Digital Forensics Meets AI Poisoning: Tracking the Invisible Threats in Training Data

Agnieszka Sawrasewicz
Southern Illinois University, School of Computing
Course: Digital Forensic 2025 CS 513
Instructor: Professor Alvi Ataur Khalil
Fall 2025
https://agnieszkasawrasewicz.com/special-char-forensics/

***Abstract*—As machine learning systems are increasingly trained on large, heterogeneous datasets, data poisoning has emerged as a potent and covert attack vector: adversaries seed training corpora with crafted examples that degrade model behavior, insert stealthy backdoors, or bias model outputs in targeted ways. This paper positions digital forensics as a complementary discipline for detecting, attributing, and mitigating poisoning in AI systems. We survey recent technical work on poisoning and adversarial attacks, synthesize forensic-relevant signals (artifact footprints, provenance metadata, model-level signatures), and propose a practical forensic framework integrating artifact collection, explainability-driven evidence, and chain-of-custody practices for AI models and datasets. We present an experimental plan to evaluate detection and attribution techniques across representative poisoning scenarios (backdoors, subtle perturbations, tokenization-level evasion) and discuss constraints, ethical considerations, and reproducibility artifacts intended for forensic practitioners.**

***Index Terms*—data poisoning, adversarial ML, digital forensics, backdoors, provenance, explainability, federated learning, attribution**

I. Introduction

Modern AI systems rely on massive datasets aggregated from diverse sources. While scale provides capability, it also expands the attack surface: malicious actors can insert poisoned training examples or carefully crafted inputs that later transform model behavior in predictable ways. Unlike traditional malware, poisoning can be invisible in normal operation, activated only in rare contexts or triggered by subtle inputs (e.g., token-level anomalies, special characters, or emoji sequences). Digital forensics — the discipline of evidence collection, timeline reconstruction, and attribution — is well suited to address these threats, but a systematic forensic framework tailored to AI poisoning remains underdeveloped.

This paper asks a focused research question: *can forensic methods be adapted to detect, trace, and attribute poisoning in modern AI pipelines, and how can these methods be operationalized to support incident response and legal chain-of-custody?* We answer by (1) surveying recent advances in poisoning and adversarial forensics, (2) proposing a layered forensic framework (data and artifact-level analysis, explainability-based evidence, and process-oriented chain-of-custody), and (3) defining an experimental methodology for evaluating detection and attribution techniques across attack families relevant to LLMs and other models.

Contributions

This paper makes three contributions:

1) A concise literature synthesis of recent poisoning and adversarial-forensics research (including works on jailbreaks, tokenization attacks, and backdoors) that highlights gaps relevant to forensic practice.
2) A practical forensic framework that integrates artifact collection, provenance verification, explainability-derived evidence, and chain-of-custody procedures targeted at training-data poisoning incidents.
3) An experimental plan and evaluation metrics for testing detection and attribution techniques (including a reproducible artifact list and suggested datasets), aimed at producing operational guidance for investigators and educators.

II. Background and Related Work

The literature on adversarial attacks and data poisoning is growing quickly. Representative recent works include efforts to produce transferable jailbreaks and adversarial suffixes (Zou et al., 2023; Ben-Tov et al., 2025), studies on tokenization-level attacks and special-character evasion (Sarabamoun, 2025; "Bad Characters", 2025), and several papers that examine universal or short-length adversarial attacks (Liao & Sun, 2024; Fu et al., 2025). Systematic surveys (e.g., Cinà et al., 2022) catalog a wide variety of poisoning strategies, while specialized techniques (tensor-decomposition backdoor detection, 2023) propose model-weight-level detection mechanisms.

From a forensics viewpoint, prior work such as *Poison Forensics* (Shan et al., 2021) and FLForensics (2023) are directly relevant: they propose clustering and client-attribution methods for tracing poisoned samples or malicious federated clients. However, most existing defenses focus on detection or mitigation within the ML pipeline and do not provide a full forensic workflow that supports evidence preservation, reproducible attribution, or legal defensibility. Our review identifies three recurring gaps: (1) the lack of provenance-first workflows for dataset ingestion, (2) limited integration of explainability and attribution techniques as admissible forensic evidence, and (3) sparse tooling and guidance for federated or third-party-sourced datasets.

III. Forensic Threat Model and Definitions

We assume an adversary whose objectives may include: stealthy model manipulation (backdoor insertion), targeted misclassification (label-flip or influence attacks), or supply-chain contamination (poisoned third-party data). Capabilities vary: attackers may inject data into public corpora, modify data in third-party APIs, or compromise individual contributors (insider). Our forensic model aims to detect both *direct poisoning* (poisoned samples present in training data) and *latent poisoning* (backdoors latent in weights but rarely triggered).

We assume access to some combination of the following artifacts for a forensic investigation: raw training snapshots, training logs (timestamps, batch composition), data ingestion manifests, model checkpoints, system-level logs for data pipelines, and (potentially) federated client contribution records. The availability of artifacts dictates which forensic methods are feasible.

IV. Forensic Framework: Approaches & Tools

We propose a layered framework that combines technical detection with process-oriented evidence handling.

A. Artifact-level detection

- **Provenance and dataset manifests:** capture signed manifests (hashes, metadata) for all ingested data. Use cryptographic hashing and signed manifests to create tamper-evident ingestion records.
- **Clustering & outlier analysis:** iterative clustering (e.g., as in Poison Forensics) to surface anomalous samples in embedding space or token distribution space; cluster labels can be cross-referenced with source metadata.
- **Tokenization / text-normalization signatures:** detect unusual token-frequency patterns or non-standard unicode usage (special characters, emoji) that may indicate evasion attempts.
- **Backdoor signature detection:** model-weight and activation-space analyses, including tensor-decomposition and activation clustering, to find hidden triggers or unit-level anomalies.

B. Explainability-driven evidence

- **Attribution methods:** influence functions, Shapley-based contribution scores, and gradient-based attribution to estimate training-sample influence on suspect behaviors.
- **Counterfactual testing:** targeted test inputs that probe for latent behaviors and map model responses to minimal triggering inputs — results recorded as forensic evidence.
- **Human-in-the-loop validation:** pair explainability outputs with expert review (linguistic or domain experts) to strengthen the evidentiary value for reporting.

TABLE I: Comparison of representative adversarial attack studies (2023–2025).

Paper	**Attack Type / Focus**	**Core Insight**	**Mitigation / Defense**
Zou et al. (2023)	Universal jailbreak suffixes (GCG)	Gradient-guided token optimization bypasses alignment across models.	Suggests adversarial training; highlights need for stronger filters.
Ben-Tov et al. (2025)	Attention hijacking jailbreaks	Adversarial suffixes dominate shallow `adv→chat` attention paths.	“Hijacking Suppression” reduces attack success 2–10×.
Fu et al. (2025)	Short-length adversarial training	Training on short suffixes defends against long attacks ($\sqrt{M_{test}/M_{train}}$ scaling).	Efficient short-length training keeps >50 % utility.
Sarabamoun (2025)	Special-character encoding attacks	Unicode control + homoglyph obfuscation bypass token filters.	Unicode normalization, encoding validation, adversarial fine-tuning.

C. Log & artifact forensics

- **Timeline reconstruction:** combine ingestion manifests, CI/CD logs, and checkpoint timestamps to build a tamper-evident timeline of dataset origin and model updates.
- **Cross-system correlation:** correlate host-level logs (VM/container IDs, uploader accounts) with dataset provenance to support attribution and potential legal follow-up.

D. Chain-of-custody and evidence sealing

- **Signed snapshots and sealing:** store cryptographically-signed dataset and model snapshots, with documented custody transfers (who had access, when).
- **Forensic packaging:** standardized archive format (manifest + hashes + signature + metadata) for court-ready evidence.

E. Federated Learning Forensics

- **Client fingerprinting:** statistical fingerprints of client updates and contribution patterns to identify anomalous clients.
- **Secure aggregation + accountability:** design trade-offs that preserve privacy yet allow post-hoc attribution (e.g., deferred decryption under legal request).

V. Experimental Plan and Evaluation

We propose an experimental protocol to validate detection and attribution techniques across a taxonomy of poisoning attacks.

A. Attack scenarios

Representative scenarios to evaluate:

1) **Backdoor triggers:** explicit trigger phrases, token sequences, special characters or emoji that flip model output when present.
2) **Label-flip and influence attacks:** mislabeled but plausible examples to bias supervised models.
3) **Subtle perturbation poisoning:** near-indistinguishable perturbations that change model decision boundaries.
4) **Tokenization manipulation:** attacks that exploit or alter tokenizer behavior to evade detection.

B. Datasets and data collection

We will use a mixture of:

- Public benchmark corpora (e.g., OpenWebText-like subsets, GLUE-style datasets for classification tasks) for reproducibility.
- Controlled synthetic corpora where ground-truth poisoning labels are known (to measure attribution accuracy).
- Federated learning emulation (simulated clients with benign and malicious contributions).

C. Detection baselines and metrics

Baselines:

- Clustering-based outlier detection (Poison Forensics style).
- Model-weight decomposition methods (tensor-decomposition backdoor detection).
- Explainability-derived influence scoring (influence functions, gradient-based).

Metrics:

- Detection performance: True Positive Rate (TPR), False Positive Rate (FPR), precision, recall, AUC.
- Attribution accuracy: proportion of true poisoned samples correctly identified and source attribution precision.
- Forensic utility: evidence completeness (fraction of required artifacts available), timeliness (time-to-detection), and reproducibility score (are results replicable given artifacts).

D. Analysis plan

We will evaluate detection methods across attack types and poisoning fractions, report ROC curves, and measure attribution accuracy against known ground truth. We will also conduct small user studies with domain experts to evaluate whether explainability outputs increase confidence in forensic claims.

Framework Overview and Experimental Plan Summary

1. Forensic Framework (Proposed Approach). This research introduces a structured method for investigating AI poisoning incidents—a digital forensics blueprint for machine learning systems. The framework integrates several layers:

- **Artifact-level detection:** Identify anomalies in datasets, logs, or model weights.
- **Explainability-driven evidence:** Use SHAP values, influence functions, or attention maps to reveal why a sample or model behavior appears poisoned.
- **Chain-of-custody:** Maintain cryptographically signed records and hashes so forensic evidence remains verifiable and admissible.
- **Federated forensics:** Extend the framework to distributed systems where multiple clients contribute to model training.

This layered approach enables investigators to detect, trace, and preserve AI-poisoning evidence with both technical rigor and legal defensibility.

2. Experimental Plan (Evaluation Design). To validate the framework, we define a structured testing plan that includes:

- **Attack scenarios:** Evaluate on representative cases—backdoor triggers, label-flip, subtle perturbation, and tokenization manipulation.
- **Datasets:** Use both public benchmarks and synthetic datasets with known ground truth for controlled experimentation.
- **Metrics:** Measure detection rate, false positives, attribution accuracy, and reproducibility of results.
- **Analysis:** Compare clustering, tensor-decomposition, and influence-based detection techniques across attack categories.

This combination of framework design and experimental plan forms the core of the proposed research under Option 1, ensuring both conceptual innovation and empirical evaluation.

VI. Ethics, Legal, and Reproducibility Considerations

Working with poisoned data has ethical and legal implications. Our protocol emphasizes:

- Use of researcher-controlled poisoned datasets or publicly shared testbeds; never deploying live attacks on third-party systems.
- IRB review for any human-subjects (expert validation or user studies).
- Documentation of chain-of-custody and evidence sealing to allow legal defensibility.
- Reproducibility: release of synthetic datasets, analysis scripts, and detailed artifact manifests where possible (omitting any sensitive or privacy-protected content).

VII. Implementation Roadmap and Deliverables

Planned milestones (8–12 weeks per milestone, modular):

1) **Month 1–2:** Build reproducible dataset pipeline and generate controlled poisoned corpora (backdoors, label flips, tokenization evasion).
2) **Month 3–4:** Implement detection baselines (clustering, tensor-decomposition, influence-scoring) and instrument training pipelines to capture ingestion manifests and logs.
3) **Month 5:** Run detection experiments, produce metric reports, and refine forensic packaging procedures.
4) **Month 6:** Conduct small expert validation study and prepare reproducibility artifacts (scripts, synthetic data, documentation).

VIII. Comparative Analysis of Recent Adversarial Attack Studies

Table II compares four representative papers that shaped current understanding of jailbreaks, tokenization attacks, and adversarial robustness in large language models (LLMs). Each study targets a different layer of the threat spectrum—from universal suffix attacks to character-level evasion and alignment-aware defenses.

TABLE II: Comparison of Key Adversarial Attack Papers (2023–2025)

Paper	Attack Type / Focus	Core Insight	Defense Proposed
Zou et al. (2023)	Universal jailbreak suffixes (*GCG*)	Gradient-guided token optimization creates transferable prompts bypassing safety alignment.	Suggests adversarial training; notes gaps in safety filtering.
Ben-Tov et al. (2025)	Attention hijacking in jailbreaks	Jailbreaks exploit shallow `adv→chat` attention; success scales with hijacking strength.	*Hijacking Suppression:* rescales dominant vectors, cutting attack rate by 2.5–10×.
Fu et al. (2025)	Short-length adversarial training	Short suffixes defend against longer ones ($\sqrt{M_{test}/M_{train}}$ scaling).	Efficient training retains over 50% model utility.
Sarabamoun (2025)	Encoding / homoglyph attacks	Unicode and encoding obfuscation bypass token filters; size $\neq$ robustness.	Layered defense: normalization, validation, fine-tuning, anomaly detection.

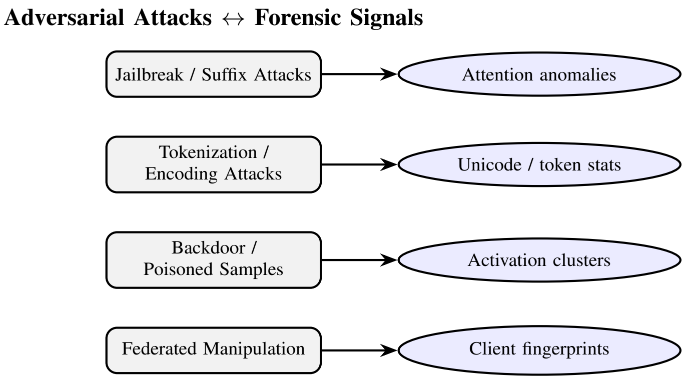
Adversarial Attacks ↔ Forensic Signals
Jailbreak / Suffix Attacks
Attention anomalies
Tokenization / Encoding Attacks
Unicode / token stats
Backdoor / Poisoned Samples
Activation clusters
Federated Manipulation
Client fingerprints

Fig. 1: Compact mapping of attack categories to primary forensic indicators. Optimized for IEEE single-column layout.

Synthesis

Collectively, these works show a continuum of vulnerabilities: Zou et al. expose the existence of universal suffixes; Ben-Tov et al. explain their internal mechanism; Fu et al. propose scalable adversarial training to resist them; and Sarabamoun (2025) broadens the threat surface to tokenization and encoding. Forensic analysts can leverage these insights to design feature extraction pipelines that monitor attention-flow anomalies, adversarial token patterns, and encoding irregularities during dataset ingestion or model audit.

IX. Forensic Implications of Jailbreak and Tokenization Attacks

The recent wave of adversarial studies extends beyond theoretical model vulnerabilities—they reveal new forensic dimensions for tracking, attribution, and evidentiary reconstruction of compromised AI systems. Jailbreak, suffix-based, and character-level attacks all leave distinctive traces in text pipelines, tokenizers, and model activations that can serve as forensic indicators.

A. Jailbreaks as Behavioral Evidence

Universal jailbreak suffixes (e.g., "`adv→chat`" patterns) manifest reproducible token-sequence artifacts that can be detected through:

- **Prompt lineage reconstruction:** forensic tracing of conversation logs to identify recurring jailbreak substrings or token co-occurrence clusters.
- **Attention-map anomalies:** abrupt redistribution of attention weights toward suffix tokens; consistent patterns can be logged as behavioral fingerprints.
- **Temporal correlation:** timestamps of anomalous prompts can reveal coordinated attack campaigns or testing events within system logs.

These features provide measurable behavioral evidence that distinguishes intentional jailbreaks from benign long-context prompts.

B. Tokenization and Encoding Forensics

Character-level and encoding-based attacks introduce forensic challenges similar to digital steganography. Potential forensic entry points include:

- **Tokenizer residue analysis:** identification of unknown or fragmented tokens caused by abnormal Unicode or Base64 sequences.
- **Unicode anomaly profiling:** histogram analysis of code-point ranges and directionality (e.g., right-to-left override, zero-width injection).
- **Decoding-path reconstruction:** cross-verification of raw input logs with post-decoding text to detect hidden payloads or automatic decoding performed by preprocessing layers.

Such indicators can be used to reconstruct attack chains and preserve adversarial artifacts as admissible forensic evidence.

C. Forensic Attribution Opportunities

The forensic perspective transforms these vulnerabilities into opportunities for traceability:

- **Signature databases:** build repositories of known jailbreak or encoding attack patterns for automated matching during investigations.
- **Provenance tagging:** embed cryptographically verifiable hashes of sanitized tokenizer vocabularies or decoder versions to link outputs to specific model builds.
- **Explainability correlation:** pair SHAP or influence-function visualizations with artifact timestamps to demonstrate causal relationships between poisoned inputs and unsafe outputs.

D. Forensic Integration in AI Pipelines

Integrating these insights into forensic workflows enables continuous monitoring:

1) **During ingestion:** detect and quarantine inputs with abnormal tokenization or encoding behavior.
2) **During inference:** log token-level embeddings and attention distributions for later forensic review.
3) **During incident response:** apply attribution and decoding reconstruction to confirm whether unsafe model outputs stem from adversarial prompts or internal corruption.

X. Discussion and Future Research Directions

The intersection of digital forensics and adversarial machine learning opens several promising avenues for interdisciplinary research. Building on our framework and the reviewed literature, we identify key directions that can extend the technical and investigative capabilities of forensic AI systems.

A. Integrating Adversarial Detection with Forensic Pipelines

Future systems should embed adversarial-prompt detection modules within forensic logging layers. A continuous auditing loop—where jailbreak or tokenization anomalies are automatically logged, clustered, and cross-referenced with provenance metadata—could enable **live forensics** of model behavior. Such systems would treat each inference as a potential forensic event, producing verifiable traces suitable for later attribution or legal inspection.

B. Explainability as Evidence

Explainable AI methods remain underutilized in formal investigations. Future research should formalize the admissibility of explainability outputs (e.g., SHAP heatmaps, attention saliency maps) as forensic artifacts. Establishing standards for **explainability-derived evidence** would allow investigators to present model behavior in a transparent and reproducible manner.

C. Hybrid Detection Models

Combining security analytics with digital forensics may yield robust hybrid detectors:

- **Attention-pattern signatures** extracted from jailbreak datasets to flag adversarial token flows.
- **Provenance-aware anomaly detection** that correlates model activations with dataset origin.
- **Cross-modal correlation** linking text, image, or code-generation anomalies to shared poisoning indicators.

Such multi-view systems could detect poisoning at both training and inference time, reducing the latency between compromise and forensic visibility.

D. Legal and Ethical Frameworks

As AI forensics matures, integration with digital-evidence standards becomes critical. Future work should explore how established frameworks such as ISO/IEC 27037 and NIST SP 800-86 can be extended to include **AI artifact preservation**, covering model checkpoints, tokenizer configurations, and adversarial test records. Cross-jurisdictional collaboration will be essential to ensure that AI forensic evidence maintains integrity and privacy compliance.

E. Longitudinal and Cross-Model Studies

Current robustness measurements (e.g., Sarabamoun 2025) show large variation across model architectures. A coordinated longitudinal study—tracking model updates, tokenizer revisions, and retraining cycles—could map how adversarial and forensic resilience evolve over time. This would provide empirical baselines for the durability of defense mechanisms.

Summary

The next phase of research should treat adversarial attacks not only as security risks but as forensic opportunities. By combining anomaly detection, provenance tracking, explainability, and legal process modeling, AI forensics can move from reactive investigation to proactive, auditable assurance.

XI. Conclusion

This paper advances a forensic-first perspective on artificial intelligence security, positioning digital forensics as a complementary discipline to adversarial defense research. Through analysis of poisoning, jailbreak, and tokenization attacks, we emphasize that the same mechanisms enabling LLM manipulation also create identifiable forensic signatures—token irregularities, encoding patterns, and behavioral traces—that can support evidence collection and attribution.

Our proposed framework integrates four pillars: provenance-aware data handling, model-level signature analysis, explainability-derived evidence, and chain-of-custody protocols adapted for AI artifacts. Together, these components form the basis for a reproducible, legally defensible approach to investigating and mitigating AI poisoning incidents.

Looking forward, future work will operationalize this framework through empirical validation, cross-model benchmarking, and forensic toolchain development. Special emphasis will be placed on hybrid detectors that fuse adversarial analytics with provenance logs, formal standards for explainability as evidence, and longitudinal monitoring of model robustness. By bridging digital forensics and AI safety, we aim to transform reactive incident analysis into proactive, auditable assurance of trustworthy machine learning systems. @articlezou2023, title = Universal and Transferable Adversarial Attacks on Aligned Language Models, author = Zou, Andy and Wang, Zifan and Carlini, Nicholas and Nasr, Milad and Kolter, J. Zico and Fredrikson, Matt, year = 2023, eprint = 2307.15043, archivePrefix = arXiv, primaryClass = cs.CL, url = https://arxiv.org/abs/2307.15043

@miscliao2024, title = AmpleGCG: Learning a Universal and Transferable Adversarial Suffix Generator, author = Liao, Zeyi and Sun, Huan, year = 2024, eprint = 2404.07921, archivePrefix = arXiv, primaryClass = cs.CL, url = https://arxiv.org/abs/2404.07921

@articlefu2025, title = "Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence, author = Fu, Shaopeng and Ding, Liang and Wang, Di, year = 2025, eprint = 2502.04204, archivePrefix = arXiv, primaryClass = cs.LG, url = https://arxiv.org/abs/2502.04204

@articlebentov2025, title = Universal Jailbreak Suffixes Are Strong Attention Hijackers, author = Ben-Tov, Matan and Geva, Mor and Sharif, Mahmood, year = 2025, eprint = 2506.12880, archivePrefix = arXiv, primaryClass = cs.CL, url = https://arxiv.org/abs/2506.12880

@articlesarabamoun2025, title = Special-Character Adversarial Attacks on Open-Source Language Models, author = Sarabamoun, Ephraiem, year = 2025, eprint = 2508.14070, archivePrefix = arXiv, primaryClass = cs.CR, url = https://arxiv.org/abs/2508.14070

@inproceedingsshan2021, title = Poison Forensics: Traceback of Data Poisoning Attacks in Neural Networks, author = Shan, Shawn and Bhagoji, Arjun Nitin and Zheng, Haitao and Zhao, Bo and Zhao, Ben Y., year = 2021, note = arXiv preprint / USENIX follow-ups, eprint = 2110.06904, archivePrefix = arXiv, primaryClass = cs.LG, url = https://arxiv.org/abs/2110.06904

@articlecina2022, title = Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning, author = Cinà, Antonio Emanuele and Grosse, Kathrin and Demontis, Ambra and Vascon, Sebastiano and Zellinger, Werner and Moser, Bernhard A. and Oprea, Alina and Biggio, Battista and Pelillo, Marcello and Roli, Fabio, year = 2022, journal = ACM Computing Surveys (accepted / arXiv), eprint = 2205.01992, archivePrefix = arXiv, primaryClass = cs.LG, url = https://arxiv.org/abs/2205.01992

@articletensor2023, title = TEN-GUARD: Tensor Decomposition for Backdoor Attack Detection in Deep Neural Networks, author = Hossain, Khondoker Murad and Oates, Tim, year = 2024, eprint = 2401.05432, archivePrefix = arXiv, primaryClass = cs.LG, url = https://arxiv.org/abs/2401.05432

@articleadvtok2025, title = Adversarial Tokenization, author = Geh, Renato Lui and Shao, Zilei and Van den Broeck, Guy, year = 2025, eprint = 2503.02174, archivePrefix = arXiv, primaryClass = cs.CL, url = https://arxiv.org/abs/2503.02174

@inproceedingsbadchars2025, title = Bad Characters: Imperceptible NLP Attacks, author = Boucher, Nicholas and Shumailov, Ilia and Anderson, Ross and Papernot, Nicolas, year = 2022, booktitle = Proceedings of the 43rd IEEE Symposium on Security and Privacy (SP), pages = 1987–2004, doi = 10.1109/SP46214.2022.9833641, url = https://doi.org/10.1109/SP46214.2022.9833641

@articleemoji2025, title = Interpreting How Emojis Trigger LLMs' Toxicity, author = Authors, year = 2025, note = preprint (arXiv) — see URL, eprint = 2509.11141, archivePrefix = arXiv, url = https://arxiv.org/abs/2509.11141

@articleflforensics2023, title = Tracing Back the Malicious Clients in Poisoning Attacks to Federated Learning, author = Jia, Yuqi and Fang, Minghong and Liu, Hongbin and Zhang, Jinghuai and Gong, Neil Zhenqiang, year = 2024, eprint = 2407.07221, archivePrefix = arXiv, primaryClass = cs.CR, url = https://arxiv.org/abs/2407.07221

REFERENCES