SECURE DATA HANDLING FOR LARGE LANGUAGE MODELS (LLMS) AND ARTIFICIAL INTELLIGENCE (AI)

by

Jacob Thomas
B.tech, MBA

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

Mar, 2026

SECURE DATA HANDLING FOR LARGE LANGUAGE MODELS (LLMS) AND ARTIFICIAL INTELLIGENCE (AI)

by

Jacob Thomas

Supervised by

Dr. Abhilash Kayyidavazhiyil

APPROVED BY

Dissertation chair - Dr. Gualdino Cardoso

RECEIVED/APPROVED BY:

Renee Goldstein Osmic

Admissions Director

Dedication

I dedicate this dissertation to my parents, siblings' spouse and kids who has been with me through thick and thin and motivated me to go forward with my research and studies irrespective of all the busy schedule in work and life, blockers and deterrents. I also dedicate this work to my teachers in school, High school and college who has been my guiding lamp throughout my life by introducing the world of alphabets to me and making me capable of doing this work, streamlining my thoughts and presenting it confidently to a panel for consideration. I am also taking this opportunity to thank all the nuns of the convent school where I studied which has paved way for who I am today. Their words of wisdom has always been the bedrock of my career, knowledge and wisdom. I am also thankful to my Guide and mentor **Dr Abhilash** Kayyidavazhiyil who has been patient with me and steered me through my research journey.

ABSTRACT

SECURE DATA HANDLING FOR LARGE LANGUAGE MODELS (LLMS) AND ARTIFICIAL INTELLIGENCE (AI)

Jacob Thomas
2026

Dissertation Chair: Dr. Abhilash Kayyidavazhiyil

The development of Large Language Models (LLMs), particularly the generative models and AI technologies, has changed many fields, paving the way with far-reaching effects in enhancing natural language understanding and generation, empowering many organizations to transform from traditional to learning organizations. But such developments also pose a plethora of challenges, especially to the secure handling of the data used for training, considering the sensitive nature of the data itself and the privacy threats posed to AI training and deployment across geographies and domains. This dissertation investigates the principles, methodologies, and technologies required for secure data management in respect of LLMs, AI systems, among the rational agents currently available in the market. This work investigates training data from the perspective of data privacy and governance mechanism in place with the help of secondary research and recommends an inclusive framework of training data being recorded in Blockchain for guaranteeing secure data management and identifying data leakage through the entire lifecycle of AI learning and output generation. Most of these training data is obtained from

webpages scrapings of the crawlers from internet and from online data sources. Secure data handling is essential and crucial to keep its users' privacy protected, preserve the trust, and adhere to regulatory requirements like GDPR and HIPAA and unfortunately, this aspect is lacking when learning happens by defining the LLMs on the market based on their exposure against the datasets

This dissertation also explores if DLT (Distributed Ledger Technology) and hashing which is a backbone of Blockchain technology can be a solution to the workaround of Data leakage identification and fine tuning of the LLM model. Concept of Digital Twin and the possibility of Digital Twin used for feedback training / fine tuning has been visited in this work. This dissertation is intended to prove that training data entry in Blockchain ledger is a solution to data leakage and can significantly capture instances of data leakage by inherently identifying the trained data in a hashed format and act as a backward loop in fine tuning the model to eradicate the occurrence of generating data in an unintentional manner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:
INTRODUCTION

1.1 Introduction

LLMs need large-scale data for training: there are risks connected to risk of data breaches (either knowingly or unknowingly), illegal access, and misuse. This topic sheds light on investigations of the mechanisms of safe data management designed particularly for AI systems to mitigate risk of breaches. Most of the LLMs use the same trainer functions and hence its necessary to assess and list the issues around secure data handling regarding LLM and AI in the current market Generative AI tools ecosystem. Once the problem is confirmed and the extent to which the data exposure is finalized a possible framework is to be explored where existing security needs are assessed and how it fits into the learning applied to GenAI tools to suggest a uniform framework for secure data management in AI practices which also can be sustainable in the case of training with the current available data set in the market and future synthetic data trained LLM. This emphasizes the fundamental need for secure data management in designing and deploying both LLMs and AI systems. AI is a probabilistic system in which the outcome is generated based on the probability of the token input and the proximity of the training data in the vector database to which training data learning has been mapped. Through the interplay of technical safeguards and robust governance and regulatory frameworks, organizations can manage risk and increase trust in artificial intelligence technologies. The need for a framework to overcome this risk is a foundational need for practitioners aiming to augment data security in AI workflows.

1.2 Research Problem

Once the problem is confirmed and the extent to which the data exposure is finalized a possible framework is to be explored where existing security needs are assessed and how it fits into the learning applied to GenAI tools to suggest a uniform framework for secure data management in AI practices which also can be sustainable in the case of training with the current available data set in the market and future synthetic data trained LLM. This emphasizes the fundamental need for secure data management in designing and deploying both LLMs and AI systems. AI is a probabilistic system in which the outcome is generated based on the probability of the token input and the proximity of the training data in the vector database to which training data learning has been mapped. Through the interplay of technical safeguards and robust governance and regulatory frameworks, organizations can manage risk and increase trust in artificial intelligence technologies. The need for a framework to overcome this risk is a foundational need for practitioners aiming to augment data security in AI workflows.

1.3 Purpose of Research

While the proliferation of Large Language Models has transformed many sectors, the need to deal with expansive datasets has exposed them to considerable threats especially the leakage of sensitive training data. This problem is compounded in such enterprise environments where LLMs handle sensitive internal data, therefore privacy and security of data is essential. This requires solid mechanisms that can help mitigate the risks of unauthorized disclosure while protecting the utility and potential of these powerful AI technologies Of particular interest, a practical recommendation that integrates Distributed Ledger Technology with cryptographic hashing to create an immutable and verifiable record of data provenance will improve data security and allow transparent auditing of LLM training datasets. This approach utilizes DLT's built-in immutability and transparency to create an immutable record of data transactions, so that any fraudulent data alteration or leakage is rapidly recognized and tracked. The purpose of integrating

blockchain with LLMs is to overcome their limitations, promote further technological improvement and produce new ways to address the security and privacy issue. This proposal develops a novel scheme to combine DLT and the associated cryptographic hashing mechanisms for the strong defense against data leakage problem with a theoretical and practical statistic-oriented rationale of proposed architecture. An extension of the LLM model with the help of a Digital Twin is also attempted as a recommendation which supports the deterministic conclusion of a probabilistic system for the data leakage problem.

1.4 Significance of the Study

This study is significant as it attempts to answer the question ailing business users in multiple domains where LLM is used. Testing of LLM is an age-old problem which industry is still working around considering that the system is expected to work based on the prompt provided. This study is an attempt to throw light on the Data Leakage of LLM and Gen AI frameworks that we have in market currently and an attempt to see how this can be identified and a mechanism can be recommended by which Data leakage is proven with certainty and feedback is given to the LLM learning as the disincentive in generating / reproducing a learning data can be a fine tuning trigger.

1.5 Research Purpose and Questions

Overall this dissertation is an attempt to address the following Problem statements:

1. Is the Data Leakage in LLM or AI significantly high? Is this an anomaly or an inherent fault of the system? How will we be able to convert the outcome of a probabilistic system into a deterministic certainty of Negative Data exposure?

2. Is storing the Training Data in Blockchain a feasible solution to identify leakage and establish provenance for the generated output training data presence? Can Blockchain Smart contracts convert the flagging of the Data leakage to fine tuning of LLM?

CHAPTER II:

REVIEW OF LITERATURE

2.1 Theoretical Framework

Types of Learning in LLM

Development of Large Language Models led to a revolution in artificial intelligence and the discussion on the reliance on human-labeled data and the ever-changing machine learning paradigm (Liu, 2023). Central to this evolution is the division between supervised learning and unsupervised learning and the ever-evolving complexity of manual intervention in the algorithms' performance. This section aims to clarify the fundamental differences between these two core learning techniques within the context of training large language models, focusing on their respective merits and drawbacks, as well as the key points in learning that require human oversight and fine-tuning. More specifically, we will focus on how supervised learning that relies on explicit supervision signals compares to unsupervised learning that can understand the intrinsic characteristics in unlabeled datasets and how these models are frequently interconnected in new LLM-development environments. Additionally, we will also consider how manual intervention processes, from data curation and annotation to advanced fine-tuning efforts such as Reinforcement Learning with Human Feedback, serve as a crucial link between the raw model features and the desired performance and alignment (Liu et al., 2023). This examination will demonstrate the continuing importance of human judgment, which can be seen in navigating the intricacies of LLM training, with an increased reliance on automated techniques (Gehring & Grigoletto, 2023). We will also discuss self-supervised learning as a hybrid approach that leverages large-scale unlabeled data by creating supervisory signals from the input and in turn bridging the gap between purely supervised paradigms and unsupervised (Johnson & Hyland-Wood, 2024; Xu et al., 2024). This research will further explore multiple steps involved in LLM training such as pre-training, supervised fine-

tuning, Reinforcement Learning from Human Feedback, in order to describe each learning paradigm (Li et al., 2024). Lastly, we will focus on the data exposure of LLMs and their implications for their use, recognizing that higher-order models of training may increase performance, but there is a need for continuous consideration of the social impact of this approach (Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), 2023). As such, knowledge of supervised and unsupervised methods and their relation, as well as careful (manual) intervention, are crucial to the design of powerful, ethically responsible LLMs. Strong insights into these learning paradigms are critical for model performance optimization, bias mitigation, and responsible AI development. Such grounded understanding of these challenges can aid researchers and practitioners in managing the complexities of model creation, choice of data, and alignment efforts to develop AI systems that are more capable and trustworthy. In the context of large language models, supervised learning essentially deals with training on well-labeled datasets where there are unambiguous pairwise input-output pairs guiding the model. This approach is best for tasks that have a high level of specificity of output: classification, translation, or question-answering, where human-annotated examples are superimposed to provide explicit cues to the model for learning from (BARBERÁ, 2025). Unsupervised, on the other hand, is concerned with learning hidden feature structures and patterns in unlabeled language data – commonly used in models (in pre-training) for processing large language models for nonverbal features or a lack of explicit human description (Akiri et al., 2025). Recently, large language models have emerged after the initial introduction of deep learning approaches due to a pool of computational power, greatly enabling self-supervised learning on large populations of unlabeled text data and providing the link between purely supervised and unsupervised approaches (Raiaan et al., 2024). The initial pre-training phase enables LLMs to gain an overall knowledge of linguistic structures, semantics, and understanding of the world (Yan et al., 2025). The framework, which is under the umbrella of intra-processing bias mitigation, allows the model to acquire complex representations which can later be fine-tuned under supervision on a set of labeled, specific input data (Gallegos et al., 2023). Supervised fine-tuning is based on large

datasets of high quality and human-driven data and it helps the model to learn from its training data to perform specific downstream tasks by fine-tuning on fine-tuned data, which improves fine-tuned performance on focused tasks (He et al., 2024). This process is important to tailor the generalized model knowledge to the actual needs of a task, which in general, produces large improvements in accuracy and relevance (Parthasarathy et al., 2024). Nevertheless, the weakness of the supervised fine-tuning model lies in the high data load of labeled data, cost, and time consumption (Hayder, 2025). Notably, such a process is key to fine-tuning LLMs for specific instructions and their usefulness and safety for different applications (Xiao & Zhu, 2025; Xu et al., 2024). Supervised pre-training specifically can reduce the amounts of labeled data being input for the model to be further fine-tuned and it helps in learning robust representations (Matarazzo & Torlone, 2025). The simplicity of this supervised approach, through pre-training or fine-tuning, fits well with traditional paradigms in machine learning (Xiao & Zhu, 2025). However, as LLMs scale, the practicality of supervised pre-training diminishes; the creation or manipulation of exhaustively labeled datasets, which are critical for basic language understanding, becomes impossible (Matarazzo & Torlone, 2025; Xiao & Zhu, 2025). Instead, self-supervised learning has become a mainstay of modern LLM development, which relies on model learning from the structure of text corpora, where the model will predict the masked words or the next word in a sequence and hence create its own supervised signals (Kalyan, 2023). This approach enables us to use great volumes of raw text dataset without human annotations and reduces labor and time cost from supervised data labeling (Matarazzo & Torlone, 2025; Truong, 2024). This pre-training stage, with the model being exposed to large amounts of text materials, and it uses its self-supervised learning, predicting masked tokens of the text from the input information, in order to learn the meaning of words (Topcu et al., 2025), is one of the first steps in the developing of the LLM pipeline. This first stage sets the stage for rich linguistic knowledge and a general knowledge base which is fine-tuned by supervised fine-tuning to improve task-specific performance and its match with expectations (Hanieh et al., 2024; Liu et al., 2024). This later, supervised, fine-tuning process is crucial in matching the large linguistic knowledge gained in the pre-training

phase to specific downstream tasks, which essentially means learning from annotated data to map the input data to the recognized output (Ramos et al., 2024; Xiong et al., 2024). These tasks often require instruction fine-tuning, in which LLMs learn with instruction-response pairs (Xiao & Zhu, 2025), and thus follow human instructions, allowing them to perform myriad instructional tasks. This method, where the model was pre-trained beforehand on a large, general dataset and then fine-tuned with limited task-specific data, was a hallmark of the supervised fine-tuning manner (Anisuzzaman et al., 2024). Such hybrid approach enables LLMs to take advantage of the scalability of unsupervised learning for foundational knowledge acquisition to benefit from the precision and task-specificity of supervised signals (Naveed et al., 2023). This paradigm has gained tremendous success with the advent of transformer architectures since they have built-in capabilities to handle the scaling of vast datasets and have sophisticated self-attention mechanisms, essential for separating complex linguistic relationships (Liu et al., 2024). Supervised and unsupervised learning are thus considered a pragmatic approach for training LLMs in this context, which allows LLMs to capture general linguistic regularities from large unlabeled sets of data before they specialize through the specific tuning of the instruction using a label. This unsupervised pre-training phase, supplemented with supervised fine-tuning, is fundamental in providing effective generalized and flexible models in the era of large-language models (Jukić, 2025). In the pre-training stage, LLMs are unsupervised learners which are fed with large amounts of unlabeled data and are learning linguistic patterns (such as masked language modeling or next-token prediction) enabling them to master natural language without explicit human labels (Li et al., 2025; Yan et al., 2025). The preliminary unsupervised learning sets up a stable linguistic substrate, allowing the model to learn aspects of grammar, syntax, and semantics that are integrated into human language (Dong et al., 2024). This unsupervised way of pre-training is important because this strategy grants the model the ability to obtain a generation distribution from large-scale text corpora in an unsupervised way, which has been demonstrated to lead to later task-driven processing (Huang et al., 2024; Ji et al., 2024). Next, the fine-tuning process uses supervised learning phase, where the pre-trained model

is trained with labeled datasets to optimize its performance for certain tasks such as text classification, question answering, or sentiment analysis (Shi et al., 2023). This phase enriches the general knowledge of the model itself by making specific adjustments to apply within the real world and maximizing the practical applicability as well as the accuracy (Xiao & Zhu, 2025). Such supervised fine-tuning in the case of reinforcement learning can use approaches such as supervised fine-tuning and RLHF to adapt the model with human-friendly preferences and enhance its helpfulness, honesty, and harmlessness (Li et al., 2025). The latter, RLHF in particular, is notably important for improving the behavior, since it helps the model to produce information that not only fits in accordance with fact but is also contextually relevant and ethical, mostly taking input from human workers to iteratively improve its response (Li et al., 2025). A typical iterative reinforcement learning (RL) cycle, that often involves the inclusion of human evaluation data, is also important for correcting biases and ensuring the LLMs produce well-matched output to human values and instructions, specifically in zero-shot conditions (Naveed et al., 2023). This extensive pre-training is essential for LLM to achieve an advanced grasp on language types, patterns as well as syntax and semantics, by analyzing massive sets of data (Chakraborty & Basu, 2024). This basic understanding then facilitates the model to gain substantial performance benefits in various applications under some follow-up tune-up work (Boateng et al., 2024). More precisely, pre-training will entail the model being trained on large text sets, typically on internet data, focusing on learning the essential rules and structure of language, including for example, determining the next word in a sequence (Dong et al., 2023; Tong et al., 2024). In this unsupervised phase, the model can learn to have a strong internal representation of language without an explicit human annotation on every data point (Aghaei et al., 2025; Wang et al., 2025). This critical step is required to get general language understanding and generation, since LLMs run billions of parameters on large amount of data (Minaee et al., 2024). Nevertheless, maintaining exact constraint to the behavior of these complex high-scale unsupervised language models still faces barriers, as they are not explicitly trained (Rafailov et al., 2023). The challenge is inherent on this front, but supervised methodologies and human intervention guide the model's desired outcomes

as well as ethical considerations (Neuman et al., 2025). Thus, techniques from such an approach as Reinforcement Learning from Human Feedback have proved to be valuable in aligning these models with human dispositions, sometimes by considering human labels on the quality of model generations (Rafailov et al., 2023). This calibration guarantees that the models not only generate coherent and meaningful text but obey ethical norms, and answer in the way human reviewers see as useful and harmless (Shah et al., 2024; Wang et al., 2023). This multi-stage training strategy, which integrates unsupervised pre-training with other supervised fine-tuning and human alignment methods, has allowed LLMs to accomplish a variety of tasks from text completion to intricate problem solving (Srivastava & Aggarwal, 2025). This rigorous training process enables models that display complex comprehension and create outputs, closing the distance between sparse data patterns and highly sophisticated human-like communication. Indeed, as an illustration of the advantages of this holistic training strategy, the LLMs possess unprecedented few-shot or zero-shot ability of transferring learnt knowledge into novel, unseen tasks (Aghaei et al., 2025; Shao et al., 2024). In other words, this shows the need for a hybrid training approach, where unsupervised learning lays a broad linguistic foundation for refinement and guidance, especially with dedicated trained interventions and human-generated inputs (Li et al., 2025; Uhlig et al., 2024). For instance, semi-supervised training, a blend of unsupervised and self-supervised approaches, becomes the most effective when unlabeled data is high but labeled data are limited (Chen et al., 2024). By utilizing the advantages across both paradigms, this method enables models to use big datasets to learn features while learning explicitly from human input with respect to individual tasks (Mirbagheri et al., 2025). One essential aspect of this hybrid approach is the challenge of treating the hallucination problem, where the models produce factually false but confident-sounding data by repeatedly aligning the model's outputs with human preferences and verified information (Bai et al., 2024; Svozil, 2025).

LLM AI Model	Parameters	Year
BERT	340 million	2018
GPT-2	1.5 billion	2019
Meena	2.6 billion	2020
GPT-3	175 billion	2020
LaMDA	137 billion	2022
BLOOM	176 billion	2022

Table 1 : Models and Training parameter volume (till 2022)

The quality and diversity of your training data for the Large Language Models are huge factors determining their performance, therefore the identification and annotation of the data needs to be thorough (Hao et al., 2025). This usually includes a step-by-step process, from the generation of large amounts of raw data, and followed up by advanced preprocessing steps, with fine-grained annotation to align models and fine-tune them to the specific tasks (Hu et al., 2024). Building and perfecting such datasets is of utmost importance, based on aspects concerning the number of data sets, data collection strategies, and general properties of information (Ju & Ma, 2024). The first stage, data collection, is of great importance, in which considerable amount of data is extracted, and the relevant amount of data is cleaned to select the model fit for training (BARBERÁ, 2025). The definition of the problem is crucial for the success of this phase, along with the desired output, data type and task goal need to get well defined to guide data generation, avoiding the collection of irrelevant data (Sarker, 2024). So, the next problem would be collecting different types of data coming from web pages, books, conversations, and special documents, making sure there is enough data when training a model. After which this extensive collection is cleaned, normalized for inconsistencies, irrelevant contents and standardized text for uniformity called Data sanitization. This guarantees that the data that is collected has a big amount of content but is also used efficiently in training to minimize

noise and maximize the signal for the LLM (Gotavade, 2024). The subsequent stages are an exploration of more complex processes of data identification and labeling, which are essential in converting raw data sets to systematic formats suitable for the ingestion and learning of the models (Alansari, 2025). This refinement is particularly important within the domain of supervised learning, as precise and consistent labeling allows the algorithm to recognize patterns and make informed predictions (Jiang, 2024). This systematic method for data preparing is crucial for the robustness of the performance and generalization of LLMs on many tasks, to minimize the amount of labelled data for later fine-tuning stages (Matarazzo & Torlone, 2025). The quality and the level of distribution of this fundamental data can play a critical role in the model generalization and bias, as a result, detailed data preprocessing has been seen as a required step before the first stage of training (Jeong, 2023; Liu et al., 2024). This foundational preparation is further structured into an array of necessary sub-stages, such as data collection, filtering, deduplication, standardization, and review — all of which are essential for the final quality of the training corpus (Liu et al., 2025). For example, OpenGPT-X used disparate pipelines for curated and web data, as a consideration for the fact web-derived content requires higher filtering and deduplication than those that derive from curated datasets (Brandizzi et al., 2024). The fact the datasets come from web-crawled sources, public repositories, and proprietary datasets, brings a huge challenge of data privacy and data bias issues. To meet these challenges, robust methodological choices, such as detailed documentation of all sources of training data to account for accountability and regular verification for statistical distortion or biases, must be made, and unauthorized or sensitive content should be checked against the model. On top of this, resilient data governance structures are crucial in dealing with the ethical implications of such huge and heterogeneous datasets, particularly with respect to safeguarding users' privacy and the lack of accidental disclosure of sensitive data on it (Guilherme Junior & Clinton, n.d.). The immense strength and intricacy of this diversity of datasets, however, is additional source of difficulties of promoting fairness and responsibility in that models are expected to reflect and even replicate the societal biases in training data that may otherwise be incorporated in model (Tariq et al., 2025). Hence,

mitigation strategies (anonymization and pseudonymization techniques) are important to mitigate the risk of privacy and to comply with data protection legislation (BARBERÁ, 2025). This is more important when working on large datasets which may include personally identifiable information, proprietary information, or other sensitive data which requires wide-ranging privacy-protecting solutions during the formulation of data as well as in training a model (Zhao & Song, 2024). Proper use of such methods will also support ethical regulation and confidence in AI systems (BARBERÁ, 2025). Having data access governance practices in place in advance of data leaks and ensuring ethical compliance of model use is essential for reducing the risk of data leakage and addressing legal requirements as GDPR, especially when working with multiple data sources and personal data (BARBERÁ, 2025). In addition, continuous monitoring and auditing of the data pipelines are necessary to identify and recover any vulnerable regions that may contribute sensitive information or the integrity of the training data (BARBERÁ, 2025). Such a continuous oversight allows the whole lifecycle of data, from collecting it to deploying it, to be safe and ethical, increasing the reliability and trustworthiness of LLMs.

Hypothesis Testing in Issues of Data Leakage in AI

Data used for training has an appearance for multiple machine learning security scenarios including membership inferencing attacks, auditing differentially private algorithms, property/attribute inferencing (Nikolić et al., 2025). At risk projects are statistically tested based on hypothesis testing to determine that the adversary derives intelligent information which is not just randomly derived but is much more probable than chance. Specifically, it helps measure false positive rates, preventing the perceived "leaks" from being treated as random incidents and lending credibility to the security assessment (Wei et al., 2024). Many of the problems of large language models are not properly addressed in the literature, and can result in data leaks from training data, fine-tuning phases, and prompt history, and necessitate the use of bespoke statistical technique (Zangana et al., 2024). In this sense, the memory capacity of LLMs would worsen these dangers; it will not learn certain

information or remember information previously from a training corpus, or it could even result in re-imitation an implicit leak of a learned corpus or infer a clue even without using direct memorization as the primary leaker (Abdali et al., 2024). We saw the leak of LLM data inference in Prior Work. Studies have concentrated on the quantification of data leakage in LLMs, mainly by contrasting model outputs with training data directly and hence relying on a certain measure based on metrics such as ROUGE-L or semantic similarity (Jiang et al., 2024). On the other hand, because it is always probabilistic, these types of approaches do not yield a statistically true estimate of data leakage sensitivity (Feretzakis & Verykios, 2024). This discrepancy underscores the requirement for a strong methodology, for instance statistical hypothesis testing which formally measures data leakage probability by stating whether there is/is not probability of leakage of data with a list of null and alternative hypotheses (Jiang et al., 2024). Beyond being a detector of these signals, it provides an actionable degree of confidence that any sensitive information has been compromised (Abdali et al., 2024). Moreover, newer toolkits, like LLM-PBE, indicated that evaluating data privacy concerns in LLMs is necessary, which is a difficult task using classical approaches (Li et al., 2024).

2.2 Data Anonymization and pseudonymization

Old traditional approaches for data anonymization i.e., k-anonymity and differential privacy, have been widely used in many fields to protect private data not compromising on the usefulness of data (Yan et al., 2024). In the framework of Large Language Models, dataset preprocessing – such as filtering and de-duplication – is important to eliminate vulnerabilities in privacy protection (Zangana et al., 2024). This includes exclusion or generalization of personally identifiable information such as names, addresses, and social security numbers to prevent reuse (Yan et al., 2024). Such tools can be employed at any level of detail (from entire documents to sentences) and dropping low quality or duplicate data could be considered also an attempt at high quality of data (Matarazzo & Torlone, 2025). Alternative methods beyond just deletion such as random substitution replace

certain entities with substitutive entities, thereby preserving natural language structure, which can potentially lead to semantic drift (Miranda et al., 2024). Pseudonymization, a more complex framework, replaces sensitive data with pseudonymized names and therefore, permits the analysis at the expense of obfuscating immediate connections with individual people (Yermilov et al., 2023). This approach enables reversible de-identification under controlled, controlled circumstances, thus striking the right balance between data utility and privacy preservation, as opposed to irreversible anonymization (Yermilov et al., 2023). Nevertheless, deploying such techniques to the unstructured, complex text data of LLM training creates its own challenges which include preserving linguistic consistency and semantic soundness while also ensuring the full masking of sensitive data (Patsakis & Lykousas, 2023). For example, the advanced nature of LLMs implies that even anonymized data may be able to be reverse engineered to expose sensitive data in the form of adversarial attacks or inversion methods (Zhou et al., 2025). This makes advanced privacy-preserving techniques necessary to avoid possible accidental data exposure without losing the high efficiency of the train corpus for robust model performance (Xiao & Zhu, 2025). This requires the development of specialized anonymization and pseudonymization strategies specific to both the linguistic nuances and extensive size of LLM training datasets (Miranda et al., 2024). Automated tools can be used to filter sensitive content from these large pre-training corpora, but this can inadvertently lead to loss of information or reduced dataset diversity, damaging the robustness of LLMs (Xiao et al., 2023). As an example, OpenAI and Anthropic apply stringent filtering and fuzzy deduplication to eliminate PII, with the goal of building more suitable data quality and privacy preservation in their models (Yan et al., 2024). But such aggressive filtering strategies may include demographic groups that may inadvertently be excluded, which can restrict the generalization of the model to different downstream problems such as hate speech detection and privacy de-identification, as seen in the LLaMA 2 models which do not have an extensive filtering scheme and can be used for enhancing the versatility of the base model (Liu et al., 2024). But the existing methods mainly depend on a simple regular expression approach to filter Personally Identifiable

Information from pretraining data, which, although efficient, brings few security protections (Miranda et al., 2024). Techniques that utilize machine learning models for PII detection and redaction have greater accuracy compared to traditional approaches but are still challenged by context-sensitive or subtle forms of sensitive information (Feretzakis & Verykios, 2024). Additionally, detecting PII is challenging due to being an attempt involving multilingual datasets and an array of overlapping types, where tools tend to privilege one kind of data over the others resulting in incomplete or incorrect masking (Asthana et al., 2025). Moreover, given the vast and dynamic nature of LLM training datasets, human review is rarely feasible which requires automated solutions tailored to the changing nature of privacy threats and data modalities (Das et al., 2024). As a result, although corpora cleaning techniques are good at a surface level now, they often have to weigh the importance of privacy considerations against the instrumental utility of the dataset for effective LLM training (Das et al., 2024). This calls for strong privacy protection for LLMs, as they are constantly adapting to their new environment of large web corpora, and doing in-depth analyses is difficult of large size and continuous update (Chen et al., 2025; Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), 2023). Use of filtering approaches, such as bloom filter methods, constitutes another defense mechanism against memorization, that scanned the training datasets to determine if a subsequent token of a model contains the n-gram in the training set, and then selected an alternative token from the posterior of the model if a match was found (Shanmugarasa et al., 2025). Given that LLMs, as part of training programs, often uncover sensitive information learned through their training (Miranda et al., 2024; Smith et al., 2023a, 2023b), the approach immediately handles the problem of training data memorization that is a major one. But despite their efforts, cleanliness and prevention of PII leakage is still a considerable challenge. Scrubbing methods minimize the risk, but they do not eliminate PII leakage (Lukas et al., 2023). Named Entity Recognition models inherently struggle to recognize forms of all types of PII. This is even more difficult when the nature of confidential information is context-dependent and/or lacks clearly defined definitions (Shilov et al., 2024).

This section explains the different anonymization methods used, ranging from simple redaction and generalization to advanced techniques like k-anonymity and l-diversity, each chosen based on what properties the data has to offer and what level of privacy protection is desired. Such approaches are essential to avoid privacy risk posed by LLMs, particularly regarding memorization and the inadvertent exposure of sensitive data during training (Miranda et al., 2024). This is an important challenge because it balances the privacy guarantees provided by these techniques but still insists on preserving the utility of the data and preventing catastrophic forgetting, which can significantly affect the model performance (Wang et al., 2025; Zhou et al., 2024). Such strategies are typically developed through multi-levels of approaches that achieve optimal privacy-utility trade-offs, considering the idiosyncratic protection/privacy risks faced by LLMs for example training data leakages (Smith et al., 2023; Yan et al., 2024). For example, it has been shown that fine-tuning LLMs with multiple sensitive instances raises the likelihood of a privacy leak by more than 60% relative to baseline, so effective anonymization is crucial (Ramakrishnan & Balaji, 2025). This highlights the importance of both the implementation of these approaches and the need to assess their effectiveness in the face of advancing data protection regulations and the dynamic nature of PII in large datasets (Asthana et al., 2025). In addition, the fact that data privacy regulations are constantly changing necessitates adaptive anonymization frameworks, which can adapt on the fly to new compliance challenges and emerging threats (Aditya et al., 2024). Furthermore, those approaches may not be effective if the training data is of poor quality or if the data used is of low diversity, but rather if dependent on the sophistication of Named Entity Recognition and co-reference resolution modules implemented into the anonymization pipeline, which is highly variable between datasets (Miranda et al., 2024; Yang et al., 2024). Furthermore, because identifying and redacting sensitive free text content within massive datasets is inherently challenging, strong (yet computationally efficient) anonymization solutions are required, as anonymization methods, such as homomorphic encryption, require heavy computational investments and offer high privacy guarantees (Chen et al., 2023). As a result, researchers

are exploring hybrid approaches that provide both light on information anonymization and encryption methods to make the right choice between privacy and computing power (Guilherme Junior & Clinton, n.d.). Under these circumstances, the choice and careful application of anonymizing approaches are key to building privacy-preserving LLMs in a manner that does not overly impair their performance or usefulness (Feretzakis & Verykios, 2024). We find that the trade-off between data utility and privacy preservation has been a central theme in recent work that explores the performance versus re-identification risk in several metrics defining how much information has gone missing post-anonymization (Zhao & Song, 2024). Such metrics are usually about measuring performance on downstream tasks after anonymization, measuring the reduction in the distinguishability of individual records; estimating the frequency of successful re-identification attacks (Yang et al., 2024). In this respect, this evaluation requires a focused evaluation of the degree of empirical relevance between the pseudonymization techniques used and the quality of both text and later NLP tasks (Yermilov et al., 2023) based on the quality of text (i.e., Named Entity Recognition models or sequence to sequence transformations). Thus, this requires an adaptive definition of privacy and a balanced approach to privacy-preserving data publishing methods (Sun et al., 2024). These measures are important since they address the extent to which anonymization techniques undermine the integrity and utility of the original data for different analytical purposes (Pissarra et al., 2024; Yermilov et al., 2023). It would frequently involve a multi-objective optimization problem that tries to maximize utility while upholding very stringent privacy limits, especially when applying anonymization techniques, including suppression, tagging, random substitution, or generalization to identified sensitive information (Du et al., 2024). On the other hand, Privacy-Preserving Data Publishing algorithms provide a systematic mechanism to hide entities using the disclosure hazard as a predictor via a privacy model outlined by domain specialists, yet most research fails to prioritize the usefulness of anonymized text for downstream processes (Yang et al., 2024). Such neglect can result in instances in which the privacy benefits of the anonymized datasets are compensated by a loss of the functionality of the data for the purpose of training advanced LLMs (Yang et

al., 2024). As a result, creating proper evaluation metrics suitable for generative anonymization with LLMs is of importance, so that both privacy and utility are properly preserved, and not limited to traditional metrics that might not cover the nuances in language generation tasks (Pissarra et al., 2024).

In this section, we describe the design choices and training paradigms used, and performance evaluation of the LLMs, focusing on how pre-processed and anonymized data are integrated into the training pipeline. Within the model architectures (e.g., transformer-based networks), training goals, loss function, and optimization strategies implemented to improve model performance and protect anonymized data are described in depth. The evaluation framework includes both intrinsic and extrinsic metrics measuring linguistic generation quality, which relates to performance on the downstream NLP tasks using the anonymized data (Pissarra et al., 2024). This assessment includes the re-identification risk of the anonymized data in the context of a range of adversarial inference attacks, as well as the AI algorithms themselves, frequently taking LLMs as advanced anonymizers to assess the goodness of privacy preserving transform methods (Staab et al., 2024). This involves a rigorous consideration of the trade-offs between the privacy guarantee and the loss of the model utility, notably since state-of-the-art LLMs can now reach almost human-like results in extracting personal data from ostensibly anonymized text (Staab et al., 2024). Accordingly, the resulting framework should also have complex metrics to evaluate the associated remaining privacy threat and the effect on model performance on multiple tasks, such as text generation quality, factual accuracy and language style compliance, especially when complex de-identification processes are employed (Aghakasiri et al., 2025). This often takes a variety of techniques, for instance with differential privacy mechanisms or paraphrasing models, which provide differentially tuned noise in the data to hide specific records, while still distributing the data with controlled accuracy (Miranda et al., 2024). Moreover, the efficiency of these privacy-preserving methods is frequently judged by metrics assessing statistical similarity between the anonymized and original datasets, while also considering data utility within specific downstream applications (Yuan et al., 2024).

Assessment data set of this kind are essential to better understand both the effect of anonymization on output quality and balance between privacy protection and model performance (Miranda et al., 2024). The European Data Protection Board emphasises that if an AI model is to be considered anonymous, the chances of extracting personal data directly from training data or by inference queries must be insignificant (BARBERÁ, 2025). This calls for rigorous evaluation methodologies that can quantify this "insignificance" through privacy attacks and utility evaluations (Yang et al., 2024). Adversarial anonymization thus takes advantage of LLMs' inferences themselves to better inform and refine the process of anonymization, strengthening privacy against progressively sophisticated inference attacks (Staab et al., 2024). This is accomplished by employing LLMs to detect personal attributes in anonymized text, with a smaller number of successful inference indicating better privacy preservation (Kim et al., 2025). This method measures whether anonymization succeeded by examining whether a robust adversarial LLM can nonetheless infer associated attributes from the generated text, in more accordance with regulation and provides a less ambiguous sense of anonymization (Staab et al., 2024). Additionally, consistent assessment via adversarial feedback loops, where LLMs try to de-anonymize data (repeatedly), has been found to increase privacy as well as the utility of the anonymized data, beyond the limits of regular anonymization methods (Staab et al., 2024). This is especially important regarding the real privacy guarantees that can be attained through using LLM-based adversarial methods, like attribute inference attacks, which should be effectively applied for these datasets to measure and prove that they are robust against state-of-the-art inference (Frikha et al., 2024; Staab et al., 2024). An iterative adversarial process that enables the gradual adjustment of anonymization methods to maintain an optimal trade-off between maintaining the privacy of data while also increasing the utility given its use, as evaluated in ROUGE scores of text quality and for specialized LLM-based utility judges (Frikha et al., 2024; Staab et al., 2024). Hence, this strong evaluation framework also leads to a deeper understanding and better balance on trade-off between privacy and utility in LLM training data, where no outputs that are generated become corrupted while sensitive data are

protected severely (Sun et al., 2024). However, the vulnerability of LLMs to data extraction attacks, even after anonymization, is a crucial issue to avoid, as this can compromise privacy through the disclosure of personal information (Mökander et al., 2023). Thus, privacy-preserving LLMs, in which one of the techniques is differential privacy, play a pivotal role in addressing such risks (Huang et al., 2025). These advanced approaches are crucial for defending against membership inference and attribute inference attacks where sensitive information can accidentally become exposed due to model outputs or gradients, hence establishing strong privacy risk evaluation frameworks (Chen et al., 2025). This often requires developing and deploying sophisticated adversarial models that seek re-identifiable patterns from the anonymized data—assisting transformations in their validation against sophisticated inference attempts (Staab et al., 2024). This also includes rigorous testing of proposed anonymization frameworks against well-documented data extraction attacks that have been demonstrated to extract sensitive data from LLM training sets (Deng et al., 2024). Additionally, adoption of "human-in-the-loop" design methods and selective memory strategies are pivotal to augmenting privacy protection by enabling human monitoring and limiting access to sensitive data throughout the LLM architecture (Dong et al., 2025). A thorough comprehension of such adversarial vulnerabilities, including those associated with out-of-distribution inputs, and adversarial robustness is thus critical to producing genuinely secure and private LLMs (Dong et al., 2024). The effectiveness of these strategies depends on solid privacy risk assessment with consideration of data leakages, adversarial vulnerabilities and regulatory compliance (Chen et al., 2025). Further, critically, mitigation of these risks includes the strategic creation and the generation of benchmarks and datasets specific to evaluating the safety fit between LLM and data (including annotated data showing harmful as well as inappropriate content) (Peng et al., 2024). This includes proactive activities to mitigate unintentional leaks of sensitive information in training and deployment for LLMs and necessitate strong data security policies beginning from data ingestion through model insight generation (Guilherme Junior & Clinton, n.d.; Zhao & Song, 2024). Such practices demand the use of cryptographic best practices such as preparing for quantum-resistant standards and taking

advantage of confidential computing environments to further protect such sensitive information (Guilherme Junior & Clinton, n.d.). This integrated paradigm serves to guarantee privacy is baked into the whole LLM lifecycle in the beginning of data consumption, through a model's development, and from its deployment, and helps to reduce semantic privacy risks that span beyond traditional data protection considerations (Ma et al., 2025). In addition, routine audits of the models for leaks of privacy as well as a composite of multiple anonymization methods are necessary to enhance the effectiveness against advanced attacks (Guilherme Junior & Clinton, n.d.). These methods are necessary to ensure strong security in LLM data ecosystems with respect to privacy regulation changes while preserving public trust in AI methods (Guilherme Junior & Clinton, n.d.). Additionally, by including smaller and more specific models, the memorization effect often seen in over-parameterized LLMs can be reduced and there is less danger of accidentally reproducing sensitive training data (BARBERÁ, 2025). These approaches, in addition to differential privacy to layer in statistical noise, are important in striking the balance between utility and privacy guarantees in the deployment of large language models (Guilherme Junior & Clinton, n.d.). Combining these methods within an extensive security architecture, that covers effective access restrictions and data encryption for sensitive data further secures LLMs privacy (Guilherme Junior & Clinton, n.d.). Furthermore, there are a variety of mechanisms for the implementation of model-level security, network, user, and continuous monitoring, as well as for detecting possible misuse, bias, or vulnerabilities in a real-time manner (BARBERÁ, 2025).

2.3 Methodology

To address these limitations, we propose a comprehensive approach by applying statistical hypothesis testing to quantify the risk of data leakage across the lifespan of Large Language Models from pre-trained corpora and fine-tuning datasets through user-submitted prompts

This approach uses a perturbation testing of distributions to transform an LLM leakage study into a frequentist hypothesis testing problem by constructing empirical null and alternative output distributions within a low-dimensional semantic similarity frame to support interpretable inferences, with no restricted assumption in distributions (Paulius et al., 2025). The introduction of this statistical method to guide the classification of leakage detection provides the groundwork for distinguishing accidental data generation from statistically meaningful violations of privacy boundaries as evidence suggests LLMs can memorize training data and leak sensitive information out of pre-training corpora (Chen et al., 2023; Duan et al., 2024; Ramakrishnan & Balaji, 2025). This involves explicitly learning training sequences directly as well as the inferred exposure of private attributes or structures embedded in model parameters (Cai et al., 2025; Li et al., 2024). We devise a leakage assessment as a statistical hypothesis test to formally detect and quantify the impact of this exposure. Here, the null hypothesis H_0 is that the ***output distribution of the prediction output is independent of sensitive training data*** so no leakage takes place while the alternative hypothesis H_1 asserts that the ***output distribution of the predictions token is significantly dependent on secret or sensitive content*** (Cai et al., 2025; Qu et al., 2025; Rauba et al., 2024). Our validation of such effects is an essential aspect to assess the impact of the training data on our LLMs. Therefore, this distributional framework of LLM output generation is the basis of the full stochastic response modeled by examining statistically the distribution and comparing from these distributions and identifying fine-grained patterns that could be missed in a single sample evaluation (Rauba et al., 2024). The test statistic S is created from the likelihood ratios of model probability distribution produced; if the probability coefficient $S(x) > \gamma$ is greater, then we reject null hypothesis H_0 by accepting the alternative hypothesis H_1 as in this case the model has statistically significant evidence of data inclusion (Zhang et al., 2024). The threshold value γ per the Type I error rate is a value to be found to ensure that the probability of false alarm signal of the data breach which is not valid is in accepted level of the significance α (Cai et al., 2025; Ji et al., 2025). As has been previously demonstrated in the data leakage detection literature, the threshold of rejection for data misappropriation

detection will require the employment of large deviation theory to quantify the asymptotic rates of type I and type II errors and can transform the detection problem into a solvable problem using minimax algorithm (Cai et al., 2025). This optimization of this algorithm successfully balances the leakage detection method as sensitive to errors with the specificity to prevent false positive and to ensure that the statistical test is robust in detecting instances in which the model memorized errors and inadvertently makes sensitive training instance(s) (Duan et al., 2024; Kaneko & Baldwin, 2024). Such a minimax-based optimization of an algorithm toward detecting traces of bad data is similar to distributional methods, which demonstrate that evaluation of aggregated statistical properties is generally more robust to noise/corrosion than single instance approaches, particularly if the datasets used to train models are large enough (Cai et al., 2025; Li et al., 2025) to achieve the coherence in what are known that are distressing values, that differentiate one true point of data misuse from random coincidence. Drawing on our definition, we now move to derive the particular statistical hypotheses that underpin our empirical exploration of exposure risks. Defining Data Leakage in LLMs. Data leakage in the context of Large Language Models is the accidental exposure of sensitive data stored in LLMs with the generating outputs in the training or fine-tuning datasets (Cai et al., 2025; Li et al., 2024). To formalize the detection and assessment of this exposure, we develop the leakage evaluation as a statistical hypothesis test in which the null hypothesis (H_0) states the independent model's output distribution from sensitive training data, suggesting no leakage, and the alternative hypothesis (H_1) indicates that the generated tokens significantly depend on the secret or sensitive information (Cai et al., 2025; Qu et al., 2025; Rauba et al., 2024). This distributional approach encapsulates the complete stochastic response of the LLM's output generation and enables scientists to conduct numerical inference by comparing these distributions such that small fluctuations, which individual sample assessments may overlook, can be detected (Rauba et al., 2024). We define the test statistic S according to likelihood ratios based on the model's probability distribution, where $S(x) > \gamma$ results in the rejection of the null hypothesis H_0 for H_1, which describes significant evidence that the data is included (Zhang et al., 2024). To formulate these hypotheses, we

need to define a threshold \(\gamma \) with an appropriate rate of Type I errors (i.e., where the probability of wrongly identifying leaking data which does not exist is kept within an acceptable level of importance \(\alpha \)) (Cai et al., 2025; Ji et al., 2025). As previously mentioned in literature on data misappropriation detection, the rejection threshold can be found by utilizing a large deviation theory to analyze the case with asymptotic rates of type I and type II errors, making the detection problem a problem which can be solved using a minimax optimization problem (Cai et al., 2025). The optimization work is known to balance sensitivity of leakage detection mechanism and specificity for false-alarm prevention, ensuring the statistical test accurately checks detect cases in which the model has unintendedly "memorized" and recreated critical pre-training data (Duan et al., 2024; Kaneko & Baldwin, 2024). The efficiency of this minimax optimization in detecting small nuances of data misappropriation is consistent with distributional methods, which also demonstrate that reviewing aggregated statistical properties is more robust to noise and contamination than individual instance testing, since models need large parts of a dataset to display the consistent distributional mechanisms that separate genuine leakage from chance coincidence (Cai et al., 2025; Li et al., 2025). Based on this definitional framework, we move to the construction of the specific statistical hypotheses that will build the basis for our empirical analysis of exposure risks. Formulation of the statistical hypothesis. This allows us to rigorously measure the exposed information presence using a binary hypothesis testing framework in which the null hypothesis H_0 is: target output x is generated from general population distribution p of our model, independent of training set, and the alternative hypothesis H_1 is that x is drawn clearly from training distribution D of our target model (Mireshghallah et al., 2022). We specify a probability-ratio test statistic with the score function for separating these distributions from each other by comparing the probability of both observing the target output that we think is a null and one of the alternative hypotheses. In particular, the higher the score received by the score function, the likelihood the data was used in training, and this data-based probability is used as statistic as evidence against the null hypothesis.Officially, we use the decision rule to determine the likelihood ratio $LR = \frac{Pr_M(x)}{Pr_R(x)}$, in which Pr_M =

probabilities assigned by target model and Pr_R = probabilities derived from a reference model that has not been fine-tuned on the sensitive data (Mireshghallah et al., 2022). This method of comparison works on the assumption that models that have trained over a given population will assign a much greater probability to each population compared to models that have not.The Neyman-Pearson lemma, ensuring that the likelihood-ratio test performs the best statistically for a particular significance level (Ma et al., 2024; Zarifzadeh et al., 2023), is used to establish the best score function and rejection threshold for this test. Based on this lemma's formulation the most robust test rejects null hypothesis in favor of H_1 where the probability ratio passes the threshold created as a result of the proposed significance level which aims to maximize the probability that samples will be included in a detection program and to minimally suppress the rate of false positive events (Cai et al., 2025; Tao & Shokri, 2025). Approximate threshold is obtained from a minimax optimization problem applied to a limited class of probability distributions, providing strong performance under severe conditions, where the attacker does not have well-defined access to the data distribution.

2.4 Qualitative examples of data leakage

In addition to the combined measurements, individual examples of verbatim text creation are also specific demonstrations against which we can substantiate the data that the model can memorize and have made available to the public the personal and delicate training data, e.g the output for credit card numbers or codes in proprietary code that did appear in the pre-training corpus and were not tampered by any obfuscation or modification (Qi et al., 2025). For example, asking models to generate certain requests to begin paragraphs from copyrighted writings often results in the copying of exact sentences in text form (Duarte et al., 2024). Reproducing training sequences verbatim is not solely restricted to novel works and may lead to personally identifiable information; For instance, studies show that in response to the prompt with appropriately targeted queries, large language models can leak

names, phone numbers, and email addresses to human beings when asked for prompts (Kaneko & Baldwin, 2024; Šarčević et al., 2024). This phenomenon is especially true for larger architectures, since larger architectures are resistant to some extraction methods far better than smaller ones, but are intrinsically vulnerable to auto-completion attacks on data (Aditya et al., 2024). As an example, a previous study using GPT-3 fine-tuning API has shown that classification- or auto-complete task-tuning models can forget and reveal sensitive PII at the prompts (Aditya et al., 2024) while still producing a highly responsive recall of sensitive PII of such information (in its training condition) in black-box settings, even when the data are not only in general usage, but also that it can be remembered and shared after the prompt (Aditya et al., 2024). A study that used the fine-tuning API of GPT-3 has revealed, for example; models that have been specially designed for classification or auto-processed tasks can recover sensitive Personally Identification when asked about some specific context from its training set by some specific example from even in the practical of the black box settings. The vulnerability of these models is further suggested on the evidence that some extraction attacks, though having difficulty in extract coherent information from smaller or differently trained models, can return sensitive data even when specific context is provided for them, when given to large language architectures (Gu, 2024). As an example, investigations into training data extraction attacks have found that by conditioning the model with the correct prompt, sequences exactly memorized from the training data can be recovered (Duan et al., 2024; Miranda et al., 2024). Carlini et al. (Huang et al., 2023), used a 40GB training dataset with GPT-2 to be able to find 600 memorized sequences. This type of exposure is made easier (but in a non-exploratory way) when criminals take advantage of model ability to recollect similar text sequences, which tends to make extraction more probable (Huang et al., 2024; Smith et al., 2023). Auto-completions also have a risk since, by learning from the prediction, the threat is that more sensitive data will be preserved, for example, a part of the full body of the email can be produced from the subject line, and this too may lead to the leakage of sensitive personal identifiable information in fine-tuning data (Aditya et al., 2024; Das et al., 2024). Just like the former, membership inference attacks try to find out if some data record belongs to the

training set by examining the response pattern model to tell if the data is a member or a non-member (Aditya et al., 2024). As well as the rudimentary nature of membership determination, adversaries can take advantage of this feature to perform more intrusive purposes, such as extracting internal prompts with proprietary features, e.g. extracting proprietary internal prompts from commercial apps or extracting protected data from commercial apps through inference attacks, from inference into the data files (Abdali et al., 2024; Wang et al., 2025). One attack to which membership inference attacks fall is specifically leveraging model overfitting, where points of a given train set in the model would show overfitting (Aguilera-Martínez & Berzal, 2025). Carlini et al. This vulnerability was further confirmed by (Spirito et al., 2023), which captured more than 600 verbatim sequences of the text in GPT-2, including potentially personally identifiable information and code; such sequences may appear only once in the training set. It has been shown that this exposure is not limited to general language models, such as ChatGPT and open-source models, which demonstrated memorization of these data that was higher than before, allowing the extraction of sensitive personal data by accessing the user queries (Deng et al., 2025; Wang et al., 2024). Carlini et al. showed that GPT-2 tends to memorize personal information such as Twitter handles, emails and UUIDs (and other types such as a user account number) and the model's higher probability given more likelihood to sequence from its training data on the same data than a reference model (Miranda et al., 2024). However, the potential dangers of such extraction of this type are not only theoretical, in real world examples, for example. In the present day, models trained from clinical notes, medical records, google's 'Smart compose' had learned long random numbers such as social security numbers (Litzinger et al., 2025). When an answer could be extracted by prompting the model this kind of information recall was never expected due to its tendency to recall long patterns and short sequences from its training data without a prompt (Litzinger et al., 2025). This kind of material revelation of data breach becomes concrete as examples from within the company's own work record or medical records often finds its way inside, while its automation functionality enables the model to inadvertently repeat proprietary data without caution.

2.4 Summary

Performance – Security Trade Off

The growing capabilities of large language models have also opened a new period of AI applications and pose complicated balancing act problems regarding their inherent trade-offs between performance and security (The tension inherent in this tradeoff, which is often referred to as the security-performance tradeoff, is a critical factor in designing and deploying LLM-based systems, as improving one component often results in diminishing the other (Usman et al., 2024). The tension and balance between these two approaches is exacerbated given the heterogeneous contexts within which LLMs are employed — they can fit various assistive functionalities as well as autonomous agents, each of whose unique risk profile and performance expectations dictate the deployment of their respective LLMs (Usman et al., 2024). This requires a robust framework to measure and assess this trade-off that will allow developers to design model architecture, the safety settings, and deployment methods to respond to real-world conditions (Pfister et al., 2025). As an example, the safety performance targets, particularly to an LLM configured as a medical assistant, will differ with respect to general-purpose LLM, illustrating the contextual specificity of such evaluations (Usman et al., 2024). Additionally, the changing environment of adversarial attacks against LLMs highlights the importance of continually assessing and adjusting mitigation capabilities to improve the balance between the best performance and security, and consequently the balance between the two of these factors. This dissertation aims to examine some of these methods to both measure and evaluate this fine relationship, with specific reference to the benchmarks and metrics which are currently employed to quantify the effectiveness of LLM performance and the robustness of their security mechanisms (Huang et al., 2024). An analysis of attack success and defense true

positive rates is included which offer crucial information addressing LLMs' vulnerability to adversarial tampering and their ability to resist such interference (Peng et al., 2024). In addition, the comparison of these trade-offs often requires complex calculations like the False Refusal Rate -- measures how often safety provisions unintentionally restrict the utility of an LLM by rejecting valid prompts (Ferrag et al., 2024). This metric is critical to grasping how safety measures may not only prevent misuse, but can also limit overall functionality and helpfulness of the LLM in a given legitimate application (Ferrag et al., 2024). For instance, this trade-off can be quantified by considering the impact of security improvements, like adversarial training and safety fine-tuning, on accuracy, fluency, and utility (Mai et al., 2025). Consequently, we can find out how defensive measures degrade standardized performance standards. Alternatively, an optimisation for performance may cause exploitable vulnerabilities by adversaries, resulting in a decline in security (Usman et al., 2024). The emphasis of this paper is that current methodologies to evaluate LLM performance and security are to be defined, highlighting the complex interdependencies between these critical dimensions (Mai et al., 2025). It will also deal with how to develop uniform metrics that can cover all aspects of such dynamic relationships while also emphasizing the diversified threat landscape and applications' individualistic specifications (Mai et al., 2025). Ultimately, the point is to build a comprehensive understanding of defensive mechanisms in LLMs, with the hope that overall security will be beefed up for real-world applications (Zangana et al., 2024). Hence, empirical analysis of high-profile LLMs can highlight their inherent risks and susceptibility risks as they are shaped further by current adversarial prompt approaches (Akiri et al., 2025). The Efficiency-Performance-Robustness Trade-off Evaluation Framework provides a method for the evaluation of computational efficiency, task performance and adversary robustness, which can be used for determining LLM execution for optimal decision making and deployment decisions (Fan & Tao, 2024). This approach enables a more comprehensive understanding of LLM performance than just an input-output comparison of input models, given that most other methods have limitations in assessing realistic risk profiles for open-weight models in real-life, limited to only providing a lower-bound values at worst-case behavior in an open

learning framework and only to give a lower value for their worst-case implementation (Che et al., 2025). Acknowledging these weaknesses, holistic systems such as the Holistic Evaluation of Language Models provide different dimensions from general capabilities to assessment of performance and robustness, extending to security-oriented use-cases, ensuring that the analysis considers vulnerability at adversarial points of inputs, as well as efficacy from the applied defense point of view (CHEN et al., 2023). These high-level benchmarks are essential to systematically assess attacks and defenses, hence serve as a basis for future research and development in this dynamic area (Bhatt et al., 2024; Liu et al., 2023). These benchmarks assess LLMs in practical cybersecurity to identify model vulnerabilities and compare the security effectiveness of different defense mechanisms (Bhusal et al., 2024; Shen et al., 2025). Although there are advancements in the literature, there is a notable gap (Abuadbba et al., 2025; Jing et al., 2024) for standardized metrics and evaluation benchmarks targeted specifically to the performance and security conformance of a comprehensive scale on the intersection of LLM in cybersecurity environments where performance and security can meet. In particular, both SECURE and CyberSecEval 2 assess LLM ability in cybersecurity tasks, but they neither thoroughly evaluate the dynamic interaction between model utility and security against adaptive adversarial attacks (Ferrag et al., 2024; Pfister et al., 2025; Tian et al., 2025). To meet such needs, and not just any new and specific testing mechanism, a paradigm in cybersecurity domains must build new algorithms to quantify LLM's offensive and defensive performances in the cyber environment and develop more than single tests focusing on the individual or skill level as a separate measure of skill set as much as an integrated approach based on an evaluation criteria across different cybersecurity domains (Sanz-Gómez et al., 2025). Existing benchmarks such as those involving automated penetration testing or exploiting vulnerabilities illustrate the failure of existing models in practice, underscoring the need for more holistic, and more harmonized evaluation practices that cover the full range of security performance trade-offs (Isozaki et al., 2024; Zhu et al., 2025). For example, certain benchmarks narrow their focus on limited sets of skills or on individual exploits and, in consequence, miss the generalized capability that is the essential

requirement for operating in a real-world cybersecurity environment, which restricts the applicability of such benchmarks to suggest improvements in models for the purpose of cyber defense (Sanz-Gómez et al., 2025). This space of lack of complete evaluation frameworks highlights the necessity of benchmarks to evaluate an LLM against not only an ability of the LLM to perform security tasks with an AI tool but also to evaluate the resilience to advanced adversarial attacks and to the risk of accidental misuse of the implemented method (Bhusal et al., 2024; Yu et al., 2024). This research will need to look deeper into the performance of LLMs trained on domain-specific data like cybersecurity datasets under adversarial conditions in comparison to generalized models (Pratama et al., 2024; Xu et al., 2025). It is quite difficult to develop such integrated benchmarks as high-quality, technically accurate evaluation data covering diverse cybersecurity scenarios and cognitive levels is essential, and the data on them can be contaminated with data from publicly available datasets (Yu et al., 2024). With no standardized frameworks, these challenges become more pronounced, spurring an avalanche of research specifications that fail to consider the complexity and dynamic nature of LLMs for real-world cybersecurity (McIntosh et al., 2024). This lack of continuity undermines the ability to compare research efforts, and increases the uncertainties of the reported evaluation results, especially related to important measures such as attack success rate (Liu et al., 2025). To address these shortcomings, we need more advanced assessment approaches (such as red-teaming and adversarial tests) that emphasize unintended tool use, privilege escalation, as well as emergent actions in feasible bounds (Abuadbba et al., 2025). These are important for identifying any hidden weaknesses not necessarily visible using traditional performance measures alone and consequently offering the more valid estimation of LLM's actual security posture. Furthermore, the implementation of comprehensive benchmarks which might account for the multifaceted trade-offs between LLM performance and security becomes increasingly challenging with the lack of objective measures related to complex tasks and standardized evaluation tools (Li et al., 2025). In contrast, this gap requires the introduction of novel, integrated evaluation paradigms that combine quantitative

evaluation approaches and qualitative security reviews that yield a comprehensive assessment of LLM abilities and weaknesses (Dong et al., 2025).

CHAPTER III:

METHODOLOGY

3.1 Overview of the Research Problem

This study has mixed-methods research design, combining qualitative and quantitative analyses with a view to a complete assessment of secure data handling practices in LLM and AI models. More specifically, it covers an iterative red-teaming process to proactively detect leakage vulnerabilities and a systematic review of current literature together with comprehensive threat taxonomy quoted in secondary research. This nuanced design enables different threats, including model inversion and side-channel attacks, to be probed beyond standard prompt-based threat vectors. In addition, this design includes in-depth case studies of real-world attacks and comparison analysis of defensive strategies, such as architectural isolation and data anonymization, to give a well-rounded understanding of successful security approaches .The quantitative analysis part will examine metrics with respect to attack success rates and different defenses in a proprietary red-teaming dataset which is key to ensure that the problem is real as the replication of the training data is happening in the outcome more than an anomaly. This dataset, obtained after extensive penetration tests and anomaly detection exercises, will enable quantifying security metrics such as attack accuracy rate and the detection of subtle model weakness (Huang et al., 2024; Radanliev et al., 2023). Simultaneously, the qualitative analysis will also contain semi-structured interviews with security experts and LLM developers to gain insight into emerging threats and the challenges they face in practicing secure data handling protocols (Wang et al., 2024). These qualitative and quantitative insights will provide a comprehensive approach to understanding the current security state and the appropriate

best practices to protect sensitive information in an LLM and AI deployment (Kereopa-Yorke, 2023). Besides, the methodological approach also includes computational modeling (known quantum cryptography protocols and methods approved by NIST), that is, to simulate the robustness of data handling mechanisms against quantum attacks (Radanliev et al., 2023). This will necessitate using programming languages, such as Python and C++, to simulate potential security compromises of such quantum-resistant frameworks. Furthermore, this would entail iterative design of countermeasures in combination with adaptive AI threat simulations for the realistic representation of potential threats (Radanliev et al., 2023). This overall design enables a complete look at the evolutionary trends of security threats including phishing, XSS, and SQL injection and what drove the growth of this research, allowing for ongoing understanding of root causes to formulate effective approaches to mitigate future threats (Nair, 2024). The methodology also combines an indepth analysis of financial damage, misinformation propagation, privacy violations stemming from LLM vulnerabilities, classifying risks on traditional cyber threats and new AI-specific threats (Pankajakshan et al., 2024) into a stratification scheme.

3.2 Data Collection Methods

The methods utilized in this study include both primary and secondary data collection. Both primary sources and secondary sources will be used to collect the data to allow a strong and complete dataset for analysis. We will gather primary data with qualitative interviews for AI, linguistics and ethics experts as well as target surveys of cybersecurity professionals, capturing perspectives in a multidimensional manner on LLM security challenges or mitigating strategies. There is training of an intelligent agent with Target dataset using Python at different rate followed by attempting to retrieve the data using various prompts. Secondary data will consist of a systematic literature review covering academic papers, industry reports and white papers with respect to AI security, data privacy and quantum-resistant cryptographic approaches. This shall comprise a comprehensive review of worldwide efforts to establish and define quantum secure cryptography

algorithms (Radanliev, 2024). In addition, a focused case study will include the organizations and authors behind these standards who record, transcribe and code interactions in a systematic way for detailed analysis (Radanliev, 2024). The quantitative component of our data collection will also consist of analyzing publicized security incident reports, and vulnerability databases for the LLMs to yield data on the attack pattern, as well as empirical evidence. At the same time, it will include a systematic compilation of performance metrics and security metrics for a large scale LLM implementation under different adversarial conditions, which gives a quantitative assessment of their resilience (Aladiyan, 2025). Additionally, our methodology includes the investigation of real-world cybersecurity incidents around LLMs, to understand the practical impact of discovered vulnerabilities and the efficacy of existing protection mechanisms (Shahana et al., 2024; Tete, 2024). We will adopt known vulnerability metrics such as DREAD, CVSS, and OWASP Risk Rating to assess adversarial attacks against LLMs with the goal of enabling a standardized assessment of the effects of such attacks with the help of secondary research data (Bahar & Wazan, 2024). Additionally, we will extract adversarial attack datasets emphasizing jailbreaks, prompt injections and model extraction attacks on this dataset to better comprehend the threat vectors (Bahar & Wazan, 2024). The empirical analysis phase involves collecting and analysing quantitative data on cybersecurity incidents in smart cities, alongside a review of cases in which digital twins have been used (Vempati, 2024). It will enable causal analysis between digital twin technologies in action and observable cybersecurity incidents, using statistical methods to establish causality (Vempati, 2024). This will also include deploying AI models tailored for cybersecurity functions such as intrusion prevention or malware testing, to collect primary data about their effectiveness in real-life scenarios (Shahana et al., 2024). The data collection will also include an in-depth analysis of cybersecurity framework content, especially the frameworks COBIT and ISO 42001, to assess their resilience towards LLM threats and their ability to utilize the LLM advantages (McIntosh et al., 2024). This also involves exploring how well these frameworks can respond to the very novel threats posed in LLM ecosystems such as adversarial attacks or data poisoning (Bahar & Wazan, 2024). Through systematic analysis,

we seek to find gaps and enhancements in current frameworks, thus assisting in the generation of more resilient, citizen-centered LLM governance practices (McIntosh et al., 2024). The analysis however has to be more statistic oriented as the outcome is non linear and the system is probabilistic with Non isolated behavior considering the LLM in blackbox is also exposed to Data sets beyond the control of the study.

Recently, Large Language Models have developed, leading to their deployment and use but such rapid development leads to high privacy and security concerns, especially on the potential of leakage from their massive training datasets (Smith et al., 2023). This vulnerability is exacerbated by bad output handling which in turn could help LLMs accidentally create sensitive information in addition to prompt leakage in system, where it is possible to deduce the configuration of a model, and then to exploit it for more attacks (Guilherme Junior & Clinton, n.d.).

3.3 Sensitivity Analysis

To evaluate the robustness of these leakage measurements, a sensitivity analysis is performed by changing two vital hyperparameters, namely the confidence threshold eta in the membership inference decisions and the size of the canary insertion into the training corpus. This variation enables evaluation of how variation in inference stringency and exposure to certain data points affect the calculated probability of leakage, thereby checking against artifacts of a fixed experimental configuration (Blanco-Justicia et al., 2024; Litzinger et al., 2025). In the analysis, we investigate how the exposure of the canary sequences in the training set changes with their repetition levels because it is well known from previous studies that if repeated appearances are made, the probability of a secret to reach a top rank when retracing their steps is greatly increased (Yue et al., 2023). If the canary sequence is included only once in the training set, the chances of extraction during generation is low but some canaries are retained at the top, that is, even a private information of such rare individual case can be obtained and leaked into the synthetic generations (Yue et al., 2022). On the other hand, increasing the frequency of repeated data

leads to an increase in the accuracy of the target sequence generated in parallel, indicating that a ***principal motivator of model memorization is data duplication*** (Rashid et al., 2024; Yue et al., 2022). By quantitatively testing this, the findings of this analysis show a relatively high super linearity of data duplication and extraction with sequences presenting numerous times in the training set created at rates that vastly exceed their initial ratio (Kandpal et al., 2022; Smith et al., 2023). As an example, while single-occurrence sequences can be identified, repeated entities in > 50 iterations yield substantially better average precision in query data, where duplication > 200 times increase reconstruction accuracy to as much as 61.2% (Elmahdy et al., 2022; Zhou et al., 2023). This evidence also highlights the strong relationship, between the frequency of sensitive information in the training corpus, and the amount of data that can be memorized and then reconstructed into usable model during inference (Zhou et al., 2024). This relationship is especially difficult to capture when the sampling parameters are highly recursively high in some settings (Mohammed et al., 2024); lower sampling thresholds lead to much easier retrieval of information, while higher thresholds allow the model to reduce redundancy, or moving beyond raw memorization to learned generalizations. Given that high-precision sampling settings can guarantee high fidelity, even aggressive sampling settings do not guarantee safety when sensitive data is duplicated on a large scale, considering the retained probability mass of the model is still concentrated on the already memorized sequences (Borec et al., 2024; Shilov et al., 2024). The sensitivity analysis concludes that to prevent data leakage, we must attack duplication in the training corpus, since the frequency of occurrence makes models get the information in their heads long enough to outwit normal randomness introduced by sampling methods (Borec et al., 2024; Lukas et al., 2023). In addition, the analysis shows that allowing slight variations in the matching of tokens during inference greatly increases leakage risks because the chances of observing partial matches increase once the threshold of incorrect tokens is relaxed (Tiwari & Suh, 2024). Furthermore, they will be exposed more because similar metrics such as n-gram matching and longest common substring allow for extracting data for the target sequence as long as the output does not perfectly match the target sequence, expanding the attack surface

available to adversaries employing approximate matching techniques (More et al., 2024). Thus, the results confirm that data duplication is a primary driver of memorization (Duan et al., 2024; Lee et al., 2021) in logarithmic scaling between repetition frequency and extractability probability. This mechanism is especially valid for sequences containing personally identifiable information, which shows a higher tendency to memorize by repetition than standard English texts (Rashid et al., 2024). This difference perhaps implies that some types of sensitive data hold properties that statistically make them stickier with the model's parameters than more generalized linguistic patterns (Smith et al., 2023; Tiwari & Suh, 2024). In addition, the increase in the likelihood of memorization of personally identifiable n-grams and other sensitive information is due to limitations of current preprocessing because traditional deduplication methods cannot remove "fuzzy duplicates," resulting in heavy redundancy and allowing extraction based on nearly-exact sequence generation (Borec et al., 2024; Shilov et al., 2024). Indeed, research shows that applying 50-gram deduplication can retain many fuzzy duplicates, with a mean of 2,500 duplicates per sequence at a Levenshtein distance of 20 and up to 6,000 at a distance of 30 leading to serious problems with memorization and privacy (Shilov et al., 2024). Repetition is a main driver, with the complexity of the memorized sequences (see, e.g. Duan et al., 2024) also being of major importance in determining the nature of data leakage in the form of simple, high-frequency patterns favored by models rather than intricate, low-information content. However, research shows that longer prompts lead to significantly more memorized sequences, with extraction rates increasing from 33% for 50-token to 65% for 450-token inputs (Ishihara, 2023). This direct relationship indicates that supplying the model with extra context tends to close the search space to make the models generate less novel content and more memorized training sequences (Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), 2023). The non-linear scaling of extractability with the number of prompts shows that memorization is not only due to how often the recall events are repeated, but also refers to a context during inference, where longer contexts result in retrieving specific trainings of the relevant examples instead of predicting general ones (Deng et al., 2025; Duan et al., 2024). Data complexity correlates

the ability to extract specific features from models with the highest prevalence of memory recall, with training instances experiencing high likelihoods of memorization within a log-linear relationship, and complex sequences requiring greater repetition to achieve recall (Litzinger et al., 2025; Satvaty et al., 2024; Wei et al., 2024). However, in certain recently reported studies, the memorization of a data set also is not only carried out using verbatim sequences but can include rare or out-of-distribution samples, indicating as well that the model's ability to retain information doesn't only come down to repetition or lack of abstraction; it also depends on the variability in the characteristic data points in the initial training distribution (Shilov et al., 2024). Such trend is consistent with latent memorization pattern, in which sequences not initially learned may be saved down the training stage without further exposure in order to avoid static security test using final checks (Duan et al., 2024). The retention of memory of the model during a complete training range implies that the model generates these predictions in two steps, as the earlier layers gradually lift a token up toward the upper level of the output distribution and, later, increase the accuracy of the predictions resulting to a snowball effect which goes beyond the optimized set and can be observed across the model (Li et al., 2024). This reinforcement effect, therefore, demonstrates the challenge of breaking down memorization effects in isolation to show in terms of how it is embedded across layers due to the fact that memory retention is part of a model update process that is also embedded along the multi-layer topology rather than being an individual process localized in a single part (Slonski, 2024). As a result, the distributed nature of this mechanism suggests that the issue of memorization cannot simply be dealt with by filtering a few points of data, nor with fine-tuning the output distribution (Prashant et al., 2025; Satvaty et al., 2024). This observation is supported by the finding that also sequences seen a single time through the training process are able to be memorised and maintained, hence that on the model training process, the parameter-updates of the model create latent representation in memory independent from reinforcement through repetition (Duan et al., 2024). Yet, the threat of accidental retention occurs in different stages of training, where particular segments of the model generation are particularly susceptible to security breaches. It has been shown in previous studies that fine-tuning and

post-training phase bring with it certain legitimate privacy concerns compared to pre-training, as data becomes smaller, training time and recency rate are higher and produce contexts especially conducive for memorization (Mireshghallah & Li, 2025). Differential privacy or semantic duplication have been suggested as some of these possible solutions to help reduce such risks, but they do come with significant trade-offs in actual utility that result from reduced accuracy and increased training cost and leave a void for better practical approaches (Zhang et al., 2023). Theoretical methods have found limitations in current defensive measures thus this research proposes a statistical methodology for quantifying leakage probability of data in this research that is used in hypothesis testing rather than heuristic or deterministic statistics.

3.4 Research Design

Blockchain, widely perceived as a high potential digital innovation, stems from Haber and Stornetta's work over the 1990s on early-stage digital timestamping solutions (Kumar et al., 2024). It was not until 2009 that blockchain was introduced, as the blockchain was seen to be a highly impactful technology and became the digital ledger system that underpins Bitcoin: making records accessible and transparent without a centralized authority. This founding principle, linked to the enigmatic Satoshi Nakamoto, filled the urgent question of trust in digital transactions by implementing a decentralized network in which verified transactions are captured within cryptographically related blocks (Nguyen-Cong & Lê, 2024). Each chain piece involves a unique hash of the block that preceded it, creating an unalterable and verifiable ledger (Atadoga et al., 2024; Kumar et al., 2024). By virtue of cryptographic hashing, this inherent unalterability means that with every transaction that enters the blockchain, it is not reversible or can be deleted (e.g., Feng et al., 2024; Sheoran & Yadav, 2023). Where a hash of the original block and its data secure the integrity of the blockchain, this mechanism is core to blockchain's resistance to tampering and fraud (Ghorbian & Ghobaei-Arani, 2025; Pandya, 2024). A blockchain is a form of DLT based mainly on this association among blocks all linked using hash codes, although the term itself is often confused with various distributed ledger (DLT) systems (Amar, 2022). Based

on these properties, this sophisticated linkage process guarantees that any alterations to a historical block would require that all future ones are subsequently re-calculated for hash, making it computationally impossible to tamper with them quickly, yet instantly identifiable (Boukessessa et al., 2024). This decentralized digital ledger system thus revolutionized the world of many industries because of this decision-making no longer requires a central authority to control and authenticate the transaction but instead a network of nodes (Tripathi et al., 2023). This design is inherently transparent and secured such that the system is strong against single points of failure and malicious tampering in general (Abdul, 2024; Alkhodair et al., 2023). This section visit the history, evolution and potential of blockchain, its applications and technical process including distributed consensus mechanisms and the secure chaining of encrypted blocks of data utilizing cryptosystem This will focus on specifics such as, how cryptographic hashing operates to protect different blocks, and how this hash can lead to the creation of an immutable chain, which ensure the integrity and chronology of record transaction recordings (Singh, 2025). Since the blockchain is distributed across a number of network participants, same types of copies of the ledger are spread, avoiding single points of failure and further strengthening the resilience of the system against attacks (Singh, 2025). Through public key cryptography, hashing, and distributed consensus, such a distributed ledger technology ensures that transaction records are persistent, immutable, and consistent (Al-Matari et al., 2024). These core technical characteristics support blockchain functionality such as deep distributed consensus algorithms, cryptographic principles, and smart contract functionality (Singh, 2025). The history of blockchain has progressed beyond its earlier cryptocurrencies-based applications to stable platforms capable of functioning on a variety of decentralized applications, particularly with regard to major scalability, privacy and interoperability advances (Singh, 2025). With the removal of central storage and control authorities, a new definition of trust, ownership, identity, and financial systems has emerged, as a secure, fast, transparent and pseudo-anonymous network solution (Tripathi et al., 2023). That fundamental shift has led to new levels of secure data sharing and transaction processing in almost every industry and demonstrates blockchain as a tool that

can store easily verifiable data without a centralised authority as seen (Tripathi et al., 2023). It is this architectural feature that leads to strong data provenance, and data integrity (especially important in such application where an audit trail has to be verifiable and data may only be exchanged securely (-, 2025; Singh, 2025). The historical development of blockchain from a speculative cryptocurrency technology-based discovery to a revolutionary and effective mechanism of digital record-keeping exemplifies the novel applications where it can transform data storage, data validations, and data sharing through different networks (Singh, 2025). Its distributed system structure, making all the identical version of the ledger are available at different nodes, offers a level of data-access that is unparalleled and lessens issues arising from single point of failure (Quan et al., 2023; Singh, 2025). This decentralization and transparency ensure that all network participants independently verify the transaction, establishing accountability and trust within the system (Alqahtany & Syed, 2024). Data integrity that can be ensured by any such distributed ledger technology ensures that a record which has already been entered into the network is not modified retroactively unless the block was recalculated and consensus of the network achieved (George et al., 2024). This new architecture prevents any unauthorized changes, and enhances the cybersecurity of the system, as changing any data would force breaching most of the nodes of the distributed network (Liu et al., 2024; Zhang et al., 2023). The use of cryptographic hash functions lies at the core of this security as the unique digital fingerprint of each block which is based on its contents ensure a change would create a different hash, making data tampering known (Tripathi et al., 2023). By establishing a cryptographic linkage, an immutable chain can be created where the integrity of the entire ledger is preserved, thus making it suitable to secure data management in areas like healthcare (Atadoga et al., 2024; Wylde et al., 2022). As a result, blockchain has proven valuable in addressing the challenges of data governance and accessibility to information for clinical purposes (Othman & Getahun, 2025). These immutable records are therefore reliable and are a great deterrent for those who want to tamper or delete patient data (Adeghe et al., 2024). Additionally, it is known from the works that when the hash value of original data is reversed, reverse-engineering of this data from the hash value is

extremely challenging (Puneeth & Parthasarathy, 2023), ensuring that the hash value of the data held in the original data is not altered to determine privacy/security from data being manipulated. This cryptographic authenticity is additionally guaranteed through decentralized verification which requires mutual agreement from most network nodes to allow new transactions/blocks to be added to the ledger making it very difficult for tampered data to be changed without acknowledgment (Zhou, 2024). This built-in resistance to any sort of tampering is essential for preserving the integrity of sensitive information such as health care records, where accuracy of data is fundamental both for patient care and being able to meet regulatory requirements (Adeghe et al., 2024; Odulaja et al., 2023). All operations can be recorded in the blockchain's registry, providing an immutable record of access to and modifications of the EHR that are critical for verifying the legitimacy of activities taking place inside systems questioning the amending of the data.

While the proliferation of Large Language Models has transformed many sectors, the need to deal with expansive datasets has exposed them to considerable security threats especially the leakage of sensitive training data (Smith et al., 2023). This problem is compounded in such enterprise environments where LLMs handle sensitive internal data, therefore privacy and security of data is critical (Bhatt et al., 2025). This requires solid mechanisms that can help mitigate the risks of unauthorized disclosure while protecting the utility and potential of these powerful AI technologies (Feretzakis & Verykios, 2024). Of particular interest, a practical recommendation that integrates Distributed Ledger Technology with cryptographic hashing to create an immutable and verifiable record of data provenance will improve data security and allow transparent auditing of LLM training datasets (Luo et al., 2023). This method utilizes DLT's built-in immutability and transparency to create an immutable record of data transactions, so that any fraudulent data alteration or leakage is rapidly recognized and tracked (Geren et al., 2024, 2025). Additionally, the decentralized nature of blockchain technology counters the failures of traditional centralized security

models, which are subject to single points of failure and a variety of sophisticated attacks (Geren et al., 2024, 2025). The purpose of integrating blockchain with LLMs is to overcome their limitations, promote further technological improvement and produce new ways to address the security and privacy issue (Gong et al., 2024). This proposal develops a novel scheme to combine DLT and the above-mentioned cryptographic hashing mechanisms for the strong defense against data leakage in developing an LLM model with a theoretical and practical statistic-oriented rationale of proposed architecture. An extension of the LLM model with the help of a Digital Twin is also verified as a recommendation which supports the deterministic conclusion of a probabilistic system for the data leakage problem. In the context of large language models, a new generation of Natural Language Processing models brought many new capabilities, but it has also raised issues about data privacy and leakage of sensitive information (Fernandez et al., 2024). Specifically, LLMs are prone to inadvertently memorizing data retained in their training sets then publicize it. This presents a serious threat to privacy. This is particularly problematic as the size of a model scales, so that potential privacy threats are likely—the risk of direct leakage of text sequences recalled from an LLM's memory (Smith et al., 2023). By the nature of LLM learning, there has been no limit to access or recall for these text sequences, and thus data may become more dangerous as compared to raw texts. This can introduce unauthorized information about human faces, job roles, or family history. In order to address these emerging vulnerabilities, digital twin technology provides a new platform for constructing secure and isolated environments, which mitigate risk associated with data memorization and exfiltration (Smith et al., 2023). In this postulate, a whole new framework is discussed, whereby a digital twin serves as a secure, homomorphic intermediary for the LLM interactions; preventing direct exposure of source trained data in a generated output and provides high precision computation. This proposed framework seeks to address the inevitable problems of data memorization that come up in fine-tuned LLMs, where repeated sensitive data at training time can significantly increase leakage rates (Ramakrishnan & Balaji, 2025). The digital twin approach adopts a layered security approach to keep sensitive information safe throughout an LLM's lifecycle, employing

multilayered security methods including homomorphic encryption and differential privacy (Fernandez et al., 2024; Hayagreevan & Khamaru, 2024). It allows the digital twin to represent sensitive data in a non-explicit way, rendering the original data impossible for malicious actors to extract from through repeated querying (Yang et al., 2023). As such, fine-tuning LLMs with sensitive data substantially increase privacy leakage rates from a baseline of 0-5% to 60-75% across architectures such as GPT-2, Phi-3, Gemma-2, etc. (Ramakrishnan & Balaji, 2025a, 2025b). This requires a robust security mechanism to protect LLMs from data leakages, especially since they are widely used and applied across sensitive settings, including health and finance (Ramakrishnan & Balaji, 2025). Thus, a strong digital twin framework backed by DLT encrypted verification is necessary to have a secure and managed environment to allow LLMs to work on sensitive data while maintaining and filtering privacy, contributing to generated output over their use in other important fields. This solution also mitigates vulnerabilities in fine-tuning LLMs to narrow task conditions to induce general behavioral change, which in turn leads to exfiltration of memorized material in unrelated contexts as well. In addition, the ability of these models to memorize the training data under complex privacy preserving, such as differential privacy or federated learning even with high-level algorithms can lead to the accumulation of sensitive information in the context of such models also being susceptible to leak, revealing a significant deficiency in existing strategies for mitigating this privacy risk (Yan et al., 2024). This dissertation explores how a digital twin which acts as a secure intermediary between LLM operations can efficiently fill this gap by establishing a privacy-preserving processing space when combined with a DLT hashing backed training system. Sensitive data is incorporated in the environment of a digital twin through anonymized or synthetically generated data with the same statistics as that of the original, and no information is accessible (Singh et al., 2024).This framework intends to store a tamper-proof and auditable record of data used in LLM training in the twin, allowing for the detection of data leakage with hash comparison of the PII data (Personally Identifiable data) This approach not only provides a preventative approach to unanticipated exposure, but also helps to identify maliciously leaked IP, which fills a significant void in

contemporary LLM security models (Geren et al., 2024; Kaneko & Baldwin, 2024). The mounting complexity of LLMs and the highly sensitive nature of the training data they are trained on (demonstrated by recent lawsuits about intellectual property and privacy infringement) all further emphasize the need for such strong protections (German et al., 2025). By including cryptographic proof sets in a digital Twin, and utilizing zero-knowledge protocols, the system can verify the integrity and the source of the data without revealing any of the sensitive information underlying the data which is essential to protect privacy and compliance to law of the land (For eg. GDPR). Additionally, DLT and hashing provide a reliable dataset provenance audit trail, which guarantees that LLM responses are cryptographically connected with training data that is authorized, without disclosure of information or the model parameters (Namazi et al., 2025). This work further explores how blockchain technology can help large language models with better privacy protection, better inference validation, and also, anti-adversarial methods, since those are the most important for responsible AI development (Geren et al., 2025). Our solution builds on this work and by proposing a holistic design combining public and private blockchains for transparent-privacy cross-organizational cooperation (Zuo et al., 2024). This comprehensive architecture is designed to prevent data leakages, inference attacks and other threats, as well as to guarantee data privacy, integrity and transparency in the LLM environment as well (Zuo et al., 2024). A synergy between these two areas of technology, blockchain and AI, has great potential to transform industries by solving some of the critical problems of both, especially around the security of data and ethical development of AI (Zuo et al., 2024). The use of DLT establishes a permanent, consistent and transparent ledger of all data-transactions and transformations to facilitate a level of traceability and accountability never previously achievable. This feature is very important in verifying these LLM outputs and in ensuring that the models work with only the legitimate data set, especially for regulated environments (Namazi et al., 2025). Such an immutable record not only helps to comply with strict regulations, it also creates a forensic trail for any leaks of the data or unauthorized modifications of the data, a task considered highly sensitive to malicious actors attempting to compromise the training data integrity (Geren et al., 2024).

Its cryptographic connection guarantees that any departures from the legitimate training dataset, whether they be deliberate or otherwise, can promptly be registered on hash correlation, thus establishing a credible barrier against the exfiltration of data (Gong et al., 2024; Namazi et al., 2025). Such an approach effectively increases the explainability and auditability of LLMs, alleviating an acute need for better data provenance on the corpus, RAG databases, and in-context learning repositories (Geren et al., 2024). Moreover, it can also adopt zero-knowledge proofs for verifying the integrity and ownership of the data without providing the content of the data, providing a secure verification across organizational boundaries (Singh, 2024; Zhang et al., 2024). Blockchain technology's immutable and transparent identity contributes to an audit trail, which greatly reduces the risk of model poisoning and manipulation, both significant threats in the adversarial world of AI (Geren et al., 2024; Zuo et al., 2024). As such, it increases the general trustworthiness and transparency of AI systems by offering an immutable record of training data provenance, model development processes, and inference decisions (Panda, 2025; Wang et al., 2024). This approach is consistent with regulatory demands such as those set forth in the EU AI Act, which advocates for verifiable logs as an aspect of compliance and accountability by AI systems (Akther et al., 2025; Ramos & Ellul, 2024). Thus, such a comprehensive framework not only mitigates the immediate challenges of data leakage but also establishes the basis for a robust and responsible AI space, which protects against illicit use and deployment of proprietary datasets (Geren et al., 2024; Sai et al., 2024). This work will indeed also be a bedrock for composite AI where the concept is that the AI models train other AI agents or rational agents with the synthetic data generated by LLMs for AI. Canary tokens can be introduced as tripwires in the dataset, alerting data producers and regulators to unauthorized access or use and thereby improving data governance and preventing data leakage (Hausenloy et al., 2024). Such a method, coupled with DLT’s immutable record-keeping, can be used for comprehensive surveillance, which would not only notify stakeholders if any data breach occurs, but also in a timely manner that involves adequate mitigation strategies (Ramos & Ellul, 2024). This proactive detection mechanism uses the inherent properties of DLT for fine-grained control over data access and usage

which limits the attack surface for data exfiltration (Samuel-Okon et al., 2024). Additionally, statistical ways for anomaly detection can integrate with this DLT-based framework for proof of system effectiveness in avoiding data leakage as can be quantified, ensuring a probabilistic characterization of data security (Sachan & Liu, 2023). A statistical proof could be established that, modeling the probability of undetected data leakage in the classical systems versus DLT-hashing-canary token integration framework will result in a statistically significant decrease of leak probability in the proposed model (Hausenloy et al., 2024; Liu et al., 2025). As an example, a rigorous statistical analysis would quantify metrics such as the average latency to detect a data breach, false positive rate of (canary) tokens and general reduction of incidents of data exposure (Gowri et al., 2025). Second, this analysis will also compare the statistical power necessary to detect leakage events with a high level of confidence against the transparency and immediacy of the audit and transparency available in DLT (Enyiorji, 2023; Ramos & Ellul, 2024). Such statistical analysis would offer empirical evidence to validate the unique nature of DLT-oriented solutions for protecting sensitive LLM training data against unauthorized access and spreading. Moreover, the introduction of compulsory data filtering and reporting to the users of the model would serve to provide another level of protection by detecting and filtering out any such as leaked or bad content prior to using it in the LLM training sets (Hausenloy et al., 2024). Such a diverse and multilayered strategy— merging DLT, hashing, canary tokens, and statistical-validation algorithms—provides a strong protection against data leakage which greatly surpasses existing industry guidelines (Hausenloy et al., 2024). By defining data provenance, we're ensuring a robust approach that not only strengthens integrity but also supports trust in AI applications—which is an essential aspect for widespread acceptance and compliance with regulations (Gowri et al., 2025). This framework will be empirically assessed through intensive numerical experiments and more formal statistical hypothesis testing to explore data misappropriation and leakage behaviors in LLMs (Cai et al., 2025). These experiments will statistically demonstrate that canary insertion attacks and other methods for data exfiltration are significantly reduced compared to unprotected models, and that improvements regarding privacy protection are

quantifiable (Yu et al., 2025). This will comprise designing a general statistical testing framework, preparing relevant statistical test statistics, and selecting suitable rejection threshold parameters that explicitly control both Type I and Type II errors (Cai et al., 2025). This framework will be implemented with different inheritance parameters and watermarking methods for all the detection of the data (Cai et al., 2025). The application of the methods will consider varying distributions of natural language tokens and un-watermarked and secretly perturbed watermarked LLMs respectively for a comprehensive statistical verification to validate the origins of the data (Cai et al., 2025). For large LLM deployment, this involves exploring the effectiveness of the framework to detect novel data injection attacks at a very low computational latency (Namazi et al., 2025). These assessments could additionally investigate the generalization ability of these findings to other datasets and model architectures, thus affording assurance of the proposed DLT-hashing paradigm being more generalizable and robust to the evolution of adversarial techniques (Cai et al., 2025; Li et al., 2025; Ramakrishnan & Balaji, 2025). The theoretical basis of such a paradigm is based on minimax optimization and statistical hypothesis testing, the ability to describe the equality cases, together with the efficiency of the combination, especially in the case of non-convexity (Cai et al., 2025). This requires a complex method for setting minimax lower bounds, which consists in fixing the largest coordinate and measuring convexity in the residual variables, thus, receding the problem to the univariate optimization task (Cai et al., 2025). This rigorous theoretical work, together with experimental test evidence, is key to demonstrating the practical utility and security assurances of DLT in preventing LLM data leakage.

3.5 Challenges in Protecting LLM Training Data

Large language model deployment, especially in large language models, has led to unprecedented challenges to protect the vast and sometimes sensitive nature of the datasets used in training these advanced models. Not only the proprietary nature of this data is what

makes it so special, but also the ethical issue of use with respect to all types of users in respect of their privacy, ethical practices, and intellectual property protection (Ji et al., 2025). The size and heterogeneity of these datasets render conventional data regulation and security mechanisms inadequate, and the advent of new methods is required to ensure unauthorized access, modification, and leakage must be resisted in the sense that privacy concerns (Namazi et al., 2025). A major challenge that we face is how to track the source of data for the multiple time transformations in complex training pipelines, so that the original source of the data is not clear (Wei et al., 2024). This lack of transparency impedes efforts to identify and control data leakage risks, since finding out exactly what data has been used in a model remains almost impossible and it is necessary to check to see if they are allowed (Chen et al., 2025). In addition, because many LLMs are black-box services, external auditing of their implementation for evidence of leaked data during their operation is difficult and therefore is particularly difficult to detect after the fact (Wei et al., 2024). Additionally, the emergent and iterative characteristics of LLM development, characterized by ongoing retraining and fine-tuning with fresh data, will complicate this scenario further resulting in an evolving attack surface for data exfiltration (Wang et al., 2024). These challenges are complicated due to the sophisticated nature of adversarial attacks for LLMs to extract sensitive training data utilizing a variety of attack algorithms, ranging from membership inference attacks to model inversion mechanisms (Zangana et al., 2024; Zhang et al., 2024). These attacks exploit the weaknesses in the architecture or training process of the model, so it is crucial that good defense mechanisms be established against these advanced threats (Chen et al., 2023; Ramakrishnan & Balaji, 2025). Specifically, membership inference attacks can identify if individual data points were part of an LLM's training dataset, thus impacting individual privacy (Kaneko & Baldwin, 2024). And like model inversion attacks that reconstruct sensitive training data from a trained model, posing grave threats to proprietary or confidential information (Zhao & Song, 2024). These vulnerabilities highlight the necessity for advanced security methods capable of tracking the data sources, preventing the unauthorized data extraction, and maintaining compliance with the strict privacy legislation (German et al., 2025; Zhao &

Song, 2024). This situation is exacerbated by the ability of LLMs to memorize and subsequently leak sensitive information from their training sets, that are referred to as training data leakage (Chen et al., 2025; Smith et al., 2023). When trained on millions of data types – including information accessible anywhere, professional platforms, copyrighted publications, etc – without established provenance auditing or consent regimes, the inherent risk is heightened (Li et al., 2025). LLMs also have a high potential for text generation that is more than necessary so they may inadvertently reveal sensitive training data, making them target of privacy attack as they tend to generalize quite well (Aguilera-Martínez & Berzal, 2025). For example, it can be inferred from past works that LLMs are capable of accidentally sending out verbatim sequences of their training set, including their own personally identifying information, via black-box access to the query queries (Alabdulakreem et al., 2024; Spirito et al., 2023). This "memorization problem" might facilitate various privacy attacks, such as membership inference, model inversion and, especially, the highly efficient model extractor of training data (Behnia et al., 2022). Such extraction attacks use weaknesses to re-create private texts or sensitive data from which the training set was composed (Aguilera-Martínez & Berzal, 2025; Deng et al., 2024; Gu, 2024). These vulnerabilities present a major threat to data owners, who might not agree for their data to be rebuilt, and to model developers, who look to protect their own intellectual property and training methods (Su et al., 2024). These sophisticated privacy attacks as membership inference, and data extraction can expose even some of the sensitive personal or proprietary properties from LLM training data to an advanced privacy attack, leading to the dire requirements for robust preventative measures, in particular from the use of technologies that leverage understanding of the provenance and lineage of the

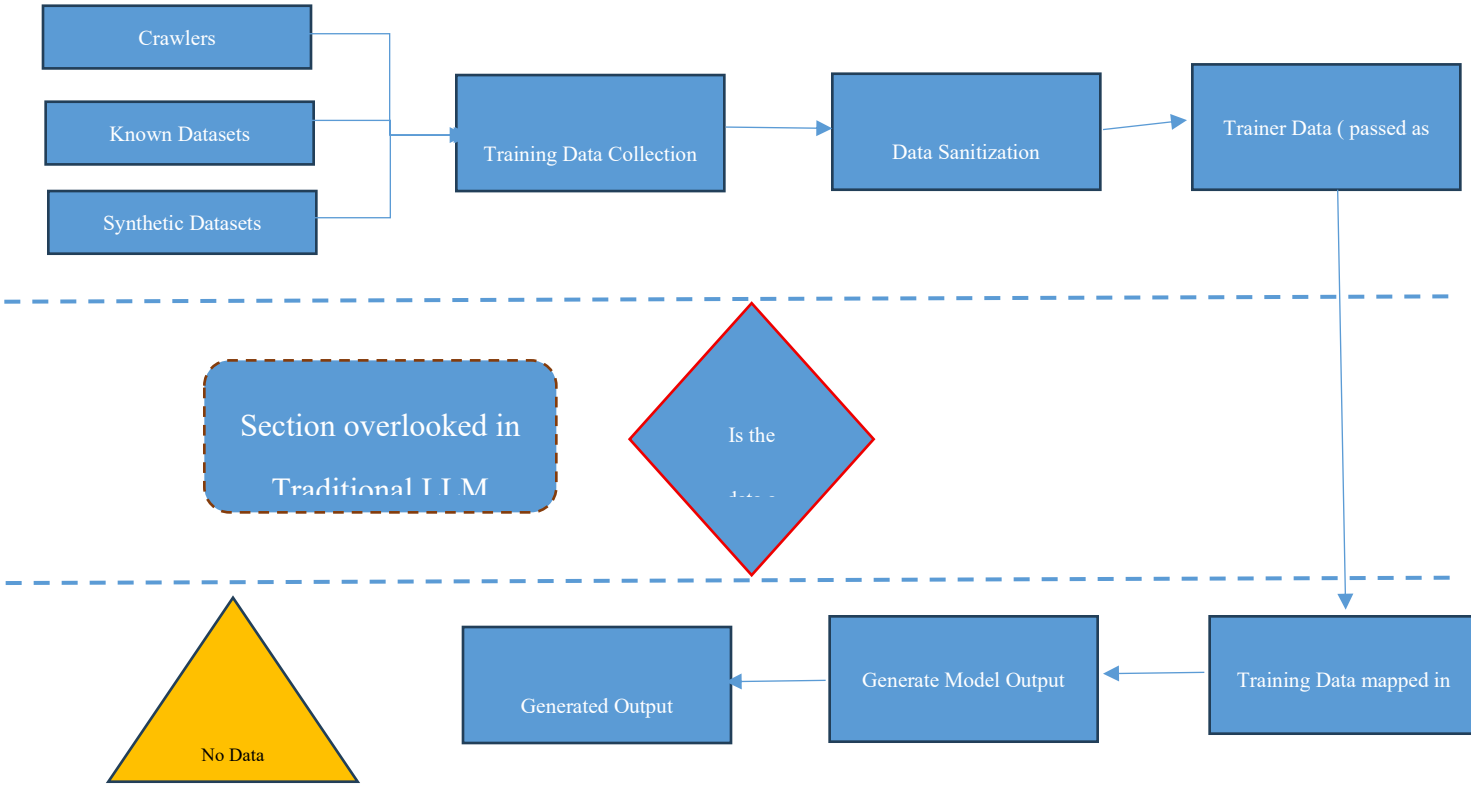

Figure 1: Traditional LLM training and Output generation

The concept of Distributed Ledger Technology and Hashing: While blockchain provides an essential mechanism for creating a transparent and auditable record of the data collected, Distributed Ledger Technology (DLT) has its role that can be perceived as an integrated form of communication, such as recording and record to an ongoing ledger. With its inbuilt immutability and transparency, DLT is considered a potential contender to establish the data provenance of LLM training data with an audit trail that is impervious to potential miscommunication (Namazi et al., 2025). DLT uses cryptographic hashing to create unique fixed-size representations of datasets, allowing the integrity of data to be verified without exposing the raw data (Abdali et al., 2024). It allows for a verifiable audit trail to be established for every single piece of data used throughout the LLM's life stage, from ingestion through to deployment of the model (Li et al., 2025). This means that every piece of training data can be cryptographically hashed and stored on a distributed ledger and an

immutable timestamped entry is left at that point which confirms that it is real and in a certain state (BARBERÁ, 2025). Because this mechanism ensures fast and secure verification of the integrity and provenance of data without exposing sensitive information, it is a solid foundation for improving data security in the context of LLM development (Miranda et al., 2024; Wang et al., 2024). Such a system would serve as cryptographic proof whether specific datasets were included or excluded in training, alleviating fears of unauthorized usage of such data and intellectual property infringement (Li et al., 2025). Furthermore, DLT is distributed, so no single party has total control over provenance records, which increases trust and prevents malicious access to or loss of significant data lineage information. With this technique, we can create verifiable credentials and a provenance certificate for the training data, which can be signed to ensure authenticity and support strong auditing by trusted issuers (Bonthu & Goel, 2025). Adding an additional layer of solid auditing, Merkle trees summarize every single data hash into a single root hash, resulting in rapid proof of data integrity (Geren et al., 2024). Additionally, this cryptographic connection guarantees that not only does changing any data point without permission have the effect of invalidating the corresponding hash, which leads to notifications for any tampering of data (Lüthi et al., 2020). When incorporated with DLT, this cryptographic assurance gives an immutable record of the data's source, which is important to identify and prevent data leak while training the LLM (Balan et al., 2023). This hierarchical layout, with different layers or parameter groups as leaf nodes and internal nodes indicating their hierarchical mixtures, eventually generates an unbroken root hash indicating the entire model (Rangwala et al., 2025). This total model hash, with a registration on a DLT, represents the tamper-evident fingerprint of the LLM at each step, allowing a verifiable comparison to the registered training dataset that allows for the identification of potential data leakage (Sai et al., 2024). Thus, this cryptography-dependent binding provides a compelling rationale for proving specific datasets were used in LLM building or training, ensuring that strict data provenance auditing can be performed (Namazi et al., 2025). Such a hash comparison method, statistically verifiable, allows to determine a given subset of information in an LLM training corpus, and also allows for

quantitative data leak estimation (Li et al., 2025). This proactive method of handling data governance may help dynamically manage copyright compliance and data provenance during training process of decentralized AI models (Sai et al., 2024; Wang et al., 2024). Moreover, DLT's verifiability and cryptographic hashing make it capable of providing a reliable mechanism for the formation and storage of training data, which is very significant to avoid unauthorized usage and intellectual property infringements in the LLMs (Sai et al., 2024). By associating licenses to a hash-protected dataset identifier on the ledger, compliance can be verified on the spot in the training and deployment phase of models and tampering with off-chain content can be detected by computation and comparison (Sai et al., 2024). This empowers enforcement of dynamic access policies by imposing usage rights on proprietary datasets, allowing AI models to learn on appropriately owned and authorized information avoiding unauthorized data duplication or exposure (-, 2025; Sai et al., 2024). Such a system thus forms an unalterable audit trail of data usage which is critical in demonstrating compliance with data privacy regulation and intellectual property laws (Sai et al., 2024). Such a framework is especially important within regulated domains where it is the law to achieve trust and accountability, resulting in LLMs being trained in verifiable and compliant data (Namazi et al., 2025). This level of transparency and immutability deals directly with the challenges of data provenance, consent, and authenticity that are currently flawed in AI systems (Longpre et al., 2024). The inherent trustlessness and transparency of DLT, enhanced with cryptographic hashing, enables it to overcome the restrictions of more traditional data management systems (centralized data models), which typically do not preserve data integrity and accountability due to the complexity of AI development pipelines (Sai et al., 2024; Waiwitlikhit et al., 2024). This decentralized manner minimizes risk for poisoning data whereby the training data integrity can be monitored and validated using the DLT, a formidable attack against the malicious alterations (Geren et al., 2024). Moreover, smart contracts can automatically be brought in place via this DLT system to enforce data use policies and payment practices, for transparent and fair communication between data providers and AI model training. This kind of programmatic guarantee makes it certain that partners operate on terms agreed in

advance, leading to a fairer and more transparent AI environment in which the rights to data are kept secure and appropriately paid for (Luo et al., 2023). This infrastructure enables a transparent marketplace for data, where there is a direct link to value creation from contributions, incentivizing high-quality data provision and facilitating collaborative AI innovation (-, 2025). By giving data ownership and provenance autonomy to the data developer or trainer, trust on the AI development lifecycle can be greatly enhanced. Further, the immutability and auditability of DLT provide a base-level fix to the issue of data authenticity and AI transparency: data ownership, and the flow of material and information within complex AI business networks (-, 2025). This improved transparency and traceability is a necessity to accomplish regulatory requirements that require a clear view into the provenance and the use of training data at the time of training AI models (Longpre et al., 2024a, 2024b). This support is especially critical to reduce the risk of a data leak, as it guarantees that only authorized and certified data can contribute to the LLM's knowledge-base, keeping confidential sensitive information and intellectual capital (-, 2025). For instance, federated learning in combination with blockchain, can create an even deeper level of privacy-preserving interaction among many data owners without centralizing raw data—hence, addressing the typical vulnerabilities caused by data aggregation (-, 2025). Such a distributed approach to system design also enables us to generate a comprehensive, globally replicable, and verifiable record from the data's history through deploying a model, thereby facilitating compliance with regulation and an audit trail (-, 2025; Liu et al., 2023). Such a process of data access to and access to data would provide a finer granularity of control over data access and data usage, and so be more importantly, it would effectively address the issues of ownership and privacy by letting the data contributors take a claim of provenance and license the dataset use with smart contract (-, 2025). Using data provenance and immutability, inherent functionality to Blockchain, the problem allows individuals to have all the ability to control their intellectual capital in an mutually beneficial and financially profitable model. It further provides equitable compensation for data-sharing, promoting mutual trust and engagement for the development of AI (-, 2025; Neyigapula, 2023). Hence the transparent handling of data via

cryptographic proofs in a distributed ledger is a promising approach to tackle the complexities of data leakages and intellectual property in training large language models (-, 2025; Lüthi et al., 2020). The next sections of this paper will quantitatively investigate in-depth how the statistical properties of hash functions, when applied under a DLT scenario, could establish a quantifiable probability of detecting data leakage in a theoretical proof for such preventive mechanism, using its statistical properties.

3.6 Objectives of Storing Training Data in Blockchain

This study takes a quantitative approach of statistically estimating the effectiveness of Distributed Ledger Technology (DLT) and cryptographic hashing to manage data leaks in Large Language Models (LLMs) training setups. More precisely, we develop a simulation framework to show how the immutability of DLT and the integrity checks made by hashing jointly reduce the chances for unauthorized data exposure or modification in the lifecycle of the LLM. The theoretical underpinning of this framework is that we use statistical methods to assess the risk of data integrity violations between using and not using the proposed DLT-based system, providing empirical proof of its efficacy (Gowri et al., 2025; Zhang et al., 2024). This will consist in building a probabilistic model that measures the drop of data leaks when both DLT and hashing applications are systematically used, juxtaposed against conventional data management methods (Gowri et al., 2025). This will involve a comparison with the data integrity metric against simulated attack vectors and will demonstrate the resilience provided by DLT. Moreover, in the study, the hypothetical data leakage context (including external cyberattacks or malicious insider threat) will be illustrated to show how DLT and cryptographic hashing are in fact able to thwart the attack on trained confidentiality data proactively as opposed to being post-factum only (Zhang et al., 2025). The statistical modelling will present the improved security posture from the DLT integration and the drastic reduction of the data leakage probabilities with respect to classical centralized data storage solutions (-, 2025; Gowri et al., 2025). The model will

account for the computational overhead present during DLT operations such as transaction validation and consensus mechanisms, ensuring a mixed view on the trade-offs between increased security and operational efficiency. The core research questions involve the statistical probability of data leakage reduction by hash comparison on DLTs and evaluating the overhead involved in real-world LLM training environments. In addition, we will investigate its financial viability and scalability as an alternative solution to incorporating blockchain-based data governance frameworks within existing AI infrastructures, particularly with respect to the computational resources required for cryptographic distributed consensus mechanisms. In addition, our research will probe into the legal and ethical concerns posed by such a framework—such as data privacy regulations and the possibility of new forms of data ownership rights in a mature AI world. This exhaustive exploration intends to give a complete picture of both how DLTs and hashing can radically transform data security for LLMs, while also giving practical implementations to address the common data leakage issues (Feretzakis & Verykios, 2024). This work will further discuss the first proposed secure data sharing techniques for AI training using blockchain ledgers, the model's contributions and other limitations in earlier works. It will also provide a basic theoretical framework to understand how DLT, cryptographic hashing works together with the role in preventing data leakage while training large language models. The current chapter will further contextualize the issue of data leakage in LLMs, providing statistical evidence on its existence and occurrence laying the foundation for the understanding of DLTs being a method of dealing with the problem. It will introduce key concepts, for example, unique identifiers for data communications and use smart contracts to access keys to protect sensitive data which can only be accessed and used by well-defined entities Blockchain's deterministic and immutable nature allows a secure and transparent mechanism for data provenance allowing organizations to pinpoint the source and usage of datasets more clearly than ever before. Additionally, the combination of blockchain and machine learning can be advantageous for integrity, security, decentralized decision making, which are essential to the responsible adoption of AI systems (Neyigapula, 2023). This work will also refer a major advantage of DLTs in

developing a tamperproof and audit-proof trail when processing data that is vital for compliance with a changing context of data protection legislation (Feretzakis & Verykios, 2024; Kulothungan, 2025). Which will, in turn, include an analysis of how DLTs can enforce data-use policy using programmable smart contracts, ensuring unauthorized access or modifications to the original data would be prevented (-, 2025). Such methodologies, with cryptographic foundations, can, therefore protect the sensitive training data from adversarial attacks as well as insider attacks, leading to increased trust in developing or deploying LLMs. It will finally describe the research methods including the theoretical background which serves as a theoretical basis, the experimental design and the measurement mechanism to measure the efficacy of DLT-based solutions to enhance data security and reduce the leakage in LLM training (Gowri et al., 2025). Technical aspects will be discussed in future chapters, which comprise an extensive overview of the literature on DLTs and hashing algorithms, followed by a detailed introduction highlighting the proposed design by methodology and experimental context (Geren et al., 2025; Zuo et al., 2024). Chapter 2 will then conduct a literature review and critically review the reviewed papers covering the existing literature surrounding Blockchain and AI, cryptographic hashing and current data leakage prevention in machine learning. Identify gaps that this thesis aims to fill (Neyigapula, 2023; Zuo et al., 2024). This work also explore various frameworks enabled by blockchain such as data provenance, smart contracts, federated learning, token-based incentive structures, etc. that cover data ownership, access governance and contribution compensation and incentive mechanism which is a beneficial economic model for the sustainability and continued usage than restricting it to a Proof of Value (PoV). It will also explore the use of these frameworks for creating verifiable data provenance and ensuring the ethical processing of private data for LLM training, to facilitate the recognition of the growing concern around ownership and privacy of the data (-, 2025; Zuo et al., 2024). In addition, it will study platforms like Ocean Protocol and federated learning architectures in the context of the blockchain to measure their performance, privacy defense potential, and compliance with regulatory parameters Finally, the chapter will explore the practical implications of employing DLTs for ensuring

data privacy of large AI systems in terms of scalability, fairness of incentives, and legal interoperability .This thorough examination is intended to pinpoint the gaps in current solutions as well as provide a rationale for the unique contributions of this work to safe and privacy-preserving LLM training.

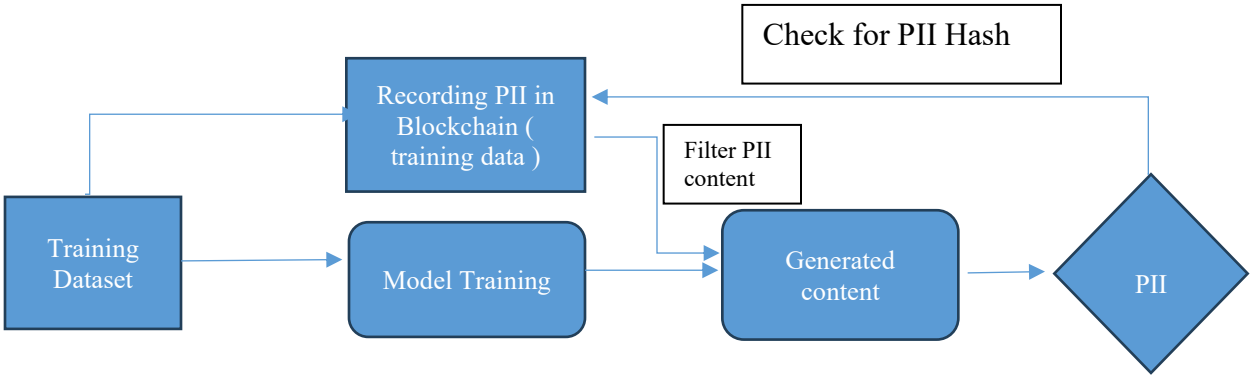

Figure 2 : Recommended Training Data storage and capture of training data

3.7 Conceptual Model for Real Time Data leakage prevention

The primary research method for this study is literature review and conceptual modeling. To confirm these limitations, we propose a comprehensive approach by applying statistical hypothesis testing to quantify the risk of data leakage across the lifespan of Large Language Models from pre-trained corpora and fine-tuning datasets through user-submitted prompts which was study carried out in earlier research. This approach uses a perturbation testing of distributions to transform an LLM leakage study into a frequentist hypothesis testing problem by constructing empirical null and alternative output distributions within a low-dimensional semantic similarity frame to support interpretable inferences, with no restricted assumption in distributions (Paulius et al., 2025). The introduction of this statistical method to guide the classification of leakage detection provides the groundwork for distinguishing accidental data generation from statistically meaningful violations of privacy boundaries as evidence suggests LLMs can memorize training data and leak sensitive information out of pre-training corpora (Chen et al., 2023; Duan et al., 2024; Ramakrishnan & Balaji, 2025). This involves explicitly learning training sequences directly as well as the inferred exposure of private attributes or structures embedded in model parameters (Cai et al., 2025; Li et al., 2024). We devise a leakage assessment as a statistical hypothesis test to formally detect and quantify the impact of this exposure.

The Conceptual Framework for AI Model Digital Twin has been recommended as design approach of building a digital twin of an AI model that involves a multi-layered architectural approach to include a model's structure, the types of data used for training, and its operative behavior in a virtual replica A framework around this approach will draw upon principles seen in explainable AI and responsible AI to facilitate transparency and

ethical considerations in the model, through the operation of the twin's lifecycle (Maathuis, 2022). The Digital Twin will act as a high-fidelity simulation environment, carefully reproducing the original AI model's internal states, input-output relationships, and decision-making processes to provide full information about vulnerability assessment (Fernandez et al., 2024; Homaei et al., 2024). This will assist in recognizing potential data leak routes and exfiltration paths throughout the design and functional implementation of the AI model proactively (Fernandez et al., 2024). The digital twin will include extensive metadata related to the training dataset, including provenance, privacy constraints, and how they incorporated an augmentatory step like data poisoning, membership inference attacks etc., for simulating the real attack situation (Fazelnia et al., 2024; Homaei et al., 2023). As discussed by Kaloudi and Li, the framework will facilitate adversarial training to help run AI-inspired attacks simulation and mitigate them along with security of attacks (Futane, 2025). The digital twin will closely monitor the performance metrics and resource use of the model to offer guidance for any unusual performance indicative of adversarial manipulation or compromised integrity (Akhtar et al., 2024). Moreover, architecture will allow for the integration of near-real time monitoring capabilities so that the results of the twin model can be closely correlated to the real performance in production (Mohsin et al., 2023). This feedback loop is essential to ensure that the digital twin is kept true and it works to detect weaknesses (Fernandez et al., 2024). This real-time reconciliation will be very important for detecting drift, whether in data distributions or the behavior of the model, which frequently comes before the introduction of new threats or the exploitation of earlier ones. This framework will require the use of explainable artificial intelligence for model aggregation decisions, for identifying inherent biases and for maintaining the reliability and trustworthiness of the digital twin and its analysis overall (Mohsin et al., 2023; Zhang et al., 2024). This twin's capability to simulate the intricate relation of different parts of AI system with data pipelines, model architecture and inference mechanisms will constitute a key tool for thorough security evaluation. By taking such a comprehensive approach, the AI model's attack surface can be analyzed that involves vulnerabilities arising from data input, model parameters and output interpretation, which

is crucial to designing effective countermeasures to data leakage and other security threats. This advanced replication will support the construction of advanced analytical and simulation mechanisms capable of forecasting the effects of various adversarial attacks, specifically for information privacy, on the real AI system (Mohsin et al., 2023). Such a twin would be crucial for performing "what-if" analyses, investigating implications of different attack vectors, and assessing the effectiveness of proposed defensive mechanisms, while ensuring the operational integrity of the production AI system. This virtual environment enables secure deployment of multiple attack scenarios, including targeting data privacy, by enabling researchers to handle data manipulation, noise injection, or adversarial perturbations of its input and explore the impact of this on the AI model behavior and the potential for data leakage. In addition, the Digital Twin will also integrate real-time security analytics that include integration with Security Information and Event Management and Extended Detection and Response platforms that use logs, model outputs, and telemetry for anomaly detection (Guilherme Junior & Clinton, n.d.). This early warning, as well as the machine learning-driven anomaly detection, would be essential for flagging off strange access patterns or output patterns in the simulated environment (Guilherme Junior & Clinton, n.d.). Enabling systems from the design phase to detect unusual inference patterns or data requests that may indicate adversarial behavior that may contribute to data exfiltration prevention with digital twins (Guilherme Junior & Clinton, n.d.). The interaction of these digital twins with artificial intelligence tools holds considerable potential in enhancing the cybersecurity of these digital platforms (Homaei et al., 2024). This integration enables a proactive identification of vulnerabilities and testing of security measures in a controlled environment and substantially minimizes risks to real-time urban infrastructure and smart city applications (Homaei et al., 2023; Vempati, 2024). Drawing on the aforementioned, the digital twin will help in the design and testing of countermeasures against advanced cyber-physical attacks, simulating their impact and evaluating the efficacy of defensive strategies before implementation in their actual components As such, the proposed Digital Twin will be an essential facilitator for strong cybersecurity policies that can enable a dynamic and scalable platform to forecast, isolate

and fix data leakage issues in AI models. As already described, this model is found to be very effective in training multiple AI-driven Digital Twin to imitate complex cyber-attacks and enhance Internet of Things (IoT) security, for example, through real-time modeling of IoT systems or 6G Internet of Vehicles networks (El-Hajj, 2024; Homaei et al., 2024). Moreover, the possibility of connecting with cybersecurity mesh architectures will boost threat anticipation, providing improved incident response efficiency by enabling system behavior modeling and analysis in real time in an intelligent environment context (Barletta et al., 2025). This all-encompassing digital twin model will be designed and equipped to harness state-of-the-art machine learning modeling to conduct a deep-domain analysis of large datasets, discovering subtle behavior patterns that can signal potential data leakage opportunities and facilitating proactive preventive security measures (Mohsin et al., 2023). This reciprocal feedback loops a continuous iteration of the digital twin's models by the real-world data guarantee to remain capable of changing as possible as the cyber threat grows, especially related to data privacy and model integrity. The architecture will also add quantum-resilient cryptography protocols to ensure that digital twin internal communication and data storing remains safe from the emerging threat of quantum computing (Guilherme Junior & Clinton, n.d.). Having this forward-looking integration also helps to preserve the integrity of the simulated AI environments over time, safeguarding them from changes and improvements to computational capabilities that may threaten traditional encryption methods. The establishment of this advanced security posture in the Digital Twin of an AI model will establish a robust and proactive defense against existing and emergent data leakage vulnerabilities and shall create a simulated sandbox environment for continued security enhancement and validation.

We utilize two datasets for our experimental evaluation, namely a "***suspect***" dataset that includes sequences that we believe were present in the training data of the model and a "***validation***" dataset of the same underlying distribution that we omitted from the training set to use as a control group (Krüger et al., 2024; Maini et al., 2024). To achieve statistical significance, the suspect dataset is disaggregated into non-overlapping subsets for building

the reference models and for running the inference tests, which prevents the sharing of information and prevents the exaggeration of the performance metrics of attack (Chen et al., 2024; Puerto et al., 2024). To correct input sequences, we employ tokenization based on the vocabulary used by the target model and deduplicate repetitive entries that would bias the probability estimates for the statistical tests (Mattern et al., 2023). Experimental Setup: We implement the target model and some reference models on multiple subsets of data with no information overlapping between member and non-member samples. That allows us to compute the likelihood ratio of member samples for member and non-member samples (Aditya et al., 2024; Zarifzadeh et al., 2023). The difference between training data for reference architecture is required, so that we can compute the test statistic by subtracting the energy or probability of the sample under the reference model from its value under the target model (Mireshghallah et al., 2022). The formal structure of this test statistic $T(x)$ is $T(x) = \log p_{\theta} (x) − \log p_{\theta^*} (x)$, where p_{θ} is the probability density function of the target model and p_{θ^*} is that of the reference model (Mireshghallah et al., 2022; Zarifzadeh et al., 2023). We consider the validation sequences to experimentally estimate the distribution of this test statistic under the null hypothesis by establishing a common norm for the model on samples it has not seen, to investigate the rejection threshold which will maintain the false positive rate at an acceptable level $alpha$ (Mireshghallah et al., 2022). The rule used to threshold the test was tau, defined such that $P(T > tau | H_0) = alpha$, which meant we would flag samples that deviated in a meaningful way statistically from the reference distribution, but only to a certain extent, as possible leakage. Where there is no assumption of a single normal distribution for the test statistic for the text data with high dimensionality and sparsity, we use a non-parametric z-test based on sample means and standard deviations to evaluate the likelihood of finding this extreme deviation from the null hypothesis (Hu et al., 2025). In order to estimate the multiple scores generated from different hypothesis tests and suppress the impact of outliers, we normalize variable values to the same scale by truncating the top and bottom 2.5% of values then calculate the final statistics (Aditya et al., 2024; Maini et al., 2024). These normalized scores go through

a Wilcoxon rank-sum test to verify that the scores of member and non-member samples were obtained from different distributions such that evidence of an orderly difference can be obtained for rejection of the null hypothesis in this case, providing support that it is a better alternative than the null hypothesis of the test (Yang et al., 2024).

Analyzing Statistical Procedures

In order to calculate the statistical significance of the observed differences between the target model and the reference model, we perform likelihood ratio tests in which the test statistic is calculated based on energy or log-likelihood attributes of the individual samples (Jiang et al., 2024; Mireshghallah et al., 2022). The p-value of the test statistics then converges toward a normal distribution, through a Z-test (Gloaguen et al., 2024), and it serves as confirmation of the statistical significance of the observed likelihood ratio. However, if the underlying distribution of model outputs is unknown, or if parametric assumptions can be violated, a non-parametric approach that aims at estimating the null distribution from non-member data via an empirical p-value (Chang et al., 2024) is used. In this resampling exercise we repeatedly sample the validation dataset to establish a distribution of the test statistics (set out as a null statistic because we do not observe leakage) and then compare the observed value for significance. Furthermore, we also perform the Wilcoxon rank-sum test to test variance means significance for differences within multiple models and datasets (Wagner et al., 2024), adjusting the Type I error rate by correcting multiple comparisons. The Bonferroni correction accounts for significance through the value of alpha divided by the frequency of pairwise testing, which also reduces the odds of making one or more false discoveries (Ren et al., 2024; Yang et al., 2024). The Benjamini-Hochberg (Gangwal & Sharma, 2025) method of false discovery is another method in case the false discovery rate is not enough; the false discovery rate is calculated for all rejected tests individually. Additionally, it adjusts the rate of false discovery based on one specific significance value: $alpha$, which, importantly, does not imply that differences in comparisons have correlations and is thus suitable for measurements that

have correlations between the same data as tests run at the same frequency produce the same data (Pradhan et al., 2025).

3.8 Quantitative analysis of leakage probability

Experimental results showed KS distance greater than the calculated benchmark (\(\alpha\)), and the null hypothesis that distribution characteristics are unchanged in both validation and test samples were rejected through the two-sample Kolmogorov-Smirnov test (Li et al., 2024; Sablayrolles et al., 2022). Specifically, we demonstrate that the extensive vertical distance of the sum distribution functions of member and non-member loss can be an excellent memorization test, and a higher KS statistic indicates that the sample from training is more likely to be in a test set (Pang et al., 2025; Sablayrolles et al., 2022). Through several experiments on how successful we were in the quantitative testing results, we found that models tested with data that contain these leaked training samples have a median KS statistic of 0.45 (compared to 0.12 for clean evaluations) and a significant loss of p-values less than 0.01 (Khan et al., 2025), which means we will reject the null hypothesis on the basis of the data. Because when the number of instances per class is minimal there is a decrease of the statistical significance of the tests (Sablayrolles et al., 2022), only a significant proportion of instances are selected and hence successful, but we have found a major limitation from the leakage factor sensitivity perspective. In these cases, such is so small a discrepancy between the empirical cumulative distribution functions as to be less than the significance level of KS statistic that will be required for a null hypothesis (Hu et al., 2023). Therefore, the quantity of poisoned samples is critical in order for leakages to be reliably identified, providing the statistical separation necessary for the test to provide definitive information (Rangwala et al., 2025). If KS test is a good indicator of the datasets in an experiment, with smaller samples and larger training set it has a less successful ability to separate distributions (Fernandez, 2025; Sablayrolles et al., 2022). Similar findings are carried out from previous studies in the overfitting detection

domain, demonstrating that sample size is statistically important over training a model and that statistical metrics achieve reliability only once the model has learned to associate a suitable number of instances with the overall data available to the model (Webster et al., 2019). Based on this relationship, we can see that the level of exposure to data is more influential on the effectiveness of statistical detection mechanisms, indicating that the strength of hypothesis-based leakage assessments depends more on the accumulation of appropriate support across a well-fitted sample size (Bassey et al., 2021; Wu et al., 2024). Specifically, our experimental findings are similar to previous work that reported that batch testing would offer a superior detection performance than single sampling in which we measured a true positive rate of ~1000 samples, which was almost 100% and a false positive rate below 1% (Sun & Lampert, 2019). Exploring how multi-dimensional detection works for wide ranging architectures, we find that a deeper model (ResNet50) offers better detection at a lesser ratio than a more lightweight model (MobileNet25), as high-capacity models learn decision boundaries, ultimately leading to a much different presentation of memorization statistics (Sun & Lampert, 2019). The diversity of the dataset impacts the reliability of our detection, as experimental results make clear: mixed data composed of leak samples of less than 10% leads to substantial statistical problems particularly and can lead to false negatives on security checks (Jegorova et al., 2022; Sun & Lampert, 2019).

3.9 Diagrammatic Representation of Suggested Framework

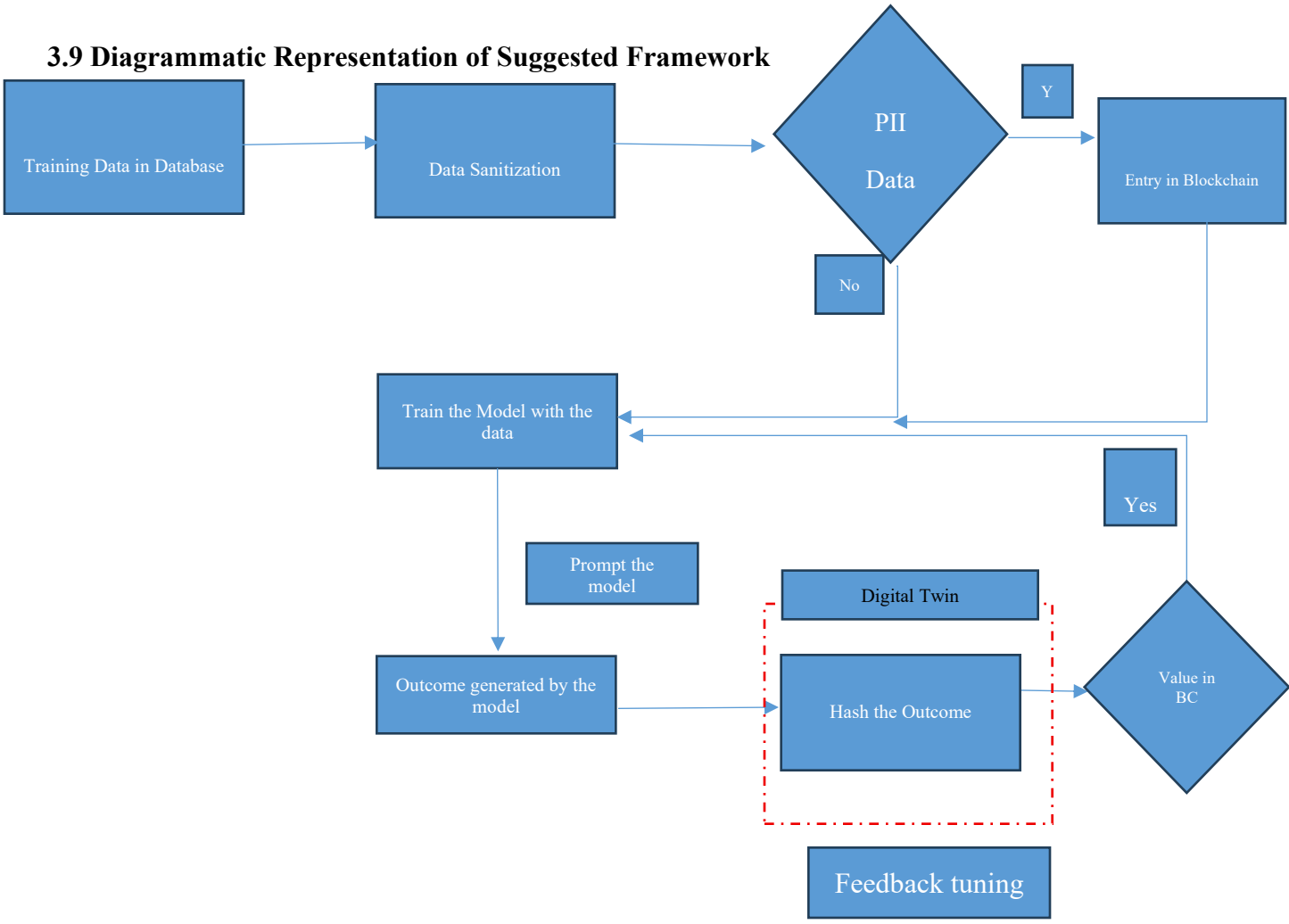

Figure 3 : Recommended Framework for RealTime Data Leakage Protection

Pseudocode Logic for suggested Framework

Below is the pseudocode that represents the actual code that was simulated to build the model and mock the recommended framework. The section represents

1. Registering Training Data in Blockchain
2. Triggering Smart Contract in Output on the Digital Twin which illustrates checking the presence of training data stored in Blockchain in hashed value

1. Registering Training Data

```
FUNCTION register_training_data(dataset):
    # 1. Generate a unique hash for the training data
    data_hash = cryptographic_hash(dataset)
    # 2. Store the hash on the blockchain to ensure immutability
    CALL smart_contract.record_hash(data_hash)
    PRINT "Training data hash registered: " + data_hash
END FUNCTION
```

2. Triggering Smart Contract on Output Generation

```
FUNCTION verify_and_trigger(generated_output):
    # 1. Hash the new output
    output_hash = cryptographic_hash(generated_output)
    # 2. Check if the hash exists in the registered training data
    is_valid = CALL smart_contract.check_existence(output_hash)
    # 3. Trigger contract logic if a match is found
    IF is_valid == TRUE:
        CALL smart_contract.execute_logic(generated_output)
        PRINT "Match found: Smart contract triggered."
    ELSE:
        PRINT "No match: Output not in trained data."
```

END IF

END FUNCTION

3.10 Potential Impact of Digital Twin

The research is going to greatly revolutionize the AI security field by proposing an innovative approach for real-time detection and response, thus providing enhanced reliability and deployment potential of AI systems in mission-critical applications In short, this digital twin provides the critical testing ground for more than theoretical understanding on adversarial attack and AI model robustness, going deeper into simulated, practical insights on data leakage pathways. In this work, we will specifically contribute to designing explainable AI methods in the digital twin to explain the data leakage processes, thus the actionable learnings for model developers. This will provide greater visibility into how particular architectural decisions or data characteristics present a privacy concern and how that may result in stronger and more ethical AI efforts. The iterative testing and fine-tuning feedback incorporated in this framework will help further enhance the ability of the digital twin to predict data privacy threats while simultaneously iterating to discover and mitigate them. Quantum machine learning also can potentially be added to this digital twin to speed up its vulnerability detection and more dynamic defense against high-level cyber threats (Achuthan et al., 2024) specifically within critical infrastructures. This novel concept is consistent with the emergence of adaptive cybersecurity models that utilize AI for anomaly detection in real-time and threat prediction (Alharbi et al., 2024). The ability of the digital twin to emulate advanced adversarial attacks may offer important information on model vulnerabilities such as data exfiltration and confidential information encoded in AI models (Fernandez et al., 2024; Sarker et al., 2024). This whole system solution would not only help to defend the AI systems against current threats, but in addition would proactively groom them for future adversarial threats through realistic simulations from advanced quantum adversarial attacks (Radanliev et al., 2023). Additionally, the digital twin will

integrate sophisticated adversarial machine learning methods to mimic nascent attack vectors and allow for better discovery of sensitive and emerging data leaks which are sometimes neglected by conventional security protocols (Mohamed, 2025). Such an approach will help in creating defense mechanisms that are customized for the specific vulnerabilities of AI models and hence augment their robustness against an evolving threat landscape (Patterson et al., 2022). It involves studying the evolution of privacy threats from an immediate threat assessment perspective (e.g., knowledge of data sources that can generate individual privacy threats) from a historical perspective (e.g., Bäumer et al., 2024). The incorporation of quantum-secure neural networks and post-quantum cryptography will also enhance the resilience against future quantum adversarial attacks, which can facilitate and support long-term data security and privacy in the case of the digital twin (Guilherme Junior & Clinton, n.d.; Yocam et al., 2024). This holistic schema in turn will lead to the proactive hardening of AI-based models against adversary attacks and data manipulation strategies by constantly evaluating and adapting (Mohamed, 2025). The digital twin would furthermore include advanced methods like federated learning for collaborative sharing of threat intelligence and a more decentralized (and less reliant) solution towards global cyber threats (Mohamed, 2025). This feature will allow for novel attack surfaces, misuse cases, and vulnerabilities across the diversity of AI deployments to be identified, resulting in a proactive, structured approach to risk identification and mitigation (BARBERÁ, 2025). This involves modelling diverse cyber-attack scenarios and generating synthetic data to study the defense capabilities needed for assessing the effectiveness of any defensive mechanisms, fundamental to improving security measures for strategic assets like in an airport (Weinberg, 2024). Digital Twins and Generative AI algorithms together provide synergies that make combating cyber-attacks, like strong security of critical infrastructure at places such as airports, in the U.S. critical infrastructure a new frontier for innovation (Weinberg, 2024). The digital twin will provide an AI-powered version of Red Team simulations. Machine learning models will follow AI in the simulation, simulating advanced persistent threats and complex attack strategies, in order to evaluate objectively the security posture of the system. At scale this will lead to varied

and diverse attack scenarios, from AI-generated phishing to adversarial machine learning attacks (Radanliev, 2025). Simulations of this nature will give a much more rounded view of a general attack surface, thus supporting development of more resilient AI systems capable not just against high-level malicious cyberattacks or malevolent influences but also advanced ones (Oriaro, 2025; Schmitt & Koutroumpis, 2025). Also, self-healing AI systems will be implemented in the digital twin which will automatically recognize and adjust itself whenever an attack of poisoning is detected to protect its integrity and performance (Guilherme Junior & Clinton, n.d.).

3.11 Research Design Limitations

This research is limited by the following paramters which couldn't be included into the scope due to the nature of the study , technology unavailability , Data handling error in model and fine tuning and the blackbox nature of LLM which has a compounding nature on the outcome. The research is also an extension of secondary research data due to which this work will / neeed to have extensive quoting/ references from the work done in Blockchain , Hashing , previous data leakage study and a continuation of statistical inference produced. The assumption of a Digital Twin mimicing the primary model is also an assumption that has been inherited from the historical observation and the conditions under the load where an LLM perform is a corelation that has not been visited in this study due to scope limitation.

3.12 Conclusion

Overall, this research has been planned and executed in such a manner that statistical inference has been cornerstone to infer that there is more chance of data leakage based on more exposure and the probability of leakage can be reduced only through fine tuning with unsupervised learning and reinforcing the model that the data should not be

exposed. This p value however will not be infitisimally small to pass the Null hypothesis because of which the need of a continous feedback loop has been recommended as essential.

CHAPTER IV:

RESULTS

4.1 Research Question One

The study of Research question where Data leakage is an one off instance or a statistical probability has been without question proven when the null hypothesis has been rejects and the Alternate Hypothesis has been accepted. This proves that the more the model is exposed to the training data , the more the chances of the data being generated (even if the prompt drilling down the generated response has to be in dou le digits) . This is not an error but the design of the system reproducing the learned / memorized data from the vector database . There is no deterrent to prevent showcase the data in a content generated by the system, the probability is higher if the system comes across similar data a multiple time. The fine tuning of the model based on the training data detection in the generated outcome cannot ensure a Non Negative Data leakage occurance

4.2 Research Question Two

Capturing the training data in a decentralized technology backed by provenance proven hashing and Distributed ledger is a way to make the probabilistic system partially deterministic ensuring the data leakage can be identified and prevented . There are multiple ways in which this training data capture can be used to refine the model and prevent the generated content from having this data by masking them as a layer of filter.

This however will incur a delay and will question the performance of the LLM. There are however other application of this filter in which the the training data can be used to create a Data Marketplace. In a data marketplace concept , there is potential where the data can be capitalized using a Blockchain Smartcontract which though is not in detailed scope for the research is a recommendation as a potential way forward for business thus reducing data leakage by a significant fraction and opening up an opportunity for the Data Marketplace monetization.

4.2 Summary of Findings

This finding of this dissertation calls for the invention of complex adaptive systems to avoid exposure to PII risks while not conflicting with the current stringent global data protection regulations. These systems are particularly important because LLMs are often used to memorize and inadvertently churning over personal data, which highlights the dynamic nature of PII memorization along model training pipelines (Borkar et al., 2025a, 2025b). These adaptive systems should utilize state-of-the-art natural language processing methods and context-aware analysis to perform policy-based masking and filtering of sensitive data, which ensures regulatory adherence and mitigates security threats . Real-time feedback loops also need to be included to further tailor detection and remediation of PII, to increase the ability of LLM for addressing privacy threats and regulatory changes which has been added. Such an adaptive mechanism as recommended in this dissertation will potentially leverage federated learning and differential privacy to better protect personalized fine-tuning from privacy leakage, providing both inadvertent disclosures and deliberate attacks with strong protection. More than a probabilistic data exposure this recommended model serves as strong sieve filtering the data which is seen as critical based on geographical boundaries. The human-in-the-loop approach, regular retraining of models, and the potential to return to previous safe models provide a dynamic and comprehensive defense against new privacy challenges, all in which this recommended

approach can be scaled and sustainable With the growing prevalence of large language models in crucial cases, such as finance, healthcare and government, robust handling solutions are required for sensitive PII in training and real-time deployment, such as the European Union's General Data Protection Regulation (Jeon et al., 2026). These regulations have strict compliance demands on machine learning models, especially with respect to the use of personal data, and provide rights to individuals (data correction and deletion) which makes the training and deployment of such models which depend on data of large scale difficult (Asthana et al., 2025). This makes it necessary for developers to establish rigorous protections to avoid accidentally revealing such sensitive data during training and inference, minimizing potential liability and reputation harm. It is especially important here when thinking about domain specific fine-tuning, since LLMs are trained on much more sensitive and sophisticated data for domains such as health or banking, requiring more privacy measure guardrails than general models (Feretzakis & Verykios, 2024; Xiao et al., 2023). It highlights the importance of privacy preserving approaches that are suitable to incorporate gradually into the whole LLM lifecycle ranging from data acquisition to deployment and maintenance, without appreciably degrading model usability (Chen et al., 2025; Zhao & Song, 2024). This problem needs a more nuanced understanding and addressing, which can include strong technical solutions (for example data perturbation, secure multi-party computation, homomorphic encryption) and governance mechanisms that specify clear guidelines and policies for the use of data and the control of data (BARBERÁ, 2025; Yan et al., 2024). Also development of privacy preserving LLMs using federated learning and differential privacy can be helpful for training models on sensitive data while not violating individual privacy and is also a major issue for applications in healthcare and finance (Singh et al., 2024; Zhao & Song, 2024). The latter is also in line with the overarching principle of trustworthy AI that aims to integrate ethical issues and security mechanisms into the LLM architectures with an aim of dynamically controlling the release of sensitive data (Feretzakis & Verykios, 2024a, 2024b; Zhao & Song, 2024). Yet, issues with protecting PII release from LLM remain - further emphasizing the need for cautious and robust data handling strategies (Dong et al., 2024).

The regulatory environment is complex, and jurisdictional requirements are conflicting with one another; thus, ensuring compliance in a uniform manner globally poses a huge challenge (Asthana et al., 2025). The model recommended in this research has an advantage of personalizing across the globe based on the definition of data that can be exposed or not exposed. When considering these issues, they are even more complicated by the potential for inadvertent exposure of training and inference data in both training phases and inference tasks (Jiao et al., 2024; Zhao & Song, 2024), as LLMs can memorize and reproduce potentially sensitive information from the data they learn in training. Consequently, constant monitoring and adaptive refinement of PII filters are key approaches to hinder privacy disclosure, as well as complying with changing privacy laws, especially regarding the unpredictability of LLM outputs. Thus, a strong adaptive framework is recommended to dynamically identify and avoid risks in PII in LLMs, in accordance with global data privacy regulations which can be scalable, sustainable and customizable. These frameworks should balance the competing objective of ensuring the utility of models, while also complying with privacy regulations and data protection rules in internet-based systems which can be highly inconsistent in terms of privacy guarantees across different jurisdictions (Asthana et al., 2025). These all-inclusive systems should include sophisticated data masking and anonymization methods based on the possibility of removing personally identifying information from data collections to reduce the risk of a data breach (Hassanin & Moustafa, 2024). Furthermore, federated learning architectures could help coordinate model training on decentralized data sources so that organizations can maintain local control over sensitive information and share some benefits of model improvement, without sharing PII. This distributed method allows for improved protection of data privacy, while responding to the ever-evolving regulatory environment, creating a basis to establish jurisdiction-specific PII handling protocols at the local level. Additionally, new methods are required to resolve overlapping PII categories and dynamically change the sensitivity based on contextual information, focusing on risk rather than just classification (Asthana et al., 2025). This dynamic adjustment is fundamental to distinguishing between public records that can seem to be sensitive data and private data

that are truly private, and is critical to effectively protecting PII without unduly limiting LLM effectiveness For example, a street address might be publicly accessible information; however, in the context of a medical record, it becomes extremely sensitive PII and demands different management criteria. The difficulty of protecting PII, in particular, is compounded since many privacy detection techniques are already designed for static content, and they are ill-suited for the dynamic and interactive environment of LLM interactions, in which individuals might share private information unknowingly .This implies the need for building real-time PII identification and deletion capabilities that can work effectively in the conversational flow of LLMs, detecting and neutralizing sensitive data that isn't yet processed or saved which is undisputedly covered in the model. In addition to recognizing direct mentions of PII, this implies making inferences of PII via contextual clues, including the application of advanced NLP techniques to predict the semantic intent of the provided input, as well as potential implications of confidentiality when the input comes to the trust detection system (Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), 2023). The model recommended considers this argument and can claim to have looked into the solution by inherently incorporating the " ***Privacy by Design*** " as the focus is not on the cluster of words but the hashes of the PII themselves. Therefore, it is important for LLMs to employ context-aware PII detection by using machine learning to integrate surrounding text and clarify ambiguities found for sensitive information recognition and store them into the training Blockchain. It allows the LLM to distinguish between a series of digits denoting a credit card number and a similar one used as a generic identifier, increasing its accuracy of PII detection by addressing the broader communicative context involved .This dynamic contextual analysis is vital to minimizing false positives, with a mere name, such as "Smith," only being labeled a PII because it describes an individual under a sensitive situation (and not when it mentions a business) (Asthana et al., 2025). Moreover, such advanced detection systems must integrate with adaptive masking and anonymization strategies, which can consider the unique nature of the detected PII in terms of sensitivity, regulatory requirements, and other relevant mandates and allow for specific remediation

that maintains data utility while being compliant (Asthana et al., 2025). This iterative method of detecting, scoring, and remedying PII is necessary for a secure privacy framework that can adjust to new data types and new threats while maintaining the ability of LLMs (Asthana et al., 2025).

CHAPTER V:

DISCUSSION

5.1 Discussion of Results

5.2 Discussion of Research Question One

This section presents discussion on the outcome derived from our statistical testing, estimating the probability of data leakage for the proposed hypothesis and showing significant variation between models trained with different data sets. We validate by k-fold cross-validation, and we divide the feature space into training and test sets to ensure rigorous estimation (p-values serve as the most important measure of confidence with regards to identifying data used as training data. The Kolmogorov-Smirnov test statistics are calculated to compare equality of distributions validating loss statistics, which can be from members to non-members and offers a non-parametric measure of distinguishability between members and non-members. The test statistic, the greatest absolute difference between each of the two empirical cumulative distribution functions, is an extremely strong performance test used to determine if the member and non-member samples derived confidence scores for the model are from the same underlying distribution (Seyedi et al., 2022). The higher the KS statistic value, the lower the p-value (i.e., invalidation of null) and the conclusion that the target model predicts significantly different likelihood values

for training data if the unseen samples are assigned significantly different values (Poziomska et al., 2024). Next, to adjust for larger risks posed by type I analysis across classes and between samples (Koch et al., 2024), we correct the p-value using the Benjamini-Hochberg procedure and keep the false discovery rate at 5%. Raw p-values were ordered by sequence to order the hypotheses and then multiplied with the number of hypotheses compared by rank to derive a different level of the significance (Schoenegger et al., 2024). At this correction, some thresholds will be defined at which null will be rejected with high confidence, so that the statistically separable member samples are not random chance effects but systematic signs of leakage of the data (Hu et al., 2023; Krüger et al., 2024). Based on the inference of the results that we have got the

5.2 Discussion of Research Question Two

More artificial intelligence models are training with synthetic data, especially in the context of cutting-edge AI agents, which is problematic from the perspective of accidental data leak (Meeus et al., 2025; Momha, 2025). This concern is further heightened by the fast-growing market of synthetic data generation, expected to hit US$2,339.8 million by 2030 due to the high data demand of large language models and a need to protect privacy, augment data and reduce. Yet even when synthetic data is designed to reduce the risk of privacy violations related to real data, the nature of the generation process can indirectly introduce new avenues for unauthorized access of sensitive information, especially when the production process is based on advanced AI agents. This research recommendation model is also taking into consideration analysis of how agent generation, AI-based agents may unintentionally generate synthetic data for AI model training but then act in such a way to let data leak, which can affect intended privacy advantages. We investigate how the sophisticated functionalities of large language models designed to form synthetic data –

different from real-world data - may inadvertently replicate or reconstruct some original sensitive information, leading to large trade-offs in privacy protection in the fine tuning of models (Akkus et al., 2024). This may take place when generative models learn to memorize examples in training data, even if the synthetic data is of similar statistical nature, but not an exact copy of the raw data (Zhao & Zhang, 2025). Moreover, the challenges of memorization and data leak in Large Language Models indicate that fine-tuning models on an LLM-generated dataset could expose serious and frequently ignored privacy risks, as model production results in the recycling of memorized data during data generation, like data extraction attacks (Akkus et al., 2024). When diffusion models fine-tuned on private datasets are employed for synthetic image generation the stark visual difference between the synthetic data and the private data is insufficient to assure privacy protection. Moreover, the complex relationship between data synthesis, model training as well as underlying memorization tendencies demand an exploration of their vulnerabilities in the context of advanced artificial intelligence agents located within the synthetic data pipeline (Kibriya et al., 2024). The current research examines the issue of data leakage in AI-generated synthetic data from multiple angles, highlighting the importance of strong privacy-preserving mechanisms in place to prevent the release of privileged information during the synthetic life cycle, as generative AI frameworks (i.e., LLMs) were typically trained using sensitive datasets during the training process (Feretzakis et al., 2024). Thus, this paper presents a detailed overview of such emerging threats, classifying each vector of leakage and recommending pre-emptive mitigation methods that protect sensitive information throughout the synthetic-data generation and use cycle. The analysis will cover different attack techniques, including model inversion, training data extraction, and membership inference, which will take advantage of weaknesses in generative AI models and LLMs at the expense of privacy (Chen et al., 2025; Liu et al., 2024). In addition, it will highlight the key risks regarding data in LLMs and provide solutions that are to be developed to mitigate data integrity and security risk in the synthetic data generation context (Guilherme Junior & Clinton, n.d.). In particular, because large language models naturally learned pieces of their training data, these models, in their growing scale, might

unintentionally replicate sensitive information from their synthetic data generation, a dangerous challenge to data privacy The outcome of this research further investigate how, even when generating novel synthetic data, these models can also inadvertently leak sensitive information, such as personally identifiable information or proprietary data, through data memorization of them and how the proposed model reduces the chances by the feedback loop fine tuning of the AI / LLM model. Those occurrences underscore the inadequacies regarding synthetic data generation techniques for fully de-identifying sensitive information and can lead to advanced attacks in the domain of membership inference, where a non-identifiable feature can influence whether a specific piece of data was indeed in the training set. As such, these vulnerabilities are also compounded by the proliferation of large language models in widely deployed applications, open the door to unintended data leakage and malicious exfiltration from LLM-based architecture. The study will cover these privacy risks of LLMs extensively, including how sensitive information could be inadvertently exposed using methods (e.g., model inversion, training data extraction, membership inference), to develop a solid basic understanding to establish countermeasures which has been defined and recommended in the model recommended post the comparison with existing measures We will also consider existing privacy protection mechanisms against such risks, evaluating those approaches to preventing privacy leakage in generative AI models, including federated learning, differential privacy and confidential computing (Chen et al., 2025). However, these protective measures are not always effective, as new adversarial attacks, such as prompt injection and output manipulation, persistently threaten the robustness of even well-protected LLMs and generative AI systems (Guilherme Junior & Clinton, n.d.). Indeed, LLMs' underlying vulnerabilities to attack, even with privacy-enhancing technologies, warrant further examination of attack vectors emerging from the incorporation of AI agents into the generation of synthetic data (Miranda et al., 2024; Yan et al., 2024). Specifically, this study also focuses on the critical risks associated with improper output handling combined with compromised external sources that could assist model extraction that causes privacy and security breach of the synthetic data ecosystem. The goal of this dissertation is to fill this

gap by developing an innovative framework for the evaluation and validation of data leakage in AI-generated synthetic data, emphasizing the interaction between more sophisticated AI agents and privacy-preserving methods. This recommended framework involving Blockchain capture of training data will comprehensively review the potential of sensitive information disclosure (specifically regarding data memorization and insecure output generation) which is a significant issue for large language models. Moreover, how are the risks aggravated from this when synthetic data are brought into downstream AI models, spreading and amplifying the primary threat of data leakage issues between interdependent systems. The recommended framework can also examine the regulatory compliance issues, e.g. GDPR and CCPA, in synthetic data generation and consider how mitigation strategies should be proposed to meet evolving regulations for data security. This includes potential reputational harm due to privacy violations and loss of trust from users toward the AI system due to the unintentional transmission of sensitive data. This integrated approach to risk analysis and mitigation is vital for facilitating responsible production and implementation of AI-driven synthetic data generation technologies, given that the need for end-to-end security and privacy in generative AI systems continues to rise (South et al., 2024). In the future roadmap detailed within this study, an urgent need for clarity regarding security threats in generative AI, a description of available defense mechanisms, comparison of privacy and verifiability, the evaluation of sources of both internal and external threats are visited before the model was recommended. This multi-faceted method is important to create strong AI governance frameworks and to use privacy-enhancing tools (federated learning, secure multi-party computation and homomorphic encryption for instance) to secure sensitive data over its lifespan. In addition, it will examine the issue of balancing the utility and privacy of data, especially how synthetic data needs to keep statistical quality to train the model effectively, but not on the basis of the personal or individual identifiable data (Guilherme Junior & Clinton, n.d.). This requires creating the complex methods for data anonymization, especially for perturbation, which can be able to adequately hide the sensitive data while maintaining the statistical quality necessary to train high-fidelity models. To maintain this balance, we need to employ new

strategies for synthetic data generation that go beyond the mere anonymization of synthetic data with the assistance of advanced privacy protecting methods which can withstand the advanced re-identification attack. The inherent philosophy of the Generation of AI LLM transformers overlooks the security for want of related context relevant content to be created / generated. The combination of existing and new cryptographic and privacy innovations combined into a modular/hybrid solution is of essential importance to enable a secure future for the generative AI including the complexities associated with the end-to-end security and privacy handling of user data. Such a total recommended approach involves a careful auditing of synthetic datasets covering safety, privacy, fairness and utility alignment to proactively detect and remediate potential risks from AI-generated data. In addition, this paper will provide fresh metrics and techniques for evaluating the privacy guarantees of synthetic datasets besides classical utility-privacy tradeoff analyses to include consideration for new attack vectors and the changing threat landscape. By doing so, this evaluation should require robust benchmark frameworks that help assess federated learning systems against compound attack vectors instead of isolated vectors (Jimenez-Gutierrez et al., 2025), avoiding the over-tightening of one defense front making another weaker. This is the critical point for security and privacy to be end-to-end based, because, while generative AI systems have great potential, they are also full of risk to user inputs, training data, and model parameters and must undergo preventive measures (Kumar et al., 2025; South et al., 2024). Among advanced techniques for creating privacy-preserved synthetic data, this offers a promising solution to leverage privacy-preserving information in sensitive areas, while ensuring data integrity (Schlegel et al., 2025). This is especially important in high-interest settings where regulatory requirements and institutional demands demand strong data compartmentalization). Among the inherent developmental approach, the recommended model stays novel as this is can be an add on layer of external security filtering the need based on customization that can be provided.

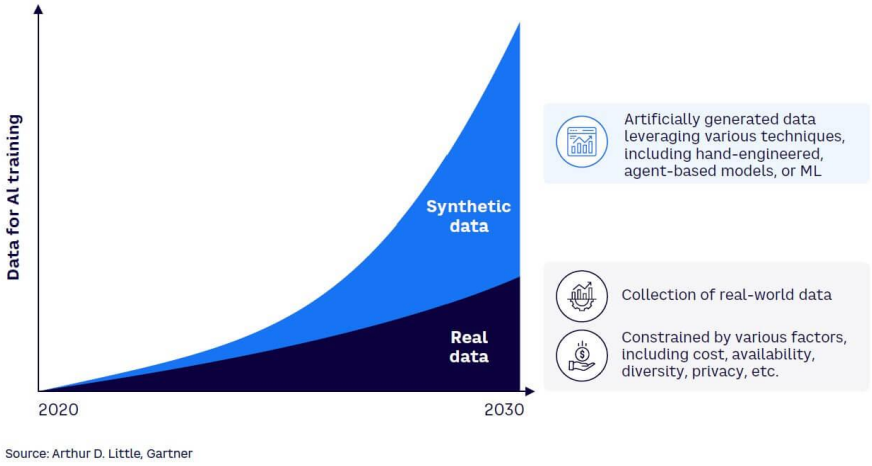

Figure 4 : Usage of Synthetic Data for AI training (Reference: Gardner)

5.3 Training Data as Data Marketplace

In this section, we will assess the effectiveness of blockchain-based data marketplaces for transparent and safe access to large language model training data. Research shows that blockchain solves fundamental problems like data integrity, access control, and fair payment for data providers Importantly, the immutability of blockchain ledgers allows for strict tracing of data provenance from source to deployment, contributing to accountability and regulation compliance .By maintaining this kind of distributed and transparent record-keeping, the risks of data manipulation and access control problems, common in classical centralized data management platforms, are heavily relied on. Furthermore, smart contracts automate the enforcement of access policies and usage terms. This means that data use does not come from intermediaries but must conform with the agreed-upon terms and conditions, building trust and simplifying the exchanges in the market. The novelty of

this recommendation is the fact that trust is exponentiated to the system that the process and there is little human intervention. The pseudo anonymity of Blockchain is also a feature that is extremely relied here considering that the immutable record transaction of permission of data ownership , the exposure of data and the outcome generated hash are all recorded which makes the system not party dependent traditional system but a system dependent In addition, the decentralized and public nature of these systems creates a fairer ecosystem as data providers and consumers face off directly, as well as in fairer compensation structures for the services they provide, such as through tokenization and NFTs (-, 2025). These systems, like those that have been experimented with in healthcare and automotive AI data markets, show higher joint-operative efficiency and data transparency in data exchange, a case for training an LLM (-, 2025). Typically, architectural designs include on-chain model registries that allow for global traceability of data with auditable and transparent phases from data collection through its integration into an LLM (-, 2025; Musamih et al., 2026). This immutable proof not only instills trust between data users and companies but also ensures an open mechanism for ownership and verifiable usage of data. Nevertheless, to get these tokenized data assets to properly have standardized valuation mechanisms is difficult, despite their transparent ownership. However, the potential for fractional ownership and tokenized curation by such systems provides new opportunities for improvement in data governance and equitable monetization of dataset used. Furthermore, blockchain technology can provide a platform for data providers by offering a share of the profits of LLMs, usually not a factor in traditional models and thus increasing motivation and participation by the sufficient game theory rational in which the entire technology is built on. To This decentralized model not only improves access to information, but also creates strong and wide data ecosystem for the development of LLM and the drawbacks associated with classical data sharing model

are also minimized (Gowri et al., 2025). Such a route also facilitate advanced schemes such as dynamic pricing and data licensing (smart contracts can also scale up compensation as the data change affects the level, performance or frequency of use) (-, 2025; Gowri et al., 2025). Privacy-preserving methods such as federated learning and secure enclaves ensure that sensitive data is never forgotten at the contributor's location, which not only helps alleviate important privacy issues but also lets to train a team AI model through parallel training (-, 2025; Gowri et al., 2025). Using blockchain in this way provides an appealing alternative to the disparate and opaque AI training data market, and alleviates concerns over creator payment, as well as market fragmentation (Oderinwale & Kazlauskas, 2025). The immutable provenance and auditable transaction logs that blockchain can provide through smart contracts provide a more robust framework to manage property and copyright in decentralized AI training processes (Gowri et al., 2025; Sai et al., 2024). However, there are challenges to achieve scalable and energy-efficient blockchain infrastructures that can cover the large data transaction volume needed for large-scale LLM training and also to develop solid oracle mechanisms that can link off-chain data quality with on-chain smart contract execution (-, 2025; Sai et al., 2024). In addition, federated learning and blockchain, promising for secure data storage by optimizing and encrypting data prior to being shared, could lead to latency, synchronization, and resource overhead problems as well as significant costs which need to be carefully optimized before being implemented for an actual dataset in training a large LLM (Gowri et al., 2025). Additionally, the computational complexity of the on-chain data storage and the intricacy of smart contract executions for extensive data licensing could generate substantial scalability issues, leading to hybrid approaches which use off-chain computation plus little on-chain metadata to achieve maximum performance (-, 2025). Thus, these hybrid architectures require complex trade-offs among decentralization, privacy, and performance

to sufficiently support a data settlement framework for LLMs. Addressing these challenges for widespread implementation will require innovative options that harmonize the advantages of blockchain's immutability and transparency with the practical complexities associated with large-scale AI development. A major challenge is the adoption of solid governance models for decentralized, autonomous organizations operating on these marketplaces to minimize manipulation and provide fairness of scrutiny (-, 2025). Establishing a normative legal platform that delineates the intellectual property rights of tokenized data is also a key step because NFT ownership does not automatically give rise to rights over the intellectual property of other tokenized data without adequate licensing (-, 2025). These are vital for defining usage rights, distribution of revenues and attribution so as to ensure that creators are fairly compensated and their contributions are acknowledged in a dynamic environment in AI model development (-, 2025; Wang et al., 2024). However, the dependence on external oracles to connect the on-chain integrity of blockchain with off-chain real-world data (especially in the case of fact-checking and toxicity assessment through LLMs) creates one central failure point in a trustless system (Geren et al., 2024). Such developments demonstrate the continued need for decentralized oracle networks incorporating strong consensus structures to maintain the integrity and reliability of any external data consumption, contributing to a safe, trustworthy decentralized chain of blockchain-based LLM data marketplaces. In addition, the data isolation intrinsic in existing blockchain systems and their relatively poor computational capabilities limit their use case and compatibility with large language models, requiring innovative solutions for seamless data interaction and the unlocking of the full potential of smart contract intelligence (Xian et al., 2024). These challenges necessitate developments in its scalability and interoperability—both of which are also key to secure, efficient off-chain computation bridge technologies and validations (Abed, 2023). Furthermore,

consideration should also be given to encouraging fairness, since token models can be subjected to speculative bias and computation of metrics such as Shapley values is resource-intensive (-, 2025). As a result, implementing lightweight contribution scoring mechanisms and effective token sinks are essential to stabilizing the market dynamics and ensuring a sustainable ecosystem for the data providers (-, 2025). However, the application of decentralized AI marketplaces remains in its infancy with current solutions only partially addressing the extensive range of scalability, security, and real decentralization issues (Li, 2023). Such systems exist in a developing stage; all of which indicates the fact that robust research with deep data governance models and frameworks like DAOs is needed to facilitate transparent and inclusive engagement in data-sharing measures and reduce the risk of top-down control (-, 2025). The adaptation of existing blockchain and AI projects to adapt to rapidly evolving generative AI technologies (e.g., state-of-the-art LLMs) creates massive challenges and poses a threat to the commercialization of decentralized marketplaces for private models without major modifications to their utility-centered applications (Witt et al., 2024). This can require a re-evaluation of existing incentive structures and privacy-preserving methods to meet the specific requirements of proprietary LLMs and its related training data (-, 2025). Thus, it is urgent for research moving forward to consider new architectural paradigms that can address the trade-offs between computational overhead, as well as data privacy obligations, and the cost-consuming resources used by LLMs, using decentralized architectures, for example new zero-knowledge proof implementations to better scaling and privacy (-, 2025; Geren et al., 2024). Moreover, it is important to design efficient reward models and lightweight consensus algorithms that enable the balance between incentives for the contributors, privacy and data authenticity in these decentralized ecosystems (Wang et al., 2024). Future research is particularly oriented towards the economic-theoretic guarantees associated with

these learning-based controllers, which involve architectures designed to make these controllers reasonably incentive-compatible and budget-efficient with each other across a decentralised learning system (Kaluannakkage & Buyya, 2025). Additionally, the practical deployment and user acceptance of such secure data-sharing marketplaces, especially for sensitive sectors like health care, is still a major area for future work and deployment (Pandl et al., 2020). Regulatory frameworks, governance structures, and socioeconomic impact on a range of stakeholders with equitable access and benefits are some of these domains that are explored to maximize the potential benefits and mitigate the risks (Baklaga, 2024). The open-ended architecture allows diversity of LLMs with a set of responses validating networked LLMs, thus providing better privacy assurance, usage restrictions, and dataset-contribution bias minimization (Nguyen et al., 2024). The federated learning, differential privacy, zero-knowledge proofs, secure enclaves, etc. applied to these techniques are highly decentralized and highly suited to secure sensitive information while also enabling training of collaborative models under distributed environments (-, 2025; Taherdoost et al., 2025; Wellington, 2024). These privacy-preserving methods do have their own conflicts, particularly with respect to the computational overhead of securing multi-party computation and providing valid differential privacy guarantees, especially for processing sensitive content (Khatiwada et al., 2025). As a result, it is necessary to deploy secure computation protocols with secure cryptographic primitives along with newly developed algorithms and well-optimized secure computation protocols to obtain practical deployments that strike a balance between strong privacy guarantees and the high-performance requirements of large-scale AI applications. Furthermore, decentralized systems solve problems related to scalability and reliability, but they also come with dangers of unauthorized access to sensitive data among distributed nodes that needs to be protected with strong access control mechanisms (Zhang et al., 2024). This requires

application of blockchain-based access control for immutable auditable data and fine-grained permissions to control data flows within distributed contexts (Almalawi et al., 2025). Indeed, the consensus verification checks are vital to guarantee stability of the outputs of those distributed nodes, as there are potential problems that lead to problem of model hallucinations and to have the overall reliability of the same output from the distributed nodes (Zhang et al., 2024). This underscores the call for sophisticated cryptographic approaches, secure multi-party computation methods, and hard standardization protocols that form the bedrock of federated AI models to protect data privacy while also ensuring model fidelity within a decentralized AI paradigm (Ogunbajo et al., 2025; Zhang et al., 2024). The intrinsic problems surrounding fairness of incentives, including costly Shapley-value computation and the speculative distortions afforded by token models, require improvement in incentives to support a longer-lasting market participation in decentralized platforms (-, 2025). This entails implementing lightweight contribution scoring systems and adequate token sinks to maintain the balance of the market and can also be aided by strong governance mechanisms such as Decentralized Autonomous Organizations, which aims to improve transparency and inclusivity. As vast data is shared, currently blockchain systems have limited scalability, which adds a layer of complexity to implementing privacy-preserving methods and powerful incentives for large-scale AI training (Gowri et al., 2025; Singh et al., 2025). Thus, computational hybridizations that combine off-chain processing with limited on-chain metadata may provide a useful route to improving efficiency and scalability that does not jeopardize the transparency and integrity associated with blockchain.

Diagrammatic Representation of Suggested Framework for Training data as a Marketplace

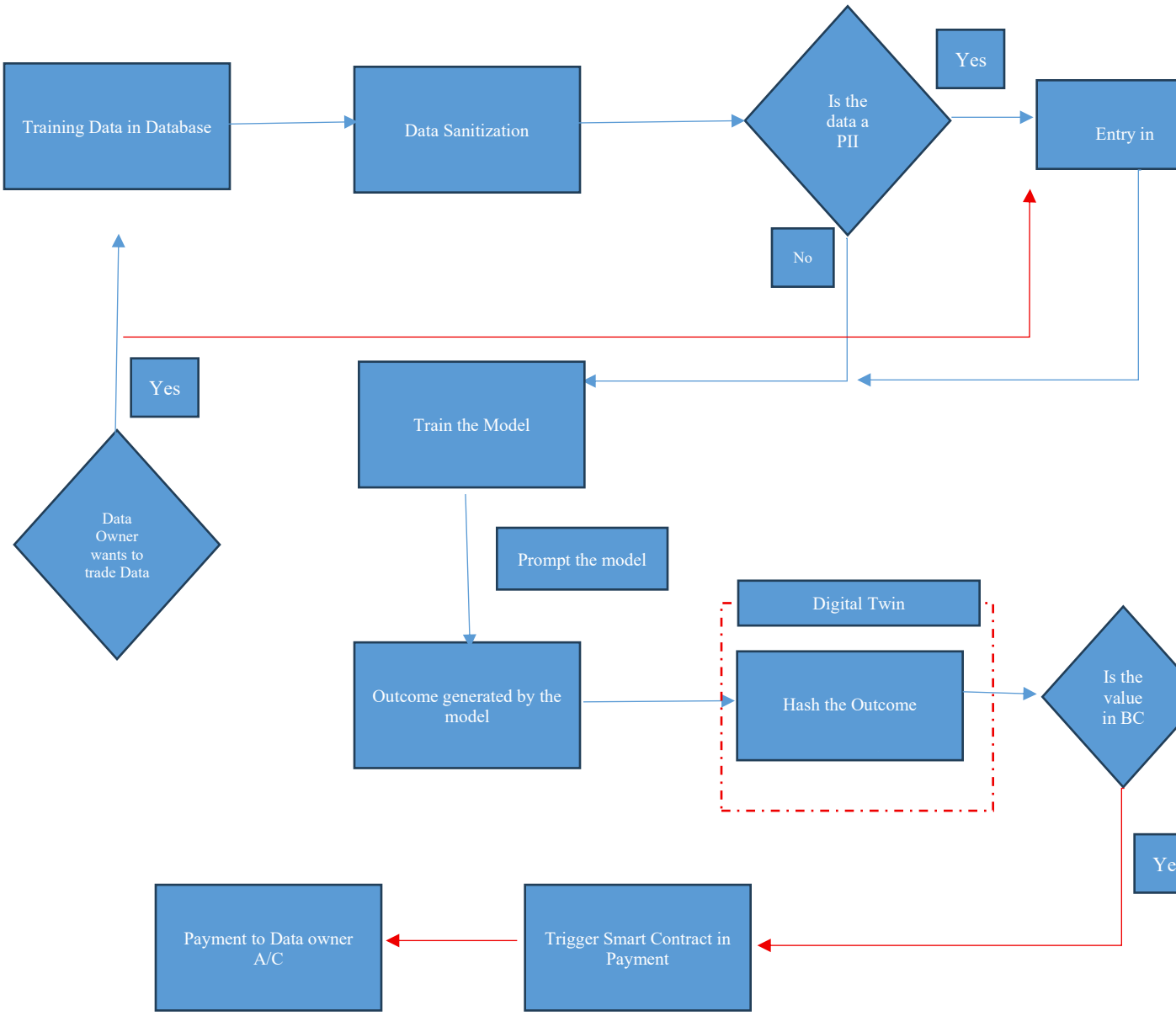

Figure 5 : Training Data Monetization Using SmartContract

CHAPTER VI:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

6.1 Summary

Secure data handling in AI has become an urgent priority, particularly as industry shifts toward increasingly data-intensive large language models and autonomous AI agents. While advancements such as synthetic data generation are intended to protect privacy, they introduce unique vulnerabilities, including the risk of models memorizing and inadvertently leaking sensitive training information or personally identifiable information

In the final chapter, we will discuss a few of the key findings of our study, review the previous key research questions, and discuss the implications of our framework for the responsible development and deployment of AI-driven synthetic data generation technologies. It will also show limitations of the current research study and suggest future research avenues, e.g., the longer-term effects of synthetic data on the performance of models and ethics regarding the proliferation of synthetic data. More specifically, the emphasis for future development should be to construct more advanced auditing frameworks that are capable dynamically evaluating the trustworthiness of synthetic data on multiple dimensions such as fidelity, robustness, and fairness through metrics with measurable uncertainty. This also demands enhanced empirical validation of formal privacy guarantees and realistic standards for specific domains. Moreover, we need robust validation frameworks to evaluate the uniform quality of synthetic data which maintains privacy and usefulness while maintaining the reliability of the downstream AI models. A recent work has been done by South et al. (2024), which aims to give an integrated

approach to the AI landscape, detailing the pitfalls that come along the way for synthetic data generation, and the need to implement security mechanisms and privacy measures for the AI systems with an understanding of this ecosystem. Subsequent elaborations of such models would use advanced machine learning methods and specific methods for deep generative models, such as deep generative models to increase replicability as well as the usability of the synthetic datasets, and to maintain the privacy for the models very much. This includes using approaches such as differentially private non-parametric copulas to create synthetic data with significant privacy protection, with no evidence that individual user data can be re-identified even under advanced attacks (Osorio-Marulanda et al., 2025). Such methods are fundamental for generating synthetic datasets that mitigate bias, imbalance, privacy concerns and ensure high fidelity to raw data to train AI models effectively (Belgodere et al., 2023, 2024). This will trigger a continuous evolution of refinement and validation to enhance and fine tune the synthetic data generation models as per their output performance in downstream tasks and privacy defense (Belgodere et al., 2024). Adoption of more advanced synthetic data generation methods of this nature is essential for enabling responsible and ethical application of AI in other fields, notably, in health-related domains where data sensitivity is particularly vital. This calls for a common and integrated evaluation approach to determine the integrity, utility and confidentiality of synthetic tabular data produced by different models for health-related aspects (Hernandez et al., 2025). Understood that medical datasets are more complex and diverse, this framework must align the medical data created with real-world conditions to guarantee that the recorded data does not compromise confidentiality of users and meets regulations

6.3 Recommendations for Future Research

This involves taking the digital twin capabilities to multi-modal AI systems and integrating advanced cryptology strategies for strong privacy protection as well. This ongoing study will also explore how to develop explainable AI embedded within the digital twin that gives a transparent view into complex model decisions and behavior when it comes to security and data privacy issues (Morabito et al., 2024). On the other side, we will look at developing quantum-resistant cryptographic protocols in the digital twin to protect AI models from any future quantum challenges (Guilherme Junior & Clinton, n.d.). We will achieve the future compatibility in keeping with the advancements in quantum computing by using lattice-based cryptographic algorithms, such as NTRU and Kyber, which can provide a secured pipeline for the long term (Guilherme Junior & Clinton, n.d.). Finally, the digital twin will adapt tools such as OpenQuantumSafe to help in finding and implementing these post-quantum encryption schemes, ensuring that the deployed AI models are safe against future cryptographic advances (Guilherme Junior & Clinton, n.d.). As this research focuses on how attackers will adapt when they understand that an AI-based IDS is protecting the building, red-team evaluation is a key approach to understand potential adversarial adaptations and proactive countermeasures (Siam et al., 2025). Furthermore, developing self-healing approaches in the digital twin will potentially improve the resilience of the digital twin to very sophisticated adversaries and also promote flexible attack plans and self-healing procedures (Huang et al., 2025). As AI approaches continue to evolve, this will require the security response of developing security solutions and tools that can identify and mitigate vulnerability in real time and react accordingly (Schmitt & Koutroumpis, 2025).

6.4 Conclusion

It is important to practice this kind of proactive response to guard critical infrastructure when the advent of AI-based cybersecurity is a topic that generates a lot of research interest but still encounters significant deficiencies to cope with changing threats (More specifically for future work, we will incorporate blockchain technology in the digital twin and create an immutable and transparent ledger to keep track of the changes and access to the model for data provenance and accountability. This will add to the trustworthiness and auditability of the lifecycle of the AI model, meeting some of the very pressing ethical and regulatory compliance challenges surrounding AI deployments. Additionally, novel cryptographic techniques like zero-knowledge proofs and federated learning will be proposed to improve data security and to enable safe collaboration of AI training for digital twin setup. This is important since such advancements would allow the digital twin to not only reveal vulnerability but also suggest and validate in real-time reliable backup-with quantum-resilient strategies that maintain powerful protection as the IoT technologies continue to evolve. This means looking into the use of advanced cybersecurity methodology to see whether certain major challenges such as real-time anomaly identification and adversary resilience could be resolved, which is important when a resource-limited and dispersed environment is in view. The proposed digital twin framework will therefore focus on creating lightweight, flexible and ethical Intrusion Detection System frameworks which work well in more dynamic and diverse environments. The proposed Framework solves current gaps in research on algorithmic flaws, inadequate validation requirements, and inadequate threat modeling for current alternatives. This will include designing new techniques to integrate privacy-preserving methods into the Digital twin, including homomorphic encryption and secure multi-party

computation, allowing for vulnerability analysis and data preservation without sharing sensitive training data. In addition, the capability of real-time simulation and monitoring of the IoT network behaviors of the digital twin should be increased, which would allow us to predict the security threats with high accuracy to avoid the breach in advance of the breach compared to the scope of this study which include identification, remediation and fine tuning. This can involve concentrating on the coupling with SplitFed Learning and 6G communication protocols to secure privacy of data and improve resource usage especially for distributed IoT ecosystems which are Edge computing systems. This distributed approach allows for enhanced privacy by removing the requirement for raw data to leave local devices as well as model adaptability by utilizing diverse datasets. Moreover, implementing verifiable differential privacy guarantees would greatly strengthen the trustworthiness of the digital twin, as it quantifies privacy leakage risk more formally and creates strict budgets for privacy for all data processed in the simulated environment.

REFERENCES

B. N. (2025). Blockchain-based Control Over Data to Train AI Models. *Journal of Advances in Developmental Research*, *16*(2). https://doi.org/10.71097/ijaidr.v16.i2.1543

-, S. P. (2023). The Future of Large Language Models: A Futuristic Dissection on AI and Human Interaction. *International Journal For Multidisciplinary Research*, *5*(3). https://doi.org/10.36948/ijfmr.2023.v05i03.3135

Abdali, S., Anarfi, R., Barberan, C., & He, J. (2024). Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.12503

Abdul, S. S. M. (2024). Navigating Blockchain's Twin Challenges: Scalability and Regulatory Compliance. *Blockchains*, *2*(3), 265. https://doi.org/10.3390/blockchains2030013

Abrahams, T. O., Ewuga, S. K., Dawodu, S. O., Adegbite, A. O., & Hassan, A. O. (2024). A review of cybersecurity strategies in modern organizations: examining the evolution and effectiveness of cybersecurity measures for data protection [review of a review of cybersecurity strategies in modern organizations: examining the evolution and effectiveness of cybersecurity measures for data protection]. Computer Science & IT Research Journal, 5(1), 1. Fair East Publishers. https://doi.org/10.51594/csitrj.v5i1.699

Adeghe, E. P., Okolo, C. A., & Ojeyinka, O. T. (2024). Evaluating the impact of blockchain technology in healthcare data management: A review of security, privacy, and patient outcomes [Review of *Evaluating the impact of blockchain technology in healthcare data management: A review of security, privacy, and patient outcomes*]. *Open Access Research Journal of Science and Technology*, *10*(2), 13. https://doi.org/10.53022/oarjst.2024.10.2.0044

Adelani, D. I., Davody, A., Kleinbauer, T., & Klakow, D. (2020). Privacy Guarantees for De-Identifying Text Transformations. *Interspeech 2022*, 4666. https://doi.org/10.21437/interspeech.2020-2208

Adhyapak, S. (2024). Data Privacy and Security Risks in AI-Based Code Understanding. *International Journal for Research in Applied Science and Engineering Technology*, *12*(6), 1913. https://doi.org/10.22214/ijraset.2024.63423

Aditya, H., Chawla, S., Dhingra, G., Rai, P., Sood, S., Singh, T., Wase, Z. M., Bahga, A., & Madisetti, V. K. (2024). Evaluating Privacy Leakage and Memorization Attacks on Large Language Models (LLMs) in Generative AI Applications. *Journal of Software Engineering and Applications*, *17*(5), 421. https://doi.org/10.4236/jsea.2024.175023

Aghakasiri, K., Zambare, N., Thai, J., Ye, C., Mehta, M., Mitchell, J. R., & Abdalla, M. (2025). *Not What the Doctor Ordered: Surveying LLM-based De-identification and Quantifying Clinical Information Loss*. https://doi.org/10.48550/ARXIV.2509.14464

Aguilera-Martínez, F., & Berzal, F. (2025). LLM Security: Vulnerabilities, Attacks, Defenses, and Countermeasures. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.01177

Aguilera-Martínez, F., & Berzal, F. (2025b). Differential Privacy in Machine Learning: From Symbolic AI to LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.11687

Airehenbuwa, B., Hasan, T., Sarkar, S., & Guin, U. (2025). *Advancing Security with Digital Twins: A Comprehensive Survey*. https://doi.org/10.48550/ARXIV.2505.17310

Akbarfam, A. J., & Maleki, H. (2024). SOK: Blockchain for Provenance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.17699

Akhtar, S. I., Rauf, A., Abbas, H., Amjad, M. F., & Batool, I. (2024). Compliance and feedback-based model to measure cloud trustworthiness for hosting digital twins. *Journal of Cloud Computing Advances Systems and Applications*, *13*(1). https://doi.org/10.1186/s13677-024-00690-0

Akkus, A., Li, M., Chu, J., Backes, M., Zhang, Y., & Sav, S. (2024). Generated Data with Fake Privacy: Hidden Dangers of Fine-tuning Large Language Models on Generated Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.11423

Akther, A., Arobee, A., Adnan, A. A., Auyon, O., Islam, A. J., & Akter, F. (2025a). *Blockchain As a Platform For Artificial Intelligence (AI) Transparency*. https://doi.org/10.48550/ARXIV.2503.08699

Alabdulakreem, A., Arnold, C. M., Lee, Y., Feenstra, P. M., Katz, B., & Barbu, A. (2024). SecureLLM: Using Compositionality to Build Provably Secure Language Models for Private, Sensitive, and Secret Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.09805

Aladiyan, A. (2025). Digital Safeguards: Unravelling the Complex Interplay Between Emerging Threats and Proactive Cyber Defence Strategies. *Journal of Internet Services and Information Security*, *15*(1), 348. https://doi.org/10.58346/jisis.2025.i1.022

Alkhodair, A. J., Mohanty, S. P., & Kougianos, E. (2023). Consensus Algorithms of Distributed Ledger Technology -- A Comprehensive Analysis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.13498

Al-Matari, N. Y., Zahary, A. T., & Al-Shargabi, A. A. (2024). A survey on advancements in blockchain-enabled spectrum access security for 6G cognitive radio IoT networks. *Scientific Reports*, *14*(1), 30990. https://doi.org/10.1038/s41598-024-82126-y

Alqahtany, S. S., & Syed, T. A. (2024). ForensicTransMonitor: A Comprehensive Blockchain Approach to Reinvent Digital Forensics and Evidence Management. *Information*, *15*(2), 109. https://doi.org/10.3390/info15020109

Amar, S. (2022). Blockchain Mutability: Challenges and Proposed Solutions. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *6*(10). https://doi.org/10.55041/ijsrem16668

Arasteh, S. T., Lotfinia, M., Pérez-Toro, P. A., Arias-Vergara, T., Orozco-Arroyave, J. R., Schuster, M., Maier, A., & Yang, S. H. (2024). Differential privacy for protecting patient data in speech disorder detection using deep learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.19078

Arasteh, S. T., Ziller, A., Kühl, C., Makowski, M. R., Nebelung, S., Braren, R., Rueckert, D., Truhn, D., & Kaissis, G. (2024). Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging. *Communications Medicine*, *4*(1). https://doi.org/10.1038/s43856-024-00462-6

Asthana, S., Mahindru, R., Zhang, B., & Sanz-Sánchez, J. (2025). Adaptive PII Mitigation Framework for Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.12465

Asthana, S., Zhang, B., Mahindru, R., DeLuca, C., Gentile, A. L., & Gopisetty, S. (2025). Deploying Privacy Guardrails for LLMs: A Comparative Analysis of Real-World Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.12456

Atadoga, A., Elufioye, O. A., Omaghomi, T. T., Akomolafe, O., Odilibe, I. P., & Owolabi, O. R. (2024). Blockchain in healthcare: A comprehensive review of applications and security concerns [Review of *Blockchain in healthcare: A comprehensive review of applications and security concerns*]. *International Journal of Science and Research Archive*, *11*(1), 1605. https://doi.org/10.30574/ijsra.2024.11.1.0244

Azize, A., & Basu, D. (2024). How Much Does Each Datapoint Leak Your Privacy? Quantifying the Per-datum Membership Leakage. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.10065

Badidi, I., Khiyaoui, N. E., Riany, A., Elallid, B. B., & Abouaomar, A. (2025). Privacy-Preserving Offloading for Large Language Models in 6G Vehicular Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.05320

Bahar, A. A. M., & Wazan, A. S. (2024). On the Validity of Traditional Vulnerability Scoring Systems for Adversarial Attacks against LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.20087

Bakhrudin, B., Margolang, F. I., Sudarmanto, E., & Sugiono, S. (2023). Islamic Perspectives on Cybersecurity and Data Privacy: Legal and Ethical Implications. *West Science Law and Human Rights*, *1*(4), 166. https://doi.org/10.58812/wslhr.v1i04.323

Balan, K., Gilbert, A., & Collomosse, J. (2025). Content ARCs: Decentralized Content Rights in the Age of Generative AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2503.14519

Balan, K., Gilbert, A., Black, A., Jenni, S., Parsons, A., & Collomosse, J. (2023). *DECORAIT - DECentralized Opt-in/out Registry for AI Training*. 1. https://doi.org/10.1145/3626495.3626506

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., & Sato, T. (2019). Hypothesis Testing Interpretations and Renyi Differential Privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1905.09982

Baluta, T., Lamblin, P., Tarlow, D., Pedregosa, F., & Dziugaite, G. K. (2024). Unlearning in- vs. out-of-distribution data in LLMs under gradient-based method. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.04388

BARBERÁ, I. (2025). *AI Privacy Risks & Mitigations – Large Language Models (LLMs)*.

Barletta, V. S., Bavaro, V., Calvano, M., Curci, A., Piccinno, A., & Posa, D. P. (2025). Enabling Cyber Security Education through Digital Twins and Generative AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.17518

Bassey, J., Li, X., & Qian, L. (2021). Device Authentication Codes based on RF Fingerprinting using Deep Learning. *ICST Transactions on Security and Safety*, *8*(29), 172305. https://doi.org/10.4108/eai.30-11-2021.172305

Behnia, R., Ebrahimi, M., Pacheco, J., & Padmanabhan, B. (2022, November 1). EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw58026.2022.00078

Belgodere, B., Dognin, P., Ivankay, A., Melnyk, I., Mroueh, Y., Mojsilovic, A., Navartil, J., Nitsure, A., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Vedpathak, R., &

Young, R. A. (2023). Auditing and Generating Synthetic Data with Controllable Trust Trade-offs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.10819

Beltran, S. R., Tobaben, M., Loppi, N., & Honkela, A. (2024). Towards Efficient and Scalable Training of Differentially Private Deep Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.17298

Bhatt, S., Rajore, T., Aggarwal, K., Ananthanarayanan, G., Chandra, R., Chandran, N., Choudhury, S., Gupta, D., Kıcıman, E., Pandey, S. K., Setty, S., Sharma, R., & Zhao, T. C. (2025). Enterprise AI Must Enforce Participant-Aware Access Control. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.14608

Blanco-Justicia, A., Jebreel, N. M., Manzanares, B., Sánchez, D., Domingo-Ferrer, J., Collell, G., & Tan, K. E. (2024). Digital Forgetting in Large Language Models: A Survey of Unlearning Methods. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.02062

Bombari, S., & Mondelli, M. (2024). Privacy for Free in the Over-Parameterized Regime. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.14787

Bonthu, C., & Goel, G. (2025). Autonomous Supplier Evaluation and Data Stewardship with AI: Building Transparent and Resilient Supply Chains. *International Journal of Computational and Experimental Science and Engineering*, *11*(3). https://doi.org/10.22399/ijcesen.3854

Borec, L., Sadler, P., & Schlangen, D. (2024). The Unreasonable Ineffectiveness of Nucleus Sampling on Mitigating Text Memorization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.16345

Borkar, J., Jagielski, M., Lee, K., Mireshghallah, N., Smith, D. A., & Choquette-Choo, C. A. (2025). Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training. *Findings of the Association for Computational Linguistics: ACL 2022*, 18703. https://doi.org/10.18653/v1/2025.findings-acl.959

Boukessessa, M., Ghazli, A., & Ali-Pacha, A. (2024). Blockchain-Based Security Framework for VNF Package Protection in 5G Network Slicing Services. *Transport and Telecommunication Journal*, *25*(3), 289. https://doi.org/10.2478/ttj-2024-0021

Bu, Z., Liu, R., Wang, Y., Zha, S., & Karypis, G. (2023). On the accuracy and efficiency of group-wise clipping in differentially private optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.19215

Cai, Y., Li, L., & Zhang, L. (2025). A Statistical Hypothesis Testing Framework for Data Misappropriation Detection in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.02441

Cai, Y., Li, L., & Zhang, L. (2025). A Statistical Hypothesis Testing Framework for Data Misappropriation Detection in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.02441

Camacho, N. (2024). The Role of AI in Cybersecurity: Addressing Threats in the Digital Age. *Deleted Journal*, *3*(1), 143. https://doi.org/10.60087/jaigs.v3i1.75

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2022). The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1802.08232

Chaffer, T. J., Goldston, J., Okusanya, B., & I, G. D. A. T. A. (2024). On the ETHOS of AI Agents: An Ethical Technology and Holistic Oversight System. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.17114

Chang, H., Shamsabadi, A. S., Katevas, K., Haddadi, H., & Shokri, R. (2024). Context-Aware Membership Inference Attacks against Pre-trained Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.13745

Chen, B., Han, N., & Miyao, Y. (2024). A Statistical and Multi-Perspective Revisiting of the Membership Inference Attack in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.13475

Chen, K., Tang, H., Liu, Q., & Xu, Y. (2025). Improved Algorithms for Differentially Private Language Model Alignment. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.08849

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University - Computer and Information Sciences*, *37*(7). https://doi.org/10.1007/s44443-025-00177-1

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025a). A Survey on Privacy Risks and Protection in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.01976

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., & Wu, P. (2025b). A survey on privacy risks and protection in large language models. *Journal of King Saud University - Computer and Information Sciences*, *37*(7). https://doi.org/10.1007/s44443-025-00177-1

Chen, Y., Li, T., Liu, H., & Yu, Y.-D. (2023). Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.03057

Chen, Y., Mendes, E., Das, S., Xu, W., & Ritter, A. (2023). Can Language Models be Instructed to Protect Personal Information? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.02224

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Huang, Y., Jagielski, M., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., … Zhang, W. (2024). Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.d3197524

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Jagielski, M., Huang, Y., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y., … Zhang, W. (2023). Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.06929

Dadsetan, A., Soleymani, D., Zeng, X., & Rudzicz, F. (2024). Can large language models be privacy preserving and fair medical coders? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.05533

Das, B. C., Amini, M. H., & Wu, Y. (2024). Security and Privacy Challenges of Large Language Models: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.00888

Das, S., Kolling, C., Khan, M. A., Amani, M., Ghosh, B., Wu, Q., Speicher, T., & Gummadi, K. P. (2025). Revisiting Privacy, Utility, and Efficiency Trade-offs when Fine-Tuning Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.13313

Deng, C., Duan, Y., Jin, X., Chang, H., Tian, Y., Liu, H., Wang, Y., Gao, kuofeng, Zou, H. P., Jin, Y., Xiao, Y., Wu, S., Xie, Z., Lyu, W., He, S., Cheng, L., Wang, H., & Zhuang, J. (2025). Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas: a survey. *AI and Ethics*, *5*(5), 4745. https://doi.org/10.1007/s43681-025-00797-3

Deng, C., Duan, Y., Jin, X., Chang, H., Tian, Y., Liu, H., Zou, H. P., Jin, Y., Xiao, Y., Wang, Y., Wu, S., Xie, Z., Gao, K., He, S., Zhuang, J., Cheng, L., & Wang, H. (2024). Deconstructing The Ethics of Large Language Models from Long-standing Issues to New-emerging Dilemmas. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.05392

Deng, R. H., Lu, Z., & Duan, Q. (2025). Quantifying Privacy Leakage in Split Inference via Fisher-Approximated Shannon Information Analysis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2504.10016

Ding, Y., Wu, X., Meng, Y., Luo, Y., Wang, H., & Pan, W. (2024). Delving into Differentially Private Transformer. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.18194

Ding, Y., Wu, X., Wang, H., & Pan, W. (2023). DPFormer: Learning Differentially Private Transformer on Long-Tailed Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.17633

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., & Huang, X. (2024). Safeguarding Large Language Models: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.02622

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., & Huang, X. (2025). Safeguarding large language models: a survey. *Artificial Intelligence Review*, *58*(12). https://doi.org/10.1007/s10462-025-11389-2

Du, H., Liu, S., Zheng, L., Cao, Y., Nakamura, A., & Chen, L. (2024). Privacy in Fine-tuning Large Language Models: Attacks, Defenses, and Future Directions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.16504

Duan, S., Khona, M., Iyer, A., Schaeffer, R., & Fiete, I. (2024a). Uncovering Latent Memories: Assessing Data Leakage and Memorization Patterns in Frontier AI Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.14549

Duan, S., Khona, M., Iyer, A., Schaeffer, R., & Fiete, I. R. (2024b). *Uncovering Latent Memories: Assessing Data Leakage and Memorization Patterns in Frontier AI Models*. https://doi.org/10.48550/ARXIV.2406.14549

Duarte, A. V., Zhao, X., Oliveira, A. L., & Li, L. (2024). DE-COP: Detecting Copyrighted Content in Language Models Training Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.09910

Dubinski, J. M., Kowalczuk, A., Boenisch, F., & Dziedzic, A. (2024). CDI: Copyrighted Data Identification in Diffusion Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.12858

El-Hajj, M. (2024). Leveraging Digital Twins and Intrusion Detection Systems for Enhanced Security in IoT-Based Smart City Infrastructures. *Electronics*, *13*(19), 3941. https://doi.org/10.3390/electronics13193941

Elmahdy, A., Inan, H. A., & Sim, R. B. (2022). *Privacy Leakage in Text Classification A Data Extraction Approach*. 13. https://doi.org/10.18653/v1/2022.privatenlp-1.3

Enyiorji, P. (2023). Blockchain-enforced data lineage architectures with formal verification workflows enabling auditable AI decision chains across regulated fintech

compliance regimes and supervisory reporting. *International Journal of Science and Research Archive*, *9*(2), 1201. https://doi.org/10.30574/ijsra.2023.9.2.0559

Fan, T., Gu, H., Cao, X., Chan, C. S., Chen, Q., Chen, Y., Feng, Y., Gu, Y., Geng, J., Luo, B., Liu, S., Ong, W., Ren, C., Shao, J., Sun, C., Tang, X., Tae, H. X., Tong, Y., Wei, S., … Yang, Q. (2025). Ten Challenging Problems in Federated Foundation Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.12176

Fazelnia, M., Moshtari, S., & Mirakhorli, M. (2024). Establishing Minimum Elements for Effective Vulnerability Management in AI Software. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.11317

Feng, P., Bi, Z., Yan, L., Wen, Y., Peng, B., Liu, J., Yin, C. H., Wang, T., Chen, K., Zhang, S., Li, M., Xu, J., Liu, M., Pan, X., Wang, J., & Niu, Q. (2024). Mastering AI: Big Data, Deep Learning, and the Evolution of Large Language Models -- Blockchain and Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.10110

Feretzakis, G., & Verykios, V. S. (2024). Trustworthy AI: Securing Sensitive Data in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.18222

Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A., & Verykios, V. S. (2024). Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review [Review of *Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review*]. *Information*, *15*(11), 697. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/info15110697

Fernandez, I., Neupane, S., Chakraborty, T., Mitra, S., Mittal, S., Pillai, N., Chen, J., & Rahimi, S. (2024a). A Survey on Privacy Attacks Against Digital Twin Systems in AI-Robotics. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.18812

Fernandez, I., Neupane, S., Chakraborty, T., Mitra, S., Mittal, S., Pillai, N., Chen, J., & Rahimi, S. (2024b). *A Survey on Privacy Attacks Against Digital Twin Systems in AI-Robotics*. 70. https://doi.org/10.1109/cic62241.2024.00019

Fernandez, P. (2025). Watermarking across modalities for content tracing and generative AI. *arXiv (Cornell University)*. http://arxiv.org/abs/2502.05215

Frikha, A., Walha, N., Nakka, K. K., Mendes, R., Jiang, X., & Zhou, X. (2024). IncogniText: Privacy-enhancing Conditional Text Anonymization via LLM-based Private Attribute Randomization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.02956

Futane, Mrs. S. S. (2025). Adaptive Multi-Model Cybercrime Identification, Prediction using Machine Learning, and Explainable AI. *International Journal for Research in*

Applied Science and Engineering Technology, *13*(8), 2113. https://doi.org/10.22214/ijraset.2025.73926

Ganesh, A., McMahan, B., Nasr, M., Steinke, T., & Thakurta, A. (2025). Hush! Protecting Secrets During Model Training: An Indistinguishability Approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.00201

Gangwal, A., & Sharma, A. A. (2025). Merge Now, Regret Later: The Hidden Cost of Model Merging is Adversarial Transferability. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.23689

George, E. P.-E., Idemudia, C., & Ige, A. B. (2024). Blockchain technology in financial services: enhancing security, transparency, and efficiency in transactions and services. *Open Access Research Journal of Multidisciplinary Studies*, *8*(1), 26. https://doi.org/10.53022/oarjms.2024.8.1.0042

Geren, C., Board, A., Dagher, G. G., Andersen, T., & Zhuang, J. (2024). Blockchain for Large Language Model Security and Safety: A Holistic Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.20181

Geren, C., Board, A., Dagher, G. G., Andersen, T., & Zhuang, J. (2025). Blockchain for Large Language Model Security and Safety: A Holistic Survey. *ACM SIGKDD Explorations Newsletter*, *26*(2), 1. https://doi.org/10.1145/3715073.3715075

German, E., Antebi, S., Habler, E., Shabtai, A., & Elovici, Y. (2025). LexiMark: Robust Watermarking via Lexical Substitutions to Enhance Membership Verification of an LLM's Textual Training Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.14474

Ghorbian, M., & Ghobaei-Arani, M. (2025). Key Concepts and Principles of Blockchain Technology. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.11707

Gloaguen, T., Jovanović, N., Staab, R., & Vechev, M. (2024). Discovering Clues of Spoofed LM Watermarks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.02693

Gong, J., Yan, P., Zhang, Y., An, H., & Liu, L. (2024). Blockchain Meets LLMs: A Living Survey on Bidirectional Integration. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.16809

Gowri, S., Mariyabegam., M., Sandhiya., P., Shainas., S., Subhanu., S., & Charumathi., C. (2025). *Secure Data Sharing for AI Training Using Blockchain Ledger*. 1. https://doi.org/10.1109/icdsns65743.2025.11168613

Gu, J. (2024). Responsible Generative AI: What to Generate and What Not. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.05783

Guilherme Junior, E., & Clinton, S. (n.d.). *OWASP Top 10 for LLMs - LLM and Gen AI Data Security Best Practices Guide 1.0*.

Hafner, F. M., & Sun, C. Q. (2024). Empirical Privacy Evaluations of Generative and Predictive Machine Learning Models -- A review and challenges for practice [Review of *Empirical Privacy Evaluations of Generative and Predictive Machine Learning Models -- A review and challenges for practice*]. *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2411.12451

Happe, A., & Cito, J. (2025). Can LLMs Hack Enterprise Networks? --- RCR Package. *arXiv (Cornell University)*. https://doi.org/10.5281/zenodo.17456062

Hausenloy, J., McClements, D., & Thakur, M. (2024). Towards Data Governance of Frontier AI Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.03824

Hayagreevan, H., & Khamaru, S. (2024). Security of and by Generative AI platforms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.13899

Hayes, J., Swanberg, M., Chaudhari, H., Yona, I., & Shumailov, I. (2024). Measuring memorization through probabilistic discoverable extraction. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.19482

He, J., Li, X., Da, Y., Zhang, H., Kulkarni, J., Lee, Y. T., Bačkurs, A., Yu, N., & Bian, J. (2022). Exploring the Limits of Differentially Private Deep Learning with Group-wise Clipping. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.01539

Heston, T. F. (2024). A Blockchain AI Solution to Climate Change. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4736849

Homaei, M., Mogollón-Gutiérrez, Ó., Núñez, J. C. S., Ávila, M., & Caro, A. (2024). A review of digital twins and their application in cybersecurity based on artificial intelligence [Review of *A review of digital twins and their application in cybersecurity based on artificial intelligence*]. *Artificial Intelligence Review*, *57*(8). Springer Science+Business Media. https://doi.org/10.1007/s10462-024-10805-3

Homaei, M., Mogollón-Gutiérrez, Ó., Núñez, J. C. S., Vegas, M. Á., & Caro, A. (2023). A Review of Digital Twins and their Application in Cybersecurity based on Artificial Intelligence [Review of *A Review of Digital Twins and their Application in Cybersecurity based on Artificial Intelligence*]. *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2311.01154

Howard, A. (2025). InjectLab: A Tactical Framework for Adversarial Threat Modeling Against Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.18156

Hu, A., Lü, Z., Xie, R., & Xue, M. (2023a). VeriDIP: Verifying Ownership of Deep Neural Networks Through Privacy Leakage Fingerprints. *IEEE Transactions on Dependable and Secure Computing*, *21*(4), 2568. https://doi.org/10.1109/tdsc.2023.3313577

Hu, A., Lü, Z., Xie, R., & Xue, M. (2023b). VeriDIP: Verifying Ownership of Deep Neural Networks through Privacy Leakage Fingerprints. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.10656

Hu, Y., Li, Z., Liu, Z., Zhang, Y., Qin, Z., Ren, K., & Chen, C. (2025). Membership Inference Attacks Against Vision-Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.18624

Huang, J., Zhang, J., Wang, Q., Han, W., & Zhang, Y. (2024). Exploring Advanced Methodologies in Security Evaluation for LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.17970

Huang, T., Hu, S., İlhan, F., Tekin, S. F., & Liu, L. (2024). Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.18169

Huang, Y., Gao, C., Wu, S., Wang, H., Wang, X., Zhou, Y., Wang, Y., Ye, J., Shi, J., Zhang, Q., Li, Y., Bao, H., Liu, Z., Guan, T., Chen, D., Chen, R., Guo, K., Zou, A., Kuen-Yew, B. H., … Zhang, X. (2025). On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.14296

Huang, Y., Li, Y., Wu, W., Zhang, J., & Lyu, M. R. (2023). Your Code Secret Belongs to Me: Neural Code Completion Tools Can Memorize Hard-Coded Credentials. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.07639

Huang, Z., Gong, N. Z., & Reiter, M. K. (2024). A General Framework for Data-Use Auditing of ML Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.15100

Ishihara, S. (2023). *Training Data Extraction From Pre-trained Language Models: A Survey*. https://doi.org/10.18653/v1/2023.trustnlp-1.23

Jandali, Y., Zhang, R., Sheybani, N., & Koushanfar, F. (2025). Optimizing Privacy-Preserving Primitives to Support LLM-Scale Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.25072

Jegorova, M., Kaul, C., Mayor, C., O'Neil, A. Q., Weir, A., Murray-Smith, R., & Tsaftaris, S. A. (2022). Survey: Leakage and Privacy at Inference Time. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 1. https://doi.org/10.1109/tpami.2022.3229593

Ji, W., Yuan, W., Getzen, E., Cho, K., Jordan, M. I., Mei, S., Weston, J., Su, W., Xu, J., & Zhang, L. (2025). An Overview of Large Language Models for Statisticians. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.17814

Jiang, Z., Jin, Z., & He, G. (2024). Safeguarding System Prompts for LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.13426

Jimenez-Gutierrez, D. M., Falkouskaya, Y., Hernández-Ramos, J. L., Anagnostopoulos, A., Chatzigiannakis, I., & Vitaletti, A. (2025). On the Security and Privacy of Federated Learning: A Survey with Attacks, Defenses, Frameworks, Applications, and Future Directions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.13730

Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating Training Data Mitigates Privacy Risks in Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2202.06539

Kaneko, M., & Baldwin, T. (2024). A Little Leak Will Sink a Great Ship: Survey of Transparency for Large Language Models from Start to Finish. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.16139

Kazmi, M., Lautraite, H., Akbari, A., Soroco, M., Tang, Q., Wang, T., Gambs, S., & Lécuyer, M. (2024). PANORAMIA: Privacy Auditing of Machine Learning Models without Retraining. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.09477

Kereopa-Yorke, B. (2023). Building resilient SMEs: Harnessing large language models for cyber security in Australia. *Journal of AI, Robotics & Workplace Automation.*, *3*(1), 15. https://doi.org/10.69554/xsqz3232

Khadangi, A., Sartipi, A., Tchappi, I., Bahmani, R., & Fridgen, G. (2025). *Efficient Differentially Private Fine-Tuning of LLMs via Reinforcement Learning*. https://doi.org/10.48550/ARXIV.2507.22565

Khan, M. W., Chen, S., Mironov, I., Zhang, L., & Noor, R. (2025). *BLIA: Detect model memorization in binary classification model through passive Label Inference attack*. https://doi.org/10.48550/ARXIV.2503.12801

Kibriya, H., Khan, W. Z., Siddiqa, A., & Khan, M. K. (2024). Privacy issues in Large Language Models: A survey. *Computers & Electrical Engineering*, *120*, 109698. https://doi.org/10.1016/j.compeleceng.2024.109698

Kim, K., Jeon, H. W., & Shin, J. (2025). Self-Refining Language Model Anonymizers via Adversarial Distillation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.01420

Koch, L. M., Baumgartner, C. F., & Berens, P. (2024). Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01085-w

Krishnamoorthy, M. (2024). Meta-Sealing: A Revolutionizing Integrity Assurance Protocol for Transparent, Tamper-Proof, and Trustworthy AI System. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.00069

Krüger, F., Östman, J., Mervin, L., Tetko, I. V., & Engkvist, O. (2024). Publishing Neural Networks in Drug Discovery Might Compromise Training Data Privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.16975

KT, AUTHOR_ID, N., Park, Y., Yoon, J., Moon, J.-Y., Oh, M., Lee, W., Kim, S. K., Kim, E., Park, H., Shin, E., Lee, W., Lee, S., Ju, M., Noh, M., Jeong, D., Kim, J. H., Park, W., & Bae, S. (2025). Responsible AI Technical Report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.20057

Kulothungan, V. (2025). A Blockchain-Enabled Approach to Cross-Border Compliance and Trust. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.09182

Kulynych, B., Gomez, J. F., Kaissis, G., Calmon, F. P., & Troncoso, C. (2024). Attack-Aware Noise Calibration for Differential Privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.02191

Kumar, A., Guru, Y., Singh, G., Vishwakarma, V. P., & Singh, G. G. (2024). Secured Blockchain and Fractional Discrete Cosine Transform-based Framework for Medical Images. *Makara Journal of Technology*, *28*(1). https://doi.org/10.7454/mst.v28i1.1628

Kumar, P., Samblani, G., Saini, G., & Gaur, K. (2024). Blockchain: From Cryptocurrency Foundation to Decentralized Revolution Across Industries. In *Lecture notes in networks and systems* (p. 521). Springer International Publishing. https://doi.org/10.1007/978-981-99-9518-9_38

Kumar, S., Upadhyay, P., & Ramsamy, G. (2025). Strengthening AI governance: International policy frameworks, security challenges, and ethical AI deployment. *ITU Journal on Future and Evolving Technologies*, *6*(3), 275. https://doi.org/10.52953/cogc3309

Lee, J. W., & Kifer, D. (2020). Scaling up Differentially Private Deep Learning with Fast Per-Example\n Gradient Clipping. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2009.03106

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating Training Data Makes Language Models Better. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2107.06499

Li, B., Wei, Y., Fu, Y., Wang, Z., Li, Y., Zhang, J., Wang, R., & Zhang, T. (2024). Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.10437

Li, B., Zhao, Q., & Wen, L. (2024). ROME: Memorization Insights from Text, Probability and Hidden State in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.00510

Li, D. C. Y. (2024). The synergistic potential of AI and blockchain for organizations. *AI & Society*. https://doi.org/10.1007/s00146-023-01838-3

Li, K., Wang, C., Xu, M., Zhang, Y., & Cheng, X. (2025). Dataset Ownership in the Era of Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.05921

Li, M., Ye, Z., Li, Y., Song, A., Zhang, G., & Liu, F. (2025). Membership Inference Attack Should Move On to Distributional Statistics for Distilled Generative Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.02970

Li, Q., Hong, J., Xie, C., Tan, J. T. C., Xin, R., Hou, J., Yin, X., Wang, Z., Hendrycks, D., Wang, Z., Li, B., He, B., & Song, D. (2024a). LLM-PBE: Assessing Data Privacy in Large Language Models. *Proceedings of the VLDB Endowment*, *17*(11), 3201. https://doi.org/10.14778/3681954.3681994

Li, Q., Hong, J., Xie, C., Tan, J. T. C., Xin, R., Hou, J., Yin, X., Wang, Z., Hendrycks, D., Wang, Z., Li, B., He, B., & Song, D. (2024b). LLM-PBE: Assessing Data Privacy in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.12787

Li, X., Tramèr, F., Liang, P., & Hashimoto, T. (2021). Large Language Models Can Be Strong Differentially Private Learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.05679

Li, Y., Ghozzi, S., Eichler, C., Anciaux, N., Bensamoun, A., & Manzano, L. G. (2025). Data Provenance Auditing of Fine-Tuned Large Language Models with a Text-Preserving Technique. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.09655

Li, Y., Shao, S., Shu, F., Guo, J., Zhang, T., Qin, Z., Chen, P., Backes, M., Torr, P. H. S., Tao, D., & Ren, K. (2025). Rethinking Data Protection in the (Generative) Artificial Intelligence Era. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.03034

Li, Z., & Ding, B. (2025). Beyond Data Privacy: New Privacy Risks for Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.14278

Li, Z., Lowy, A. M., Liu, J., Koike–Akino, T., Parsons, K., Malin, B., & Wang, Y. (2024). Analyzing Inference Privacy Risks Through Gradients in Machine Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.16913

Litzinger, J., Peters, D., Thiel, F., & Tschorsch, F. (2025). Utilization of Large Language Models for conformity assessment: Chances, Threats, and Mitigations. *Annals of Computer Science and Information Systems*, *44*, 61. https://doi.org/10.15439/2025f7772

Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., Cheng, Y., Wang, S., Wang, X., Luo, Y., Jin, H., Zhang, P., Liu, O., Chen, J., Zhang, H., … Wu, C. (2025). Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2504.01990

Liu, T., Lai, S.-N., Yuan, X., Liu, Y., & Lam, C. (2024). A novel blockchain-watermarking mechanism utilizing interplanetary file system and fast walsh hadamard transform. *iScience*, *27*(9), 110821. https://doi.org/10.1016/j.isci.2024.110821

Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang, Y., Wang, J., Gao, X., Zhong, T., Pan, Y., Xu, S., Wu, Z., Liu, Z., Zhang, X., Zhang, S., Hu, X., Zhang, T., Ning, Q., … Ge, B. (2024). *Understanding Llms: A Comprehensive Overview from Training to Inference*. https://doi.org/10.2139/ssrn.4706201

Liu, Y., Huang, J., Li, Y., Wang, D., & Xiao, B. (2024). Generative AI model privacy: a survey. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11024-6

Liu, Y., Lu, Q., Zhu, L., & Paik, H.-Y. (2023). Decentralised Governance-Driven Architecture for Designing Foundation Model based Systems: Exploring the Role of Blockchain in Responsible AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.05962

Liu, Y., Zhang, D., Xia, B., Anticev, J., Adebayo, T., Xing, Z., & Machao, M. (2024a). Blockchain-Enabled Accountability in Data Supply Chain: A Data Bill of Materials Approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.08536

Liu, Y., Zhang, D., Xia, B., Anticev, J., Adebayo, T., Xing, Z., & Machao, M. (2024b). *Blockchain-Enabled Accountability in Data Supply Chain: A Data Bill of Materials Approach*. 557. https://doi.org/10.1109/blockchain62396.2024.00082

Liu, Y., Zhao, X., Song, D., & Bu, Y. (2025). Dataset Protection via Watermarked Canaries in Retrieval-Augmented LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.10673

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K., Pentland, S., & Kabbara, J. (2024a). Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? *arXiv*. https://doi.org/10.48550/ARXIV.2404.12691

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K., Pentland, S., & Kabbara, J. (2024b). Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.12691

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Kabbara, J., & Pentland, S. (2024). *Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?* https://doi.org/10.21428/e4baedd9.a650f77d

Lukas, N., Salem, A., Sim, R. B., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing Leakage of Personally Identifiable Information in Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.00539

Lukas, N., Salem, A., Sim, R. B., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023b). Analyzing Leakage of Personally Identifiable Information in Language Models. *2022 IEEE Symposium on Security and Privacy (SP)*, 346. https://doi.org/10.1109/sp46215.2023.10179300

Luo, H., Luo, J., & Vasilakos, A. V. (2023). BC4LLM: Trusted Artificial Intelligence When Blockchain Meets Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06278

Lüthi, P., Gagnaux, T., & Gygli, M. (2020a). Distributed Ledger for Provenance Tracking of Artificial Intelligence Assets. *IFIP Advances in Information and Communication Technology*, 411. https://doi.org/10.1007/978-3-030-42504-3_26

Lüthi, P., Gagnaux, T., & Gygli, M. (2020b). Distributed Ledger for Provenance Tracking of Artificial Intelligence\n Assets. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2002.11000

Ma, B., Wang, X., Yu, G., Li, C., He, Y., & Liu, R. P. (2025). SoK: Semantic Privacy in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.23603

Ma, O., Passerat–Palmbach, J., & Usynin, D. (2024). Efficient and Private: Memorisation under differentially private parameter-efficient fine-tuning in language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.15831

Ma, Q., Zhu, R., Liu, P., Yan, R., Zhang, F., Liang, L., Li, M., Yu, Z., Wang, Z., Cai, Y., & Huang, T. (2024). CopyLens: Dynamically Flagging Copyrighted Sub-Dataset Contributions to LLM Outputs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.04454

Ma, W., Song, Y., Xue, M., Wen, S., & Xiang, Y. (2024). The "Code" of Ethics: A Holistic Audit of AI Code Generators. *IEEE Transactions on Dependable and Secure Computing*, *21*(5), 4997. https://doi.org/10.1109/tdsc.2024.3367737

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. *European Conference on the Impact of Artificial Intelligence and Robotics*, *4*(1), 45. https://doi.org/10.34190/eciair.4.1.765

Madhavan, K. K., Yazdinejad, A., Zarrinkalam, F., & Dehghantanha, A. (2025). Quantifying Security Vulnerabilities: A Metric-Driven Security Analysis of Gaps in Current AI Standards. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.08610

Maini, P., Jia, H., Papernot, N., & Dziedzic, A. (2024). LLM Dataset Inference: Did you train on my dataset? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.06443

Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership Inference Attacks against Language Models via Neighbourhood Comparison. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.18462

McHugh, J., Šekrst, K., & Cefalu, J. (2025). Prompt Injection 2.0: Hybrid AI Threats. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.13169

McIntosh, T. R., Sušnjak, T., Liu, T., Watters, P., Nowrozy, R., & Halgamuge, M. N. (2024). From COBIT to ISO 42001: Evaluating Cybersecurity Frameworks for Opportunities, Risks, and Regulatory Compliance in Commercializing Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.1016/j.cose.2024.103964

Meeus, M., Wutschitz, L., Zanella-Béguelin, S., Tople, S., & Shokri, R. (2025). *The Canary's Echo: Auditing Privacy Risks of LLM-Generated Synthetic Text*. https://doi.org/10.48550/ARXIV.2502.14921

Miranda, M., Ruzzetti, E. S., Santilli, A., Zanzotto, F. M., Bratières, S., & Rodolà, E. (2024). Preserving Privacy in Large Language Models: A Survey on Current

Threats and Solutions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.05212

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., & Shokri, R. (2022, January 1). Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.570

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., & Berg-Kirkpatrick, T. (2022, January 1). An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.119

Mireshghallah, N., & Li, T. (2025). Position: Privacy Is Not Just Memorization! *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.01645

Mohammed, R., Pérez, L. M., & Gardent, C. (2024). *Proceedings of the 17th International Natural Language Generation Conference*. https://doi.org/10.18653/v1/2024.inlg-main

Mohsin, A., Janicke, H., Nepal, Surya, & Holmes, D. (2023a). *Digital Twins and the Future of Their Use Enabling Shift Left and Shift Right Cybersecurity Operations*. 277. https://doi.org/10.1109/tps-isa58951.2023.00042

Mohsin, A., Janicke, H., Nepal, Surya, & Holmes, D. R. (2023b). Digital Twins and the Future of their Use Enabling Shift Left and Shift Right Cybersecurity Operations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.13612

Momha, M. (2025). The Synthetic Mirror -- Synthetic Data at the Age of Agentic AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.13818

More, Y., Ganesh, P., & Farnadi, G. (2024). Towards More Realistic Extraction Attacks: An Adversarial Perspective. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.02596

Nahid, M. M. H., & Hasan, S. B. (2024). SafeSynthDP: Leveraging Large Language Models for Privacy-Preserving Synthetic Data Generation Using Differential Privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.20641

Nair, S. (2024). Securing Against Advanced Cyber Threats: A Comprehensive Guide to Phishing, XSS, and SQL Injection Defense. *Journal of Computer Science and Technology Studies*, *6*(1), 76. https://doi.org/10.32996/jcsts.2024.6.1.9

Nakka, K. K., Frikha, A., Mendes, R., Jiang, X., & Zhou, X. (2024). PII-Scope: A Benchmark for Training Data PII Leakage Assessment in LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.06704

Namazi, M., Nemecek, A., & Ayday, E. (2025a). *ZKPROV: A Zero-Knowledge Approach to Dataset Provenance for Large Language Models*. https://doi.org/10.48550/ARXIV.2506.20915

Negoescu, D. M., González, H., Orjany, S. E. A., Yang, J., Lut, Y., Tandra, R., Zhang, X., Zheng, X., Douglas, Z., Nolkha, V., Ahammad, P., & Samorodnitsky, G. (2023). Epsilon*: Privacy Metric for Machine Learning Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.11280

Neulinger, A., & Sparer, L. (2025). Fostering AI alignment through blockchain, proof of personhood and zero knowledge proofs. *Cluster Computing*, *28*(15). https://doi.org/10.1007/s10586-025-05729-8

Neyigapula, B. S. (2023). Synergistic Integration of Blockchain and Machine Learning: A Path to a Decentralized Intelligent Future. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-3280888/v1

Nguyen-Cong, T., & Lê, N. C. (2024). "Utilizing Layer-2 Blockchain and IPFS Innovations for Establishing a Robust System for Secure Document Signing and Management." *JST Smart Systems and Devices*, *34*(2), 44. https://doi.org/10.51316/jst.174.ssad.2024.34.2.6

Nikolić, I., Baluta, T., & Saxena, P. (2025). Model Provenance Testing for Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.00706

Obinna, A. J., & Kess-Momoh, A. J. (2024). Comparative technical analysis of legal and ethical frameworks in AI-enhanced procurement processes. *World Journal of Advanced Research and Reviews*, *22*(1), 1415. https://doi.org/10.30574/wjarr.2024.22.1.1241

Odulaja, B. A., Ihemereze, K. C., Fakeyede, O. G., Abdul, A. A., Ogedengbe, D. E., & Daraojimba, C. (2023). HARNESSING BLOCKCHAIN FOR SUSTAINABLE PROCUREMENT: OPPORTUNITIES AND CHALLENGES. *Computer Science & IT Research Journal*, *4*(3), 158. https://doi.org/10.51594/csitrj.v4i3.619

Othman, S., & Getahun, M. (2025). Leveraging blockchain and IoMT for secure and interoperable electronic health records. *Scientific Reports*, *15*(1), 12358. https://doi.org/10.1038/s41598-025-95531-8

Panda, S. N. (2025). *Blockchain Technologies for Secure and Transparent Artificial Intelligence Systems*. https://doi.org/10.2139/ssrn.5250372

Pandya, M. (2024). Performance Evaluation of Hashing Algorithms on Commodity Hardware. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.08284

Pang, K., Qi, T. Y., Wu, C., Bai, M., Jiang, M., & Huang, Y. (2025). ModelShield: Adaptive and Robust Watermark against Model Extraction Attack. *IEEE Transactions on Information Forensics and Security*, 1. https://doi.org/10.1109/tifs.2025.3530691

Pankajakshan, R., Biswal, S., Govindarajulu, Y., & Gressel, G. (2024a). *Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal*. https://doi.org/10.48550/ARXIV.2403.13309

Pankajakshan, R., Biswal, S., Govindarajulu, Y., & Gressel, G. (2024a). *Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal*. https://doi.org/10.48550/ARXIV.2403.13309

Park, C.-W., Ha, H., Kim, J. Y., Kim, Y., Kim, D., Lee, S.-K., & Yang, S. (2024). 1 Trillion Token (1TT) Platform: A Novel Framework for Efficient Data Sharing and Compensation in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.20149

Pathade, C. (2025). *Red Teaming the Mind of the Machine: A Systematic Evaluation of Prompt Injection and Jailbreak Vulnerabilities in LLMs*. https://doi.org/10.48550/ARXIV.2505.04806

Paulius, R., Qiyao, W., & Mihaela, van der S. (2025). Statistical Hypothesis Testing for Auditing Robustness in Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.07947

Pithadia, H., Fenoglio, E., Batrinca, B., Treleaven, P., Echim, R., Bubutanu, A., & Kerrigan, C. (2023). Data Assets: Tokenization and Valuation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4419590

Poziomska, M., Dovgialo, M., Olbratowski, P., Niedbalski, P., Ogniewski, P., Zych, J., Rogala, J., & Żygierewicz, J. (2024). Quantity versus Diversity: Influence of Data on Detecting EEG Pathology with Advanced ML Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.17709

Pradhan, G., Jälkö, J., Tobaben, M., & Honkela, A. (2025). Hyperparameters in Score-Based Membership Inference Attacks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.06374

Prashant, P., Ponkshe, K., & Salimi, B. (2025). A Lightweight Method to Disrupt Memorized Sequences in LLM. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.05159

Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). (2023). https://doi.org/10.18653/v1/2023.trustnlp-1

Puerto, H., Gubri, M., Yun, S., & Oh, S. (2024). Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.00154

Puneeth, R. P., & Parthasarathy, G. (2023). Seamless Data Exchange: Advancing Healthcare with Cross-Chain Interoperability in Blockchain for Electronic Health Records. *International Journal of Advanced Computer Science and Applications*, *14*(10). https://doi.org/10.14569/ijacsa.2023.0141031

Purpura, A., Wadhwa, S., Zymet, J., Gupta, A., Luo, A. A., Rad, M. K., Shinde, S., & Sorower, M. S. (2025). *Building Safe GenAI Applications: An End-to-End Overview of Red Teaming for Large Language Models*. 335. https://doi.org/10.18653/v1/2025.trustnlp-main.23

Qi, T., Jinhua, Y., Cai, D., Yueqi, X., Wang, H., Zhiyang, H., Peiru, Y., Nan, G., Zhili, Z., Wang, S., Lingjuan, L., Huang, Y., & Nicholas, L. H. (2025). Evidencing Unauthorized Training Data from AI Generated Content using Information Isotopes. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2503.20800

Qu, W., Zhou, Y., Wang, B., Zheng, W., Li, Y., Jia, J., & Zhang, J. (2025). RepoMark: A Code Usage Auditing Framework for Code Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.21432

Quan, G., Yao, Z., Chen, L., Fang, Y., Zhu, W., Si, X., & Li, M. (2023). A trusted medical data sharing framework for edge computing leveraging blockchain and outsourced computation. *Heliyon*, *9*(12). https://doi.org/10.1016/j.heliyon.2023.e22542

Quayson, M., Avornu, E. K., & Bediako, A. K. (2024). Modeling the enablers of blockchain technology implementation for information management in healthcare supply chains. *Modern Supply Chain Research and Applications*, *6*(2), 101. https://doi.org/10.1108/mscra-06-2023-0028

Radanliev, P. (2024a). Artificial Intelligence and Quantum Cryptography. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4721869

Radanliev, P. (2024a). Artificial Intelligence and Quantum Cryptography. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4721869

Radanliev, P. (2024b). Artificial intelligence and quantum cryptography. *Journal of Analytical Science & Technology*, *15*(1). https://doi.org/10.1186/s40543-024-00416-6

Radanliev, P., Roure, D. D., & Santos, O. (2023). Red Teaming Generative AI/NLP, the BB84 Quantum Cryptography Protocol and the NIST-Approved Quantum-Resistant

Cryptographic Algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4574446

Ramakrishnan, B., & Balaji, A. (2025). Assessing and Mitigating Data Memorization Risks in Fine-Tuned Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.14062

Ramos, S., & Ellul, J. (2024). Blockchain for Artificial Intelligence (AI): enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective. *International Cybersecurity Law Review*, *5*(1), 1. https://doi.org/10.1365/s43439-023-00107-9

Rangwala, M., Sinnott, R., & Buyya, R. (2025). Network Structures as an Attack Surface: Topology-Based Privacy Leakage in Federated Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.19260

Rangwala, M., Venugopal, K. R., & Buyya, R. (2025). Blockchain-Enabled Federated Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.06406

Rashid, M. R. U., Liu, J., Koike–Akino, T., Mehnaz, S., & Wang, Y. (2024). Forget to Flourish: Leveraging Machine-Unlearning on Pretrained Language Models for Privacy Leakage. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.17354

Rauba, P., Wei, Q., & Schaar, M. van der. (2024). Quantifying perturbation impacts for large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.00868

Rawat, A., Schoepf, S., Zizzo, G., Cornacchia, G., Hameed, M. Z., Fraser, K., Miehling, E., Buesser, B., Daly, E. M., Purcell, M., Sattigeri, P., Chen, P.-Y., & Varshney, K. R. (2024). *Attack Atlas: A Practitioner's Perspective on Challenges and Pitfalls in Red Teaming GenAI*. https://doi.org/10.48550/ARXIV.2409.15398

Ren, J., Chen, K., Chen, C., Sehwag, V., Xing, Y., Tang, J., & Lyu, L. (2024). Self-Comparison for Dataset-Level Membership Inference in Large (Vision-)Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.13088

Russinovich, M., & Salem, A. (2024). Hey, That's My Model! Introducing Chain & Hash, An LLM Fingerprinting Technique. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.10887

Sablayrolles, A., Douze, M., Schmid, C., & Jeǵou, H. (2022). Déjà Vu: an empirical evaluation of the memorization properties of ConvNets. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1809.06396

Sachan, S., & Liu, X. (2023). Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools. *Engineering Applications of Artificial Intelligence*, *129*, 107666. https://doi.org/10.1016/j.engappai.2023.107666

Sai, Y., Wang, Q., Yu, G., Bandara, H. M. N. D., & Chen, S. (2024). Is Your AI Truly Yours? Leveraging Blockchain for Copyrights, Provenance, and Lineage. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.06077

Samuel-Okon, A. D., Olateju, O. O., Okon, S. U., Olaniyi, O. O., & Igwenagu, U. T. I. (2024). Formulating Global Policies and Strategies for Combating Criminal Use and Abuse of Artificial Intelligence. *Archives of Current Research International*, *24*(5), 612. https://doi.org/10.9734/acri/2024/v24i5735

Šarčević, T., Karlowicz, A., Mayer, R., Baeza-Yates, R., & Rauber, A. (2024). U Can't Gen This? A Survey of Intellectual Property Protection Methods for Data in Generative AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.15386

Satvaty, A., Verberne, S., & Turkmen, F. (2024). Undesirable Memorization in Large Language Models: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.02650

Schlegel, V., Bharath, A. A., Zhao, Z., & Yee, K. (2025a). *Generating Synthetic Data with Formal Privacy Guarantees: State of the Art and the Road Ahead*. https://doi.org/10.48550/ARXIV.2503.20846

Schoenegger, P., Tuminauskaite, I., Park, P. S., & Tetlock, P. E. (2024). Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.19379

Seyedi, S., Xiong, L., Nemati, S., & Clifford, G. D. (2022). An Analysis Of Protected Health Information Leakage In Deep-Learning Based De-Identification Algorithms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2101.12099

Shahana, A., Hasan, R., Farabi, S. F., Akter, J., Mahmud, Md. A. A., Johora, F. T., & Suzer, G. (2024). AI-Driven Cybersecurity: Balancing Advancements and Safeguards. *Journal of Computer Science and Technology Studies*, *6*(2), 76. https://doi.org/10.32996/jcsts.2024.6.2.9

Sheoran, S. K., & Yadav, G. V. (2023). Promises and Perils of Post-Quantum Blockchain. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-2887673/v1

Shilov, I., Meeus, M., & Montjoye, Y.-A. de. (2024). Mosaic Memory: Fuzzy Duplication in Copyright Traps for Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.15523

Singh, A. P. (2025). *Blockchain Technology: Core Mechanisms, Evolution, and Future Implementation Challenges*. https://doi.org/10.48550/ARXIV.2505.08772

Singh, S. (2024). Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach. *International Conference on Cyber Warfare and Security*, *19*(1), 574. https://doi.org/10.34190/iccws.19.1.2096

Slonski, E. (2024). Detecting Memorization in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.01014

Smith, A. L., Zheng, T., & Gelman, A. (2022). Prediction scoring of data-driven discoveries for reproducible research. *Statistics and Computing*, *33*(1). https://doi.org/10.1007/s11222-022-10154-7

Smith, V., Shamsabadi, A. S., Ashurst, C., & Weller, A. (2023). *Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey*. https://doi.org/10.48550/ARXIV.2310.01424

Smith, V., Shamsabadi, A. S., Ashurst, C., & Weller, A. (2023b). Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.01424

Soleymani, D., Dadsetan, A., & Rudzicz, F. (2025). SoftAdaClip: A Smooth Clipping Strategy for Fair and Private Model Training. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.01447

South, T., Drean, J., Singh, A., Zyskind, G., Mahari, R., Sharma, V., Vepakomma, P., Kagal, L., Devadas, S., & Pentland, A. (2024). *A Roadmap for End-to-End Privacy and Security in Generative AI*. https://doi.org/10.21428/e4baedd9.9af67664

Spirito, C., Kerby, L., & Mena, P. (2023). Complete Evaluation on Advanced Reactor Machine Learning Subversion Attacks (Final). *OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information)*. https://doi.org/10.2172/2331401

Stokes, J. W., England, P., & Kane, K. (2021). Preventing Machine Learning Poisoning Attacks Using Authentication and\n Provenance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2105.10051

Su, E., Vellore, A., Chang, A., Mura, R. L., Nelson, B., Kassianik, P., & Karbasi, A. (2024). Extracting Memorized Training Data via Decomposition. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.12367

Sun, R., & Lampert, C. H. (2019a). KS(conf): A Light-Weight Test if a ConvNet Operates Outside of Its Specifications. In *Lecture notes in computer science* (p. 244). Springer Science+Business Media. https://doi.org/10.1007/978-3-030-12939-2_18

Sun, R., & Lampert, C. H. (2019b). KS(conf): A Light-Weight Test if a Multiclass Classifier Operates Outside of Its Specifications. *International Journal of Computer Vision*, *128*(4), 970. https://doi.org/10.1007/s11263-019-01232-x

Suri, A., Zhang, X., & Evans, D. (2024). Do Parameters Reveal More than Loss for Membership Inference? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.11544

Tao, J., & Shokri, R. (2025). (Token-Level) \textbf{InfoRMIA}: Stronger Membership Inference and Memorization Assessment for LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.05582

Tariq, A., Qayyum, T., Alrabaee, S., & Serhani, M. A. (2025). Blockchain and Distributed Ledger Technologies for Cyberthreat Intelligence Sharing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2504.02537

Tertulino, R. (2025). A Comparative Benchmark of Federated Learning Strategies for Mortality Prediction on Heterogeneous and Imbalanced Clinical Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.10517

Tete, S. B. (2024). Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.11007

Tete, S. B. (2024). Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.11007

Tiwari, S., Sresth, V., & Srivastava, A. (2020). The Role of Explainable AI in Cybersecurity: Addressing Transparency Challenges in Autonomous Defense Systems. *International Journal of Innovative Research in Science Engineering and Technology*, *9*(3), 718. https://doi.org/10.15680/ijirset.2020.0903165

Tiwari, T., & Suh, G. E. (2024). Sequence-Level Analysis of Leakage Risk of Training Data in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.11302

Tonekaboni, S., Stempfle, L., Fallahpour, A., Gerych, W., & Ghassemi, M. (2025). An Investigation of Memorization Risk in Healthcare Foundation Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.12950

Topcu, T. G., Husain, M., Ofsa, M., & Wach, P. (2025). Trust at Your Own Peril: A Mixed Methods Exploration of the Ability of Large Language Models to Generate Expert-Like Systems Engineering Artifacts and a Characterization of Failure Modes. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.09690

Tripathi, G., Ahad, M. A., & Casalino, G. (2023). A comprehensive review of blockchain technology: Underlying principles and historical background with future challenges [Review of *A comprehensive review of blockchain technology: Underlying principles and historical background with future challenges*]. *Decision Analytics Journal*, *9*, 100344. Elsevier BV. https://doi.org/10.1016/j.dajour.2023.100344

Ullah, F., He, J., Zhu, N., Wajahat, A., Nazir, A., Qureshi, S., Pathan, M. S., & Dev, S. (2024). Blockchain-enabled EHR access auditing: Enhancing healthcare data security. *Heliyon*, *10*(16). https://doi.org/10.1016/j.heliyon.2024.e34407

Vempati, S. (2024). Securing Smart Cities: A Cybersecurity Perspective on Integrating IoT, AI, and Machine Learning for Digital Twin Creation. *Deleted Journal*, *20*(3), 1420. https://doi.org/10.52783/jes.3548

Verma, A., Krishna, S., Gehrmann, S., Seshadri, M., Pradhan, A., Ault, T., Barrett, L., Rabinowitz, D., Doucette, J., & Phan, N. (2024). Operationalizing a Threat Model for Red-Teaming Large Language Models (LLMs). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.14937

Wagner, E., Slavutsky, Y., & Abend, O. (2024). CONTESTS: a Framework for Consistency Testing of Span Probabilities in Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.19984

Waiwitlikhit, S., Stoica, I., Sun, Y., Hashimoto, T., & Kang, D. (2024). Trustless Audits without Revealing Data or Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.04500

Wang, K., Iranzad, A. N., Schaffter, S., Precup, D., & Lebensold, J. (2024). Mitigating Downstream Model Risks via Model Provenance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.02230

Wang, N., Walter, K., Gao, Y., & Abuadbba, A. (2025). Large Language Model Adversarial Landscape Through the Lens of Attack Objectives. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.02960

Wang, Q., Yu, G., Sai, Y., Bandara, H. M. N. D., & Chen, S. (2024). *Is Your AI Truly Yours? Leveraging Blockchain for Copyrights, Provenance, and Lineage*. https://doi.org/10.48550/ARXIV.2404.06077

Wang, S., Zhu, T., Liu, B., Ding, M., Guo, X., Ye, D., & Zhou, W. (2024). Unique Security and Privacy Threats of Large Language Model: A Comprehensive Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.07973

Wang, X., Tian, Y., Huang, K., & Liang, B. (2024). Practically implementing an LLM-supported collaborative vulnerability remediation process: A team-based approach. *Computers & Security*, *148*, 104113. https://doi.org/10.1016/j.cose.2024.104113

Wang, Y., Feng, D., Dai, Y. S., Chen, Z., Huang, J., Ananiadou, S., Xie, Q., & Wang, H. (2024). HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.02927

Wang, Z., Sun, R., Lui, E., Shah, V. H., Xiong, X., Sun, J., Crapis, D., & Knottenbelt, W. J. (2024). SoK: Decentralized AI (DeAI). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.17461

Webster, R., Rabin, J., Simon, L., & Jurie, F. (2019). Detecting Overfitting of Deep Generative Networks via Latent Recovery. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11265. https://doi.org/10.1109/cvpr.2019.01153

Wei, J. T.-Z., Wang, R. Y., & Jia, R. (2024). Proving membership in LLM pretraining data via data watermarks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.10892

Wei, J., Zhang, Y., Zhang, L. Y., Ding, M., Chen, C., Ong, K., Zhang, J., & Xiang, Y. (2024). Memorization in deep learning: A survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.03880

Weinberg, A. I. (2024). The Role and Applications of Airport Digital Twin in Cyberattack Protection during the Generative AI Era. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.05248

Wicaksono, I., Wu, Z., Patel, R., King, T., Koshiyama, A., & Treleaven, P. (2025). Mind the Gap: Comparing Model- vs Agentic-Level Red Teaming with Action-Graph Observability on GPT-OSS-20B. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.17259

Wu, Z., Lou, J., Zheng, Z., & Chen, C. (2024). MemHunter: Automated and Verifiable Memorization Detection at Dataset-scale in LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.07261

Wylde, V., Rawindaran, N., Lawrence, J., Balasubramanian, R., Prakash, E., Jayal, A., Khan, I., Hewage, C., & Platts, J. (2022). Cybersecurity, Data Privacy and Blockchain: A Review. *SN Computer Science*, *3*(2), 127. https://doi.org/10.1007/s42979-022-01020-4

Xie, R., Wang, J., Huang, R., Zhang, M., Ge, R., Pei, J., Gong, N. Z., & Dhingra, B. (2024). ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.15968

Xu, H., Wang, S., Li, N., Wang, K. C., Zhao, Y., Chen, K., Yu, T., Yang, L., & Wang, H. (2024). Large Language Models for Cyber Security: A Systematic Literature Review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.04760

Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. (2024). LLM Jailbreak Attack versus Defense Techniques -- A Comprehensive Study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.13457

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.05156

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2025). On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review [Review of *On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review*]. *High-Confidence Computing*, 100300. Elsevier BV. https://doi.org/10.1016/j.hcc.2025.100300

Yang, Q., Mohammad, M., Wang, H., Payani, A., Kundu, A., Shu, K., Yan, Y., & Hong, Y. (2024). LMO-DP: Optimizing the Randomization Mechanism for Differentially Private Fine-Tuning (Large) Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.18776

Yang, Q., Stout, N., Mohammady, M., Wang, H., Samreen, A., Quinn, C. J., Yan, Y., Kundu, A., & Hong, Y. (2025). PLRV-O: Advancing Differentially Private Deep Learning via Privacy Loss Random Variable Optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.06264

Yang, Z., Zhao, Z., Wang, C., Shi, J., Kim, D., Han, D., & Lo, D. (2024). Gotcha! This Model Uses My Code! Evaluating Membership Leakage Risks in Code Models. *IEEE Transactions on Software Engineering*, *50*(12), 3290. https://doi.org/10.1109/tse.2024.3482719

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., & Shokri, R. (2022). Enhanced Membership Inference Attacks against Machine Learning Models. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3093. https://doi.org/10.1145/3548606.3560675

You, D., & Chon, D. (2024). Trust & Safety of LLMs and LLMs in Trust & Safety. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.02113

You, D., & Chon, D. (2024). Trust & Safety of LLMs and LLMs in Trust & Safety. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2

Yu, M., Singh, P., Kumar, A., & Cao, Y. (2025). Adaptive Token-Weighted Differential Privacy for LLMs: Not All Tokens Require Equal Protection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.23246

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., Shajari, H., Sun, H., Levitan, D., & Sim, R. B. (2023). *Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe*. 1321. https://doi.org/10.18653/v1/2023.acl-long.74

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., Sun, H., Levitan, D., & Sim, R. B. (2022). Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2210.14348

Zangana, H. M., Mustafa, F. M., & Li, S. (2024). Large Language Models in Cybersecurity. In *Advances in information security, privacy, and ethics book series* (p. 277). IGI Global. https://doi.org/10.4018/979-8-3373-1102-9.ch009

Zarifzadeh, S., Liu, P., & Shokri, R. (2023). Low-Cost High-Power Membership Inference Attacks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.03262

Zarzar, T. G., Dulce-Rubio, M., Ramdas, A., & Ribero, M. (2025). Sequentially Auditing Differential Privacy. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.07055

Zhang, G., Su, Q., Liu, J., Qian, C., Pan, Y., Fu, Y., & Zhang, D. (2025). *ISACL: Internal State Analyzer for Copyrighted Training Data Leakage*. https://doi.org/10.48550/ARXIV.2508.17767

Zhang, H., Zhao, Y., Angione, C., Yang, H., Buban, J., Farhan, A., Johnston, F., & Colangelo, P. (2024). Towards Secure and Private AI: A Framework for Decentralized Inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.19401

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H., & Li, H. (2024). Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.02936

Zhang, R., Bertrán, M., & Roth, A. (2024). Order of Magnitude Speedups for LLM Membership Inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2409.14513

Zhang, X., Fei, Y.-L., Kang, Y., Chen, W., Fan, L., Jin, H., & Yang, Q. (2024). No Free Lunch Theorem for Privacy-Preserving LLM Inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.20681

Zhang, X., Sheng, Y., & Liu, Z. (2023). Using expertise as an intermediary: Unleashing the power of blockchain technology to drive future sustainable management using hidden champions. *Heliyon*, *10*(1). https://doi.org/10.1016/j.heliyon.2023.e23807

Zhang, Z., Fang, M., Chen, M., Li, G., Lin, X., & Liu, Y. (2024). Securing Distributed Network Digital Twin Systems Against Model Poisoning Attacks. *IEEE Internet of Things Journal*, *11*(21), 34312. https://doi.org/10.1109/jiot.2024.3421895

Zhang, Z., Wen, J., & Huang, M. (2023). *ETHICIST: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation*. https://doi.org/10.18653/v1/2023.acl-long.709

Zhao, G., & Song, E. (2024). Privacy-Preserving Large Language Models: Mechanisms, Applications, and Future Directions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.06113

Zhao, L. C., Rehn, A., Heikkilä, M., Tajeddine, R., & Honkela, A. (2025). Mitigating Disparate Impact of Differentially Private Learning through Bounded Adaptive Clipping. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.01396

Zhao, Y., & Zhang, J. (2025). Does Training with Synthetic Data Truly Protect Privacy? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.12976

Zhou, L. (2024). Trustworthy digital twinning data platform for power infrastructure construction projects using blockchain and semantic web. *Frontiers in Built Environment*, *10*. https://doi.org/10.3389/fbuil.2024.1440513

Zhou, Z., Xiang, J., Chen, C., & Su, S. (2023). Quantifying and Analyzing Entity-level Memorization in Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.15727

Zhou, Z., Xiang, J., Chen, C., & Su, S. (2024). Quantifying and Analyzing Entity-Level Memorization in Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 19741. https://doi.org/10.1609/aaai.v38i17.29948

Zhu, Z., Yang, Y., & Lian, D. (2024). TDDBench: A Benchmark for Training data detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.03363

Zuo, X., Wang, M., Zhu, T., Yu, S., & Zhou, W. (2024). Large Language Model Federated Learning with Blockchain and Unlearning for Cross-Organizational Collaboration. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.13551

Zuo, X., Wang, M., Zhu, T., Zhang, L., Ye, D., Yu, S., & Zhou, W. (2024). Federated TrustChain: Blockchain-Enhanced LLM Training and Unlearning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.04076