

Data Security and Privacy in GPT Models: Techniques and Challenges

David Ghiurău ^{1,2,*} and Daniela Elena Popescu ^{2,*}

¹ Department of Computers and Information Technology, Politehnica University of Timisoara, 2 V. Parvan Blvd, 300223 Timisoara, Romania

² Department of Computers and Information Technology, Faculty of Electrical Engineering and Information Technology, University of Oradea, 410087 Oradea, Romania

* Correspondence: david.ghiurau@student.upt.ro (D.G.); depopescu@uoradea.ro (D.E.P.)

Abstract

The rapid advancement of Generative Pre-trained Transformer (GPT) models has led to their widespread adoption across applied domains such as healthcare, finance, education, and enterprise software engineering. However, the large-scale data requirements and generative capabilities of these models introduce significant challenges related to data security, privacy preservation, and regulatory compliance. This paper presents a systematic literature review conducted in accordance with the PRISMA 2020 guidelines, analyzing 60 peer-reviewed empirical studies published between 2020 and 2025 in Q1 and Q2 journals indexed in the Web of Science Core Collection. The review examines the evolution of GPT architectures and evaluates state-of-the-art security and privacy techniques, including encryption, differential privacy, federated learning, data anonymization, model distillation, and secure deployment mechanisms. Key challenges identified include unintended memorization of sensitive data, adversarial prompt-based attacks, and performance degradation resulting from privacy-preserving constraints, with reported accuracy reductions ranging from 5% to 20% depending on the applied technique. Additionally, the analysis highlights increased computational overhead, in some cases exceeding 30–40% training or inference cost when advanced cryptographic methods are employed. Regulatory and ethical implications are assessed in relation to frameworks such as GDPR, CCPA, HIPAA, and the proposed EU Artificial Intelligence Act. The findings emphasize the need for privacy-by-design approaches and scalable governance strategies to support secure and trustworthy deployment of GPT models in applied real-world environments.

Keywords: Generative Pre-trained Transformers; data security; data privacy; large language models; differential privacy; federated learning; AI governance; ethical AI; privacy-preserving machine learning; regulatory compliance

Academic Editor: Ephraim Suhir

Received: 23 January 2026

Revised: 11 February 2026

Accepted: 12 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Overview of Large Language Models

1.1. Introduction

The rapid evolution of large language models (LLMs) has fundamentally transformed natural language processing and its real-world applications. Models such as BERT, PaLM, LLaMA, Claude, and the Generative Pre-trained Transformer (GPT) family have demonstrated unprecedented capabilities in text generation, reasoning, and decision support. Among these, GPT-based models have achieved exceptional prominence due to

their large-scale deployment, API-based accessibility, and integration into critical domains such as healthcare, finance, education, and software engineering [1,2].

In parallel with this technological progress, a growing body of survey and review literature has emerged to analyze security and privacy challenges in LLMs. Early surveys primarily examined general threat landscapes, including data leakage, adversarial attacks, and model misuse, often adopting a model-agnostic perspective. More recent reviews have proposed taxonomies of security and privacy threats and discussed ethical and regulatory considerations. However, existing surveys typically treat LLMs as a homogeneous class or focus on isolated categories of risks, with limited attention to deployment-scale implications, systematic methodological rigor, and privacy–utility trade-offs. A systematic comparison between the present work and existing surveys is provided in Table 1.

Table 1. Systematic comparison between current surveys.

Survey	Scope	Models Covered	Methodology	Security & Privacy Focus	Regulatory Analysis	Key Distinction
Security and Privacy Challenges of LLMs: A Survey [3]	General LLMs	BERT, GPT-2/3, PaLM	Narrative review	Technical threats	Limited	Broad but non-systematic
A Survey on LLM Security and Privacy: The Good, the Bad, and the Ugly [4]	General LLMs	ChatGPT, Bard, GPT-3/4	Literature review	Attacks, misuse, defenses	No	Security lens incl. misuse and defensive use
Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions [5]	Privacy-centric	GPT-2/3/4, BERT	Systematic review	Memorization, data leakage, DP	Limited	End-to-end privacy lifecycle view
Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey [6]	Lifecycle-based	Multiple LLMs	Structured taxonomy	Pre-train → deployment threats	No	Scenario-specific threat modeling
Security Concerns for Large Language Models: A Survey [7]	LLM systems	ChatGPT-like models	Chronological survey	Prompt attacks, agents, misuse	No	Strong focus on agentic LLM risks
A Survey on Privacy Risks and Protection in LLMs [8]	Privacy-centric	GPT-style LLMs	Risk-mitigation mapping	Extraction, inversion, MIAs	Limited	Defense-oriented taxonomy
Privacy in Fine-Tuning Large Language Models [9]	Fine-tuning stage	GPT-family, open LLMs	Stage-specific review	MIAs, backdoors, DP	No	Fine-tuning as primary risk vector
Membership Inference Attacks on Large-Scale Models: A Survey [10]	Privacy attacks	LLMs and LMMs	Attack taxonomy	Membership inference	No	First MIA-focused LLM survey

1.2. GPT Models as State of the Art

Despite these valuable contributions, GPT-based models represent a uniquely impactful and well-documented case that warrants focused analysis. Their widespread industrial adoption, extensive real-world exposure, and increasing regulatory scrutiny have generated a substantial body of empirical evidence on security failures, memorization risks, prompt-based exploitation, and privacy violations. These characteristics distinguish GPT models from other LLM families and motivated a dedicated, systematic investigation.

GPT-4 was selected as a baseline model for the review because it operates in a scale regime where emergent reasoning, generalization, and robustness properties stabilize, making it a common reference point in academic evaluations of LLM capabilities and limitations. Its widespread use in benchmarking studies further supports its suitability as a canonical model for analyzing security, privacy, and safety risks [11].

GPT-4o was included to reflect the current transition from text-only LLMs to multi-modal general-purpose models. Finally, the o1 model series was selected to capture reasoning-oriented inference-time scaling, an increasingly influential design direction in LLM research. Recent studies indicate that reasoning performance can be enhanced through inference strategies without violating known scaling principles, and similar approaches are now adopted across multiple LLM families [12]. Including o1 therefore allows the review to cover not only parameter scaling but also deliberative reasoning paradigms.

Accordingly, this work positions GPT models as a representative and industrially dominant subclass of LLMs, enabling a focused yet generalizable analysis of security and privacy challenges. We conducted a PRISMA 2020-compliant systematic literature review of security and privacy techniques applied to GPT-based systems, integrating technical mechanisms with governance and regulatory frameworks. By combining quantitative privacy-utility trade-off analysis with legal and ethical considerations, this survey addresses key limitations identified in prior reviews and provides a structured foundation for future research on trustworthy, compliant, and responsible large language models.

1.3. Transformer Architecture in GPT Models

The GPT series was built by utilizing this architecture to make way for a new era in NLP. GPT (Generative Pre-trained Transformer) models are a family of model architectures, primarily developed by OpenAI, which are used to solve various tasks in the field of NLP. They are built on the basis of large-scale unsupervised learning on plentiful textual data and employed to solve a broad range of language tasks, including text generation, summarization, translation, and so on. They belong to the class of transformer-based models that have been successful in both understanding and generating text. OpenAI has developed various generations based on the transformer architecture, as outlined in Table 2 [13–15].

Table 2. GPT model innovations through generations.

Generation	Highlights	Key Innovations	Applications
GPT-4 (2023)	More reliable, nuanced understanding, better creativity, and logical reasoning	Multimodal input (via plugins, but not native), larger token window (8k–32k),	Coding assistants, legal and healthcare support (careful use), research summarization, and

		and improved safety measures	complex problem solving
GPT-4o	“Omni” model: native multimodal (text, vision, audio) capabilities; near real-time responsiveness (~232 ms latency); more emotional, human-like conversations	Unified architecture for all inputs/outputs, real-time speech, and highly emotional interaction support	Real-time multimodal assistants, accessibility support, creative media generation, and interactive agents
GPT-o1	“Open-weight” minimal GPT prototype: transparent and modular for research and reproducibility	Open architecture, micro-modular building blocks, and auditable and explainable AI	Educational purposes, transparent AI deployment, and academic research
Future Models (GPT-5 and beyond)	Autonomous multi-agent AI, long-term memory, on-device and cloud hybrid models, and native agentic behavior (task completion over days/weeks)	Persistent memory, cross-agent collaboration, tool use is deeply integrated, and highly efficient architectures	Autonomous research teams; personalized, memory-driven assistants; complex simulation modeling; and advanced automation systems

Besides the growth in use cases and their applications, the difference is also visible with regard to the capabilities, token limit, and training data used for each model in Table 3 [14,15].

Table 3. Capabilities of each GPT model by version.

Model	Capabilities	Token Limit	Speed	Training Data
ChatGPT-4 (2023)	Advanced reasoning	8k/32k tokens	Slower	Multimodal data (filtered)
GPT-4o (2024)	Native multimodal	128k tokens	Very fast	Extensive text/image/audio datasets
GPT-o1	Open-weight modular	128k tokens	Fast	Clean public datasets
Newer models	Autonomous memory agents	>256k tokens	Instant	Synthetic, human-curated, and dynamic collection

Even though the development and use of this architecture has shown lots of benefits and useful applications in simplifying human tasks, these models still face major

challenges, the main one being data privacy and security. GPT-like models memorize training data, which can lead to unintentional information leakage [16,17]. The synthetic data that are generated can also be biased and neglect ethics, and despite fine-tuning efforts, GPT models can perpetuate stereotypes and biases present in training data [18,19]. A bigger and more impactful issue is the environmental one, as training large models requires substantial computational resources, raising concerns about sustainability.

This research connects technical methods to larger, practical issues appearing in applications. It highlights the privacy rise/performance fall trade-off and the legal and ethical guidelines that are changing the expectations for the development of GPT models. Future work needs to be conducted to investigate new safe AI methodologies, identify holes in ongoing techniques, and propose more secure and privacy-preserving AI models.

The key contributions of this article are the following:

1. It provides a systematic review of data security and privacy in GPT-based models, emphasizing the evolution of large language models and their deployment across multiple application domains while examining the associated security risks and protection mechanisms.
2. It identifies and categorizes the primary security and privacy challenges in GPT architectures, including data memorization, membership inference, prompt injection, model extraction, and poisoning attacks, offering a structured understanding of the current threat landscape.
3. The paper reviews existing privacy-preserving and security-enhancing techniques, such as differential privacy, federated learning, encryption-based approaches, anonymization strategies, secure fine-tuning, and controlled inference pipelines, while analyzing their effectiveness, limitations, and performance overhead.
4. It highlights quantitative privacy–utility trade-offs reported in the literature and suggests potential technical solutions, including hybrid privacy-preserving training schemes, secure deployment architectures, and standardized evaluation metrics, outlining future research directions to address current limitations.
5. The paper emphasizes ethical, legal, and regulatory considerations related to GPT model deployment, addressing issues such as sensitive data leakage, compliance with GDPR, CCPA, HIPAA, and the EU AI Act, and the necessity of privacy-by-design and security-by-design principles for responsible generative AI systems.

The paper is structured into eight sections as follows:

Section 1: Overview of Large Language Models provides an overview of the evolution of the transformer architecture in the NLP domain, comparing the versions of OpenAI models, and on top of that are presented the main risks of current models regarding data security and privacy, highlighting various surveys performed on different LLMs.

Section 2: Methodology explains the process and approach used to create this systematic review, mentioning the selection process and how the identification, screening, eligibility, and inclusion were conducted.

Section 3: Security and Privacy Challenges in Large Language Models in AI Models highlights the potential exploits of GPT models and explains the current data governance that is in place across the globe.

Section 4: Data Security approaches in GPTs reviews existing techniques available at the moment and used in the training of GPT models, comparing their strengths and weaknesses and highlighting the use cases of each.

Section 5: Challenges in Ensuring Data Security and Privacy details the current weaknesses of GPT models regarding data privacy and their mitigation strategies, as well as their impact on performance, user privacy, scalability, and feasibility. The challenges analyzed include data leakage, unintended memorization, adversarial attacks, performance losses in favor of privacy, computation costs, and user transparency and trust.

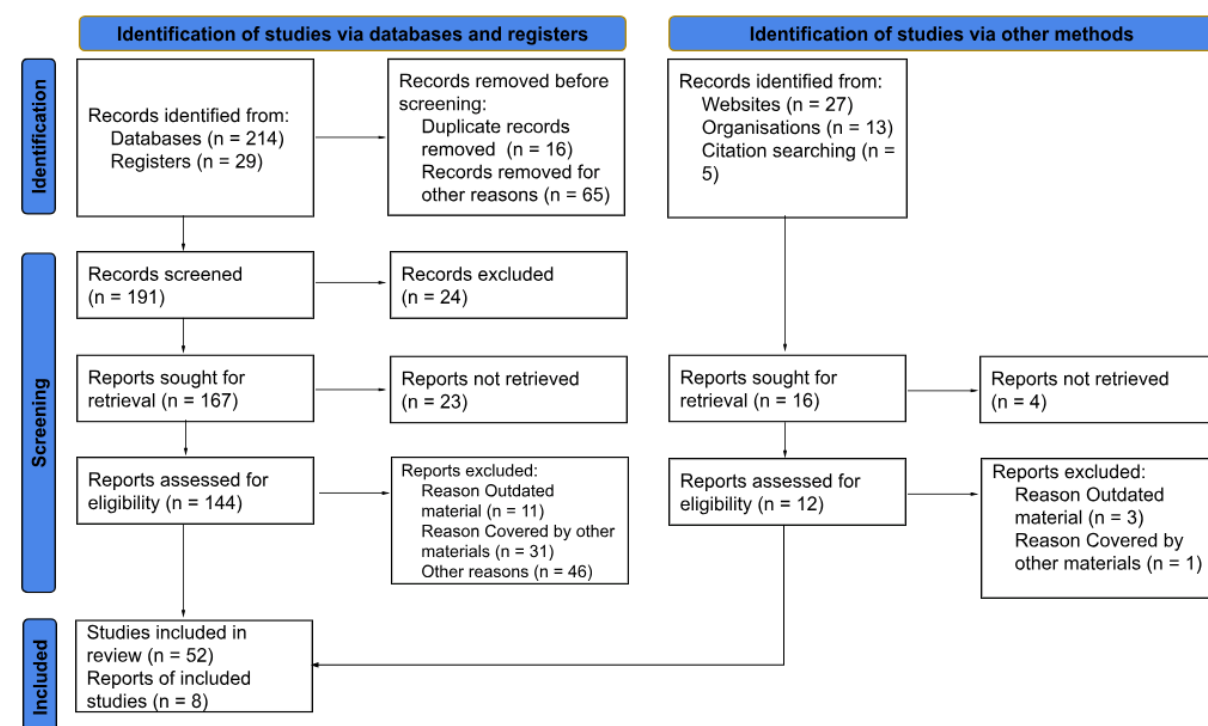
Section 6: Case Studies and Applications mentions the real-world examples that validate theories, challenges, and show the impact of the current weaknesses of the models. Each example suggests lessons that can be used for future iterations of GPT models.

Section 7: Proposed directions discusses the directions that are mandatory in the next iterations in order to improve data transparency, increase user trust, solve biased outputs, and prevent attacks. The section offers recommendations for future directions, suggesting areas where further investigation is needed.

Section 8: Conclusion summarizes the key findings of the research, drawing final conclusions based on the analysis, highlighting the current data security and privacy techniques and challenges raised by GPT models. It provides a concise summary of what has been learned from the study and its contributions to the field.

2. Methodology

This systematic literature review (SLR) was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines [20,21]. The PRISMA 2020 flow diagram is included to illustrate the study selection process (Figure 1). The review process was structured into three sequential stages, planning, conducting, and reporting, ensuring methodological rigor and transparency in synthesizing research related to data security and privacy in GPT models, with a focus on techniques and associated challenges.



From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi: 10.1136/bmj.n71.

Figure 1. PRISMA flow chart with the number of studies and reports reviewed [20].

2.1. Search Strategy and Inclusion Criteria

The review targeted peer-reviewed journal articles published between 1 January 2020 and 30 November 2025, reflecting the period of accelerated development and adoption of artificial intelligence technologies. The Web of Science (WoS) Core Collection was selected as the primary database due to its extensive coverage of high-impact journals and its

emphasis on rigorously peer-reviewed literature. Priority was given to journals indexed in SSCI Q1 and Q2 to ensure the inclusion of influential and methodologically robust studies relevant to the social, ethical, and educational dimensions of AI.

Although the use of a single database may limit the comprehensiveness of the review, WoS was selected for its recognized selectivity and alignment with the objectives of this study. Future research could incorporate additional databases, such as Scopus and ERIC, to broaden disciplinary coverage and capture complementary perspectives from educational and interdisciplinary research domains. In addition, arXiv was considered a relevant source for early-stage research on technical developments in AI; however, preprints were excluded from this review due to the lack of formal peer review.

The search strategy employed Boolean operators to combine keywords across three primary domains: privacy, security, and machine learning. Privacy-related keywords included “data privacy,” “AI data privacy,” “AI privacy,” “AI data security,” “privacy rights,” “privacy constraints,” and “privacy legislation.” Security-related terms comprised “security violations,” “data security,” “secure data,” “data breaches,” and “data usage.” Machine learning-related keywords included “GPT training data,” “machine intelligence,” “machine learning data,” and “AI data corpus.”

Explicit inclusion and exclusion criteria were defined to delimit the scope of the review and ensure the quality and relevance of the selected studies. These criteria, summarized in Table 4, guided the systematic screening and selection of research papers included in this review.

Table 4. Exclusion/inclusion criteria of research papers for this review.

Exclusion Criteria	Inclusion Criteria
Study type: Exclude non-peer-reviewed studies	Language: Articles published in English
Non-Published data: Exclude grey literature such as reports or unpublished data	Seniority: Articles published within the last five years
Lack of outcome reporting: Exclude studies that do not report on the primary outcomes of interest	Applicability: Articles that remain relevant given the latest scientific advances and policy changes
Duplications: Exclude duplicate studies or preliminary reports of already published research	Outcome relevance: Studies that investigate or report on specific outcomes for the generative models
Not scoped: Studies outside higher education or not involving AI as a core tool	Sample size: Studies that ensure a robustness of tests and analysis
Not reviewed: Proceeding papers and retracted papers	Peer-reviewed: Empirical articles published in Q1 or Q2 journals

2.2. Study Selection Process

The study selection process followed the four standard PRISMA stages: identification, screening, eligibility, and inclusion.

2.2.1. Identification

A total of 288 records were initially retrieved from the selected databases. During the metadata export stage, 81 records were excluded using predefined filters, including document type and publication status. Review articles, conference proceedings, and retracted papers were removed at this stage. Conference proceedings were excluded because they typically undergo less rigorous peer-review processes than journal articles, are often subject to shortened review timelines, and frequently report preliminary findings whose validity may not yet be fully established.

2.2.2. Screening

In the screening phase, the titles and abstracts of 207 records were assessed. A total of 63 studies were excluded because they were classified as Early Access publications or were written in languages other than English. Early Access articles were excluded to preserve the structural consistency of the dataset and to ensure reproducibility of the analysis. Although these articles have generally completed peer review, they often lack finalized bibliographic information (e.g., volume, issue, and page numbers), and their citation metadata may change as database indexing is finalized. Non-English publications were excluded to ensure consistent interpretation of content and uniform methodological assessment across all included studies.

2.2.3. Eligibility

The eligibility stage involved a full-text assessment of the remaining 144 articles. Following this evaluation, 84 records were excluded based on additional criteria, including lack of relevance to artificial intelligence, non-empirical study designs, absence of clearly described practical applications, missing impact factor indexing, or publication in journals outside the targeted quartiles. The Q1–Q2 journal filter was applied at this stage rather than during initial screening to allow for a comprehensive evaluation of both topical relevance and methodological quality.

2.2.4. Inclusion

The final sample comprised 60 empirical studies published in Q1–Q2 journals. Restricting the review to high-impact journals ensured that only methodologically robust and influential studies were included. A summary of the study selection process, including the number of records excluded at each stage, is presented in the PRISMA flow diagram (Figure 1).

2.3. Quality Assessment

To ensure that only methodologically robust and high-quality studies were included, a systematic quality assessment was conducted on the final set of 60 selected articles. A seven-criterion evaluation framework was applied to assess the methodological soundness and reporting quality of each study. The criteria included:

- (1) Clarity of research objectives;
- (2) Inclusion of a comprehensive literature review;
- (3) Clear presentation of related work to contextualize the study within existing research;
- (4) Detailed description of the methodology or model architecture;
- (5) Clear and transparent presentation of research results;
- (6) Alignment between conclusions and stated research objectives;
- (7) Explicit recommendations for future research directions [22,23].

Only studies that satisfied all seven criteria were retained for further analysis, ensuring a consistent level of methodological rigor across the reviewed literature.

2.4. Inclusion of High-Impact Conference Proceedings

This study adopted a systematic, evidence-driven research methodology that integrated high-impact conference proceedings alongside journal publications to capture the most recent and influential advancements in data security, privacy preservation, and robustness of large language models (LLMs), focused on GPT-based architectures. Given the fast-evolving nature of artificial intelligence security research, premier conferences

often precede journal publications and thus provide state-of-the-art methods, threat models, and defenses.

Conference proceedings were selected based on the following criteria:

- (1) Scientific impact and selectivity: Conferences with low acceptance rates and strong peer-review standards;
- (2) Relevance to LLM security and privacy: Focus on adversarial learning, differential privacy, federated learning, prompt attacks, and model leakage;
- (3) Community recognition: Endorsement by the AI, NLP, and cybersecurity research communities.

Based on these criteria, this methodology prioritized proceedings from:

- (1) NeurIPS and ICML (machine learning security, privacy, and federated learning);
- (2) IEEE Symposium on Security and Privacy and ACM CCS (formal security analysis and adversarial threats);
- (3) ACL and EMNLP (NLP-specific privacy, prompt attacks, and data leakage).

Relevant conference papers were identified through systematic keyword-based searches targeting LLM privacy leakage, membership inference, model inversion, prompt injection, federated learning for LLMs, and differential privacy in NLP.

This approach allowed for methodological consistency while preserving technical depth.

Recent conference studies demonstrate that federated learning alone is insufficient without complementary mechanisms such as differential privacy and secure aggregation, particularly for GPT-scale models.

Cross-validation was achieved by comparing findings across multiple conferences.

2.5. Data Extraction, Impact Stratification, and Synthesis

Data extraction was performed to systematically capture information related to current techniques and challenges associated with data security and privacy in GPT models. The extracted data focused on key dimensions, including research context, pedagogical and instructional frameworks, social and ethical implications, instructional innovation strategies, higher-order outcomes, and disciplinary scope. This structured approach enabled a comprehensive synthesis of how GPT-based systems utilize data, the associated risks, and the mitigation strategies proposed in the literature, while also facilitating the identification of emerging trends and research gaps.

To assess variations in academic influence among the included studies, an impact stratification analysis was conducted using citation data. Citation count is a widely accepted indicator of scholarly impact [24]. As all selected studies were published in Q1 or Q2 journals, journal-level metrics provided limited differentiation. Consequently, studies were categorized based on citation count quartiles: high-impact studies (≥ 36 citations; top 25%), medium-impact studies (5–35 citations), and low-impact studies (< 5 citations; bottom 25%). Citation data were collected on 12 November 2025.

3. Security and Privacy Challenges in Large Language Models

This section synthesizes the main security and privacy challenges identified in the analyzed literature. To ensure clarity and analytical rigor, the challenges are organized according to a structured taxonomy, which also serves as the organizing principle for the remainder of this review.

This taxonomy provides a coherent analytical framework that guides the organization of the review and enables a systematic synthesis of technical, operational, and regulatory challenges affecting large language models. The summary of all taxonomies and their organization in the paper can be viewed in Table 5.

Table 5. Taxonomy of security and privacy challenges in large language models.

Category	Description	Representative Threats	Impact
Data Leakage and Memorization	Unintended retention and disclosure of sensitive training data	Memorization, extraction attacks, and inference	Privacy violations
Adversarial and Poisoning Attacks	Malicious manipulation of model behavior	Prompt injection, data poisoning, and model inversion	Integrity and confidentiality loss
Privacy–Utility Trade-Offs	Performance degradation due to privacy mechanisms	Differential privacy and federated learning overhead	Reduced accuracy and utility
Computational and Energy Overhead	Increased resource consumption	Latency, scalability limits, and energy cost	Operational inefficiency
Governance and Compliance Risks	Legal and ethical constraints	GDPR, AI Act, and accountability gaps	Regulatory non-compliance

Based on the systematic literature review conducted in accordance with the PRISMA 2020 guidelines, the security and privacy challenges affecting large language models (LLMs) could be categorized into five major dimensions, spanning technical, operational, and governance-related aspects.

3.1. Data Leakage and Memorization Risks

Large language models may unintentionally memorize sensitive or personally identifiable information contained in training datasets, particularly rare or unique data instances. Such memorization can lead to unauthorized data disclosure through inference attacks, prompt-based extraction, or model probing techniques. These risks are especially critical in regulated domains such as healthcare, finance, and legal services, where data confidentiality is paramount [25–27].

3.2. Adversarial and Data Poisoning Attacks

This category encompasses malicious techniques aimed at manipulating LLM behavior, including adversarial prompting, prompt injection, training data poisoning, model inversion, and backdoor attacks. These threats exploit the generative and adaptive properties of LLMs, potentially compromising both model integrity and confidentiality during training and inference phases [28–30].

3.3. Privacy–Utility Trade-Offs

Privacy-preserving mechanisms such as differential privacy, federated learning, secure aggregation, and data sanitization are commonly employed to mitigate data exposure risks. However, these techniques often introduce performance degradation, reduced accuracy, or limited generalization capability. This category captures the inherent trade-offs between privacy guarantees and model utility, emphasizing the need for context-aware optimization strategies [31–34].

3.4. Computational, Scalability, and Energy Overhead

The integration of security and privacy-enhancing techniques frequently results in increased computational complexity, inference latency, memory usage, and energy consumption. These overheads pose significant challenges for large-scale deployment and sustainability, particularly in cloud-based and edge computing environments [35,36].

3.5. Governance, Regulatory, and Ethical Risks

Beyond technical vulnerabilities, LLMs operate within complex regulatory and ethical frameworks, including GDPR, CCPA, HIPAA, and the EU AI Act. This category addresses challenges related to accountability, transparency, explainability, auditability, and legal compliance, highlighting the necessity of aligning technical safeguards with governance and policy requirements [37–40].

4. Data Security Approaches in GPTs

The large-scale utilization of GPT models requires strong data confidentiality to defend against sensitive information leaks. The three most essential methods used for protecting data are: encryption, depending on the encryption algorithm, the data are more or less secure; data anonymization and de-identification, this method tries to make data unrecognizable to the usual processes while still helping the AI model to update or train on the data; and lastly, federated learning and data minimization, where the training is more decentralized, which lends to a less vulnerable pool of data.

The most prevalent is symmetric encryption. It only requires one key, for both encrypting and decrypting, which is efficient for large quantities of data. For example, algorithms like AES are commonly used to encrypt stored data, such as training sets or model parameters. Despite its secured encryption key, management is necessary to avoid unauthorized access [41]. Yet another possible encryption for securing data is the asymmetric one, which uses public and private key pairs, like RSA, to protect the data being transmitted. This approach can be useful for sending secrets between multiple systems, including sharing model updates between cloud services. The splitting of the encryption and decryption keys provides further security in untrusted networks [42].

Due to the property of homomorphism, encrypted computations on encrypted data can be performed without requiring decryption and so the confidentiality of the data can be maintained during processing. For instance, AI models can process encrypted user queries, such that their privacy is protected in important applications such as healthcare and finance [43]. The homomorphic encryption flowchart is visible in Figure 2.

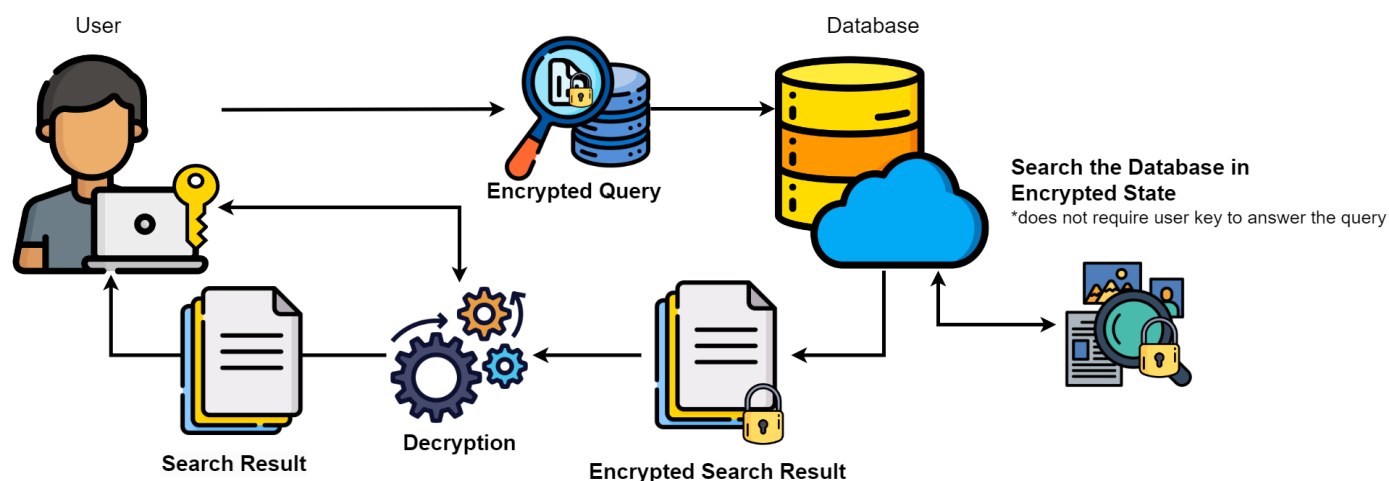


Figure 2. Homomorphic encryption flowchart [44,45].

The asterisk highlights that the query processing can be performed without necessitating direct access to the user's decryption key at execution time. This process adopts a cryptographic scheme that enables computation over encrypted data, such as searchable encryption or related privacy-preserving techniques, thereby ensuring that the database remains encrypted throughout the query lifecycle.

This is useful in a GPT fine-tuning setting since the training examples are from various organizations. Therefore, the privacy of the data is maintained without minimizing data utility. Encryption is also of vital importance for GPT model safety, as it tackles some critical security and privacy issues:

- Preserving Raw Data Training Data

With encryption, proprietary data used to train AI models are kept safe, preventing unauthorized access or data leaks. This is especially important in scenarios that deal with sensitive/private information [26].

- Protecting Model Predictions and User Actions

Applications where encryption is being applied to secure against interception or malicious use include chatbots and customer service tools where privacy regulations such as GDPR are driving enhanced trust and privacy [46].

Data anonymization and data de-identification are necessary procedures for meeting the standards for privacy regulations like GDPR and HIPAA and can reduce the risk of inadvertently exposing private information. These techniques keep user's data private by making it unidentifiable while still being useful for training and running AI systems [47].

There are various techniques for anonymization, but the most common one is data masking, a process by which sensitive data fields are substituted with artificial but realistic values. For example, names or addresses may be swapped with fake numbers in datasets in order to keep the data out of reach, but the style of the dataset persists. Such a technique is commonly employed in testing and development scenarios [48].

Another approach is pseudonymization, meaning that identifiers are replaced by artificial pseudonyms. For example, a generic name such as "John Doe" can be substituted by a random "User123," which allows for some of the data to be shared without compromising true identities. Pseudonymized data can sometimes be de-anonymized; however, the access control mechanisms make it difficult to be performed by an unauthorized user [49].

The differential privacy approach adds mathematical noise to datasets or query responses in order to hide individual data points. For instance, GPT models trained with differential privacy can generalize from the aggregate statistics of the data without being able to memorize particular facts. This kind of protection is well suited to privacy vs. utility, in particular [47].

Differential privacy can be achieved through three primary means: noise addition, where random noise is added to training data or model updates to obscure individual contributions (e.g., adding Laplace or Gaussian noise to gradients during training); privacy budgets, wherein cumulative privacy loss over repeated queries is tracked to prevent excessive data exposure; and federated differential privacy, which integrates differential privacy with federated learning to guarantee both distributed data processing and strong privacy protections.

The primary benefit is that user privacy is balanced with the ability of the model to learn meaningful patterns and reduced the risk of the model accidentally learning sensitive information, like credit card numbers or personally identifiable information [47].

On the other hand, there are limitations. The biggest issue is that model precision can deteriorate due to noise injection, and it is more expensive in terms of tuning and optimization as it needs to find the appropriate trade-off between privacy and utility.

The term synthetic data is used to describe artificially created data that imitate the statistical properties of real data but do not contain any actual sensitive or personal data. It is now widely used in the context of machine learning pipelines in order to address the legal and ethical restrictions that arise when considering real data, especially in domains such as healthcare, finance, and user behavior modeling.

Model distillation is a learning algorithm where a smaller model (student) is trained to mimic the behavior of a larger model (teacher). In this way, the student model can transfer the teacher knowledge without directly accessing the original training data, and therefore benefiting from both computational efficiency and privacy preservation sides [50].

In distillation, usually the teacher model generates output probabilities (or “soft labels”) for a given dataset. These soft labels, which are often overlooked by hard labels, contain rich information about class distributions and similarities between classes. The student model is trained to imitate the outputs, and it is successful in estimating the decision boundaries of the teacher model. This indirect learning paradigm encourages the student to generalize effectively across the training and test distribution while not having access to, and being thus urged to infer, the raw sensitive data from which the teacher is learnt [51].

In terms of privacy, model distillation is especially useful. As the student model is learning from the teacher’s outputs, it generalizes the data, hence removing any specific data patterns that might contain some private or personal information. Therefore, the fact that quantized knowledge distillation cannot be reversed and that privacy data or model components do not suffer attacks, data leakages, and model inversion make it a good component for preserving privacy in machine learning pipelines.

In addition to the privacy benefits, model distillation has practical gains. The student model is usually smaller and less complex and can be run faster and deployed on devices with limited capabilities, such as mobile phones or edge hardware, without much modification. But, as with the first example, this method is not perfect. Student models frequently fail to reproduce the performance of larger teachers in full when deep contextual knowledge or complicated reasoning is necessary. Distillation relies heavily on the quality of the teacher model and the variety of data on which the teacher is trained during the distillation period.

And the last common technique is the de-identification one, which comprises generalization and suppression. Generalization groups certain values together, for instance, instead of “29,” we may now also have “20–30.” This diminishes the granularity of information but maintains data that are relevant to an analysis [52]. Meanwhile, suppression entirely erases sensitive information, like wiping out exact locations or other identifying information from a set of data. This guarantees that the data are no more a threat to identification and can be used for compliance and analysis [53].

Each of the presented methods has its strengths and weaknesses, as well as use cases that are most suitable for the approach. Table 6 summarizes each method based on these criteria.

Table 6. Methods for data security and their use cases.

Method	Strengths	Weaknesses	Use Cases
Encryption	High confidentiality and robust against interception	Requires effective key management and is computationally intense	Data storage, model parameters, and secure transmission

Symmetric Encryption	Fast and efficient for large data volumes	Vulnerable if the key is exposed and key distribution is challenging	Secure model training and encrypted backups
Asymmetric Encryption	Secure key exchange and ideal for distributed systems	Slower than symmetric encryption and larger key sizes are required	Cloud-based model updates and secure data sharing
Homomorphic Encryption	Maintains privacy during processing and is useful for confidential tasks	Computationally expensive and slower processing times	Healthcare analytics and financial data processing
Data Anonymization	High privacy preservation and reduces risk of exposure	Risk of re-identification if improperly applied	Data sharing and machine learning training
Data Masking	Maintains dataset structure and is safe for testing and development	Data utility may be reduced and reversible if poorly implemented	Software testing and model evaluation
Pseudonymization	Enables data sharing with reduced privacy risks	Can be re-identified if additional data are exposed	Research data sharing and collaborative projects
Differential Privacy	Balances privacy and data utility and prevents memorization of sensitive data	Reduced accuracy due to noise and complex parameter tuning is required	Federated learning and decentralized AI training
Federated Learning	Minimizes data transfer risks and preserves user privacy	Computationally intense and has synchronization challenges	Mobile devices and collaborative AI models
Model Distillation	Reduces memory footprint and enhances privacy through indirect learning	May lose context from original training data and student model may underperform	Edge devices and mobile applications
Synthetic Data Generation	Bypasses privacy concerns and enhances model training diversity	May lack true representativeness and synthetic biases can be introduced	Healthcare, financial modeling, and IoT applications [54]
De-identification	Meets compliance with GDPR and HIPAA and maintains analysis relevance	Data utility can diminish with excessive generalization	Public health datasets and demographic analysis

As the deployment of GPT models scales across diverse industries, maintaining strong data confidentiality is crucial for preventing information leakage and ensuring compliance with privacy regulations. Techniques such as encryption, data anonymization, de-identification, and federated learning are pivotal in safeguarding sensitive information. Encryption secures data during storage and transmission, while anonymization and de-identification protect individual privacy by masking identities. Moving forward, enhancing these methods with advanced machine learning techniques and interdisciplinary collaboration will be key to building resilient, privacy-preserving AI systems. Successful integration of these technologies not only fortifies data protection but also instills public trust in AI-driven solutions, paving the way for broader adoption in critical sectors like healthcare, finance, and cybersecurity.

5. Challenges in Ensuring Data Security and Privacy

There are multiple layers in the quest to protect data and privacy in GPTs. These models are trained on and generate copious amounts of information, which is frequently sensitive or proprietary. Because they are applied at scale, dealing with security holes without giving up performance or trust are key concerns.

Ensuring data security and privacy in large language models (LLMs), particularly GPT-based systems, involves multiple interrelated technical, operational, and ethical challenges. These models are trained on vast amounts of data, often containing sensitive or proprietary information, and are deployed at large scale. As a result, maintaining robust security and privacy guarantees without compromising performance, usability, or user trust remains a critical concern.

5.1. Data Memorization and Leakage Risks

GPT models exhibit inherent privacy vulnerabilities stemming from their architecture and training procedures. Pre-trained models have been shown to memorize and reproduce fragments of their training data, including sensitive or personally identifiable information such as names, phone numbers, or social security numbers [55]. This risk increases with model size and dataset complexity. Empirical studies have demonstrated that models such as GPT-4 can recall verbatim sensitive training samples, leading to severe privacy violations [56].

5.2. Adversarial and Model Manipulation Attacks

Another major challenge arises from adversarial threats targeting GPT models. These include adversarial prompts designed to induce harmful, misleading, or unethical outputs, as well as model extraction attacks that aim to reconstruct model behavior through iterative querying. In addition, data poisoning attacks, where malicious actors contaminate training or fine-tuning datasets, can skew model behavior and undermine both integrity and reliability [57].

5.3. Privacy–Performance Trade-Offs

A fundamental challenge in secure model design is balancing privacy guarantees with model performance and usability. Differential privacy is widely adopted to protect individual data contributions; however, it typically introduces a measurable performance penalty. The injection of noise into training data or gradients can reduce model accuracy and generalization capability, particularly for large-scale models [58]. Similar trade-offs arise when applying other privacy-preserving techniques.

5.4. Computational and Resource Overhead

Many privacy-enhancing mechanisms incur significant computational costs. Techniques such as homomorphic encryption enable computation over encrypted data but remain prohibitively expensive for real-time inference or deployment on resource-constrained devices [59]. Secure multi-party computation and encrypted inference further exacerbate scalability limitations, restricting their applicability in large-scale systems.

5.5. Data Utility Degradation

Approaches such as data anonymization, redaction, or sanitization can reduce privacy risks but often diminish the richness and informativeness of training datasets. This loss of data utility can limit the learning capacity of models and negatively impact downstream task performance, particularly in complex reasoning or domain-specific applications.

5.6. Scalability and Infrastructure Constraints

Deploying privacy and security mechanisms at the scale required by modern GPT models presents substantial technical and economic challenges. Techniques such as federated learning, while effective in keeping data localized, introduce significant communication and synchronization overheads. Moreover, the need for high-performance computing infrastructure, such as GPUs and large-scale storage, creates accessibility and equity concerns, as many organizations lack the necessary resources [60].

5.7. Energy Consumption and Sustainability

Security primitives and privacy-preserving features often introduce additional latency and energy consumption. These overheads raise sustainability concerns, particularly for applications requiring real-time processing or operating under strict cost constraints. Increased energy usage also contributes to broader environmental impacts associated with large-scale AI deployment.

5.8. Transparency and Trust Challenges

Building user trust extends beyond technical safeguards. Users and stakeholders increasingly demand transparency regarding data usage and model decision-making processes. However, current LLM architectures are largely opaque, functioning as black boxes with limited interpretability. Even existing explainability techniques struggle to provide meaningful insights into how inputs influence outputs [61,62].

5.9. Limitations of Explainable AI Approaches

Explainable AI (XAI) methods are often proposed as solutions for improving transparency and accountability [63]. Nevertheless, current XAI techniques remain insufficient for large-scale LLMs, offering limited explanatory power and failing to fully support regulatory accountability or user understanding.

5.10. Societal Trust and Misuse Concerns

Concerns regarding AI misuse, such as the spread of misinformation, privacy breaches, or unethical applications, further erode public trust [64]. In the absence of strong governance frameworks and enforceable safeguards, these risks hinder widespread adoption of GPT-based systems.

5.11. Summary of Challenges

Overall, the challenges associated with ensuring data security and privacy in GPT-based systems are multifaceted, encompassing technical vulnerabilities, performance–privacy trade-offs, scalability constraints, transparency limitations, and trust-related issues. These challenges underscore the need for continued research, holistic system design, and the integration of technical, organizational, and governance-oriented solutions.

A summary of the identified challenges, along with their associated risks, mitigation strategies, impacts on performance, and scalability considerations, is provided in Table 7.

Table 7. Challenges with their risks and mitigation strategies.

Challenge	Description	Risks	Mitigation Strategies	Impact on Accuracy/Performance	Scalability & Feasibility
Data Leakage	Output of sensitive information seen during training	Exposure of personal/private data and legal liabilities	Dataset filtering, strong access controls, and differential privacy	Moderate decrease depending on strength of protection	Medium—requires pre-training and access controls
Unintended Memorization	Model stores and reproduces specific data points from training	Privacy breaches and trust erosion	Differential privacy, regularization, and smaller context windows	Notable trade-off in fluency and recall	Medium—depends on model size
Adversarial Attacks	Exploiting the model via adversarial prompts or poisoning datasets	Misinformation, misuse and data exposure	Adversarial training, monitoring, and input/output filters	Low-to-moderate depending on robustness techniques	High—defense needs constant updating
Performance vs. Privacy	Balancing utility with user protection	Model becomes less useful if overly private and risks compliance if too open	Differential privacy, federated learning, and secure enclaves	Significant trade-offs and often domain-specific	Low-to-medium depending on method used
Computational Cost	High resource use for secure training/inference	Barriers for small orgs and environmental impact	Efficient algorithm design, parameter sharing, and hardware acceleration	Minimal if optimized, high otherwise	Low—often favors well-funded actors
User Transparency and Trust	Lack of insight/control over data usage	Reduced adoption and ethical and legal concerns	Explainable AI (XAI), opt-in data policies, and transparency reports	Indirect impact: affects user engagement more than raw performance	Medium—current XAI tools still evolving

Ensuring data security and privacy in GPT models is a multifaceted challenge, requiring robust technical solutions, thoughtful trade-offs, and transparent practices. Addressing vulnerabilities, resource demands, and user trust issues is essential for ethical AI deployment and widespread acceptance.

6. Case Studies and Applications

Large language models, including GPT-based systems, are increasingly deployed across domains where data security and privacy are critical requirements. In such contexts, organizations must ensure that sensitive, proprietary, or regulated information is protected while maintaining model performance and usability. This section examines representative real-world case studies and applications to illustrate how privacy-preserving and security-enhancing techniques are integrated into practical deployment of GPT models and related LLMs.

6.1. Industrial Adoption of Secure GPT Systems

GPT models and other LLMs have been widely adopted in industries that handle sensitive data, including healthcare, finance, education, and customer services. As these models frequently process personal or confidential information, organizations employ multiple layers of security and privacy protection, such as data anonymization, access control, encryption, and secure deployment infrastructures. These measures enable the practical use of LLMs while complying with regulatory and ethical constraints.

6.2. OpenAI and Enterprise-Grade GPT Deployments

During the development and deployment of OpenAI's GPT models, privacy and safety considerations have been treated as core design principles. Privacy-preserving techniques, including differential privacy mechanisms, are incorporated during the fine-tuning stages to mitigate the risk of sensitive data leakage from training datasets. Additionally, Reinforcement Learning from Human Feedback (RLHF) is employed to steer model behavior away from generating harmful, biased, or privacy-violating content through human-aligned feedback loops [65].

Furthermore, OpenAI has implemented strict data handling policies to minimize the retention of personally identifiable information (PII) from user interactions. These safeguards have enabled the secure deployment of GPT models in enterprise environments, such as Microsoft's Azure OpenAI Service, thereby increasing user trust and facilitating adoption in real-world business settings.

6.3. Federated Learning in Mobile and Consumer Applications

Google's Gboard mobile keyboard application represents a prominent example of privacy-preserving LLM deployment through federated learning. In this approach, user data remains on-device, and only aggregated model updates are transmitted to central servers. By avoiding the transfer of raw user data, this architecture enhances privacy protection while maintaining high-quality personalization and predictive performance, in compliance with data protection regulations [66].

6.4. Secure Collaborative Model Training with NVIDIA NeMo

NVIDIA's NeMo open-source toolkit integrates secure multi-party computation and encrypted data pipelines to support federated training of large language models across multiple institutions. This approach enables collaborative model development without requiring organizations to share proprietary or sensitive datasets. Such solutions are particularly relevant for joint research initiatives and cross-border collaborations, where data sovereignty, confidentiality, and regulatory compliance are essential considerations [56].

6.5. Healthcare Applications and Regulatory Compliance

In the healthcare domain, Philips has deployed GPT-based models to summarize patient records and support clinical decision-making processes. To ensure compliance with healthcare data protection regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), data anonymization and pseudonymization techniques are applied prior to model training. These measures allow the models to capture clinical context and insights without exposing identifiable patient information, resulting in improved diagnostic efficiency, reduced administrative burden, and strengthened privacy guarantees [35].

6.6. Domain-Specific Impact of Secure GPT Model Applications

Secure GPT model deployments have demonstrated significant benefits across multiple application domains, including:

- **Healthcare:** Enhancement of diagnostic support systems, clinical summarization tools, and patient engagement platforms, while maintaining compliance with HIPAA and GDPR.
- **Finance:** Secure customer communication, fraud detection, and compliance-driven automation, protecting sensitive financial data through encryption and access control mechanisms.
- **Education:** Adaptive learning tools and personalized student support systems that safeguard academic records and ensure responsible data usage.
- **Customer Service:** Deployment of chatbots and virtual agents, where real-time anonymization and strict data handling policies are essential to prevent customer data disclosure.

6.7. Summary of Case Studies and Applications

The presented case studies demonstrate that privacy-preserving and secure GPT model deployment is feasible across diverse sectors when appropriate technical, organizational, and governance measures are applied. These real-world examples highlight the practical relevance of the discussed security and privacy techniques and emphasize the importance of aligning technical safeguards with regulatory requirements and domain-specific constraints.

A comparative overview of application domains, associated challenges, key use cases, applied privacy strategies, and resulting benefits is provided in Table 8.

Table 8. Sector based implications of ChatGPT.

Sector	Challenges	Key Applications	Privacy Strategies	Key Benefits
Healthcare	Sensitive patient data (e.g., EMRs, imaging) governed by HIPAA and GDPR	Clinical note summarization and virtual health assistants	Data anonymization and differential privacy	Improved patient care, reduced admin load, and preserved confidentiality
Finance	High fraud risk, stringent data security and compliance needs	Fraud detection and AI customer support chatbots	Secure multi-party computation and role-based access controls	Faster, more secure financial services and enhanced trust
Education	Student privacy under FERPA and need for safe personalized learning	Personalized tutoring and plagiarism detection	Anonymized learning data and encrypted submission analysis	Democratized learning with ethical personalization
Retail	Customer data protection (purchase history and payment info)	Product recommendations and automated support agents	Pseudonymized datasets and privacy-compliant user tracking	Stronger engagement and loyalty and secure personalization

The integration of data security and privacy measures in GPT models is critical across industries. By employing strategies like federated learning, anonymization, and differential privacy, companies have successfully deployed AI solutions that protect sensitive

information. These advancements enhance trust and utility, paving the way for responsible and effective AI adoption.

While the technical challenges discussed above highlight the limitations of current security and privacy mechanisms, their real-world applicability is fundamentally shaped by regulatory, governance, and ethical constraints, which are addressed in the following section.

6.8. Governance, Regulatory, and Ethical Frameworks for Large Language Models

The deployment of large language models (LLMs) in real-world applications is not solely constrained by technical security considerations but also by complex regulatory, governance, and ethical requirements. As these models increasingly rely on large-scale personal and proprietary data, ensuring lawful, transparent, and accountable data handling practices has become a fundamental prerequisite for trustworthy AI systems.

6.8.1. Data Protection and Privacy Regulations

Multiple regulatory frameworks have been introduced to govern the collection, processing, and storage of personal data used in AI systems. Prominent examples include the General Data Protection Regulation (GDPR) in the European Union, the California Consumer Privacy Act (CCPA), the Health Insurance Portability and Accountability Act (HIPAA), and the proposed European Union Artificial Intelligence Act. These regulations impose obligations related to data minimization, explicit consent, transparency, auditability, and the right to access or erase personal data. The main characteristics of the proposed acts are summarized in Table 9.

For LLMs, compliance with such regulations requires careful control over training datasets, including anonymization or pseudonymization techniques, secure data storage, and clear documentation of data provenance. Failure to adhere to these legal requirements can result in significant legal liabilities and erosion of public trust.

Table 9. Regulations in place for enhancing data privacy in concern of AI training data.

Regulation	Jurisdiction	Key Provisions	AI Implications
General Data Protection Regulation (GDPR) [67]	European Union (EU)	Lawfulness, fairness, and transparency; data minimization; right to access and erasure (“Right to be Forgotten”); accountability and data protection by design; and explicit consent for data processing	Train AI models on anonymized or pseudonymized data, provide transparency in AI data processing, and allow users to challenge AI-driven decisions
California Consumer Privacy Act (CCPA)	California, USA	Right to data access and portability; opt-out of data selling; right to deletion; and mandatory transparency on data usage	Honor opt-out and deletion requests, clearly disclose AI use for advertising or analytics, and manage compliance across multiple jurisdictions
Health Insurance Portability and Accountability Act (HIPAA)	USA	Safeguarding Protected Health Information (PHI), encryption and de-identification requirements, strict access controls	Ensure AI models handling healthcare data meet PHI protection standards and encrypt and de-identify patient data
Personal Data Protection Act (PDPA) [68]	Singapore	Accountability for data usage,	Secure AI training and deployment processes and obtain explicit user

		clear consent and notification requirements, and security obligations against unauthorized access	consent for data collection and processing
Artificial Intelligence Act (Proposed)	European Union (EU)	Categorization of AI applications by risk (minimal, high, and unacceptable) and transparency mandates for high-risk applications	Implement risk assessments for AI systems and ensure transparency, especially in sensitive areas like biometric identification
General Data Protection Law (LGPD) [69]	Brazil	Inspired by GDPR, emphasizes data protection rights, lawful basis for data processing, and data portability rights	Align AI practices with lawful data processing bases and facilitate data portability for users

6.8.2. Data Governance and Accountability

Data governance encompasses the policies, standards, and organizational practices that ensure responsible and ethical data management throughout the AI lifecycle. In the context of LLMs, effective governance frameworks support accountability, traceability, and compliance by enforcing access control, audit mechanisms, and documentation of model development processes.

Strong governance structures also enable organizations to demonstrate compliance with regulatory requirements and facilitate post-deployment audits. By integrating governance mechanisms early in the development process, organizations can mitigate risks related to data misuse, unauthorized access, and regulatory violations.

6.8.3. Ethical Considerations and Societal Impact

Beyond legal compliance, ethical challenges such as bias, fairness, discrimination, and misuse represent critical concerns for LLM deployment. Models trained on biased or unbalanced datasets may generate discriminatory outputs, potentially violating non-discrimination principles embedded in data protection laws.

Additionally, the opaque nature of many LLM architectures limits explainability, transparency, complicating accountability, and user trust. Ethical AI frameworks emphasize the importance of transparency, human oversight, and fairness to ensure that automated decision-making systems align with societal values.

6.8.4. Alignment Between Technical Safeguards and Policy Requirements

While technical solutions such as differential privacy, encryption, and federated learning provide important safeguards, they are not sufficient in isolation. Effective deployment of LLMs requires alignment between technical security mechanisms and regulatory obligations, often referred to as privacy-by-design and security-by-design principles.

Proactive strategies, including regular audits, transparency tools, and governance policies, are essential for bridging the gap between technical innovation and legal compliance. Integrating these approaches supports sustainable, ethical, and trustworthy AI systems capable of operating across jurisdictions and application domains.

7. Proposed Directions

Since AI technologies, such as GPT models, continue to develop, progress in privacy-preserving technology and addressing ethical and governance concerns is becoming more and more important, whether it is in the form of zero-knowledge proofs, new forms of encryption, or federated learning, or it is in the emergent ethical dialogues around responsible AI application, privacy as a fundamental right, and AI developer accountability.

In the same way that LLMs and GPT models are increasing in scale and influence, so too does the difficulty of ensuring that they run securely, ethically, and transparently. The future of AI will also be defined by advancements in privacy-preserving techniques, a stronger focus on ethical considerations, and widely accepted governance frameworks.

Among the most promising frontiers in secure AI research is that of zero-knowledge proofs (ZKPs). A key feature of ZKPs is that one party can prove the truth of a statement without disclosing the information behind it, a feature that is valuable for verifying model inputs or training data providers in a privacy-preserving manner. In AI, ZKPs can be employed to validate that federated learning updates come from real sources but without revealing raw user data. The most recent research aims to achieve better scalability and computational efficiency of ZKPs, e.g., for real-time AI inference as well as for encrypted pipelines.

At the same time, new encryption tools are emerging. Homomorphic encryption enables computation over encrypted data, preserving privacy end to end, but it still has performance challenges when it is applied to large models such as GPT models. Multi-party computation (MPC) allows collaborative model training without the need to share sensitive datasets, which makes it particularly appealing for regulated domains like healthcare or finance. Furthermore, in light of the long-term risks to classical encryption posed by quantum computing, researchers are developing quantum-safe cryptographic protocols that can be used to future-proof AI systems.

Federated learning is on the rise as a building block for distributed AI. Recent methods for personalized federated learning, based on adapting global models to local data distributions without sharing data, can achieve better accuracy while maintaining privacy [70]. Cross-device federated learning helps to efficiently scale up AI systems even where there are disparities in individual devices from processing power and bandwidth. Note that the inclusion of differential privacy in federated learning possibilities prevents an individual user from being re-identified through the introduction of controlled noise in aggregated updates.

The ethical foundations of AI, as AI systems are gaining more prominence, is more relevant than ever. Characterizing privacy as a fundamental human right is reforming global AI norms, driving developers to consider developing models with informed consent, data transparency, and user self-determinacy. The ethical development of AI, after all, means more than just safeguarding privacy; it also means respecting the dignity and choices of the users of these systems.

Fairness is also another major concern, as training data bias can result in outputs that are biased to mirror or perpetuate societal disparities, particularly in high-stakes areas such as hiring, healthcare, and criminal justice. To combat this, researchers are building fairer datasets and designing algorithms that can detect and even correct biases. Regular audits and fairness checks are crucial for determining if such systems, including GPT models, have and continue to meet such expectations across different applications.

The ethical implications of AI development are also coming under closer scrutiny. When a language model does harm, whether it creates false information, leaks sensitive data, or makes biased recommendations, it can be difficult to say exactly who is responsible. Suggestions for increasing accountability include recording detailed information about model development, sources of training data, and deployment procedures. Furthermore, explainable AI (XAI) techniques can render model decisions more interpretable, so both developers and users know how and why the model gives a certain output.

All of the factors mentioned above can be integrated into a common framework for model development, allowing an effective balance between privacy and performance and increasing user trust, removing bias, and offering transparency. The chart in Figure 3 explains the process.

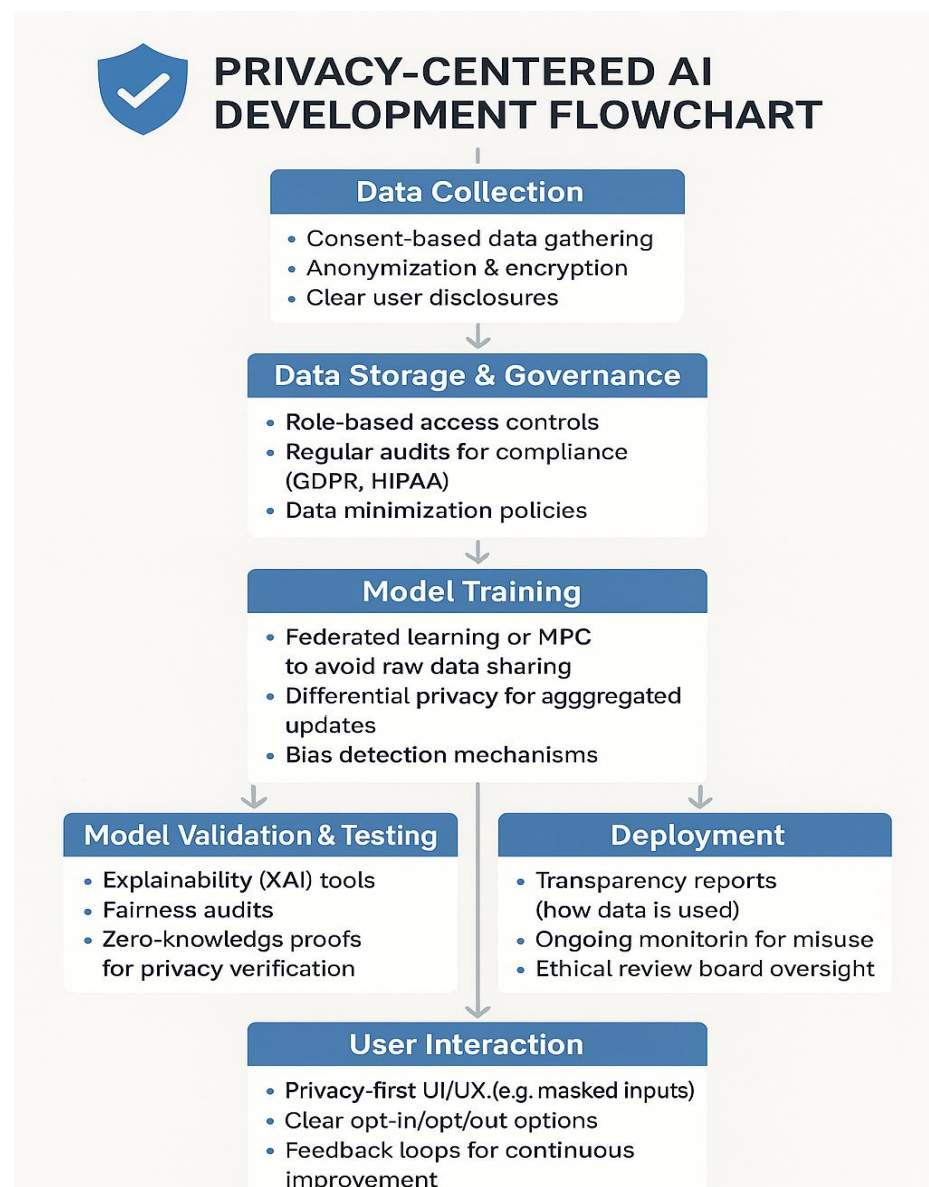


Figure 3. Flowchart of how GPT model training should occur to include data privacy and maintain performance.

Responsible use of generative AI also involves long-term ethical considerations. Policymakers and technologists are discussing how to stop AI models from being abused to spread disinformation, surveil people, or carry out cyberattacks. Just as important, however, will be the pursuit of technological innovation with social equity, thereby making sure that AI is not causing more job loss or digital divide.

According to the principle of privacy by design, privacy requirements are to be built throughout the entire lifecycle of AI, starting from data collection and pre-processing, through model training, to data use and deployment. This principle is practically implemented by enforcing strict data governance standards (including anonymization, encryption, and access control), as well as clear transparency requirements, to let users know what is being done with their data [71].

Navigating complex and fragmented regulatory environments, such as GDPR in Europe, CCPA in California, and HIPAA in the healthcare industry, is one of the key challenges for international companies. International harmonization of these rules is necessary to minimize the burden of regulation and to facilitate responsible innovation.

Organizations such as the United Nations and the OECD are in the process of developing common AI standards, such as standardized privacy metrics and benchmarking tools.

Lastly, establishing independent oversight bodies enables the possibility of guaranteeing that AI systems are compatible with ethical norms and public interest. Ethics boards can also oversee significant AI deployments, evaluate risks, and suggest mitigating strategies. One of the most important of these is also public–private partnerships, which have a role to play in supporting the development of responsible AI ecosystems, including through the sharing of datasets, co-creation of ethical benchmarking tools, and such support for transparency. To make it more convenient for entities to conform, governments might consider providing rewards, like tax deductions or research grants, for privacy-enhancing technologies, as well as cracking down on organizations that fail to maintain rigorous data protection protocols.

Innovations in privacy-preserving AI, ethical AI development, and robust policy frameworks are essential for addressing the challenges of security and privacy in AI systems. By combining advanced technologies, ethical safeguards, and comprehensive regulations, the AI industry can foster responsible innovation and build trust with users and stakeholders.

8. Conclusions

Based on a thorough examination of the main techniques, issues, and advancements in data security and privacy for GPT models and similar AI models, a number of important results are obtained. These demonstrate the interaction between technology, ethics, and policy.

Homomorphic encryption, secure multiparty computation (SMPC), and quantum-safe encryption will be essential ingredients in protecting contemporary sensitive data. These methods not only provide secure processing and storage but can be computationally expensive; therefore, improving them and implementing them during the AI development process is mandatory.

Techniques such as pseudonymization, generalization, and differential privacy prevent personal data from being traced or remembered by the model, improving data privacy. On top of that, a differential privacy trade-off between privacy and usability can be achieved by adding some control noise, but this control noise requires careful tuning to not degrade performance.

The use of decentralized training helps to avoid data exposure, ensuring security. Techniques such as secure aggregation and generalization in federated learning make this realizable for large-scale usage.

Model distillation, generation of synthetic data, or using public datasets that do not reveal secret details of entities make AI systems more efficient by lessening their reliance on raw data.

GPT models are not without risks, like memorization, adversarial poisonous data, and data leaks. All of these risks require constant model auditing and strong mitigation strategies. There are trade-offs between privacy and performance because privacy-preserving techniques commonly lower the accuracy and scalability of models.

Developments in protocols such as ZKPs and differential privacy also attempt to solve these trade-offs, but refinement is still needed. Advanced privacy operations translate to increased computations and energy consumption, resulting in barriers for deployment across the board.

Federated learning and encryption seem promising but require optimization to scale well.

Guaranteeing the transparency and explainability of the model is a necessary condition for trust-building, and it is both a technical and ethical challenge.

Privacy needs to be seen as a human right with privacy-respecting models. Reducing bias and promoting fairness in AI systems are vital to prevent harmful discriminatory behavior. Laws such as GDPR and CCPA provide the basis for data protection but require alignment across geographies for global adherence. Well-established sector-wide norms in transparency, data minimization, and responsibility are needed to make sure AI is used ethically.

The next frontier in privacy-preserving AI is zero-knowledge proofs and quantum-safe encryption. Also, having a class of methods called explainable AI (XAI) can serve as a bridge between complex models and users, providing a better explanation of how models generate content based on user data by requiring documentation, reports, and monitoring to hold organizations accountable.

Ensuring a stable trade-off between privacy, performance, and energy will be key for the sustainability of privacy-preserving AI systems in the long run.

GPT models must be integrated with state-of-the-art privacy-preserving techniques, ethical AI principles, and strong governance frameworks in order to be used in a safe and responsible manner. In a rapidly changing technology landscape, it will be important to address challenges such as the scalability of AI, how it can be held accountable, and how we can make it transparent, in order to ensure that AI benefits humanity while also respecting human rights. Further development and cooperation between technologists, ethicists, and policymakers are prerequisites to attain these goals.

This study has several limitations that should be acknowledged. First, the analysis was restricted to Q1–Q2 journal articles indexed in the Web of Science, which ensured the inclusion of high-quality and influential research but may have introduced selection bias by excluding relevant contributions from Q3–Q4 journals, regionally focused publications, or emerging studies with lower citation visibility. Although the Q1–Q2 filter was applied at the full-text assessment stage to support a more comprehensive evaluation of content quality and relevance, this decision may have introduced minor inconsistencies during the selection process.

In addition, Early Access publications were excluded to maintain dataset consistency and reproducibility, as such articles often lack finalized bibliographic metadata (e.g., volume, issue, and pagination) and exhibit unstable citation metrics across indexing databases. However, this exclusion may have resulted in the omission of recent and potentially innovative research, which is frequently disseminated in Early Access form prior to formal publication. This trade-off between data stability and timeliness represents an inherent limitation of the review.

Second, the reliance on secondary data limits the ability to assess the real-time applicability and operational risks of AI systems across diverse institutional and technical contexts. While the findings are most relevant to settings with established AI infrastructures, their broader generalizability remains uncertain. Future research could address these limitations by incorporating multiple journal tiers, a wider range of educational and organizational contexts, real-world case studies, and Early Access publications, as well as by triangulating quality assessment using complementary impact and relevance metrics to better balance rigor, timeliness, and external validity.

Author Contributions: Conceptualization, D.G.; methodology, D.G.; formal analysis, D.G.; investigation, D.G.; writing—original draft preparation, D.G.; writing—review and editing, D.G.; visualization, D.G.; supervision, D.E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hang, C.N.; Yu, P.D.; Morabito, R.; Tan, C.W. Large Language Models Meet Next-Generation Networking Technologies: A Review. *Future Internet* **2024**, *16*, 365. <https://doi.org/10.3390/FI16100365>.
2. Motlagh, F.N.; Hajizadeh, M.; Majd, M.; Najafi, P.; Cheng, F.; Meinel, C. Large Language Models in Cybersecurity: State-of-the-Art. *Int. Conf. Inf. Syst. Secur. Priv.* **2025**, *2*, 98–110. <https://doi.org/10.5220/0013377600003899>.
3. Das, B.C.; Amini, M.H.; Wu, Y. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* **2024**, *57*, 51.
4. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. *High Confid. Comput.* **2024**, *4*, 100211. <https://doi.org/10.1038/s41591-023-02448-8>.
5. Miranda, M.; Ruzzetti, E.S.; Santilli, A.; Zanzotto, F.M.; Bratières, S.; Rodolà, E. Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions. *Trans. Mach. Learn. Res.* **2025**, *2025*, arXiv:2408.05212.
6. Wang, S.; Zhu, T.; Liu, B.; Ding, M.; Ye, D.; Zhou, W.; Yu, P.S. Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.* **2025**, *58*, 35.
7. Li, M.Q.; Fung, B.C.M. Security Concerns for Large Language Models: A Survey. *J. Inf. Secur. Appl.* **2025**, *95*, arXiv:2505.18889.
8. Chen, K.; Zhou, X.; Lin, Y.; Feng, S.; Shen, L.; Wu, P. A Survey on Privacy Risks and Protection in Large Language Models. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 163.
9. Du, H.; Liu, S.; Zheng, L.; Cao, Y.; Nakamura, A.; Chen, L. Privacy in Fine-Tuning Large Language Models: Attacks, Defenses, and Future Directions. In *Lecture Notes in Computer Science (LNAI)*; Springer: Singapore, 2025; Volume 15873, pp. 326–344.
10. Wu, H.; Cao, Y. Membership Inference Attacks on Large-Scale Models: A Survey. **2025**, *1*, arXiv:2503.19338.
11. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
12. OpenAI; Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; et al. GPT-4o System Card. *arXiv* **2024**, arXiv:2410.21276.
13. Rodriguez, J.D.; Hay, T.; Gros, D.; Shamsi, Z.; Srinivasan, R. Cross-Domain Detection of GPT-2-Generated Technical Text. In Proceedings of the NAACL 2022–2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 1213–1233. <https://doi.org/10.18653/V1/2022.NAACL-MAIN.88>.
14. Deprecations—OpenAI API. Available online: <https://platform.openai.com/docs/deprecations> (accessed on 1 September 2025).
15. Models—OpenAI API. Available online: <https://platform.openai.com/docs/models> (accessed on 1 September 2025).
16. Zhao, X.; Li, L.; Wang, Y.X. Provably Confidential Language Modelling. In Proceedings of the NAACL 2022–2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 943–955. <https://doi.org/10.18653/v1/2022.naacl-main.69>.
17. Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramèr, F.; Zhang, C. Quantifying Memorization Across Neural Language Models. In Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, 1–4 May 2023.
18. Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, arXiv:2306.11698.
19. Welsh, A.P.; Edwards, M. Text Generation for Dataset Augmentation in Security Classification Tasks. *arXiv* **2023**, arXiv:2310.14429.
20. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. <https://doi.org/10.1136/BMJ.N71>.
21. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; Antes, G.; Atkins, D.; Barbour, V.; Barrowman, N.; Berlin, J.A.; Clark, J.; et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann. Intern. Med.* **2009**, *151*, 264–269. <https://doi.org/10.7326/0003-4819-151-4-200908180-00135>.
22. McKenzie, J.E.; Brennan, S.E. Overviews of Systematic Reviews: Great Promise, Greater Challenge. *Syst. Rev.* **2017**, *6*, 185. <https://doi.org/10.1186/S13643-017-0582-8>.
23. Fariani, R.I.; Junus, K.; Santoso, H.B. A Systematic Literature Review on Personalised Learning in the Higher Education Context. *Technol. Knowl. Learn.* **2022**, *28*, 449–476. <https://doi.org/10.1007/S10758-022-09628-4>.

24. Tahamtan, I.; Bornmann, L. What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018. *Scientometrics* **2019**, *121*, 1635–1684. <https://doi.org/10.1007/S11192-019-03243-4>.
25. Achar, S. Data Privacy-Preservation: A Method of Machine Learning. *ABC J. Adv. Res.* **2018**, *7*, 123–130. <https://doi.org/10.18034/ABCJAR.V7I2.654>.
26. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In Proceedings of the 28th USENIX Security Symposium, Santa Clara, CA, USA, 14–16 August 2018; pp. 267–284.
27. Tirumala, K.; Markosyan, A.H.; Zettlemoyer, L.; Aghajanyan, A. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38274–38290.
28. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
29. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **2020**, *6*, 346–360. <https://doi.org/10.1016/J.ENG.2019.12.012>.
30. Jedrzejewski, F.V.; Thode, L.; Fischbach, J.; Gorschek, T.; Mendez, D.; Lavesson, N. Adversarial Machine Learning in Industry: A Systematic Literature Review. *Comput. Secur.* **2024**, *145*, 103988. <https://doi.org/10.1016/J.COSE.2024.103988>.
31. NeurIPS. Towards AI-Based Auditing of Privacy Risks in Privacy-Enhancing Technologies. Available online: <https://neurips.cc/virtual/2023/83053> (accessed on 10 February 2026).
32. Park, D.; Lee, E.; Lee, J.-W. Powerformer: Efficient and High-Accuracy Privacy-Preserving Language Model with Homomorphic Encryption. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; Volume 1, pp. 11090–11111.
33. Ferrag, M.A.; Ndhlovu, M.; Tihanyi, N.; Cordeiro, L.C.; Debbah, M.; Lestable, T.; Thandi, N.S. Revolutionizing Cyber Threat Detection with Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices. *IEEE Access* **2024**, *12*, 23733–23750. <https://doi.org/10.1109/ACCESS.2024.3363469>.
34. Huang, W.; Cheng, A.; Zhang, Z.; Wang, Y. DPF-CM: A Data Processing Framework with Privacy-Preserving Vector Databases for Chinese Medical LLMs Training and Deployment. In *Findings of the Association for Computational Linguistics: EMNLP*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 12904–12916.
35. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and Accurate Deep Learning with Electronic Health Records. *NPJ Digit. Med.* **2018**, *1*, 18. <https://doi.org/10.1038/S41746-018-0029-1>.
36. Hanrahan, G. Computational Neural Networks Driving Complex Analytical Problem Solving. *Anal. Chem.* **2010**, *82*, 4307–4313. <https://doi.org/10.1021/AC902636Q>.
37. Voigt, P.; von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 2nd ed.; Springer: Cham, Switzerland, 2024; pp. 1–339. <https://doi.org/10.1007/978-3-031-62328-8>.
38. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *Int. Data Priv. Law* **2017**, *7*, 76–99. <https://doi.org/10.1093/IDPL/IPX005>.
39. Riad, A.B.M.K.I.; Barek, M.A.; Rahman, M.M.; Akter, M.S.; Islam, T.; Rahman, M.A.; Mia, M.R.; Shahriar, H.; Wu, F.; Ahamed, S.I. Enhancing HIPAA Compliance in AI-Driven MHealth Devices Security and Privacy. In Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), Osaka, Japan, 2–4 July 2024; pp. 2430–2435. <https://doi.org/10.1109/COMPSAC61105.2024.00390>.
40. The Digital Services Act Package | Shaping Europe’s Digital Future. Available online: <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package> (accessed on 3 December 2024).
41. Daemen, J.; Rijmen, V. *The Design of Rijndael*; Springer: Berlin, Germany, 2020. <https://doi.org/10.1007/978-3-662-60769-5>.
42. Rivest, R.L.; Shamir, A.; Adleman, L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Commun. ACM* **1983**, *26*, 96–99.
43. Gentry, C. Fully Homomorphic Encryption Using Ideal Lattices. In Proceedings of the forty-first annual ACM symposium on Theory of computing (STOC '09), New York, NY, USA, 31 May–2 June 2009; pp. 169–178.
44. How Practical Is Homomorphic Encryption?—MCSI Library. Available online: <https://library.mosse-institute.com/articles/2022/06/how-practical-is-homomorphic-encryption/how-practical-is-homomorphic-encryption.html> (accessed on 2 September 2025).
45. Goldreich, O. *The Foundations of Cryptography*; Cambridge University Press: New York, NY, USA, 2004; References—Scientific Research Publishing; Volume 2, Basic Applications. Available online: <https://www.scirp.org/reference/referencespapers?referenceid=2077842> (accessed on 2 September 2025).

46. Bender, E.M.; Gebru, T.; Mcmillan-Major, A.; Shmitchell, S.; Shmitchell, S.-G. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Toronto, Canada, 3–10 March 2021. <https://doi.org/10.1145/3442188.3445922>.
47. Dwork, C.; Roth, A.; Dwork, C.; Roth, A. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. <https://doi.org/10.1561/04000000042>.
48. Samarati, P. Protecting Respondents' Identities in Microdata Release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. <https://doi.org/10.1109/69.971193>.
49. Narayanan, A.; Shmatikov, V. Robust De-Anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 18–22 May 2008.
50. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
51. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy, SP, San Jose, CA, USA, 22–26 May 2016; pp. 582–597. <https://doi.org/10.1109/SP.2016.41>.
52. Khaled, E.E. Concepts and Methods for De-Identifying Clinical Trial Data. Available online: <https://nap.nationalacademies.org/resource/18998/EIEmamandMalin-Paper.pdf> (accessed on 3 September 2025).
53. Sweeney, L. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570.
54. Sebestyen, H.; Popescu, D.E.; Zmaranda, R.D. A Literature Review on Security in the Internet of Things: Identifying and Analysing Critical Categories. *Computers* **2025**, *14*, 61. <https://doi.org/10.3390/COMPUTERS14020061>.
55. Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, Ú.; et al. Extracting Training Data from Large Language Models. In Proceedings of the 30th USENIX Security Symposium Vancouver, Canada, 11–13 August 2021; pp. 2633–2650.
56. Hardy, S.; Henecka, W.; Ivey-Law, H.; Nock, R.; Patrini, G.; Smith, G.; Thorne, B. Private Federated Learning on Vertically Partitioned Data via Entity Resolution and Additively Homomorphic Encryption. *arXiv* **2017**, arXiv:1711.10677.
57. Shah, S.P.; Deshpande, A.V. Addressing Data Poisoning and Model Manipulation Risks Using LLM Models in Web Security. In Proceedings of the International Conference on Distributed Systems, Computer Networks and Cybersecurity, ICDSCNC, Bengaluru, India, 20–21 September 2024. <https://doi.org/10.1109/ICDSCNC62492.2024.10941696>.
58. Ethicist: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation. Available online: https://www.researchgate.net/publication/372248359_Ethicist_Targeted_Training_Data_Extraction_Through_Loss_Smoothed_Soft_Prompting_and_Calibrated_Confidence_Estimation (accessed on 2 September 2025).
59. Balle, B.; Cherubin, G.; Hayes, J. Reconstructing Training Data with Informed Adversaries. In Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 23–26 May 2022; pp. 1138–1156. <https://doi.org/10.1109/SP46214.2022.9833677>.
60. Haresamudram, K.; Larsson, S. Three Levels of AI Transparency. *Computer* **2022**, *56*, 93–100. <https://doi.org/10.1109/MC.2022.3213181>.
61. Yu, H.; Guo, Y. Generative Artificial Intelligence Empowers Educational Reform: Current Status, Issues, and Prospects. *Front. Educ.* **2023**, *8*, 1183162. <https://doi.org/10.3389/FEDUC.2023.1183162>.
62. Orenstrakh, M.S.; Karnalim, O.; Suarez, C.A.; Liut, M. Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases. In Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), Osaka, Japan, 2–4 July 2023; pp. 121–126. <https://doi.org/10.1109/compsac61105.2024.00027>.
63. Hu, J.; Liang, Y.; Zhao, W.; Mcareavey, K.; Liu, W. An Interactive XAI Interface with Application in Healthcare for Non-Experts. *Commun. Comput. Inf. Sci.* **2023**, *1*, 649–670. https://doi.org/10.1007/978-3-031-44064-9_35.
64. (PDF) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Available online: https://www.researchgate.net/publication/323302750_The_Malicious_Use_of_Artificial_Intelligence_Forecasting_Prevention_and_Mitigation (accessed on 2 September 2025).
65. OpenAI (2023). ChatGPT (Mar 14 Version) [Large Language Model]—References—Scientific Research Publishing. Available online: <https://www.scirp.org/reference/referencespapers?referenceid=3704885> (accessed on 2 September 2025).
66. Chen, C.; Fu, J.; Lyu, L. A Pathway Towards Responsible AI Generated Content. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Macao, China, 29–25 August 2023; pp. 7033–7038. <https://doi.org/10.24963/ijcai.2023/803>.
67. General Data Protection Regulation (GDPR)—Legal Text. Available online: <https://gdpr-info.eu/> (accessed on 3 December 2024).
68. Liang, X.; Ye, N. Bart Custers et al. (2019): EU Personal Data Protection in Policy and Practice. *Int. J. Leg. Discourse* **2020**, *5*, 119–125. <https://doi.org/10.1515/IJLD-2020-2031>.

69. Hartzog, W.; Richards, N.M. Privacy's Constitutional Moment and the Limits of Data Protection. *Boston Coll. Law Rev.* **2020**, *61*, 1687. <https://doi.org/10.2139/ssrn.3441502>.
70. Morar, C.D.; Popescu, D.E. A Survey of Blockchain Applicability, Challenges, and Key Threats. *Computers* **2024**, *13*, 223. <https://doi.org/10.3390/COMPUTERS13090223>.
71. Morar, C.D.; Popescu, D.E.; Novac, O.C.; Ghiurău, D. Rethinking Blockchain Governance with AI: The VOPPA Framework. *Computers* **2025**, *14*, 425. <https://doi.org/10.3390/COMPUTERS14100425>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.