Open Source Model Security Risk Detection for AI Trustworthy

*

Zelalem.S
Researcher AI, Quantum
Nov. 2025

***Abstract*—Open source models are publicly available possessing risks that adversaries can exploit the model locally and then use the crafted attacks to deceive it. AI models can lead to death and life decisions. It needs a cost benefit analysis on open sources to verify and evaluate models, datasets and algorithm security risks. Attacks extract sensitive information about the training data, manipulate the model's training significantly to influence its decision.**

AI algorithms are black boxes which are difficult to understand including for the creators of a model. Mission critical agent developers need offer explainable interpretation for their model outcomes and decisions in order to be trusted and accepted claiming its internal algorithm working logic.

AI models must be interpreted in features and decision making logic easier for understanding and trust how the model reaches to its decision. In this paper, we propose a novel Anewari Adenguari intrepretable model design to verify and evaluate open source models security bench marking IBM adversary tool box, Google large language model auditor and OWASP top ten AI risks. Anewari and Adenguari are local words, the first is the one which finds model vulnerability while the second countermeasures it.

Triple security check algorithm which identifies security vulnerability from the model, justify decisions through explaining interpretation and trust worth by correcting security issues. The prototype test experiment for proposed design and algorithm demonstrates the how to apply the triple security checks.

***Index Terms*—open source, critical application, model, algorithm, AI, trustworthy, LLM, Interpret, claims**

I. Introduction

One of the primary AI security threats is adversarial attacks [6] where malicious actors manipulate models by subtly altering input data, leading to incorrect predictions and decisions. Research to develop adversarial defense mechanisms, including robust training methods, anomaly detection, and verification techniques to enhance system resilience against attacks is demanding now a days.

The reliance on AI in Cybersecurity creates risks as attackers can exploit weaknesses in algorithms. They poison it by training datasets or they use AI driven malware to bypass security defenses. Implementing AI explain-ability [7] and checking to ensure AI driven security system is trustworthier.

Government and industry bodies develop AI governance frameworks to regulate the ethical use of AI and mitigate risks [8]. European Union's AI Act and the NIST AI Risk Management Framework aims to establish guidelines [9] for AI transparency, accountability, and security.

Model theft [10] protection in the increasing deployment of AI models across various sectors ensures the confidentiality and integrity of proprietary. Attackers attempt to reverse engineer the models, steal proprietary training data, or extract valuable intellectual property. To mitigate these threats, secure model deployment methods and model protection strategies safeguard critical AI assets.

AI driven automation [11] is being used in sectors such as healthcare, finance, and transportation where security breaches can have severe consequences. AI security measures in these domains include real-time anomaly detection, robust access controls and AI driven risk assessment frameworks.

AI driven misinformation and deepfake [12] threats are posing serious security concerns. Malicious actors are leveraging AI generated content to spread misinformation, create hyper realistic deepfakes, and manipulate public perception. Developing AI powered detection tools that can identify manipulated content can be used to combat these threats.

Looking ahead the future of AI security will be defined by continuous advancements in AI driven threat intelligence [13], ethical AI practices will be essential in mitigating risks and ensuring the safe and responsible deployment of AI technologies and collaboration play a crucial role in shaping a secure AI driven future [14]. But AI complex algorithms and dependence on vast datasets introduces unique security challenges. Unlike traditional software systems, AI systems are adaptive, data driven, and capable of learning from inputs, making it difficult to predict and evaluate security risks. As AI becomes increasingly integral to decision making

processes, ensuring the security of AI systems is paramount to prevent misuse, adversarial and unintended consequences.

Evaluating and ensuring AI model algorithms weather applications are secure from external and internal threats is a priority. These tests assess the system's vulnerability to various attack vectors, including data poisoning, adversarial attacks, model inversion, and privacy breaches. In this context, understanding the challenges of AI security testing [15] is essential for ensuring AI technologies remain safe, trustworthy and ethical.

Attacks can be subtle and imperceptible, making them difficult to defend against and test for. Small changes to the input data, often imperceptible to humans, can cause AI models to misclassified or misinterpret the data leading to incorrect decisions. These attacks can be used to deceive AI models in critical applications.

Adversarial attacks are transferable, the attack that works on one AI model may also work on another model, even models trained differently. This presents a significant challenge for security testing since attackers can exploit vulnerabilities. AI models often rely on large volumes of data including sensitive personal information to learn and make decisions. AI systems become more integrated into everyday life, ensuring that data is securely collected, stored and processed is critical. Breaches or misuse of this data can lead to serious privacy violations.

Data poisoning [16] occurs when an attacker manipulates the training data, corrupting the AI model's learning process and leading it to make faulty decisions. This can have devastating consequences, especially in domains like healthcare where AI systems make life or death decisions.

Model inversion attacks [17] can reconstruct sensitive data from the trained AI model, exposing private information that was used during the training phase. AI models are black boxes due to their lack of transparency and difficulty in explaining their decision making processes. This lack of explain-ability [18] poses significant challenges for AI security as it becomes difficult to understand why a model makes certain predictions and harder to detect when the model is being attacked or manipulated. The lack of transparency makes it challenging for security to identify vulnerabilities or potential exploitation points.

Attacks could lead to a loss of confidence in AI systems, particularly in applications where reliability and trust are critical. AI systems to be widely adopted, stakeholders must trust the models' decisions. Because security flaws, undetected attacks or manipulation can undermine this trust and prevent the adoption of AI in critical areas [19].

II. RELATED WORKS

The traditional Cyber security vulnerabilities testing goal is to fix security bugs and gaps executed by Cybersecurity professionals using security tools and skills. But AI vulnerabilities will require a different set of tools, strategies and approaches. Because the assets in AI are datasets and models. Like traditional CIA triad, confidentiality attacks, integrity attacks and authenticity attacks, the AI attacks are executed to access, change or manipulate datasets and models.

The limitations in the AI algorithms is that adversaries can exploit in order to make the system fail due shortcomings of the current state of the art methods. Dataset model is the only source of knowledge, if it is corrupted by attacker, the model learns wrongly towards attackers intention that tricks it in decision making process.

Developing secure and robust AI models to ensure privacy and security of sensitive data [1] has been emphasized. The paper conducted a survey highlighting different types of AI attacks and risks. Recommended to develop robust and secure AI models that are resilient implementing new detection techniques and security measures. Improving the explain-ability and transparency of AI models, to enhance trust and accountability.

The research aim was to improve security and reliability of AI models to foster trust in AI technologies. The paper [3] used comprehensive literature review and experimental validations using the CIFAR-10. Anomaly detection, optimization strategies, and ensemble learning to identify and mitigate the effects of poisoned data model trained. Based on the research, data poisoning significantly degrades the model performance by 27 percent in image recognition tasks. And 22 percent in fraud detection models from Insurance Claims dataset. It restored accuracy levels by an average of 15-20 percent and it demonstrated that ensemble learning techniques provide an additional layer of resilience, reducing false positives and false negatives. Threats posed by data poisoning degrades AI performance both in image classification and fraud detection. But its ensemble technique creates additional layer performance headaches for the purpose of security.

The methodology is designed to contributing to AI security by systematically assessing poisoning attacks and their countermeasures. Experiment conducted in controlled poisoning attacks in image classification and fraud detection models. The main limitation claimed is dataset specific findings. Rather, a generic applicable agent that is interpretable with trust bases decision is our research focus

Logical decision assistance is one of still unanswered cybersecurity issues [2]. Technology understanding while making decisions, the how and why helps to know risks and correct accordingly. So that interpretation management significantly improve systems' cybersecurity capabilities

minimizing risk costs.

Here under are the gaps in related works [3], [2], [4], [5].The research gap of which this research focus are, i. Identify open source model security risks for mission critical AI agent applications. ii. Design security mechanism to evaluate open source AI models and datasets for critical agents. iii. Show a prototype of implementation in triple check security to embed in critical agents. iv. Evaluate by interpreting and justification on decision process of models.

Proposing a novel research on open source model triple check of 'Anewari', 'Adenguari' (A1, A2) in which A1 finds security error claims and A2 correct the claim to boost critical agent trust confidence.

III. OPEN SOURCE MODEL SECURITY RISKS

Unverified dataset source Untrustworthy sources or datasets that are not verified are at higher risk of being compromised. An adversary could poison data and publish it.

Unverified model source Adversaries can publish models for victims to use. Later use their knowledge of the model to perform attacks on it. Models from suspicious sources be pre trained on poisoned data.

Pipelines Insider in depth knowledge about its architecture, weights, and training data. Knows about the security mechanism to protect the model. Malicious insiders then serve as a threat to the ML pipeline.

Evasion: Attacks input data to cause a trained model to classify it.

Extraction: Adversary attempts to build a model that is similar or identical to a victim model by accessing to the original model.

Inference: Attacks generally aim at reconstructing to train the victim model. Have access to data used to train the model.

Poisoning: To corrupt the victim model during training. The Fig 2 illustrates adversarial risks in addition to the OWASP.

Transfer Learning	Misleading, incorrect results. Breach of sensitive information in the training dataset.	Secure storage and sharing of pre trained models, Proper data protection measures for the pre-trained models and training dataset.
Model Skewing	Lead to incorrect decisions on the output of the model	Accurately reflect the underlying distribution of the training data
Output Integrity	Loss of confidence in the model's predictions and results	Proper authentication and authorization measures to ensure the integrity of the inputs and outputs, Adequate validation and verification of inputs and outputs
Model Poisoning	Model's predictions manipulated, Confidential information can be extracted	Access controls to the model's code and parameters, Proper secure coding practices.

Risk	**Impact**	**Recommended Security**
Input Manipulation	input data to mislead the model	Adversarial training, Use robust designed models, Input validation
Data Poisoning Attack	model to behave undesirable	Validate and verify data before training
Model Inversion Attack	Confidential information of input data can be compromised	Access control, Input validation, Model transparency, Regular monitoring, Model retraining
Membership Inference	Incorrect model predictions, Loss of confidentiality and privacy	Model training on randomized or shuffled data, Model Obfuscation, Regularization
Model Theft	Reputation of organization	Secure model development
AI Supply Chain	Machine learning project compromised	Not relying on untrusted third party code

Fig. 1. owasp top 10 risk

IV. DESIGN

A. Benchmarks

1. IBM Adversarial Robustness Toolbox

Open source Python framework for machine learning security. Supports for tasks of classification, regression and generation working for all data types. MNIST Modified National Institute of Standards and Technology 60,000 samples and a test set of 10,000. The dataset is divided into training and testing sets, making it suitable for training and evaluating machine learning algorithms.

2. Google Large Language Model LLM Auditor

This is designed to evaluate and enhance the factual grounding of responses generated by Large Language Models (LLMs).It has verifiable claims within the text and internal knowledge to determine their accuracy. The agent evaluates and improves the factual grounding of responses generated by LLMs using its critic agent and reviser agent. We context ancient Ethiopian builders termed in Amharic as 'Anewari and Adenguari' mechanism. Anewari is the one who finds security error and Adenguari is who gives a correction on it. Both are different skills to be used in the prototype test experiment.

3. OWASP AI Model Top Security Risks

The image in Fig 1 shows OWASP [20] AI model security risks.

B. Architecture

The architecture shows the pre-processing security, interpretation, training, model and dataset tasks while applying open sources in mission critical application agents. Have a look in Fig 3.

V. VERIFICATION AND EVALUATION

A. Trustworthy

AI security testing incorporates to interpret and explain models. Techniques help to explain how AI models make

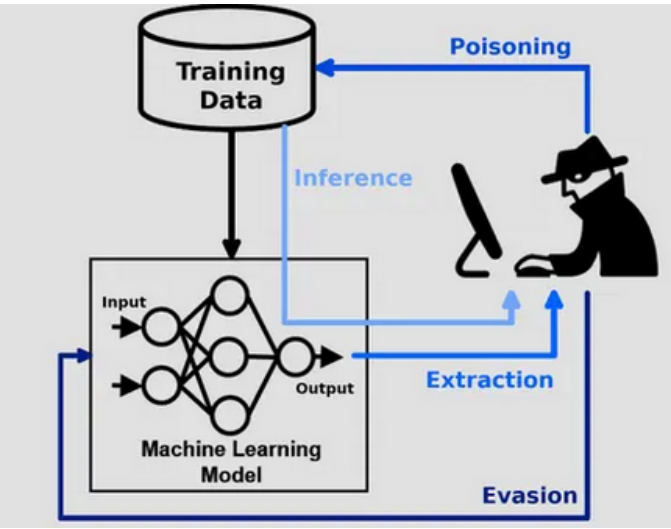

Fig. 2. Risk

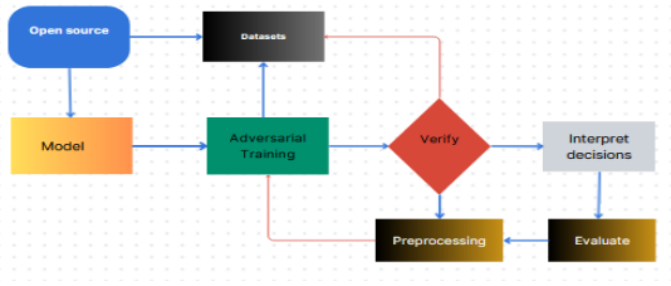

Fig. 3. Vrification, evaluation,interpretation

predictions through adopting model audit and monitoring practices to ensure model's behavior is aligned with expectations to be examined for potential risks. Robustness testing conducted to evaluate the model's ability to handle novel, adversarial, or evasive inputs. It is implemented in systems to detect anomalies for model's prediction to indicate tampering through ensuring the trustworthiness and reliability over time.

Mission critical developers focus on open source dataset and algorithm to enhance model performance or accuracy, security even not yet considered while validating models. This greatly impacts applications operation making wrong decisions like a doctor's decision in saving the patient's life. But detection and evaluation of security is becoming as critical as enhancing performance and accuracy in era of AI.

B. Verification Through Explainable Interpretation

A1 (Anewari) always finds security errors against the open source data, model or algorithm and justify it by interpretation. A2 (Adenguari) corrects in such a way it should behave reasoning the claim adding trust confidence. A1 and A2 are in relations with adversary detection, dataset and decision process evaluation through simulated prototype illustrated in Fig 4. To implement a prototype of A1,A2 demonstrating by practical scenario cases in CNN, MNIST and google LLM bench marks. The A1 traces the why or why not and how reached for decisions to be trusted justifying its security from adversarial manipulation not to behave out of its intention. Our algorithm finds security problems, justify through interpretation of claims and then correct in such a way a model must behave.

Interpretation is understanding the models inner working used to correlate training data to check by detection. And then explain to develop trust providing useful explanations for users of the critical agent that are dynamic to adapt for security demands. Let us assume the scenario of Google large language model reasoning process. In our case, let us context it as benchmark.

Claim: Earth is further away from the Sun than Mars.
A1:Identify and verify the claim as Interpreted justification: The average distance of Earth from the Sun is about 150 million kilometers, while the average distance of Mars from the Sun is about 228 million kilometers. Thus, Mars is farther from the sun than the Earth.
The claim made in the answer is demonstrably false based on available astronomical data searched.
A2: Mars is further away from the Sun than Earth. Corrected through searched facts. Likewise,

How the model performs on what it was trained is adversarial training test of vulnerabilities.
Claim 1: Model from a reputable source, others have used or reviewed it Claim 2: No dangerous serialization arbitrary code is embedded

Claim 3: No suspicious patterns or backdoor triggers show up under stress testing
Claim 4: It generalizes well and doesn't "remember" specific inputs from training
Claim 5:Its license is permissive
A1: Identify and verify claims 1,2,3,4,5, justify and interpret.
A2: Claims checked and verified so that model has high confidence of trust
Data must be complete, diverse, realistic, verifiable, clean and split into sub samples. Accordingly

Claim 1: There are many nulls or anomalies missing and corrupt data
Claim 2: There are duplicate rows
Claim 3: There are outliers,inconsistencies
Claim 4: There are verifiable dataset source lists
A1:Identify claims 1, 2, 3, 4, justify and interpret
A2:Claims verified so that data has high confidence of trust.

C. Detecting By Triple Checks

Let us show codewise demonstration in Fig 6. Based on the detection and evaluation illustration in Fig 5. A1 analyzes

Inference	Misclassification, confidence	Retraining, balancing
Evasion	Low confidence adversarial input	Adversarial training
Extraction	Confidence skew (indirect)	Add noise, limit API addons
Poisoning	Class imbalance	Dataset cleansing, retraining

Fig. 4. A1A2

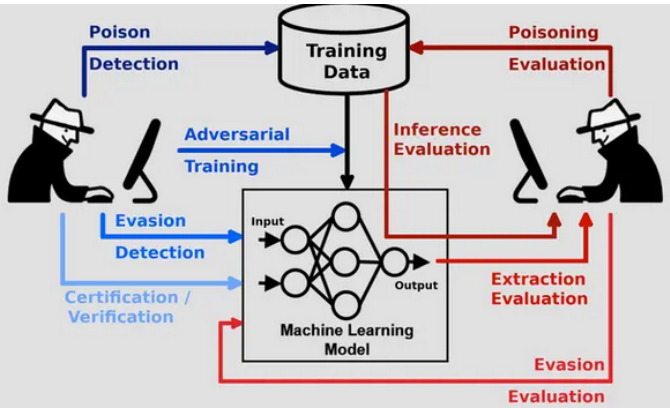

Fig. 5. Detection

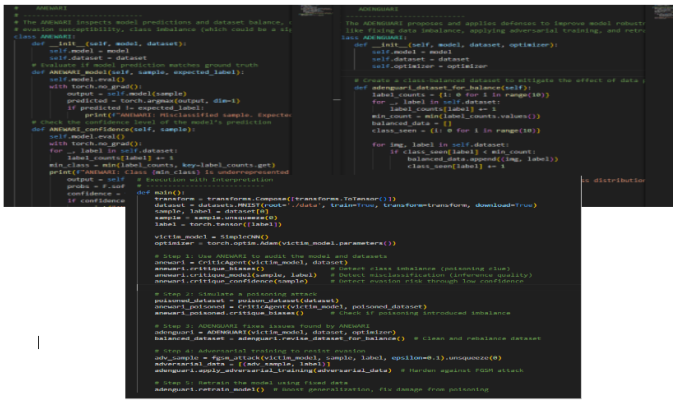

Fig. 6. Prototype

the model from the implementation for inference quality with low confidence evasion risk, and dataset imbalance thats a poisoning indicator.

A2 applies balancing it through adversarial training and retraining to mitigate attacks. Triple checking is done to verify and evaluate listed claims. Accordingly, CNN model and MNIST dataset security can be assured by A1 and A2.

Claim 1: Evaluate if model prediction matches ground truth

Claim 2: Check the confidence level of the model's prediction

Claim 3: Analyze dataset class balance identifying poisoning

Claim 4: Simulate poisoning by flipping a subset of one class to another

Claim 5: Generate adversarial samples to test model evasion vulnerability

A1:Justify the claims 1, 2, 3, 4,5 A2:Claims detected and corrected so that model strength is assured with high confidence of trust.

This triple security check algorithm implemented with mission critical AI applications. Hence open source security risks and the matter of trust worthy AI verified.

As demonstrated, A1 observes model and dataset, detects anomalies, biases, and weaknesses. And A2 applies defenses balancing classes, adversarial training, and retraining.

VI. CONCLUSION

AI agents decision is a blackbox process that are very difficult to know by a human. Unless security verified and evaluated with innovative test mechanisms, specially for critical agent applications model decisions has high probability impact while they are operational.

The research milestones careful model verification and evaluation that are applicable for critical application proposing A1 and A2 for trustworthy which can be applicable and to be used in choosing open source models while critical agent development.

Future work is to release the triple check algorithm to be embedded with critical agents. So that model algorithms and datasets are clean, interpreting and deployable in pipelines to remotely accessible.

REFERENCES

[1] Md Mostafizur Rahman, Aiasha Siddika Arshi, Md Mehedi Hasan, Sumayia Farzana Mishu, Hossain Shahriar, Fan "Security Risk and Attacks in Artificial Intelligence (AI): A Survey of Security and Privacy", June 2023.

[2] Rammanohar Das and Raghav Sandhane, "Artificial Intelligence in Cyber Security", 2021 J. Phys.: Conf. Ser. 1964 042072

[3] Halima I.Kure,Pradipta Sarkar, Ahmed B.Ndanusa, "Detecting and Preventing Data Poisoning Attacks on AI Models", 2025

[4] RIA SINHA, Artificial Intelligence In Cyber Security, IJCRT,2023

[5] MARIAM ALDHAMER, "The Impact of Artificial Intelligence on the Future of Cybersecurity", Jan 2023.

[6] Shilin Qiu, Qihe Liu, Shijie Zhou and Chunjiang Wu, "Review of Artificial Intelligence Adversarial Attack and Defense Technologies", 4 March 2019, semanticscholar

[7] Prashant Gohel, Priyanka Singh, Manoranjan Mohanty, "Explainable AI: current status and future directions",12 Jul 2021, arXiv.

[8] Liron Pantanowitz, Matthew Hanna, Joshua Pantanowitz, Joe Lennerz, Walter H. Henricks, Peter Shen, Bruce Quinn, Shannon Bennet, Hooman H. Rashidi, "Regulatory Aspects of Artificial Intelligence and Machine Learning", December 2024, ScienceDirect.

[9] Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework", 26 Jan 2024, arxiv

[10] Peihao Li, Jie Huang, Shuaishuai Zhang, Chunyang Qi, SecureEI: "Proactive intellectual property protection of AI models for edge intelligence", December 2024, ScienceDirect.

[11] Lampis Alevizos, Vinh Thong Ta, "Automated Cybersecurity Compliance and Threat Response Using AI, Blockchain and Smart Contracts", 12 Sep 2024, arxiv

[12] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, Dacheng Tao, "Deepfake Generation and Detection: A Benchmark and Survey", 26 Mar 2024.

[13] Shrit Shah, Fatemeh Khoda Parast, "AI-Driven Cyber Threat Intelligence Automation", 26 Oct 2024, arxiv.

[14] https://www.ibm.com/think/insights/artificial-intelligence-future, Accessed on Feb 2025.IBM

[15] Opiłka, F.; Niemiec, M.; Gagliardi, M.; Kourtis, M.A. "Performance Analysis of Post-Quantum Cryptography Algorithms for Digital Signature". Appl. Sci. 2024, 14, 4994. https://doi.org/10.3390/app14124994

[16] J. Fan, Q. Yan, M. Li, G. Qu and Y. Xiao, "A Survey on Data Poisoning Attacks and Defenses," 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), Guilin, China, 2022, pp. 48-55, doi: 10.1109/DSC55868.2022.00014.

[17] Zhou, Zhanke, et al. "Model Inversion Attacks: A Survey of Approaches and Countermeasures." arXiv preprint arXiv:2411.10023 (2024).

[18] RZhang Y, Tiňo P, Leonardis A, Tang K. "A survey on neural network interpretability". IEEE Trans Emerg Top Comput Intell. 2021;20:20.

[19] RNori Katagiri. (2024)," Artificial Intelligence and Cross-Domain Warfare: Balance of Power and Unintended Escalation.", Global Society 38:1, pages 34-48.

[20] OWASP Machine Learning Security Top 10 Draft release v0.3, Top 10 2023 List.