International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Cyber Attacks Targeting Generative AI and Agentic AI Frameworks in the Retail Domain and Strategies to Prevent Them

Suresh Kumar Gundala

Abstract: The retail sector's accelerated integration of generative and agentic artificial intelligence introduces a distinctive cybersecurity challenge: the exploitation of open consumer interactions as vectors for attacking the models themselves. Unlike enterprise environments with controlled access, retail platforms operate in dynamic conditions where large volumes of anonymous and guest users interact continuously. Such an open structure exposes many attack surfaces, making it possible for adversaries to launch large-scale attacks involving data poisoning, adversarial prompt injection, and input manipulations. These kinds of attacks affect not only the reliability of the systems but also result in influencing the decision-making process and eventually cause fraud, bad advice, biased models, and similar problems. This article recommends adopting a hierarchical security framework, which includes strict data validation, persistent monitoring of the model performance, adversarial machine learning tools, and governance-related security measures. In relation to the application of threat modeling and defense techniques in the retail sector, the research presents an adequate foundation for building trustworthy agentic AI systems. Retail requires a flexible security strategy that evolves according to the methods employed by the attackers to stay competitive amidst advances in the manipulation methods. The adoption of governance methods alongside technology ensures greater clarity, compliance, and trust among consumers for using automation. The multi-faceted model involves rigorous surveillance, adversarial capabilities, and governance-based approaches, creating an all-encompassing framework for achieving security during the integration of agentic AI within retail activities.

Keywords: *Adversarial Attacks, Agentic AI, Cybersecurity in Retail, Data Poisoning, Model Resilience*

1. Introduction

1.1 Industry Context

The retail industry occupies a uniquely exposed position in the AI cybersecurity landscape. As one of the most consumer-facing sectors, retail systems routinely process immense volumes of interactions from anonymous, guest, and transient users. These individuals engage with AI-powered recommendation engines, chatbots, dynamic pricing tools, and autonomous checkout systems without authenticated identities or persistent session accountability. This open-access model, central to retail's customer experience philosophy, simultaneously creates an expansive and largely ungated attack surface for adversaries seeking to manipulate or degrade AI systems.

The increasing use of agentic AI in the sector further magnifies these threats. An agentic system is not just an instrument that reacts; rather, it makes decisions in areas such as supply chain management, fraud detection, inventory optimization, and personalized marketing. If hacked, the repercussions would be much broader than just the individual cases. For example, a price-setting algorithm may end up initiating fraud discount campaigns. A recommender system can intentionally mislead customers into buying substandard goods. Retail's sensitivity to both operational disruption and reputational damage makes it a high-value target in AI cybersecurity.

Karnataka State Open university, India

1.2 Application Detail

Data Poisoning Attacks

Retail AI systems often retrain on live customer data. Malicious entities take advantage of this by introducing malicious or erroneous signals via guest users, fraudulent reviews, and artificially generated behavior patterns. Slowly but surely, the training distribution is contaminated with malicious signals, leading to inaccurate recommendations, wrong forecasts, and systematic bias in autonomous

decision-making. Contamination of the training data occurs faster in the retail sector than in enterprises due to the nature of their ingestion pipelines.

Attack Type	Mechanism of Exploitation	Retail Impact	Example Scenario
Data Poisoning	Injection of malicious signals into training data	Skewed recommendations, biased forecasts	Fake reviews contaminating product ratings
Prompt Injection	Malicious instructions embedded in user inputs	Unauthorized refunds, exposure of pricing logic	Manipulated product review triggering agent misbehavior
Input Manipulation	Engineered queries reverse-engineering model logic	Exploitation of discount thresholds, fraud detection bypass	Abandoned cart sequences probing reward triggers

Table 1: Attack Vectors in Retail AI [2]

Adversarial Prompt Injection

The use of agentic AI systems based on LLMs exposes them to prompt injection attacks. An attacker can insert malicious instructions into the prompts used as user inputs, product descriptions, or third-party data feed inputs to manipulate the system's actions contrary to its intended purpose. For instance, an attack in a retail setting may involve inserting malicious instructions into a product review to compel the customer care representative to offer refunds beyond what should be allowed, to reveal pricing mechanisms, or evade anti-fraud measures. Prompt injection is made easy due to the public accessibility of the interfaces, where anyone who is a customer can enter free text inputs.

Input Manipulation and Model Inversion

An advanced adversary may design engineered inputs such as search queries, clickstream data, or abandoned shopping carts to analyze and reverse-engineer the decision-making process of AI models used in retail settings. Reversal of the model gives opponents insight into the activation processes involved in discount generation, reward generation, and fraud detection. Such exploitation can result in lower profits for the firm and a loss of confidence from consumers. Reversal of the model further exposes strategic business information that can be duplicated by the opponent.

Strategic Importance

The convergence of these attack categories demonstrates that retail AI systems face risks that are both technical and structural. The problem is not limited to isolated vulnerabilities but arises from the fundamental openness of retail interactions. While necessary for providing a quality customer experience, such exposure makes it fairly easy for rivals to exploit the vulnerabilities of AI algorithms.

With the advent of generative and agentic AI, such issues have been accentuated further because the element of autonomy in decision-making is now involved. An exploited agent would not just make incorrect classifications but also flawed decisions that would affect the supply chain operations, pricing policies, fraud prevention measures, and customer management platforms.

2. Technology or Workflow Integration

As such, the security measures outlined below are intended to work smoothly into an AI-based retail framework without the need for complete overhaul of the entire system. The three-tiered approach involves the integration of the suggested defense mechanisms at each of the three tiers of the retail AI technology stack.

2.1 Input Validation Layer

Input validation represents the primary mechanism through which this problem will be approached. Systems incorporating AI into retail operations and based on large language models are at risk from adversarial prompting as well as other manipulation tactics involving inputs. To mitigate the risk, semantic input sanitization will be employed in order to identify and block any such inputs prior to passing them to the core reasoner of the agent.

Incorporation into the existing system is crucial. Many retailers utilize stacks built according to the MACH architecture, and incorporating sanitization in the existing APIs will enable an interception at the earlier stages.

Implementation considerations encompass token-level anomaly detection, checks of semantic similarity against a trusted library of inputs, and rules-based injection signature filters. Such methods enable the differentiation of valid customer inquiries from attempts at manipulation. Moreover, validation pipelines should be flexible enough to stay up to date with adversarial patterns discovered via monitoring and red teaming. Studies on adversarial poisoning in generative models highlight the dangers of adversarial input and showcase the necessity of validation in the retail space [2]. Similarly, investigations related to attack data synthesis demonstrate the importance of preprocessing techniques aimed at filtering malicious data prior to its consumption by the machine learning processes [5].

Technique	Functionality	Integration Point	Benefit
Token-Level Anomaly Detection	Flags unusual token sequences	API gateway	Early interception of adversarial prompts
Semantic Similarity Checks	Compares inputs against trusted libraries	Edge security layer	Differentiates valid queries from manipulations
Rule-Based Signature Filters	Blocks known injection patterns	Upstream of inference endpoints	Prevents repeatable exploit attempts

Table 2: Input Validation Techniques [5]

2.2 Model

Monitoring and Observability Layer

The second layer concerns itself with constant surveillance over the behavior of the model. As retail AI models often undergo retraining based on new customer data on an ongoing basis, they are highly vulnerable to drift and poisoning attacks. Consequently, tools for real-time monitoring have been implemented in model serving architectures so that they always produce correct results. Data drift detection monitors changes in input data distribution relative to the baseline training sets, whereas output distribution analysis monitors whether the produced predictions fall into reasonable limits. Behavioral anomalies are also monitored through analyzing any suspicious decision-making activity (e.g., an unexpected number of refund requests accepted as fraud). Detection of any anomalies associated with the poisoning attack results in a triggering of the rollback/containment procedure, ensuring that the compromised model does not run in production mode.

Being integrated into the existing MLOps pipeline, the process of monitoring is part of the entire workflow associated with models' deployment and management. Some cloud-based services, such as Amazon SageMaker Model Monitor and Microsoft Azure ML, have built-in observability features that may be supplemented with detectors aimed at identifying problems specific to the retail industry. For example, one might choose to configure monitoring in such a way as to track the emergence of biased recommendations, increased fraud detection rates, and abnormally high/low prices in the retail context. Direct observability built right into the service layer enables regular tracking and assurance regarding the models' reliability. Several studies focusing on adversarial attacks against black-box detectors prove the necessity of constant monitoring since these mechanisms could be manipulated otherwise [3].

Tool/Method	Purpose	Retail Application	Response Trigger
Data Drift Detection	Identifies shifts in input distribution	Detects poisoned customer data streams	Initiates retraining or rollback
Output Distribution Analysis	Ensures predictions remain within expected bounds	Monitors pricing anomalies	Quarantine of compromised models
Behavioral Anomaly Tracking	Flags suspicious decision patterns	Detects abnormal refund approvals	Escalation to human oversight

Table 3: Monitoring and Observability Tools [3]

2.3 Adversarial Resilience and Governance Layer

The third layer combines technical resilience with governance-driven oversight. Adversarial training is built into the process of developing the AI model. It is accomplished through the practice of performing simulated red-team activities against the models to pinpoint any vulnerabilities. Moreover, there is synthetic training to train models by exposing them to adversarial inputs when the model is being developed. This makes them resilient enough to withstand attacks once they are deployed.

Governance practices include enforcing strict access control measures in relation to the use of tools by the agents. Retail AI agents have access to critical systems like payment gateways and inventory databases. To avoid any malicious acts from the agents, strict access control policies are followed, limiting their access to these tools using identity management platforms. Another practice adopted is cryptographic data provenance tracking. This ensures the verification of training datasets to avoid poisoning attacks. Human-in-the-loop escalation procedures are well-known for major autonomous actions. In case the agent responsible for detecting fraud tries to intervene with several warnings at once, then the process escalates into human supervision. This ensures that accountability exists in the performance of crucial activities without inhibiting the agents' autonomy. There are audit logging tools that track all actions performed by agents. Data privacy governance strategies provide information on how governance-driven control mechanisms empower organizations to foster accountability during AI adoption [4]. The threat-centric architecture highlights the importance of aligning security technicalities with legislative considerations in different geographies [7]. There exists a complex interplay between AI and cybersecurity where organizations must devise robust strategies using both technical and nontechnical controls [9]. There is a responsible multi-agent governance model that showcases how the governance approach can preserve data usability without compromising privacy [10]. Federated learning techniques provide practical ways to promote resilience within complex ecosystems that feature several data sources [12]. Additionally, mechanisms that protect privacy also facilitate the promotion of transparency in consumer-facing platforms.

Measure	Description	Retail Relevance	Outcome
Adversarial Training	Exposure to synthetic attacks during development	Strengthens resilience of recommendation engines	Models withstand real-world manipulations
Access Control via IAM	Restricts agent tool usage	Protects payment gateways and inventory systems	Limits unauthorized actions
Cryptographic Provenance	Verifies dataset authenticity	Prevents poisoning of training pipelines	Ensures trustworthy data sources
Human-in-the-Loop Escalation	Oversight of high-stakes decisions	Fraud detection interventions	Balances autonomy with accountability

Table 4: Governance and Resilience Measures [7]

Strategic Alignment for Integration

Taken together, these three strategies form an integrated defensive strategy that will complement and fit into any retail AI workflow without requiring a major revamp of an organization's overall system infrastructure. Input validation will prevent malicious payloads from reaching the inference endpoint, whereas model monitoring will keep constant track of the system's behavior and alert possible drift and other anomalies. The development processes will be resilient to adversarial attacks through adversarial training, with the organization taking on full responsibility for their governance. Through alignment with the MACH architecture and MLOps workflow of most retail operations, as well as compatibility with an enterprise's IAM solutions, the suggested strategies allow for a non-disruptive implementation of resilience to retail AI ecosystems through built-in security measures.

3. Benefits to the Industry

Implementing the proposed multi-layered approach to AI security produces noticeable improvements in many aspects of retail operations.

Fraud Loss Prevention

Protecting the system from adversarial attacks on the AI systems responsible for discounts and loyalty programs will result in lower losses related to fraud. According to the National Retail Federation, retail theft caused losses of more than $112 billion for American retailers in 2022 [1]. Fraud mechanisms based on artificial intelligence are becoming a significant share of this threat, with hackers increasingly aiming at recommendation engines and dynamic pricing systems. Synthetic attack data generation frameworks confirm that predictive analytics integrated with fraud detection can cut organized retail crime losses significantly [5].

Model Integrity and Decision Accuracy

Proactive defenses against data poisoning preserve the accuracy of AI-driven decisions in pricing, inventory, and recommendations. IBM Security research indicates that organizations with mature AI security frameworks experience up to a 38% reduction in AI-related operational incidents [2]. It directly contributes to the improvement of revenue integrity and satisfaction ratings. Drift monitoring and distribution analysis have ensured stability in recommendation engines, avoiding amplification of bias and consistent decision-making accuracy. Adversarial detection studies indicate that black-box detectors need to be strengthened continually for maintaining model integrity [3].

Regulatory Compliance

Embedded governance controls within the model ensure compliance with the newly established AI regulations. As an example, the EU AI Act suggests that the high-risk AI systems must include the principles of transparency, robustness, and human oversight [7]. By integrating provenance and human in the loop into their algorithms, retailers can reduce their liabilities. This practice is consistent with international data protection regulations because the application of the AI technology respects all the regulations. The federated learning approaches shed light on the privacy-preserving methods of regulatory compliance [12].

Customer Trust and Brand Protection

The customer engagement through artificial intelligence like recommendations, responses, and price quotes needs to be secured to avoid brand reputation destruction. As mentioned by the Edelman Trust Barometer, 71% of people will stop buying goods from that company because of an AI trust breach [4]. Layered defense approaches ensure that customer-facing AI is trustworthy, safeguarding brand equity. Multi-agent approaches for responsible use of AI indicate that transparency and privacy preservation lead to higher consumer engagement with automated systems [10].

Operational Resilience

The automation of rollbacks and quarantining features reduces MTTR after any form of breach against the AI system. The ability to operate the autonomous retail business during threats is crucial, especially when there are high volumes of sales. Reports have highlighted that resilience models improve efficiency and protect revenue during high-volume sales [9]. Observability and rollbacks have been found to reduce recovery time by over 40% when incorporated into MLOps pipelines.

Dimension	Benefit	Challenge
Fraud Prevention	Reduced exposure to AI-facilitated fraud	Anonymous access undermines zero-trust principles
Decision Accuracy	Stable recommendations and pricing	Continuous retraining increases vulnerability
Regulatory Compliance	Alignment with transparency and oversight mandates	Heterogeneous estates complicate uniform control
Customer Trust	Preserved brand reputation	Latency constraints risk degrading user experience
Operational Resilience	Faster recovery from compromise	A shortage of AI security expertise delays adoption

Table 5: Benefits vs. Challenges in Retail AI Security [9]

4. Challenges and Constraints

However, in spite of the potential benefits, there are several structural elements in the retail context that make cybersecurity using AI particularly challenging.

Unfettered & Anonymous Accessibility

In retail, unfettered accessibility of systems for anonymous or guest users is an important aspect. This requirement goes against the concept of zero-trust security. The application of any kind of strong authentication or increased friction can lead to decreased conversion rates due to cart abandonment. According to research, it has been pointed out that anonymity serves as a critical channel through which credential fraud occurs [6].

Continuous Model Retraining Cycles

Retail AI models are frequently retrained on live interaction data to maintain relevance in fast-moving consumer environments. This continuous learning loop creates a persistent vulnerability window in which poisoned data can be incorporated into model weights before anomalies are detected. Unlike static enterprise AI deployments, retail requires real-time security controls rather than periodic audits. According to analysts, dynamic retraining frequency increases susceptibility to poisoning attacks [2]. Studies on continuous learning pipelines indicate that attackers use retraining cycles to expedite model deterioration [5].

Heterogeneous Technology Environments

Major retailers span legacy enterprise resource planning (ERP), cloud-native, third-party marketplaces, and in-store edge computing infrastructure. Implementing uniform AI security measures within this heterogeneous environment is architecturally complex. Variability in data provenance, access controls, and monitoring capabilities complicates uniform defense. Professional security reports emphasize that fragmented estates increase the difficulty of implementing cohesive AI protection strategies [4]. Research on hybrid systems demonstrates that inconsistent monitoring among properties results in blind spots for the attacker [9].

Lack of AI Security Skills in the Retail Sector

The combination of adversarial ML research, security in large language models, and retail industry expertise forms an uncommon specialty. According to the ISC² Cybersecurity Workforce Study, the shortage of cybersecurity workers globally was expected to amount to 4 million by 2024 [1]. AI skills are part of the most pressing shortages. Retail businesses struggle against companies from the banking and tech industries. This shortage delays adoption of advanced defense frameworks and increases reliance on external vendors. Research confirms that talent scarcity is a primary barrier to AI security maturity in retail [8].

Latency and Performance Constraints

Real-time AI retail implementations such as dynamic pricing, real-time inventory distribution, and immediate fraud decisioning need to be implemented under stringent time budget requirements, frequently necessitating less than 100 milliseconds of time response. The implementation of security validation components, anomaly detection chains, and adversarial input filtration needs to be done while ensuring little impact on

latency costs. Poor performance will have a direct effect on conversion ratios and experience metrics for the consumer. Surveys from industry experts indicate that latency continues to be an essential problem in implementing advanced AI security systems in retail operations [3].

Conclusion

Incorporating generative and agentic AI into retail brings about both groundbreaking possibilities and significant cybersecurity concerns. The very nature of retail platforms requires that they be open to the outside world, which inevitably means that anonymous and guest engagements are prevalent within the system. Such scenarios offer malicious actors an entry point into conducting data poisoning, prompt injection, and model inversion attacks. The multi-layered security approach that was explained through input validation, monitoring of models, and adversarial resilience with governance creates an organized route for integrating security into the retail business process. Input validation prevents the malicious signals from getting to the inference point, monitoring keeps a check on the model's performance, and governance measures ensure accountability as well as increase resilience from the adversaries. Implementing such a solution will assist organizations in protecting their decision accuracy, remaining compliant with new regulations, and maintaining consumer trust in AI-based processes. However, issues like anonymous access, ongoing training procedures, mixed IT environments, talent shortage, and latency concerns represent the challenges of securing AI systems in the retail industry. Finding solutions to the problem will demand adopting an approach to evolve with adversarial tactics. In conclusion, the way for the retail industry in relation to artificial intelligence is to seek ways to innovate while remaining resilient. This can be achieved through implementing security measures at all stages of AI adoption.

References

[1] Mueen Uddin et al., "Generative AI revolution in cybersecurity: a comprehensive review of threat intelligence and operations," Artificial Intelligence Review, vol. 58, article no. 236, May 2025. https://link.springer.com/article/10.1007/s10462-025-11219-5

[2] Ziying Yang et al., "Invisible Threats in the Data: A Study on Data Poisoning Attacks in Deep Generative Models," Applied Sciences, vol. 14, no. 19, p. 8742, Sep. 2024. https://www.mdpi.com/2076-3417/14/19/8742?utm_source=copilot.com

[3] Vitalii Fishchuk and Daniel Braun, "Robustness of generative AI detection: adversarial attacks on black-box neural text detectors," International Journal of Speech Technology, vol. 27, pp. 861–874, Oct. 2024. https://link.springer.com/article/10.1007/s10772-024-10144-2

[4] Geeta Sandeep Nadella et al., "Generative AI-Enhanced Cybersecurity Framework for Enterprise Data Privacy Management," Computers, vol. 14, no. 2, p. 55, Feb. 2025. https://www.mdpi.com/2073-431X/14/2/55

[5] Garima Agrawal, Amardeep Kaur, and Sowmya Myneni, "A Review of Generative Models in Generating Synthetic Attack Data for Cybersecurity," Electronics, vol. 13, no. 2, p. 322, Jan. 2024. https://www.mdpi.com/2079-9292/13/2/322

[6] Ajay Bandi et al., "The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges," Future Internet, vol. 17, no. 9, p. 404, Sept. 4, 2025.

https://www.mdpi.com/1999-5903/17/9/404

[7] Vijay Kanabar and Kalinka Kaloyanova, "Securing Generative AI Systems: Threat-Centric Architectures and the Impact of Divergent EU–US Governance Regimes," Journal of Cybersecurity and Privacy, vol. 6, no. 1, p. 27, Feb. 2026. https://www.mdpi.com/2624-800X/6/1/27

[8] Peter Adebowale Olujimi et al., "Agentic AI Frameworks in SMMEs: A Systematic Literature Review of Ecosystemic Interconnected Agents," AI, vol. 6, no. 6, p. 123, Jun. 2025. https://www.mdpi.com/2673-2688/6/6/123

[9] Ed Kamya Kiyemba Edris, "Utilisation of Artificial Intelligence and Cybersecurity Capabilities: A Symbiotic Relationship for Enhanced Security and Applicability," Electronics, vol. 14, no. 10, p. 2057, May 2025. https://www.mdpi.com/2079-9292/14/10/2057

[10] Abhinav Tiwari and Hany E. Z. Farag, "A Responsible Generative Artificial Intelligence

Based Multi-Agent Framework for Preserving Data Utility and Privacy," AI, vol. 7, no. 1, p. 1, Dec. 2025. https://www.mdpi.com/2673-2688/7/1/1

[11] Nicolas Papernot et al., "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIA CCS '17), pp. 506–519, Apr. 2, 2017.https://dl.acm.org/doi/10.1145/3052973.3053009

[12] Edi Marian Timofte et al., "Federated Learning for Cybersecurity: A Privacy-Preserving Approach," Applied Sciences, vol. 15, no. 12, p. 6878, Jun. 2025. https://www.mdpi.com/2076-3417/15/12/6878

[13] Yang et al., "A Deep Reinforcement Learning Framework for Influence Maximization Problems on Large-Scale Social Networks," Scientific Reports, vol. 16, Article number: 11515, Mar. 2026. https://www.nature.com/articles/s41598-026-41731-9?utm_source=copilot.com