Issues in Information Systems
Volume 25, Issue 4, pp. 433-442, 2025

DOI: https://doi.org/10.48009/4_iis_2025_135

Data poisoning 2018–2025: A systematic review of risks, impacts, and mitigation challenges

Frank Hartle III *Robert Morris University hartle@rmu.edu*
Steve Mancini *Robert Morris University mancini@rmu.edu*
Emily Kerry *National Cyber Forensics and Training Alliance ekerry@ncfta.net*

Abstract

Data poisoning attacks represent a critical threat to machine learning (ML) and artificial intelligence (AI) systems, with consequences across any sector employing an AI solution. As AI grows and is adopted into our personal lives and the industries we work for, the threat of manipulation may be unknown to those who adopt it and undervalued by those who may profit from it. This paper attempts, through a meta-analysis, to synthesize findings from select studies published between 2018 and 2025, evaluating the technical, ethical, and sector-specific impacts of data poisoning. Key findings reveal that even minimal adversarial disturbances (as low as 0.001% of training data) can degrade model accuracy by up to 30%, distort decision boundaries in safety-critical systems (e.g., autonomous vehicles and healthcare diagnostics), and enable targeted attacks like backdoor triggers in generative AI. Sector-specific analyses demonstrate financial losses in algorithmic trading, misdiagnoses in medical imaging, and vulnerabilities in large language models (LLMs) trained on poisoned datasets like The Pile. Mitigation strategies, including adversarial training and knowledge graph-based verification, show partial efficacy but fail to address scalability challenges. This study underscores the urgent need for robust, multi-layered defenses and interdisciplinary collaboration to safeguard AI ecosystems.

Keywords: AI, Machine learning, Large language models, Data poisoning, Adversarial training data, Label flipping

Introduction

Data poisoning is the act of intentionally sending false or misleading data inputs, which can influence the model's behavior, typically with negative consequences (Korada, 2024). There are six different types of data poisoning attacks: targeted attacks, non-targeted attacks, label poisoning/backdoor poisoning, training data poisoning, model inversion attacks, and stealth attacks.

Targeted attacks look to exploit specific hardware or software within a system, leading to the model misinterpreting the signals it is receiving. Whereas a targeted attack focuses on a certain misconfiguration, a non-targeted attack targets the system as a whole. For example, Researchers showed the insertion of malware into AI systems can manipulate outcomes without necessarily breaking the system (Shen & Xia, 2020). In this instance, a Trojan virus was installed into real Go which ultimately 'manipulated' the AI behavior (Shen & Xia, 2020). As such, targeted attacks can allow for the customization of malware specifically designed to alter AI to effectively run without the user's awareness.

Label/backdoor poisoning targets the data that the model is trained on. The data on which the model is trained is poisoned, allowing for the exploitation of the model when it is asked to make inferences on

different data (Korada, 2024). This type of data poisoning likely would require a threat actor to have direct access to the LLM and the data it is trained on, either as an insider threat or through exploitation tactics. To add, the problem becomes even more challenging as we continue to leverage AI to generate code. For example, the implementation of AI code generators allows for the direct manipulation of code generation software as most AI code generation is training on, "large amounts of data, often collected from un-sanitized online sources" (Cotroneo et al, 2024, p. 280). As such, it becomes easy to manipulate the AI training by introducing large amounts of code into code repositories wherein vulnerabilities exist (Cotroneo et al, 2024)

Training data poisoning is similar to label/backdoor poisoning; training data poisoning targets internal datasets, but targets the training examples. Usually, the manipulations are minor but can cause disruptions and affect predictions in the future. In fact, Koh et al (2022) noted in their research, adding just 3% poisoned data can result in increases in test error from 3 to 24%. The other poisoning methods target the system or datasets directly, causing disruptions or incorrect inferences; however, model inversion attacks try to extract the data through queries of the model, and the threat actor attempts to recreate the training samples the model was trained on.

Stealth attacks involve the threat actor attempting to evade detection while still exploiting the system and potentially causing harm (Korada, 2024). A stealth attack can apply to any of the previous injections as long as that injection cannot be detected by systems. Furthermore, as noted earlier, the introduction of customized malware to alter AI behavior can occur in such a way as to be unnoticed by the user (Shen & Xia, 2020).

The proliferation of AI systems in critical domains, from healthcare to finance, may heighten vulnerabilities to data poisoning, a form of adversarial attack where malicious actors manipulate training data to corrupt model behavior. Unlike traditional cyberattacks, data poisoning exploits the inherent trust in training data, making detection challenging and consequences severe. Alber et al. (2025) demonstrated that poisoning just 0.001% of medical training tokens in LLMs increased harmful outputs by 4.8%, while Huang et al. (2020) achieved 41% attack success rates in code-generating models with 3% poisoned data.

It is clear AI manipulation can occur in any stage of the process which generates the usable AI model. To limit the analysis, this paper focuses on the systematic review of data poisoning consequences across three dimensions. From this vantage point, we attempt to show the implications across all AI-connected systems. We focused on the three dimensions: technical impacts, sector-specific risks, and mitigation limitations.

Synthesis of Findings

Technical Impacts

The technical impacts of data poisoning attacks could be profound and complex. At their core, these attacks may degrade model accuracy, distort decision boundaries, and amplify vulnerabilities within affected systems. For example, even minimal poisoning, such as corrupting 0.001% of tokens in a medical dataset, can increase the incidence of harmful outputs by nearly 5%, making it difficult for human evaluators to distinguish between legitimate and poisoned content (Alber et al., 2025). In neural machine translation models used for code generation, poisoning as little as 3% of the training data can result in a 12–41% poisoning attack success rate, producing code with vulnerabilities that would otherwise not occur (Korada et al., 2024). In addition, "poisoned data does not have to look anomalous; if the poisoned points are carefully coordinated" (Koh et al, 2020, p. 4). These technical disruptions may not be limited to a single model type or application as they could potentially propagate through updates and retraining, compounding over time and undermining the reliability of AI systems across domains (Alber et al., 2025; Korada et al., 2024)

Cotroneo, Improta, Liguori, & Natella (2024) explore how neural machine translation (NMT) AI code generators can be poisoned to produce vulnerable code. Current LLMs utilize open-source data from code repositories including GitHub, HuggingFace, and StackOverflow. As the datasets are pulled from these sources, they could potentially contain poisoned data. There are few restraints or guidelines on sanitizing the data pulled from these repositories, and most are trusted without any checks. Cotroneo et al. wanted to test how secure AI code generators are. Three different NMT models were poisoned, and the results of the poisoned data were recorded. According to Contoneo et al., Seq2Seq, CodeBERT, and CodeT5+. NMT models are considered the best solution for AI-based code generation. The study's findings revealed that if 3% of the data were poisoned, it would affect code generation. When increasing the amount of the poisoned data to 6%, all NMT models contained more vulnerable code. Pre-trained models can be targeted with poisoning attacks, and it will not affect the performance of the models. The attack's success depended on the amount of poisoned training data and the model architecture.

Sector Specific Risks

Sector-specific data poisoning risks, as demonstrated by case studies in healthcare, finance, autonomous systems, and generative AI, may be equally vulnerable. In healthcare, public datasets such as The Pile and PubMed may be susceptible to poisoning. Potential attackers could inject misinformation at scale for relatively little cost. This may directly impact clinical decision-making, as models trained on these datasets may produce unsafe recommendations that clinicians cannot reliably identify as erroneous (Alber et al., 2025). In the financial sector, poisoning just 1% of training data in fraud detection or trading algorithms can lead to significant economic losses and increased false positives, undermining trust in automated systems (Korada et al., 2024). Autonomous systems, such as self-driving vehicles, could be equally vulnerable to misclassification of critical objects like road signs, which can result in catastrophic safety failures. Generative AI models may also be at risk, especially those relying on in-context learning. Targeted poisoning could decrease accuracy by up to 30%, with open-source models being easily accessible and susceptible (Li et al., 2024).

Medical large language models may be vulnerable to data-poisoning attacks that utilize the dataset known as The Pile, which is known for LLM development (Alber, et al., 2024). As with other LLM's public data can be used to train different models. Within the medical realm, these public databases include Common Crawl, PubMed, and Project Gutenberg (Alber, et al., 2024). These platforms have a lack of oversight which could lead to potentially vulnerable datasets if poisoned. Researchers performed an in-depth analysis on The Pile, as it is the most widely used dataset for LLMs and has the least vulnerable medical content (Alber, et al., 2024). Of the data that was contained within The Pile, 27.4% of it was considered a vulnerable subset, with more than half of the vulnerable data originating from a public dataset known as the Common Crawl (Alber, et al., 2024).

Alber et al. (2024) created a simulated data poisoning attack against The Pile. The researchers created 150,000 misinformation articles using OpenAI GPT-3.5 turbo API and used these articles to corrupt information in The Pile. Two different types of parameters were tested: a broad targeting technique with a parameter of 1.3 billion across many different concepts, where 0.5% and 1% of the data were poisoned, and a smaller targeting technique across single concepts with a parameter of 1.3 billion and 4 billion, where 0.001% of the dataset was poisoned (Alber, et al., 2024).

Fifteen clinicians were tasked with determining the poisoned response and the baseline response; the reviewers were unable to determine the difference between the two results. The data in the 1.3 billion parameters had the p-values 0.0314 and 0.00484 between the 0.5% and 1.0% poisoned data, respectively (Alber, et al., 2024). When the concept-specific data was poisoned, at 0.001%, there was a 4.8% increase

in harmful content. The fake data that was created was created for less than $100 USD, it is predicted by the researchers that if this were formed at scale, it would remain under $1,000 USD to train a model with 15 million injected tokens (Alber, et al., 2024).

Data Poisoning for In-context Learning (ICL) analyzed how in-context learning could be poisoned, leading to less accurate responses using different datasets and seven different models (He, et al., 2024). The datasets used in this experiment were Stanford Sentiment Treebank (SST2), Corpus of Linguistic Acceptability (Cola), Emo dataset, AG's new (AG) corpus, and Poem Sentiment (Poem). The models used to test against the datasets were Llama2-7B, Pythia, Falcon-7B, GPT-J-6B, and MPT-7B. API only models included GPT-3.5 and GPT-4.

ICL allows models to make predictions based on the prompt information, leading to more relevant predictions by the model. For this study to occur, the researchers assumed that a threat actor had access to either the full data set or a portion of the dataset (He, et al., 2024). The researchers created an ICL poisoning technique, ICLPoison, which leveraged hidden states within the ICL model. This attack vector was optimized and categorized into three main sections: synonym replacement, character replacement, and adversarial suffix (He, et al., 2024). Synonym replacement replaced a limited number of words with synonyms of that word to avoid detection while still maintaining the meaning of the statement sent to the LLM. The researchers also implemented a greedy system only allowing for replacement of a limited number of words to prevent detection. Using GloVe, the researchers find synonyms using word embedding (He, et al., 2024). Character replacement is similar to synonym replacement; however, this method replaces characters to assist in detection evasion (He, et al., 2024). Adversarial suffix adds tokens that are imperceptible to humans at the end of a prompt. As with synonym replacement, the adversarial suffix is limited to the extent to which it is allowed to replace (He, et al., 2024).

The results of the He et al. (2024) study revealed that open-source models that were not poisoned were accurate over 88% of the time (He, et al., 2024). When performing an ICLPoison attack, the ICL accuracy dropped 10% and over 30% in some instances (He, et al., 2024). Depending on the model and dataset that was used, the effectiveness of the attack was influenced; of the different ICLPoison attack vectors, synonym replacement and adversarial suffix have the largest impact on decreasing ICL accuracy. Different models had lower tolerance to poison in different datasets. API-only models employed using Llama2-7B were used to model GPT-3.5 and GPT-4 since access to direct models was unavailable (He, et al., 2024). ICL accuracy was reduced by 10% using ICLPoison techniques (He, et al., 2024).

He, et al. (2024) wanted to observe if a poisoned dataset could be transferred across different models. It was found that using the ICLPoison technique that "there was over a 30% decrease in accuracy for open-source models" (He, et al., 2024). It was found that API based models and larger models are more resistant to the poisoned code. Partial poisoning was also performed to determine how much information could be poisoned and still have an impact on the dataset. At 10% poisoned data, there was a 7% decrease in performance, and a 15% decrease at 20%.

Mitigation Limitations

Machine unlearning, a process designed to remove specific data from trained models, has been shown to be largely ineffective against sophisticated poisoning attacks. Even when allocating generous computational resources to 10% of training computing, none of the tested unlearning algorithms could fully remove poisoned data, and some attacks left model performance virtually unchanged (Nguyen, Huynh, Pham, & Tran, 2023). Adversarial training has been shown to be circumvented by novel attack strategies and often leads to substantial increases in computational cost without guaranteeing complete protection (Korada et al., 2024). Real-time monitoring and heuristic defenses, such as accuracy thresholds, have been

shown to fail to detect stealthy or reiterative poisoning attacks where poisoned data can quickly propagate and degrade model performance (Nguyen, Huynh, Pham, & Tran, 2023; Li et al., 2024). As a result, there may be a pressing need for more robust, adaptive, and multi-layered defense mechanisms to safeguard AI systems against the evolving threat of data poisoning.

Pawelczyk et al. (2025) demonstrated that there has been an increase in the need to take data out of machine learning models to be compliant with different international privacy protection laws. According to Pawelczyk et al. (2025), the most effective way to perform machine unlearning is to recreate the model, making sure the data is removed; however, this can be impractical due to the large nature of different machine learning models. Different unlearning algorithms have been made with an effort not to influence the model. The study aimed to find if a machine learning model could use unlearning algorithms to ignore data poisoning attacks (Pawelczyk, et al., 2025). The researchers came up with a different way to test poisoned data sets using Gaussian noise. This new model poisons the dataset and compares the poisoned datasets to the original one to determine the separation between the poisoned model and the original. Gaussian noise uses visually undetectable signals inside the corrupted training data (Pawelczyk, et al., 2025). This type of data poisoning had no impact on the model's performance in a significant way.

Pawelczyk, et al. (2025) allowed for up to 10% of the training computer to be used to perform unlearning utilizing different unlearning algorithms. The researchers acknowledged that 10% is considered generous, and anything more than 10% would not be practical for unlearning. To measure the effectiveness of unlearning, the researchers measured the model's performance post-unlearning compared to the performance of a non-poisoned model to determine the unlearning ability (Pawelczyk, et al., 2025).

None of the models that were tested removed all of the poisoned data completely (Pawelczyk, et al., 2025). The ability of an unlearning algorithm to work depends on the type of data poisoning that has occurred. Some models were able to mitigate some data poisoning attacks while not being able to unlearn others. The researchers have two hypotheses for why unlearning algorithms fail to remove poisons: approximate unlearning is unable to complete all of the unlearning with a reasonable computational budget (Pawelczyk, et al., 2025).

In the Alber, et al. (2024) study, using known mitigation strategies, the amount of poisoned data remained unchanged, and the researchers developed a different approach that performs cross-references between the "LLM output and biomedical knowledge graphs for medical misinformation". This model does not rely on another LLM to verify the information, but uses a separate dataset which captures over 90% of the misinformation from the poisoned LLM (Alber, et al., 2024).

Methodology

This systematic review was conducted to comprehensively identify, evaluate, and synthesize current research on the risks, impacts, and mitigation challenges of data poisoning in AI systems.

Research Questions: This review seeks to answer:

1. What are the primary types and mechanisms of data poisoning attacks targeting AI systems reported in the literature?
2. Which sectors are identified as particularly vulnerable, and what are the specific consequences observed or simulated in these domains?
3. What mitigation strategies have been proposed and evaluated, and what are their reported strengths and limitations?

4. What are the emerging ethical, societal, and governance challenges associated with AI data poisoning?

A systematic search of literature published between January 2018 and May 2025 was performed across multiple electronic databases: IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, arXiv, and Google Scholar. Search queries combined keywords such as: ("data poisoning" OR "adversarial training data" OR "backdoor attack" OR "label flipping" OR "model poisoning") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "LLM" OR "large language model") AND ("impact" OR "risk" OR "vulnerability" OR "threat") AND ("mitigation" OR "defense" OR "detection" OR "prevention"). Reference lists of identified key articles and relevant reviews were also manually scanned for additional studies.

Discussion

The synthesized findings reveal a persistent and evolving threat landscape for AI systems. The technical impacts, ranging from significant accuracy degradation to the generation of harmful or insecure outputs, are consistently reported across diverse model types and application domains.

Ethical and Societal Ramifications

A critical theme emerging from this review is the profound ethical and societal ramifications of data poisoning. Beyond performance metrics, poisoned AI systems can perpetuate biases, spread misinformation (as highlighted by studies on LLMs trained on datasets like The Pile), and erode public trust in AI. The ease with which attackers could generate harmful medical outputs (Alber et al., 2024) or vulnerable code (Cotroneo et al., 2024) underscores the potential for real-world harm. The work by organizations like CheckPoint Research (2025) highlights concerns about retrieval poisoning and the malicious modification of LLMs, further emphasizing these risks. Additionally, Shiferaw et al. (2024) noted in their research how leveraging popular AI tools, like ChatGPT, resulted in different answers to the same questions. In addition, the type of question also mattered. For example, they demonstrated there were differences in accuracy between, “what”, “why”, and “how” questions (Shiferaw et al., 2024). Consider the implications of accuracy and efficacy with the employment of AI with life-or-death situations. As such, explainable AI (XAI), which “refers to a set of methods that support humans in understanding how AI algorithms map certain inputs” becomes the process in which implementors of AI systems are able to demonstrate/explain the learning process directly to the datasets (Senoner et al, 2024, p. 1). Considering that XAI can be labeled as either interpretable or not due to complexity (Senoner et al, 2024), Hartog et al. (2024) show that explainable artificial intelligence (XAI) methods generate degrees of uncertainty and subjectivity in their interpretation. This is critical as XAI models are used to produce more human understandable interpretations of the data (Hartog et al., 2024, Senoner et al, 2024). So, data poisoning begins to assure that the model will present inaccurate results but this is coupled with the already demonstrated probability that the models are already, in some cases, producing suspect responses. Therefore, data poisoning becomes a very serious liability to AI models especially in the case where AI algorithms are unable to be explained due their complexity.

Vulnerabilities in Critical Sectors

Overall implications for various critical infrastructure sectors cannot be understated. Kovacevic et al. (2024) highlight how rapid advancements in AI are presenting, “new opportunities for enhancing efficiency and economic competitiveness across various industries, especially in banking” (p. 1). With a push for implementing AI in an effort to improve operations, it is clear, like all technology, AI represents a new vector in which adversaries can introduce threats. For example, Reserve Bank of India Deputy Governor M Rajeshwar Rao noted how data bias was one of three critical areas of concern (NBFC, 2024). Kovacevic

et al. (2024) demonstrated the dual-use possibilities wherein viable AI solutions can also be used for malicious reasons. While much research focuses on finance and healthcare, the literature increasingly points to significant vulnerabilities in government and critical infrastructure (Rosiek, 2025). Data poisoning in these areas, as discussed by various security analyses (e.g., Delinea, 2025; Certes, 2025), could disrupt essential services and undermine national security.

Current Best Practices and Their Limitations

The review identified several best practices, including data validation and sanitization, red teaming, and secure data handling, as advocated by Korada (2024) and employed by major AI providers. However, the limitations of current mitigation techniques, such as the ineffectiveness of some machine unlearning algorithms against sophisticated attacks (Pawelczyk et al., 2025) and the scalability challenges of adversarial training, remain a significant concern.

Four large generative AI companies were analyzed for their best practices regarding how they keep their datasets secure: OpenAI, Microsoft, Google, and Meta.

OpenAI analyzes the data sources that it pulls from and intermittently observes the responses of the large language model (LLM) to determine if something has happened to the dataset (Korada, 2024). Microsoft uses cryptographic authentication and safeguards the internal components of its AI model. Cryptographic authentication prevents threat actors from poisoning internal datasets or training models. As a result of Microsoft safeguarding the internal components, it makes it harder for threat actors to gain access to the system. Google leverages academic research to counter new and existing problems that may be emerging within the AI security field. Utilizing Zero Trust Content Disarm and Reconstruction (CDR) is also a way that Google attempts to keep its data secure. A CDR is used to validate, repair, and destroy any malicious content that may be uploaded to a dataset. Meta only utilizes patented CDRs to protect its data (Korada, 2024).

General best practices outlined by Korada include data validation and sanitization, red teaming, secure data handling, negative testing, and benchmark testing (Korada, 2024). Data validation and sanitization ensure that the information that is being used by the LLM is not malicious, preventing threat actors from poisoning a dataset or training data. Red teaming is when a team attempts to hack the LLM and make it do things it was not intended to. This allows developers to observe how their current model could be leveraged maliciously. Secure data handling allows only authenticated users with the correct clearances access to the model and training data. Negative testing is when the LLM is given poor data and is observed to see how that data affects the model. Benchmark testing analyzes how the LLM compares to other LLMs of the same magnitude (Korada, 2024).

He, et al., (2024) outlined two different potential defenses that could be used to prevent ICLPoison technique attack: detection-based defense perplexity filter, and preprocessing defense paraphrasing (He, et al., 2024). If a prompt is more complex, the perplexity increases; if there is a high enough perplexity, it could indicate that someone is trying to prompt the LLM maliciously (He, et al., 2024). In terms of the processes used for ICLPoison, synonym replacement was the least likely to increase perplexity, and adversarial suffix was the most likely to increase perplexity. Preprocessing defense paraphrasing is a defense method that rewrites the prompt in a way that is safe for the model to process (He, et al., 2024). When tested for accuracy against ICLPoison it was found to be very effective in targeting adversarial suffix methods but decreased detection of synonym replacement.

Conclusion and Future Research Directions

This systematic review has synthesized current knowledge on AI data poisoning, revealing its multifaceted nature and significant challenges in mitigation. Key themes emerging from the literature include the increasing sophistication of attack vectors, the severe impacts across critical sectors, and the current limitations of defensive strategies. While research has proposed various countermeasures, like Gaussian noise, (Pawelczyk, et al., 2025) a consistent finding across multiple studies (e.g., Pawelczyk et al., 2025; Nguyen et al., 2023) is that no single solution is foolproof, and many defenses struggle with scalability or adaptive attackers. One could argue, this creates 'another chapter' in the challenges of securing systems. However, as highlighted by recent analyses (Delinea, 2025; Certes, 2025; MDPI, 2025 - Enhanced Blockchain-Based Data Poisoning Defense Mechanism), there's a need for defenses that can withstand novel attack strategies and operate effectively in real-world, large-scale systems. For example, this includes exploring blockchain-based integrity verification and advanced identity management for data pipelines. Considering the overall implications on training models with corrupt data, it is clear existing solutions must be rethought.

Explainable AI (XAI) allows for a clear and transparent understanding of how AI models come to conclusions, explicitly stating how the model came to a solution. Through the use of XAI people are able to test the model and how the algorithms are impacting its decision making. (Sultan, 2025). Swarming systems are autonomous systems that make decentralized decisions. As a result of the decentralized nature of swarming systems they are prone to data poisoning attacks (Asadi, M., Rădulescu, R., & Nowé, A. 2025). The researchers proposed a new system for analysis of the autonomous systems, the PADEX framework. The XAI model compares a benign autonomous system to a potentially data poisoned system. This framework simulates the autonomous swarms and uses XAI to analyze the decisions that were simulated, providing insight into what the data poisoning may be impacting (Asadi, M., Rădulescu, R., & Nowé, A. 2025). This framework needs to be researched more to determine its' ability to benefit all sectors and advanced systems. For example, LLMs increasingly accessing real-time online information, strategies to prevent "retrieval poisoning" are crucial (Check Point Research, 2025). Thus, developing stronger ethical guidelines and governance frameworks for data handling and AI model development, as suggested by the OWASP AI Security and Privacy Guide and the EU AI Act discussions (Galileo AI, 2025), is paramount. This includes addressing the creation and distribution of "Dark LLMs" (Check Point Research, 2025). Finally, with respect to longitudinal studies on defense efficacy, more research is needed on the long-term effectiveness of defenses and how they are circumvented over time. As organizations begin to employ AI solutions as part of their daily operations, there is a possibility that they create an acceptance based on an assumed state of data purity. Therefore, in addition to ensuring that data poisoning is not occurring, further research is needed to verify that even if the data is accurate, the model is training and behaving as expected. As such, the entire AI ecosystem from algorithm design, to data entry, to data training, must have methods in place to ensure accuracy is achieved prior to assuming the output created by AI tools is correct. Any failure along this pipeline will result in inaccuracies.

AI Statement

In preparing this research, we utilized advanced Artificial Intelligence (AI) tools to assist in the research process. AL was employed to discover, organize, and synthesize relevant scholarly literature for inclusion in our study to enable a comprehensive review of the existing research. Additionally, AI-based tools were used to check for grammatical accuracy and proper formatting.

References

Alber, D. A., Yang, Z., et al. (2025). Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine, 31*, 618–626. https://doi.org/10.1038/s41591-024-03445-1

Asadi, M., Rădulescu, R., & Nowé, A. (2025). Explainable AI Based Diagnosis of Poisoning Attacks in Evolutionary Swarms. *arXiv preprint arXiv:2505.01181*.

Certes. (2025, April 9). Data poisoning: The hidden threat that could kill your AI. Certes Blog. https://certes.ai/2025/04/09/data-poisoning-threat-kill-ai/

Check Point Research. (2025). AI security report 2025: Understanding threats and building smarter defenses. Check Point Blog. https://blog.checkpoint.com/research/ai-security-report-2025-understanding-threats-and-building-smarter-defenses/

Cotroneo, D., Improta, C., Liguori, P., & Natella, R. (2024). Vulnerabilities in AI Code Generators: Exploring Targeted Data Poisoning Attacking. *Association for Computing Machinery*, 280-292.

Delinea. (2025). The rising danger of AI poisoning: When data turns toxic. Delinea Blog. https://delinea.com/blog/ai-poisoning-when-data-turns-toxic

Galileo AI. (2025, January 17). Safeguarding the Future: A Comprehensive Guide to AI Risk Management. Galileo AI Blog. https://www.galileo.ai/blog/ai-risk-management-strategies

Hartog, P. B. R., Krüger, F., Genheden, S., & Tetko, I. V. (2024). Using test-time augmentation to investigate explainable AI: inconsistencies between method, model and human intuition. *Journal of cheminformatics*, *16*(1), 39. https://doi.org/10.1186/s13321-024-00824-1

He, P., et al. (2025). Multi-faceted studies on data poisoning can advance LLM development. *arXiv*. https://arxiv.org/abs/2502.14182

He, P., Xu, H., Xing, Y., Liu, H., Yamada, M., & Tang, J. (2024, March 28). *Data Poisoning for In-context Learning.* Retrieved from arXiv: https://arxiv.org/pdf/2402.02160

Huang, W. R., Geiping, J., et al. (2020). MetaPoison: Practical general-purpose clean-label data poisoning. *NeurIPS*. https://proceedings.neurips.cc/paper/2020/file/8ce6fc704072e351679ac97d4a985574-Paper.pdf

Koh, P. W., Steinhardt, J., & Liang, P. (2022). Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111(1), 1-47. https://doi.org/10.1007/s10994-021-06119-y

Korada, S., et al. (2024). Vulnerabilities in AI code generators: Exploring targeted data poisoning attacks. *Journal of Cybersecurity, 12*(3), 45–67.

Korada, L. (2024). Data Poisoning -what is it and how it is being addressed by the leading Gen AI providers? *ResearchGate*, 105-109.

Kovacevic, A., Radenkovic, S. D., & Nikolic, D. (2024). *Artificial intelligence and cybersecurity in banking sector: opportunities and risks*

Li, M., Lian, Y., Zhu, J., Lin, J., Wan, J., & Sun, Y. (2024). A Sampling-Based Method for Detecting Data Poisoning Attacks in Recommendation Systems. Mathematics, 12(2), 247. https://www.mdpi.com/2227-7390/12/2/247

MDPI. (2025, April 7). Enhanced Blockchain-Based Data Poisoning Defense Mechanism. Applied Sciences, 15(7), 4069. https://www.mdpi.com/2076-3417/15/7/4069 (Note: This is a specific example of a recent defense study)

NBFC: RBI Dy Governor Flags AI Concerns In Financial Sector. (2024, Feb 23). *Banking Frontiers,* https://reddog.rmu.edu/login?url=https://www.proquest.com/magazines/nbfc-rbi-dy-governor-flags-ai-concerns-financial/docview/2932962770/se-2

Nguyen, D.M., Huynh, T. N., Pham, T., & Tran, A. T. (2023, February 13). *COMBAT: Alternated Training for Near-Perfect Clean-Label Backdoor Attacks.* Retrieved from OpenReview: https://openreview.net/pdf?id=Udho-Hry4RZ

Nguyen, T. T., Quoc Viet hung, N., Nguyen, T. T., Huynh, T. T., Nguyen, T. T., Weidlich, M., & Yin, H. (2024). Manipulating recommender systems: A survey of poisoning attacks and countermeasures. *ACM Computing Surveys, 57*(1), 1-39. https://doi.org/10.1145/3677328

OWASP. (n.d.). OWASP AI Security and Privacy Guide. OWASP Foundation. https://owasp.org/www-project-ai-security-and-privacy-guide/

Pawelczyk, M., Di, J. Z., Lu, Y., Sekhari, A., Kamath, G., & Neel, S. (2025, April 1). *Machine Unlearning Fails to Remove Data Poisoning Attacks.* Retrieved from arXiv: https://arxiv.org/pdf/2406.17216

Rosiek, T. (2025, January 21). AI Data Poisoning, Wiper Malware, Critical Infrastructure Attacks Could Increase in 2025, Impacting Government Cyber Resilience. GovLoop. https://www.govloop.com/community/blog/ai-data-poisoning-wiper-malware-critical-infrastructure-attacks-could-increase-in-2025-impacting-government-cyber-resilience/

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human-AI collaboration. *Scientific reports*, *14*(1), 31150. https://doi.org/10.1038/s41598-024-82501-9

Shen J., & Xia M. (2020). *AI Data poisoning attack: Manipulating game AI of Go.* Retrieved from arXiv: arxiv: https://arxiv.org/ftp/arxiv/papers/2007/2007.11820.pdf

Shiferaw, M. W., Zheng, T., Winter, A., Mike, L. A., & Chan, L. N. (2024). Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. *BMC medical informatics and decision making*, *24*(1), 404. https://doi.org/10.1186/s12911-024-02824-5

Sultan, M. (2025, May 6). *Data Privacy and Security Challenges in AI Systems.* Retrieved from Authorea: https://doi.org/10.22541/au.174655377.75454191/v1