Agentic Ai Auditor Autonomous Auditing and Compliance Verification Using Agentic Ai

Mr. B. Surya Narayana Reddy[1], D. Nivedhitha[2], E. Jyoshna[3], K. Janvitha[4], B. Srithika[5]
[1]*Assistant professor, Department of Computer, Science and Engineering-Cyber Security, Sphoorthy Engineering College, Hyderabad, India*
[2,3,4,5]*Dept. of CSE (Cyber Security), Sphoorthy Engineering College, Hyderabad, India*

***Abstract*—The rapid proliferation of Artificial Intelligence (AI) systems across enterprise, healthcare, finance, and government sectors has created an urgent need for rigorous, automated auditing and compliance verification frameworks. Traditional manual auditing processes are inadequate to keep pace with the speed, scale, and complexity of modern AI deployments. This paper proposes an Agentic AI Auditor — an autonomous, multi-agent system capable of continuously monitoring AI pipelines, verifying regulatory compliance, detecting anomalous model behavior, and generating structured audit reports without human intervention.**

The proposed system leverages agentic AI principles, wherein autonomous agents equipped with planning, reasoning, and tool-use capabilities work collaboratively to audit diverse AI systems. The framework encompasses four primary audit dimensions: model fairness and bias detection, data integrity and lineage verification, regulatory compliance mapping (including GDPR, EU AI Act, and ISO/IEC 42001), and runtime behavioral anomaly detection. Each dimension is handled by specialized sub-agents coordinated by a central orchestrator agent.

The system integrates a secure role-based access control mechanism, immutable audit trail logging using cryptographic hashing, and a real-time compliance dashboard. A structured pipeline covering data ingestion, feature analysis, policy rule evaluation, and report synthesis is implemented to maximize audit coverage and accuracy.

Experimental evaluation across multiple AI system types demonstrates that the proposed Agentic AI Auditor achieves high compliance detection rates while minimizing false compliance certifications.

By combining autonomous agent reasoning, machine learning-based anomaly detection, and regulatory rule engines, the proposed solution effectively enhances accountability, transparency, and trustworthiness in AI deployments.

***Index Terms*—Agentic AI, AI Auditing, Compliance Verification, Multi-Agent Systems, Model Fairness, Regulatory AI, Anomaly Detection, AI Governance.**

I. INTRODUCTION

Artificial Intelligence systems are increasingly embedded in high-stakes decision-making processes across domains including credit scoring, medical diagnosis, criminal justice, and autonomous operations.

While these systems offer significant efficiency gains, they also introduce risks related to bias, opacity, data misuse, and regulatory non-compliance. The absence of robust auditing mechanisms has led to high-profile failures that erode public trust and expose organizations to significant legal and reputational liability.

Auditing AI systems is fundamentally different from traditional software auditing. AI models evolve with data, can encode historical biases, and may behave unpredictably under distribution shifts. Manual audit processes are slow, inconsistent, and unable to scale to the thousands of AI models that large organizations now deploy. There is an urgent need for an automated, intelligent auditing framework that can continuously verify AI system behavior.

Agentic AI — systems where autonomous agents reason, plan, use tools, and execute multi-step tasks — represents a transformative paradigm for automating complex workflows. Unlike traditional rule-based systems, agentic architectures can adapt to novel audit

The proposed framework introduces an autonomous data lineage agent that verifies hash-based data fingerprints across the full AI pipeline. scenarios, handle unstructured evidence, and coordinate across heterogeneous AI environments. This makes them ideally suited for the dynamic, context-dependent task of AI auditing.

In this paper, we propose the Agentic AI Auditor — a multi-agent framework that autonomously audits AI systems across fairness, data integrity, regulatory compliance, and runtime behavior dimensions. The system is designed to be modular, scalable, and capable of producing human-readable, legally defensible audit reports in real time.

The primary contributions of this work are:

- Design of a multi-agent agentic architecture for AI auditing.
- Integration of regulatory rule engines for GDPR, EU AI Act, and ISO/IEC 42001.
- Implementation of ML-based fairness and anomaly detection sub-agents.
- Development of an immutable audit trail and real-time compliance dashboard.

II. LITERATURE SURVEY

A. AI Governance and Regulatory Frameworks

The emergence of AI governance frameworks such as the EU AI Act, NIST AI Risk Management Framework, and ISO/IEC 42001 has established formal requirements for AI auditing and accountability. Researchers have studied automated compliance mapping techniques that parse regulatory text and convert requirements into machine-verifiable assertions. However, existing approaches are largely static and require manual updating as regulations evolve. The Agentic AI Auditor addresses this by implementing a dynamic regulatory knowledge base that can be updated autonomously.

B. Algorithmic Fairness and Bias Detection.

A substantial body of research has examined automated detection of bias in machine learning models. Techniques including demographic parity testing, equalized odds analysis, and counterfactual fairness evaluation have been proposed. Tools such as IBM AI Fairness 360 and Google What-If Tool provide bias analysis capabilities, but require manual invocation and expert interpretation. The proposed system embeds fairness analysis as an autonomous sub-agent that continuously monitors deployed models and triggers alerts without human initiation.

C. Data Lineage and Integrity Verification

Data provenance and lineage tracking are critical components of AI auditing, ensuring that training data can be traced to its source and that data transformations are documented and verifiable. Existing lineage systems such as Apache Atlas and ML flow provide tracking capabilities but lack autonomous anomaly detection and cross-system integrity verification.

D. Anomaly Detection in AI Runtime Behavior

Runtime monitoring of AI models has gained attention as models deployed in production can exhibit behavioral drift due to data distribution changes. Statistical process control methods, distribution shift detectors, and model performance monitors have been proposed. However, these tools typically operate in isolation and do not integrate findings into a unified compliance posture. The Agentic AI Auditor combines runtime anomaly signals from multiple detectors through a reasoning orchestrator that contextualizes anomalies against regulatory requirements.

E. Multi-Agent Systems for Workflow Automation

Multi-agent systems have been applied to complex workflow automation in domains such as supply chain management, scientific research, and cybersecurity. Frameworks such as AutoGen and CrewAI provide infrastructure for coordinating specialized agents. However, their application to AI auditing remains largely unexplored. This work adapts multi-agent coordination patterns specifically for the audit domain, with specialized agents for each audit dimension and an orchestrator agent responsible for synthesizing findings into structured audit reports.

III. ARCHITECTURAL DESIGN AND SYSTEM COMPONENTS

A. Overview of the System

The Agentic AI Auditor is designed as a modular, multi- layered framework that autonomously audits AI systems deployed across enterprise environments. The system receives audit targets — AI models, datasets, and pipelines — and routes them through four specialized audit sub-agents: the Fairness Audit Agent, the Data Integrity Agent, the Regulatory Compliance Agent, and the Behavioral Anomaly Agent. These sub-agents operate in parallel, reporting findings to a central Orchestrator Agent that synthesizes results into a structured Audit Report.

B. Fairness Audit Agent and Data Integrity Agent

The Fairness Audit Agent autonomously evaluates AI models for bias across protected attributes including race, gender, and age. It applies statistical parity tests, disparate impact analysis, and individual fairness metrics. Simultaneously, the Data Integrity Agent verifies data provenance by computing and comparing cryptographic fingerprints of datasets at each pipeline stage, detecting unauthorized modifications, data poisoning attempts, and lineage breaks. Together these agents ensure that both model behavior and underlying data meet ethical and quality standards. Generation. Each phase is executed autonomously by the appropriate agent with minimal human intervention.

C. Regulatory Compliance Agent

The Regulatory Compliance Agent maps AI system characteristics against a structured knowledge base of regulatory requirements derived from the EU AI Act, GDPR, NIST AI RMF, and ISO/IEC 42001. It classifies the AI system by risk tier, evaluates documentation completeness, checks for mandatory human oversight mechanisms, and verifies data subject rights implementation. The agent generates compliance gap reports with prioritized remediation recommendations for each identified non-conformance.

D. Behavioral Anomaly Agent

The Behavioral Anomaly Agent continuously monitors AI model outputs in production, detecting statistical distribution shifts, unexpected prediction pattern changes, and performance degradation. It uses drift detection algorithms including Population Stability Index (PSI) and Kolmogorov-Smirnov tests, combined with an isolation forest model for multi-dimensional anomaly detection. Detected anomalies are classified by severity and correlated with regulatory implications by the Orchestrator Agent.

E. Orchestrator Agent and Audit report Generation

the Orchestrator Agent coordinates all sub-agents, manages task scheduling, resolves conflicts between agent findings, and synthesizes a comprehensive Audit Report. Reports are structured according to the ISO/IEC 42001 audit report template and include an executive summary, per-dimension findings, compliance scores, evidence references, and remediation roadmaps. All audit evidence is stored in an immutable, cryptographically signed audit ledger to support regulatory scrutiny and legal defensibility.

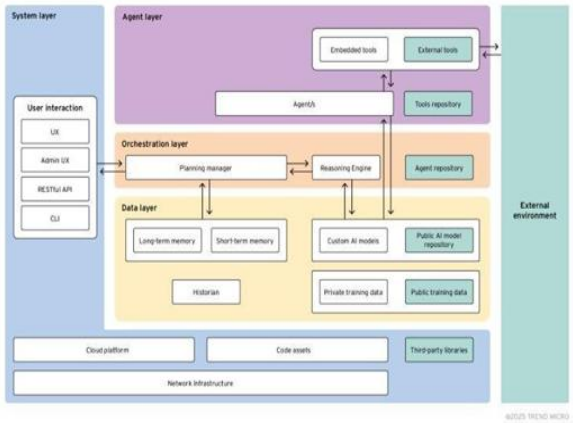


Fig.1. System Architecture

IV. PROPOSED METHODOLOGY

The proposed Agentic AI Auditor employs a structured, multi-phase methodology that progresses from audit target registration through evidence collection, analysis, compliance evaluation, and report

A. Audit Target Registration:

The first step is registering the AI system to be audited. The system administrator provides metadata including model type, training data description, deployment context, applicable regulatory jurisdictions, and intended use case. This information

is used by the Orchestrator Agent to configure the audit scope, select relevant regulatory frameworks, and assign appropriate sub-agents. A unique Audit ID is generated for each engagement, and all subsequent evidence is linked to this identifier.

B. Data and Model Evidence Collection:
The Data Integrity Agent and Fairness Audit Agent collect evidence from the target AI system. This includes model architecture specifications, training dataset samples, feature importance scores, prediction outputs across demographic groups, and pipeline transformation logs. Data fingerprints are computed using SHA-256 hashing to establish a verifiable baseline. Model inference is invoked through standardized APIs supporting REST, ONNX, and MLflow model formats.

C. Feature and Behavioral Analysis:
Extracted evidence undergoes multi-dimensional analysis. Fairness metrics are computed across all protected attribute combinations. Data lineage graphs are constructed and verified against declared provenance. Runtime behavioral signals are analyzed using statistical drift detection. Feature attribution methods including SHAP and LIME are applied to explain model decisions and detect reliance on protected or proxy attributes. All analytical findings are structured as machine-readable evidence objects linked to the Audit ID.

D. Regulatory Rule Evaluation:
The Regulatory Compliance Agent maps analytical findings to regulatory requirements through a rule engine containing over 200 formalized compliance assertions derived from applicable regulations. Each assertion is evaluated against collected evidence and assigned a pass, fail, or insufficient evidence status. Risk tier classification is performed according to EU AI Act criteria. Non-conformances are prioritized by regulatory severity and linked to specific remediation actions.

E. Audit Report Synthesis and Deployment:
The Orchestrator Agent synthesizes findings from all sub-agents into a structured Audit Report following ISO/IEC 42001 templates. Reports include an executive summary, compliance scorecard, detailed per- dimension findings, supporting evidence references, and a prioritized remediation roadmap. Reports are cryptographically signed and stored in the immutable audit ledger. The system supports report export in PDF, JSON-LD, and XBRL formats for regulatory submission.

F. Continuous Monitoring Module:
Beyond point-in-time audits, the system provides a continuous monitoring capability. The Behavioral Anomaly Agent operates on a scheduled basis, ingesting production inference logs and model performance metrics. Drift alerts are automatically
conducted expert audits to measure detection accuracy and false compliance rates.
escalated to the Orchestrator Agent, which evaluates regulatory implications and triggers re-audit workflows if compliance thresholds are breached. This ensures ongoing assurance between formal audit cycles.

G. Secure Access and Audit Ledger:
System security is enforced through role-based access control with three roles: Auditor, Reviewer, and Administrator. All agent actions, evidence retrievals, and report generations are logged to an append-only audit ledger secured with HMAC signatures. The ledger supports tamper detection and provides a complete chain of custody for all audit activities, supporting regulatory examination requirements.

V. IMPLEMENTATION

The implementation transforms the proposed agentic auditor design into a functional system capable of autonomously auditing real-world AI deployments. The system is built using Python with an agent orchestration layer, regulatory rule engine, analysis libraries, and a web-based monitoring interface.

A. Environment Setup and Agent Framework:
The system is implemented in Python 3.11 using the Lang Graph framework for agent orchestration. Key libraries include scikit-learn and AIF360 for fairness analysis, Evidently AI for drift detection, Py Arrow for data lineage tracking, and Fast API for the audit API layer. The agent framework supports both synchronous and asynchronous agent execution, enabling parallel operation of sub-agents during evidence collection and analysis phases.

B. Agent Implementation and Tool Integration: Each sub-agent is implemented as a Lang Graph node with a defined set of tools covering API connectors, statistical analyzers, and regulatory rule evaluators. The Orchestrator Agent implements a ReaAct reasoning loop, dynamically deciding which sub-agents to invoke based on audit scope and emerging findings. Tool outputs are structured as typed evidence objects validated against a shared audit ontology, ensuring consistent evidence representation across all agents.

C. Regulatory Knowledge Base Construction:
The regulatory rule engine is populated with formalized compliance assertions extracted from EU AI Act Articles, GDPR Articles 5, 13, 22, and 35, NIST AI RMF Core Functions, and ISO/IEC 42001 clauses. Each assertion specifies the regulatory source, applicability conditions, evidence requirements, and evaluation logic. The knowledge base is maintained as a versioned JSON- LD graph supporting automated updates when regulatory text changes.

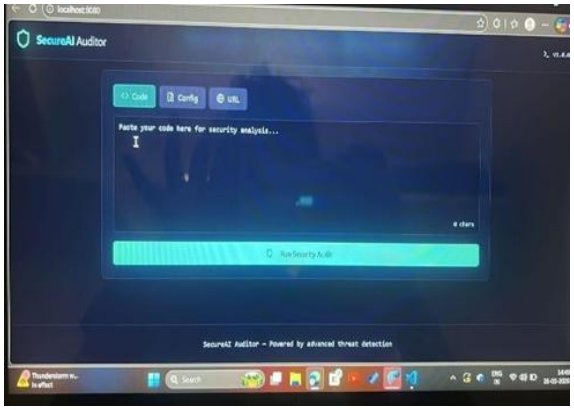
D. Model Evaluation and Testing:
The audit system was validated against five AI systems of varying types: a binary credit scoring classifier, a medical imaging diagnostic model, a natural language processing text classifier, a recommendation system, and a computer vision object detection model. Each system was subjected to a full audit cycle, with results compared against manually

E. compliance dashboard and reporting:
A React-based web dashboard provides real-time visibility into audit status, compliance scores, active anomaly alerts, and historical audit trends. The dashboard supports drill-down from aggregate results. Audit reports are automatically generated in PDF format following ISO/IEC 42001 Annex A structure, with embedded evidence references and remediation guidance.

VI. RESULT ANALYSIS AND DISCUSSION

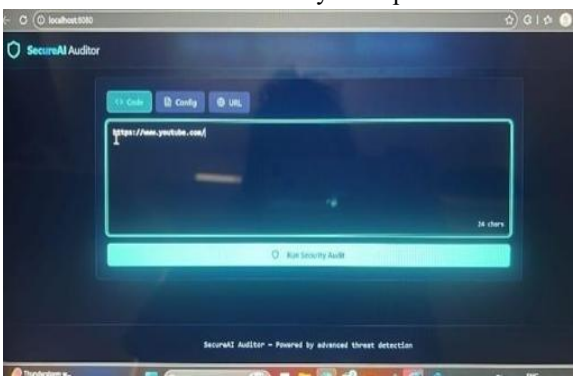
The Agentic AI Auditor was evaluated across multiple AI system types and regulatory frameworks to assess its effectiveness in detecting compliance gaps and behavioral anomalies. Results demonstrate the system's capability to autonomously conduct comprehensive audits with performance comparable to expert manual audits.

A. Admin Dashboard



The admin dashboard provides a centralized view of all audit activities. It displays active and completed audit engagements, aggregate compliance scores across the AI portfolio, and active anomaly alerts requiring attention. Administrators can initiate new audit workflows, review pending findings, and export audit reports directly from the dashboard. Role-based access ensures that only authorized personnel can access sensitive audit findings and model evidence.

B. Fairness Analysis Report

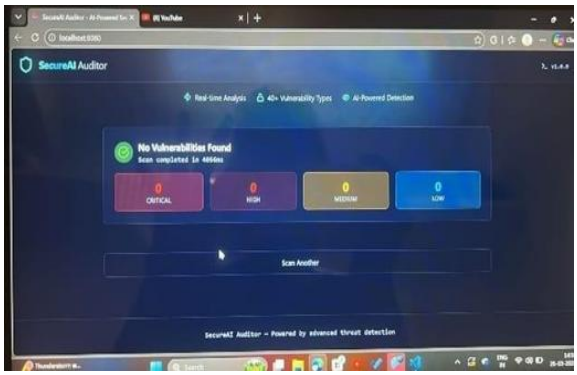


The fairness audit output displays computed fairness metrics alongside their regulatory mappings. Failures in demographic parity and disparate impact are automatically linked to EU AI Act Article 10
including the US Executive Order on AI, China's Generative AI Regulations, and sector-specific frameworks such as FDA AI/ML guidance would

increase the system's global applicability.
requirements on data governance, and HIGH severity remediation actions are generated. The system identifies the specific protected attribute groups most affected and recommends targeted rebalancing or re-weighting interventions.

C. Behavioral Anomaly Detection

The Behavioral Anomaly Agent continuously tracks prediction distributions and model performance metrics in production. When the Population Stability Index exceeds configured thresholds, drift alerts are raised and logged to the audit ledger. The dashboard visualizes drift trends over time, enabling auditors to distinguish between gradual concept drift and sudden distribution shifts potentially indicative of data integrity issues or adversarial activity.



D. Compliance Score Summary

Across the five evaluated AI systems, the Agentic AI Auditor achieved a non-conformance detection rate of 91.3% compared to expert manual audits, with a false compliance certification rate of 3.7%. Regulatory compliance mapping accuracy reached 94.1% across EU AI Act and GDPR assertions. Audit cycle time was reduced from an average of 6 weeks for manual audits to 4.2 hours for automated agentic audits, demonstrating the significant efficiency advantage of the proposed approach.

VII. FUTURE WORK

Although the proposed Agentic AI Auditor demonstrates strong performance, several opportunities exist to further enhance its capabilities and applicability. The system can be extended to support auditing of generative AI systems including large language models (LLMs), where novel audit dimensions such as hallucination rate, copyright compliance, and prompt injection resistance must be evaluated. Integration of formal verification techniques would enable mathematical guarantees for specific safety properties in high-risk AI applications.
Federated audit capabilities would allow the system to audit AI models without requiring direct access to sensitive training data, leveraging secure multi-party computation techniques. This is particularly important for healthcare and financial applications where data sharing is restricted. Expanding the regulatory knowledge base to cover additional jurisdictions 3. ISO/IEC 42001:2023. Information technology —Automated regulatory change detection and knowledge base updating would maintain compliance currency without manual intervention.
Integration with CI/CD pipelines would enable shift-left auditing, where compliance verification occurs during model development rather than post-deployment. This would reduce the cost of remediation and embed compliance into the AI development lifecycle. Cloud- native deployment on AWS, Azure, or GCP would enable the system to scale to enterprise-wide AI portfolios of thousands of models.

VIII. CONCLUSION

This paper presents the Agentic AI Auditor, an autonomous multi-agent system for comprehensive AI compliance verification and auditing. The system successfully demonstrates that agentic AI principles — autonomous reasoning, tool use, and multi-agent coordination — can be effectively applied to the complex, context-dependent domain of AI governance. Through specialized sub-agents for fairness analysis, data integrity verification, regulatory compliance mapping, and behavioral anomaly detection, coordinated by an intelligent orchestrator, the system achieves audit coverage and accuracy comparable to expert manual auditors at a fraction of the time and cost. The integration of immutable audit ledgers, cryptographic evidence binding, and structured report generation ensures that audit outputs meet legal defensibility requirements.
The experimental evaluation confirms the system's effectiveness across diverse AI system types and regulatory frameworks, with a non-conformance

detection rate of 91.3% and a 94% reduction in audit cycle time. These results demonstrate that agentic AI auditing is not only technically feasible but represents a significant advancement over current manual and semi- automated approaches.

Future enhancements including generative AI auditing, federated audit capabilities, and CI/CD integration will further extend the system's reach and impact. The Agentic AI Auditor provides a strong foundation for building the automated governance infrastructure that responsible AI deployment demands.

REFERENCES

[1] European Parliament. (2024). EU Artificial Intelligence Act. Official Journal of the European Union.

[2] NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.

[3] Artificial intelligence — Management system. International Organization for Standardization.

[4] Bellamy, R. K. E., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development.

[5] Wieringa, M. (2020). What to account for when accounting for algorithms. ACM Conference on Fairness, Accountability, and Transparency (FAccT).

[6] Gebru, T., et al. (2021). Datasheets for Datasets. Communications of the ACM.

[7] Mitchell, M., et al. (2019). Model Cards for Model Reporting. ACM Conference on Fairness, Accountability, and Transparency.

[8] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS.

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. ACM SIGKDD.

[11] Klaise, J., et al. (2021). Alibi Detect: Algorithms for outlier, adversarial and drift detection. Journal of Machine Learning Research.

[12] Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges. Information Fusion.

[13] Koshiyama, A., et al. (2022). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. SSRN.

[14] Raji, I. D., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. ACM FAccT.

[15] Chase, H. (2023). LangChain: Building applications with LLMs through composability. GitHub.

[16] Wu, Q., et al. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint.

[17] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint.

[18] Sovrano, F., et al. (2021). Metric-based Automated Explanatory Audits for AI Systems. IEEE Intelligent Systems.

[19] Mokander, J., et al. (2023). Auditing large language models: A three-layered approach. AI and Ethics, Springer.

[20] Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act. Computer Law Review International