabilities to protect privacy, as demonstrated by the ZeroLeak framework [60]. Developing unified standards for data sharing and exploring technologies such as Federated Learning and blockchain are crucial for addressing privacy challenges [75]. Refining training data diversity and improving detectors to identify harmful language are vital for enhancing LLM safety and reliability [48]. Expanding research to include additional attack types and diversifying datasets will strengthen privacy-preserving strategies [9]. Establishing standardized guidelines for flaw reporting and integrating interdisciplinary expertise into adjudication processes will promote open communication between AI developers and the community, enhancing trust and transparency in LLM applications [8]. Moreover, optimizing secure computation techniques for privacy-preserving machine learning is a promising research avenue [7].

7.2 Addressing Vulnerabilities in LLMs

Addressing vulnerabilities in LLMs is critical, particularly in sensitive applications where security and reliability are paramount. Sophisticated attackers can exploit inherent weaknesses in LLMs, such as in cybersecurity applications like honeypots where scripted behaviors can be detected [1]. Strategies to enhance LLM robustness against adversarial attacks are essential, including improving detection and prevention mechanisms for adversarial inputs designed to manipulate outputs or extract information. Adversarial training can significantly bolster LLMs' capacity to counter cyber threats, including misinformation [89, 22]. Implementing anomaly detection systems to monitor LLM behavior for deviations can enhance security by identifying potential vulnerabilities and malicious activities [63, 1]. Techniques such as differential privacy ensure models do not inadvertently disclose sensitive information under sophisticated attacks. Employing ensemble methods, where multiple models operate in parallel, enhances LLM resilience by reducing the likelihood of exploiting a single point of failure. This strategy improves decision-making and increases resistance to adversarial strategies, with techniques like high-entropy soft labels and regularization maintaining accuracy while minimizing privacy breach risks [9, 90, 88].

7.3 Exploring Ethical and Societal Implications

Exploring the ethical and societal implications of LLMs and MLLMs requires understanding potential biases and impacts across domains. As LLMs are deployed in sensitive areas like healthcare, addressing gaps in understanding their long-term impacts is essential [91]. The integration of LLMs in healthcare raises ethical considerations, including biases in training data and models, which can perpetuate inequalities and lead to unintended consequences [5]. Developing evolving ethical frameworks is vital to keep pace with technological advancements in LLMs, emphasizing continuous evaluation and interdisciplinary collaboration to embed ethical considerations in their design and deployment [41]. Robust evaluation frameworks are needed to assess the ethical implications of MLLMs, addressing unresolved questions about biases and societal impacts [2]. Decentralized identity solutions offer a promising approach to enhancing trust in machine learning applications, providing a pathway to address privacy concerns while maintaining data integrity and security [39]. Leveraging such solutions can foster greater transparency and accountability in LLM deployment, mitigating ethical and societal risks associated with these models.

8 Conclusion

The examination of large language models (LLMs) underscores the pressing need to tackle privacy issues as these technologies become increasingly integral to various sectors. Notable security vulnerabilities, such as those identified in systems like OpenAI's GPT-4, highlight the critical requirement for robust security frameworks within the LLM landscape. Differential privacy has gained prominence as an effective approach, facilitating significant data analysis while safeguarding individual privacy. Models like CADP-LM have set new standards by enhancing privacy protection

without compromising functionality, demonstrating the feasibility of privacy-preserving language models.

The WaterJudge framework exemplifies the delicate balance between quality detection and watermarking in LLMs, emphasizing the necessity of addressing privacy concerns to maintain the integrity of LLM outputs. The exploration of LLM app ecosystems has unveiled considerable security risks, with many applications containing harmful content and privacy breaches, underscoring the need for stringent regulatory oversight and improved security measures.

Ethical considerations are paramount; evidence suggests that while platforms such as Replika claim GDPR compliance, their practices raise significant concerns about user data privacy and the potential exploitation of susceptible groups. The persistent issue of hallucinations in LLMs, despite detailed surveys on mitigation strategies, remains a challenge. Additionally, the focus on dataset creation and specialized evaluations predominantly in English reveals substantial deficiencies in non-English representation, necessitating a more inclusive approach.

The risk posed by LLM-generated phishing emails is substantial, with a notable proportion of recipients susceptible to these advanced threats, underscoring the need for enhanced detection and prevention mechanisms. The LLM Factoscope offers a promising method to improve the reliability and transparency of LLMs through internal state analysis, thereby advancing the accuracy of factual outputs.

References

[1] Hakan T. Otal and M. Abdullah Canbaz. Llm honeypot: Leveraging large language models as advanced interactive honeypot systems, 2024.

[2] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[3] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[4] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

[5] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[6] Mathav Raj J, Kushala VM, Harikrishna Warrier, and Yogesh Gupta. Fine tuning llm for enterprise: Practical guidelines and recommendations, 2024.

[7] Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pages 19–38. IEEE, 2017.

[8] Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, Will Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, Michelle Hoang, Sven Cattell, and Nouha Dziri. To err is ai : A case study informing llm flaw reporting practices, 2024.

[9] Boyang Zhang, Zheng Li, Ziqing Yang, Xinlei He, Michael Backes, Mario Fritz, and Yang Zhang. Securitynet: Assessing machine learning vulnerabilities on public models, 2023.

[10] William Agnew, Harry H. Jiang, Cella Sum, Maarten Sap, and Sauvik Das. Data defenses against large language models, 2024.

[11] Chaochao Chen, Jamie Cui, Guanfeng Liu, Jia Wu, and Li Wang. Survey and open problems in privacy preserving knowledge graph: Merging, query, representation, completion and applications, 2020.

[12] Marco Cremonini. A critical take on privacy in a datafied society, 2023.

[13] Le Yang, Miao Tian, Duan Xin, Qishuo Cheng, and Jiajian Zheng. Ai-driven anonymization: Protecting personal data privacy while leveraging machine learning, 2024.

[14] Leif W. Hanlen, Richard Nock, Hanna Suominen, and Neil Bacon. Private text classification, 2018.

[15] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024.

[16] Zhuowen Yuan, Fan Wu, Yunhui Long, Chaowei Xiao, and Bo Li. Secretgen: Privacy recovery on pre-trained models via distribution discrimination, 2022.

[17] Xianlong Zeng, Yijing Gao, Fanghao Song, and Ang Liu. Similar data points identification with llm: A human-in-the-loop strategy using summarization and hidden state insights, 2024.

[18] Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. Remark-llm: A robust and efficient watermarking framework for generative large language models, 2024.

[19] Chenhao Fang, Derek Larson, Shitong Zhu, Sophie Zeng, Wendy Summer, Yanqing Peng, Yuriy Hulovatyy, Rajeev Rao, Gabriel Forgues, Arya Pudota, Alex Goncalves, and Hervé Robert. Ingest-and-ground: Dispelling hallucinations from continually-pretrained llms with rag, 2024.

[20] Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification, 2024.

[21] Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. Privacy adhering machine un-learning in nlp, 2022.

[22] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. Can large language models provide security privacy advice? measuring the ability of llms to refute misconceptions, 2023.

[23] Kongyang Chen, Zixin Wang, Bing Mi, Waixi Liu, Shaowei Wang, Xiaojun Ren, and Jiaxing Shen. Machine unlearning in large language models, 2024.

[24] Àlex Pujol Vidal, Anders S. Johansen, Mohammad N. S. Jahromi, Sergio Escalera, Kamal Nasrollahi, and Thomas B. Moeslund. Verifying machine unlearning with explainable ai, 2024.

[25] Mahtab Sarvmaili, Hassan Sajjad, and Ga Wu. Towards understanding the feasibility of machine unlearning, 2024.

[26] Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile, 2024.

[27] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Wei Zhao. Towards efficient target-level machine unlearning based on essential graph, 2024.

[28] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios, 2023.

[29] Marcos Fernández-Pichel, Juan C. Pichel, and David E. Losada. Search engines, llms or both? evaluating information seeking strategies for answering health questions, 2024.

[30] Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, and Yoshitaka Ushiku. Where is the answer? investigating positional bias in language model knowledge extraction, 2024.

[31] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[32] Stefan Rass, Sandra König, Jasmin Wachter, Manuel Egger, and Manuel Hobisch. Supervised machine learning with plausible deniability, 2021.

[33] Wenhao Wang, Xiaoyu Liang, Rui Ye, Jingyi Chai, Siheng Chen, and Yanfeng Wang. Knowledgesg: Privacy-preserving synthetic text generation with knowledge distillation from server, 2024.

[34] Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence mitigation of steganographic collusion in llms, 2024.

[35] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey, 2024.

[36] Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. Seekr: Selective attention-guided knowledge retention for continual learning of large language models, 2024.

[37] Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. Huref: Human-readable fingerprint for large language models, 2025.

[38] Li Zeng, Yingyu Shan, Zeming Liu, Jiashu Yao, and Yuhang Guo. Fame: Towards factual multi-task model editing, 2024.

[39] Will Abramson, Adam James Hall, Pavlos Papadopoulos, Nikolaos Pitropakis, and William J Buchanan. A distributed trust framework for privacy-preserving machine learning, 2020.

[40] Marco Anisetti, Claudio A. Ardagna, Nicola Bena, and Ernesto Damiani. Rethinking certification for trustworthy machine learning-based applications, 2023.

[41] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[42] Joni-Roy Piispanen, Tinja Myllyviita, Ville Vakkuri, and Rebekah Rousi. Smoke screens and scapegoats: The reality of general data protection regulation compliance – privacy and ethics in the case of replika ai, 2024.

[43] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning, 2022.

[44] Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning, 2023.

[45] Ruinan Jin, Minghui Chen, Qiong Zhang, and Xiaoxiao Li. Forgettable federated linear learning with certified data unlearning, 2024.

[46] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[47] Mazal Bethany, Athanasios Galiopoulos, Emet Bethany, Mohammad Bahrami Karkevandi, Nishant Vishwamitra, and Peyman Najafirad. Large language model lateral spear phishing: A comparative study in large-scale organizational settings, 2024.

[48] Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. Detectors for safe and reliable llms: Implementations, uses, and limitations, 2024.

[49] Rick Du, Huilong An, Keyu Wang, and Weidong Liu. A short review for ontology learning: Stride to large language models trend, 2024.

[50] Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. Toolsword: Unveiling safety issues of large language models in tool learning across three stages, 2024.

[51] R. Patrick Xian, Alex J. Lee, Satvik Lolla, Vincent Wang, Qiming Cui, Russell Ro, and Reza Abbasi-Asl. Assessing biomedical knowledge robustness in large language models by query-efficient sampling attacks, 2024.

[52] Zelin Li, Kehai Chen, Lemao Liu, Xuefeng Bai, Mingming Yang, Yang Xiang, and Min Zhang. Tf-attack: Transferable and fast adversarial attacks on large language models, 2024.

[53] Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Ranking manipulation for conversational search engines, 2024.

[54] Hanbo Huang, Yihan Li, Bowen Jiang, Lin Liu, Bo Jiang, Ruoyu Sun, Zhuotao Liu, and Shiyu Liang. Position: On-premises llm deployment demands a middle path: Preserving privacy without sacrificing model confidentiality, 2025.

[55] Khaoula Chehbouni, Jonathan Colaço-Carr, Yash More, Jackie CK Cheung, and Golnoosh Farnadi. Beyond the safety bundle: Auditing the helpful and harmless dataset, 2024.

[56] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. Policygpt: Automated analysis of privacy policies with large language models, 2023.

[57] Zikui Cai, Yaoteng Tan, and M. Salman Asif. Unlearning targeted information via single layer unlearning gradient, 2024.

[58] Andrew Bell and Joao Fonseca. Output scouting: Auditing large language models for catastrophic responses, 2024.

[59] Tao Huang, Ziyang Chen, Jiayang Meng, Qingyu Huang, Xu Yang, Xun Yi, and Ibrahim Khalil. Machine unlearning with minimal gradient dependence for high unlearning ratios, 2024.

[60] M. Caner Tol and Berk Sunar. Zeroleak: Using llms for scalable and cost effective side-channel patching, 2023.

[61] Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models, 2024.

[62] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

[63] Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024.

[64] Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. The effect of scaling, retrieval augmentation and form on the factual consistency of language models, 2023.

[65] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025.

[66] Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z. Pan. Trustscore: Reference-free evaluation of llm response trustworthiness, 2024.

[67] Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. Reasoning over public and private data in retrieval-based systems, 2022.

[68] Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. On the robustness of editing large language models, 2024.

[69] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. Are large language models reliable judges? a study on the factuality evaluation capabilities of llms, 2023.

[70] Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts, 2024.

[71] Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu. Label smoothing improves machine unlearning, 2024.

[72] Shaofei Shen, Chenhao Zhang, Yawen Zhao, Alina Bialkowski, Weitong Tony Chen, and Miao Xu. Label-agnostic forgetting: A supervision-free unlearning in deep models, 2024.

[73] Jinwen He, Yujia Gong, Kai Chen, Zijin Lin, Chengan Wei, and Yue Zhao. Llm factoscope: Uncovering llms' factual discernment through inner states analysis, 2024.

[74] Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, 2025.

[75] Neel Kanwal, Emiel A. M. Janssen, and Kjersti Engan. Balancing privacy and progress in artificial intelligence: Anonymization in histopathology for biomedical research and education, 2023.

[76] Fei Liu, Nicole Lee Fella, and Kexin Liao. Modeling language vagueness in privacy policies using deep neural networks, 2018.

[77] Mete Keltek, Rong Hu, Mohammadreza Fani Sani, and Ziyue Li. Boosting cybersecurity vulnerability scanning based on llm-supported static application security testing, 2024.

[78] Ghazaleh Shirvani, Saeid Ghasemshirazi, and Behzad Beigzadeh. Federated learning: Attacks, defenses, opportunities, and challenges, 2024.

[79] Lars Krupp, Jonas Bley, Isacco Gobbi, Alexander Geng, Sabine Müller, Sungho Suh, Ali Moghiseh, Arcesio Castaneda Medina, Valeria Bartsch, Artur Widera, Herwig Ott, Paul Lukowicz, Jakob Karolus, and Maximilian Kiefer-Emmanouilidis. Llm-generated tips rival expert-created tips in helping students answer quantum-computing questions, 2024.

[80] Xinyi Hou, Yanjie Zhao, and Haoyu Wang. On the (in)security of llm app stores, 2024.

[81] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[82] Rongting Zhang, Martin Bertran, and Aaron Roth. Order of magnitude speedups for llm membership inference, 2024.

[83] Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators, 2023.

[84] Maojun Sun. Llamacare: A large medical language model for enhancing healthcare knowledge sharing, 2024.

[85] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models?, 2023.

[86] Roma Shusterman, Allison C. Waters, Shannon O'Neill, Phan Luu, and Don M. Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice, 2024.

[87] Panagiotis Koletsis, Panagiotis-Konstantinos Gemos, Christos Chronis, Iraklis Varlamis, Vasilis Efthymiou, and Georgios Th. Papadopoulos. Entity extraction from high-level corruption schemes via large language models, 2024.

[88] Joseph Pedersen, Rafael Muñoz-Gómez, Jiangnan Huang, Haozhe Sun, Wei-Wei Tu, and Isabelle Guyon. Ltu attacker for membership inference, 2022.

[89] Hanxiang Xu, Shenao Wang, Ningke Li, Kailong Wang, Yanjie Zhao, Kai Chen, Ting Yu, Yang Liu, and Haoyu Wang. Large language models for cyber security: A systematic literature review, 2024.

[90] Zitao Chen and Karthik Pattabiraman. Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction, 2023.

[91] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.