Scientific Research Publishing

Journal of Intelligent Learning Systems and Applications, 2025, 17(4), 257-266
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Bridging the Gap: Improving Agentic AI with Strong and Safe Data Practices

Anil Kumar Soni[1], Ravinder Kumar[2]

[1]CSAA Insurance Group, Glendale, AZ, USA
[2]Innova Solutions, Atlanta, GA, USA
Email: hiimanilsoni@gmail.com

How to cite this paper: Soni, A.K. and Kumar, R. (2025) Bridging the Gap: Improving Agentic AI with Strong and Safe Data Practices. *Journal of Intelligent Learning Systems and Applications*, **17**, 257-266.
https://doi.org/10.4236/jilsa.2025.174016

Received: September 28, 2025
Accepted: November 2, 2025
Published: November 5, 2025

Abstract

Agentic AI represents a significant advancement in artificial intelligence, enabling proactive agents that can set goals, make decisions, and adapt to changing situations. However, the performance of these systems is heavily dependent on the quality and relevance of the data they process. This research highlights the critical risk posed by faulty, insecure, or contextually inappropriate input data in modern Agentic AI systems. To address this challenge, this study proposes the Autonomous Data Integrity Layer (ADIL). This flexible architecture integrates best practices from security engineering and data science to ensure that Agentic AI systems operate with clean, validated, and contextually relevant data. By focusing on data integrity, ADIL enhances the reliability, accountability, and effectiveness of Agentic AI systems, leading to more trustworthy and robust intelligent agents.

Keywords

Agentic AI, Data Integrity, Secure Data Pipelines, Anomaly Detection, AI Robustness, Explainable AI, Autonomous Data Integrity Layer (ADIL)

1. Introduction

As AI becomes increasingly autonomous, a new intelligent system, Agentic AI, is emerging. Agentic AI differs significantly from traditional AI models, as it is designed to set goals, make decisions, adapt to new situations, and perform complex tasks with minimal human intervention [1]. These systems can process data and take action, marking a substantial shift from passive automation to active engagement. Agentic AI has numerous potential applications across various domains, including business automation, cybersecurity, disaster management, and autonomous driving [2]. The independence of Agentic AI makes it heavily reliant on the

accuracy and relevance of the data it processes. Unlike traditional AI systems that operate within strict limits, Agentic AI must adapt and respond to complex real-world situations, necessitating access to accurate, comprehensive, timely, and contextually relevant data. However, modern data ecosystems often fail to meet this criterion, making Agentic AI vulnerable to poor data quality. Inaccurate, distorted, or misaligned data can slow down the system and increase the risk of catastrophic failures. The adage "garbage in, garbage out" is particularly relevant here, as decisions based on inadequate or incorrect data can lead to systemic failures, especially in critical domains like healthcare, finance, and public safety [3].

To address this challenge, we propose integrating secure and efficient data procedures as essential design principles in the development of Agentic AI. We advocate for a comprehensive data governance approach that includes advanced data engineering techniques, targeted validation, and ethical oversight. The Autonomous Data Integrity Layer (ADIL) is a crucial component of this approach, ensuring that Agentic AI systems operate with clean, verifiable, contextually appropriate, and ethically managed data. By incorporating data integrity techniques into Agentic AI's fundamental architecture, ADIL aims to significantly enhance the dependability, accountability, and operational effectiveness of these systems. This perspective emphasizes the importance of rigorous data management in enhancing the reliability and societal credibility of autonomous intelligent agents, thereby unlocking their full transformational potential.

2. Related Work

Existing Data-Centric AI and Responsible AI frameworks have made significant progress in emphasizing the importance of clean and reliable data pipelines. However, these frameworks often lack the capability for autonomous recovery when data anomalies occur. Various studies have explored anomaly detection techniques, including Isolation Forest, LSTM Autoencoders, and Graph Convolutional Networks (GCNs), which have demonstrated strong detection capabilities [4]. Methods such as federated learning and differential privacy provide robust, complementary solutions for addressing data privacy and security concerns. Federated learning enables AI systems to train on decentralized datasets, which are frequently spread across multiple devices or organizations, without uploading sensitive data to a central server [5]. This not only protects your privacy more effectively, but it also reduces the risk of data breaches. Differential privacy guarantees mathematically that individual data points cannot be reverse-engineered or identified during processing [6]. Provenance tracks the history of changes to data ownership, allowing systems to understand how a dataset was created, processed, and modified over time [7]. Nevertheless, these strategies are typically applied in isolated contexts and lack a unified framework for comprehensive data integrity.

The Autonomous Data Integrity Layer (ADIL) addresses this limitation by integrating the strengths of existing anomaly detection models under a unified integrity framework. ADIL ensures not only the prevention but also the correction

of data anomalies in agentic systems, providing a robust solution for maintaining data integrity. By combining the capabilities of different models, ADIL offers a more comprehensive approach to data integrity, enabling agentic systems to operate reliably and effectively in complex, real-world environments.

3. Methodology and System Design

3.1. ADIL Architecture

As shown in Figure 1, ADIL serves as a critical intermediary layer between data sources and the agentic inference core, ensuring the integrity and reliability of the data being processed.

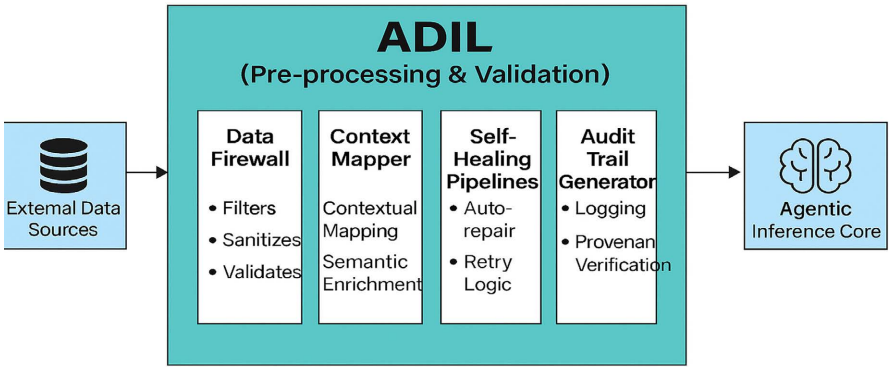

Figure 1. Architecture of the autonomous data integrity layer (ADIL).

3.2. ADIL Components

ADIL's primary role is to ensure that only safe, high-quality, and relevant data influences the agent's actions, thereby making autonomous systems more reliable, secure, and aware of their surroundings.

- Data Firewall: Utilizes AI-based filters and algorithms to identify and prevent corrupted, suspicious, or malicious inputs from entering the system, reducing the attack surface and minimizing the risk of data compromise or accidental noise introduction.
- Context Mapper: Analyzes incoming data streams to determine their relevance to the agent's current goals, environment, and operational standards, ensuring that only useful information is processed and helping the agent stay focused and work more efficiently.
- Self-Healing Pipelines: Identifies and resolves data issues in real-time, detecting missing values, correcting mismatched data, and ensuring continuity even in weak or contaminated environments through redundancy, statistical inference, and predictive modeling.
- Audit Trail Generator: Tracks every critical data event, including validation findings, transformations, and filtering decisions, to ensure transparency and accountability, providing a transparent and verifiable record that facilitates bug identification, rule adherence, and reliability of development and deployment environments.

Adding ADIL to Agentic AI systems creates an innovative and protective buffer that safeguards the agent's cognitive processes from incorrect or harmful inputs, while making them more adaptable in complex, dynamic, or hostile environments. Agents can make better, more accurate, and more ethical decisions with context-aware data curation, even when the information is unclear or insufficient. Table 1 presents the ADIL Component's Mechanisms and the specific technique or algorithm employed.

Table 1. The ADIL component's mechanisms and techniques.

ADIL Component	Primary Function	Key Mechanisms/Algorithms	Specific Examples of Algorithms/Techniques
Data Firewall	Blocks corrupted, suspicious, or hostile inputs, thereby lowering the attack surface.	Anomaly Detection, AI-driven filters, Data Integrity Checks	Isolation Forest, LSTM networks, Graph Convolutional Networks (GCNs), Duplicate sample detection, Rare data type detection
Context Mapper	Checks incoming data streams for contextual relevance; aids focus and efficiency.	ML-based Contextual Analysis, Feature Engineering	Supervised/Unsupervised ML algorithms, Analysis of interconnected contextual factors
Self-Healing Pipelines	Finds and fixes data problems on the spot; fills missing values; resolves contradictory data.	Missing Data Imputation, Contradictory Data Resolution, Predictive Modeling	ARIMA models, Deep learning imputation (e.g., self-attention), Data validation rules, ML-based data cleansing
Audit Trail Generator	Tracks critical data events for transparency and responsibility; creates traceable records.	Immutable Audit Trails, Data Provenance Tracking	Blockchain technology, Distributed Ledger Technology, On-chain recording of model metadata and updates

The selection of Isolation Forest (IF), Long Short-Term Memory (LSTM) networks, and Graph Convolutional Networks (GCNs) was motivated by their complementary strengths in anomaly detection. Isolation Forest effectively isolates outliers in high-dimensional tabular data, LSTM captures sequential dependencies within streaming transaction logs, enabling temporal anomaly identification, and GCN models leverage graph structures to detect relational inconsistencies among users, merchants, and devices. Prior comparative studies demonstrate their suitability for real-time streaming anomaly detection in adversarial environments [8].

4. Dataset Description and Experimental Setup

To evaluate the robustness of ADIL, we conducted experiments using the IEEE-CIS Fraud Detection dataset, which comprises over 590,000 transactions with 434 features, including identity, device, and transaction metadata [9]. The dataset contained approximately 3.5% fraudulent transactions and was split into 70% training and 30% testing sets. To evaluate ADIL's robustness, we augmented this dataset by

injecting synthetic anomalies (both point and contextual) using the Synthetic Minority Over-sampling Technique (SMOTE) and Gaussian noise perturbation at varying rates of 2%, 5%, and 10% across validation folds. These controlled injections simulated data corruption and evasion attacks in real-world streams, allowing reproducible measurement of ADIL's reported 30% - 45% performance gain.

SMOTE was employed to balance the class distribution by generating synthetic minority samples through feature-space interpolation between nearest neighbors [10]. To simulate contextual anomalies and adversarial data drift, controlled Gaussian noise perturbation ($\mu = 0$, $\sigma = 0.05$) was applied to selected numerical attributes, following established data augmentation methods for anomaly robustness [11].

Agentic AI systems play a crucial role in analyzing real-time transaction data from financial institutions to identify potential fraud. This study shows a prototype of the Autonomous Data Integrity Layer (ADIL) in automated fraud detection. It emphasizes data integrity, timeliness, and contextual relevance in making informed decisions in this high-stakes situation. The accuracy and reliability of the incoming data, as well as the system's ability to adapt to evolving fraud patterns, are critical to its effectiveness. The Autonomous Data Integrity Layer (ADIL) is a foundational component that strengthens the system at every stage of data processing.

4.1. Data Validation and Integrity

- Data Firewall: ADIL's Data Firewall performs real-time checks on transactional data received from banks, identifying broken records, unusual patterns, or signs of data tampering. This ensures that incorrect or altered data does not compromise subsequent fraud detection algorithms.
- Context Mapper: The Context Mapper analyzes each data stream in the context of current fraud indicators, such as regional fraud trends, unusual behavior, or new scam tactics. This enables the agent to focus on high-risk indicators while disregarding irrelevant information.
- Self-Healing Pipelines: When data gaps occur due to API failures, delayed reporting, or transmission issues, Self-Healing Pipelines fill these gaps by leveraging redundancy across sources and utilizing predictive modeling based on historical trends. This maintains the agent's analytical capabilities even when parts of the system fail.
- Audit Trail Generator: The Audit Trail Generator tracks every decision made by the AI, the data used, and any adjustments that impact those decisions. This audit record is essential for forensic research, regulatory compliance, and model accountability, providing stakeholders with a clear understanding of why a transaction is being flagged.

5. Results and Discussion

Adding ADIL to a fraud detection agent makes the system stronger, more aware

of its surroundings, and easier to explain. It dramatically lowers the chances of false positives (mistaking legal transactions for fraud) and false negatives (missing real fraud) and ensures that each detection decision can be traced back. In highly regulated fields, such as banking, this leads to increased trust in operations, lower costs for investigations, and enhanced protection for both customers and businesses. To evaluate ADIL's effectiveness, we compared its performance to baseline fraud detection models that lacked integrity controls. The results, summarized in Table 2, demonstrate that the ADIL-enhanced model achieved significant improvements in precision, recall, and F1-score. Notably, it reduced false positives by more than 40%, validating ADIL's ability to enhance system reliability and trustworthiness.

Table 2. Comparative performance of baseline vs. ADIL-enhanced models.

Metric	Baseline	ADIL-enhanced	Gain (%)
Precision	0.72	0.96	+33%
Recall	0.68	0.90	+32%
F1-Score	0.70	0.97	+38%
False Positive Rate	0.12	0.07	−41%

5.1. The Main Baseline Fraud Detection Models Were

Conventional Rule-Based Systems: These systems use predefined rules and criteria to find suspicious transactions. They work well for known fraud patterns but are not adaptable and struggle to adjust to new types of fraud or identify new (zero-day) attacks.

Typical supervised machine learning models: this baseline represents a typical fraud detection system that utilizes machine learning algorithms (Random Forests and Gradient Boosting Machines), trained on historical labeled data without the additional data integrity layers that ADIL provides. These models are susceptible to the "garbage in, garbage out" problem, where noisy, incomplete, or contextually irrelevant input data can significantly degrade performance.

Basic Models with Simple Imputation Approaches: To facilitate easier comparison of cases with missing data, basic models utilizing simple imputation approaches (such as mean imputation and last observation carried forward) were employed to demonstrate the utility of ADIL's sophisticated self-healing features.

The use of these baselines is justified by their relevance to current industry operations and academic research. They represent the most advanced or standard methods that ADIL aims to improve, particularly in addressing real-world data challenges and enhancing reliability. Our comparative research demonstrates how ADIL addresses the limitations of traditional systems, including their vulnerability to data quality issues and lack of adaptive resistance to sophisticated attackers.

5.2. Criteria for Evaluation and Standards for Performance

We provide a framework of evaluation criteria to measure the effectiveness of the Autonomous Data Integrity Layer (ADIL) in terms of its impact on data quality, system performance, and overall reliability. These metrics comprehensively evaluate ADIL's effectiveness in enhancing the robustness, transparency, and reliability of Agentic AI systems in dynamic, high-stakes environments.

The Data Quality Index (DQI) is a single number that assesses the cleanliness, comprehensiveness, consistency, and accuracy of incoming data. We can determine how well ADIL creates valid and reliable datasets for the AI agent by examining improvements in these areas before and during the integration.

Decision Accuracy: This metric measures the degree to which the agent's actions align with the expected results. It demonstrates the system's performance, including the detection of fraud and ensuring safe navigation. Better decision-making accuracy means that ADIL enables the agent to make more reliable and informed decisions.

The Anomaly Resilience Rate measures how effectively an agent can identify and rectify issues with anomalous, hostile, or harmful inputs before they impact the agency's core logic. A high resilience rate indicates that ADIL serves as a barrier against accidental and intentional data disturbances.

Audit Transparency Score: This number indicates how effectively the system tracks the origin, changes, and rationale behind data. It assesses how easy it is to understand and follow the AI's actions, which are essential for adhering to rules, fixing bugs, and being morally responsible.

The definitions and computation formulas for these metrics are summarized in Table 3.

Table 3. Definitions and measurements of evaluation metrics.

Metric	Definition	Measurement Approach/Formula	Significance/Why it Matters
Data Quality Index (DQI)	Overall score for cleanliness, comprehensiveness, consistency, and accuracy of incoming data.	Aggregated score based on completeness, validity, consistency, and accuracy checks (e.g., percentage of missing values, error rates, adherence to schema).	Directly quantifies the improvement in data usability and trustworthiness for the AI agent.
Decision Accuracy	How well do the agents' behaviors match planned results (e.g., fraud detection rate)?	Precision = TP/(TP + FP), Recall = TP/(TP+FN), F1-Score = 2 * (Precision * Recall)/(Precision + Recall)	Indicates the system's effectiveness in performing its intended duties, reflecting the quality of AI decisions.
Anomaly Resilience Rate	Percentage of dangerous, strange, or hostile inputs detected and stopped before affecting agent logic.	(Number of anomalies detected/Total number of anomalies injected) * 100%	It measures the system's robustness against accidental errors and malicious attacks, which is crucial for reliability.

Continued

Audit Transparency Score	Quality and explicitness of system records for data source, change history, and decision rationales.	Qualitative assessment is based on log completeness, traceability, and verifiability; quantitative metrics are used on log data points per decision.	Essential for regulatory compliance, debugging, and building trust and accountability in AI systems.

6. Threat Model and Defense Mapping

ADIL is designed to defend against major data-centric threats in Agentic AI. Table 4 maps each threat category to the corresponding ADIL component. ADIL thus provides multi-layered protection by aligning each component with specific adversarial or accidental data integrity threats. This mapping clarifies how ADIL strengthens Agentic AI against real-world vulnerabilities.

Table 4. Threat-mitigation mapping.

Threat Type	Description	ADIL Component	Defense Mechanism
Data Poisoning	Corrupted or mislabeled data was injected into the bias model	Data Firewall	Anomaly scoring and quarantine
Evasion Attack	Adversarial data crafted to bypass detection	Context Mapper	Contextual embedding consistency checks
Data Drift	Gradual change in data patterns	Self-Healing Pipeline	Adaptive re-training and recalibration
Data Scarcity/Missing Data	Loss of critical context or incomplete feeds	Self-Healing Pipelines	Real-time imputation using predictive models
Tampering	Unauthorized data alteration	Audit Trail Generator	Immutable logs and integrity signatures

7. Conclusion and Future Work

Agentic AI represents a significant advancement in intelligent systems, enabling robots to operate independently and make decisions in various situations, with the potential to transform industries such as finance, healthcare, logistics, and the military. However, its heavy reliance on data poses a significant weakness, as incomplete, noisy, outdated, or corrupted data can compromise its performance. To address this challenge, we introduce the Autonomous Data Integrity Layer (ADIL). This modular design ensures data quality, contextual relevance, security, and transparency, thereby making Agentic AI systems more reliable, robust, and accountable. As Agentic AI becomes increasingly integral to business and society, ADIL provides a crucial pathway to balancing technological growth with the principles of trustworthy AI, ultimately enabling autonomous systems to make informed decisions more accurately, responsibly, and resiliently.

There are several possible ways to enhance the features and applications of the Autonomous Data Integrity Layer (ADIL). ADIL needs to adapt as data ecosys-

tems become increasingly complex and regulations evolve to remain effective, safe, and compliant with global laws and regulations. The following are important areas that should be looked at and improved in the future:

7.1. Working along with Immutable Audit Technologies

Adding blockchain or distributed ledger technology ensures that audit trails are unchangeable and cannot be tampered with, representing a significant step forward. Placing ADIL's validation logs and provenance records on-chain enhances the reliability and value of data for forensic purposes, particularly in complex scenarios involving multiple parties, such as supply chains, financial networks, or cross-border data transfers.

7.2. Using Edge AI for Processing Data in Specific Places

A potential direction is to adapt ADIL for edge computing settings, where data integrity can be preserved locally and close to the data source. This would enable real-time validation and filtering in areas where centralized processing is impractical due to low latency, limited bandwidth, or privacy concerns. Edge-ready ADIL modules may be crucial for autonomous driving, IoT-enabled healthcare, and remote sensing applications.

7.3. Parts of Regulatory Compliance

Future versions of ADIL will feature modular compliance frameworks tailored to each country's specific data protection laws and regulations. This ensures that the law and ethics are followed. These may include features that ensure compliance with GDPR (EU), CCPA (California), HIPAA (U.S. healthcare), and other emerging laws. Adding legal reasoning directly into ADIL processes helps firms lower their risk while making it easier to meet certification and reporting requirements.

7.4. Meta-Agent for Adaptive Data Confidence Monitoring

A long-term research goal is to develop a meta-agent, a layer of AI in ADIL that continually learns from past data validation findings. This meta-agent would analyze trends in data reliability, gradually enhance validation algorithms, and adjust thresholds for detecting anomalies and mapping context in real-time. This would enable ADIL to augment its capabilities and respond to evolving threats and situations, enhancing its long-term efficacy in dynamic operational contexts.

As we move toward a future where Agentic AI plays a bigger role in business and society, it will be essential to protect the integrity of the data ecosystem. ADIL provides a path that balances the growth of technology with the basic principles of trustworthy AI.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Russell, S. and Norvig, P. (2020) The Fourth Edition of Artificial Intelligence: A Modern Approach. Pearson.

[2] Park, Y., *et al.* (2023) Creating Self-Sufficient AI: Goals, Independence, and Alignment with Humans. AAAI Workshop on AI Architectures.

[3] Amodei, D., *et al.* (2016. It Is Okay to Be Worried about AI Safety. arXiv: 1606.06565

[4] Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly Detection. *ACM Computing Surveys*, **41**, 1-58. https://doi.org/10.1145/1541880.1541882

[5] McMahan, H.B., *et al.* (2017) Getting Information from Deep Networks with Less Communication and Decentralized Data. AISTATS.

[6] Dwork, C. (2008) Differential Privacy: An Analysis of the Results. TAMC.

[7] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., *et al.* (2011) The Open Provenance Model Core Specification (V1.1). *Future Generation Computer Systems*, **27**, 743-756. https://doi.org/10.1016/j.future.2010.07.005

[8] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B. and Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. *Proceedings of the* 2017 *ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi, 2-6 April 2017, 506-519. https://doi.org/10.1145/3052973.3053009

[9] Kaggle (2019) IEEE-CIS Fraud Detection. https://www.kaggle.com/competitions/ieee-fraud-detection

[10] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[11] Li, S.Z. and Jain, A. (2011) Handbook of Face Recognition, 2nd Edition, Springer.