JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2025, VOL 8, NO S11

Defending The AI-Powered Commerce Stack: A Security Framework For Prompt Injection, Review Integrity, And Privacy In Genai Retail Systems

Prakash Kodali

Sri Venkateswara University, India

Abstract

Generative AI is fundamentally changing digital retail through intelligent search, conversational assistants, personalized recommendations, and generating dynamic content. The use of generative AI has created significant security gaps that disrupt customer trust, regulatory compliance, and operational integrity. Prompt injection attacks exploit untrusted content in product descriptions and user-generated reviews to manipulate assistant behavior and trigger unauthorized actions. AI-generated synthetic reviews undermine rating authenticity and distort marketplace signals at an unprecedented scale. Data poisoning compromises catalog systems and vector embeddings that power recommendation engines, degrading relevance and introducing malicious content propagation. Privacy leaks emerge from over-permissioned tool access and inadequate PII protection in conversational contexts. This framework presents layered defense architectures addressing each threat vector through input isolation, provenance tracking, quarantine systems, and access minimization. Cross-cutting governance mechanisms integrate brand guardrails, policy-as-code enforcement, and human-in-the-loop controls. Observability infrastructure enables continuous monitoring through decision logging, drift detection, and executive dashboards. The framework provides actionable guidance for engineering, security, legal, and customer experience teams building resilient AI-powered commerce systems that balance innovation with protection against adversarial manipulation and privacy violations.

Keywords: Generative AI security, E-commerce threat modeling, Prompt injection defense, Review integrity management, Privacy-preserving commerce

1. Introduction: Intelligent Commerce Systems and Contemporary Security Threats

Generative AI is transforming digital retail by enabling intelligent searching, conversational assistants, personalized recommendations, and creating dynamic content. Generative AI has created significant security vulnerabilities that can break customer trust, regulatory compliance, and operational integrity. Prompt injection attacks use untrusted content present in product descriptions and user-generated reviews to manipulate assistant behavior and trigger unconsented actions. AI-generated synthetic reviews threaten the authenticity of ratings and distort marketplace signals at a scale that can only be defined as unprecedented. Data poisoning attacks can compromise catalog systems and vector embeddings that enable recommendation engines by degrading relevance and introducing malicious content into properties. Privacy leaks can arise from over-permissioned tooling access and insufficiently protected PII present in conversational contexts. The presented framework introduces layered defense architectures, addressing each of the threat vectors through input isolation, provenance tracking, quarantine systems, and access minimization. Cross-cutting governance mechanisms embed the enforcement of brand

guardrails, policies as code, and human-in-the-loop. The observability infrastructure is a means for continuous monitoring through decision logging, drift detection, and executive dashboards. The framework provides actionable routing for engineering, security, legal, and customer experience teams developing resilient AI-powered commerce while balancing operational cost with adaptive defenses against adversarial manipulation and privacy violations.

1.1 Generative Intelligence Reshaping Digital Retail

Sophisticated language models have empowered commercial applications to evolve from fixed automation to flexible agent-based systems. Today's applications are capable of semantic query understanding, multi-faceted product attribute evaluation, contextual descriptions, and multi-dimensional purchasing workflows. Documented developments have shown value in areas including support desk automation, predictive analytics for inventories, supply chain optimization, and customized product display [1]. The technical implementation moves from query systems leveraging conversational understanding to dialog-based shopping assistants, automated sentiment analysis of customer feedback, recommendation engines incorporating total product knowledge, and pricing algorithms adjusting to competition and user behavioral patterns.

1.2 Transition from Fixed Systems to Adaptive Architectures

The shift from rule-based computing frameworks to probability-driven intelligent agents marks a significant architectural departure. Earlier commercial platforms relied on explicit decision trees and predetermined process flows, while generative technologies employ contextual reasoning and produce variable outputs. Business benefits include improved transaction completion rates, higher customer satisfaction measurements, and more efficient resource utilization. However, the non-deterministic nature of these technologies, alongside their processing of diverse unstructured information sources, creates security vulnerabilities distinct from those encountered in traditional web application environments.

1.3 Emerging Vulnerability Patterns in Intelligent Retail

Generative AI integration within commercial systems introduces three distinct vulnerability categories that intensify conventional security concerns. Operational requirements compel these systems to process information from marketplace contributors, community-generated materials, external supplier databases, and publicly accessible content sources. Standard input validation approaches designed for structured, schema-validated information prove insufficient when addressing unconstrained text, multimedia elements, and mixed-format content where malicious instructions may be concealed using methods that bypass typical filtering mechanisms.

1.4 Elevated Permissions and Consequential Operations

Authority granted to intelligent agents extends past data retrieval into operations carrying direct financial implications. Platforms authorizing price modifications, discount applications, order alterations, refund processing, and customer record access operate with permissions whose misuse could generate significant monetary losses or compliance failures. As organizations connect AI agents to proprietary APIs and internal databases for seamless user interactions, they establish pathways through which corrupted agent operations may cascade into broader infrastructure failures.

1.5 Accelerated Contamination Across Technical Infrastructure

Architectural patterns in AI-enhanced commerce accelerate the spread of compromised information throughout connected systems. Product specifications and community contributions undergo conversion into mathematical representations powering search functionality and suggestion algorithms. Once these representations are corrupted, they systematically influence product visibility across countless searches and purchase decisions. Unlike conventional databases, where incorrect entries impact isolated events, tainted representations introduce ongoing distortion across all dependent operations until identification and correction occur. Feedback mechanisms within suggestion frameworks may magnify initial quality problems, as AI-influenced user actions create signals reinforcing existing anomalies or manipulations.

1.6 Affected Parties and Compliance Environment

Security weaknesses in generative AI commerce create consequences reaching beyond technical concerns to affect various stakeholder groups and regulatory spheres. Consumer trust, fundamental to digital commerce success, declines sharply when intelligent systems deliver flawed recommendations, distorted

prices, or skewed product rankings. Market participants increasingly expect AI-enhanced shopping to provide both convenience and verifiable fairness, clear operations, and strong privacy safeguards. Breaches occurring through embedded instruction attacks generating false information, fabricated reviews distorting quality indicators, or exposure incidents revealing confidential details produce lasting reputation damage extending past immediate technical fixes.

1.7 Regulatory Requirements and Organizational Alignment

Legal frameworks impose complex restrictions on AI security within commercial operations. Consumer protection legislation requires accurate product descriptions and equitable pricing practices. Privacy statutes establish strict boundaries for personal information collection, processing timeframes, and storage durations. Digital accessibility mandates ensure equal service provision across user demographics. Competition oversight examines algorithmic pricing and recommendation mechanisms for anti-competitive behaviors. Operational challenges include maintaining autonomous system compliance with diverse regulatory mandates while these systems continuously adjust to changing data patterns and situational contexts [2]. Effective security for AI commerce requires exceptional coordination across engineering, information security, legal counsel, compliance monitoring, and customer experience teams that have historically maintained separate operational scopes and specialized terminology.

1.8 Foundational Security Approach

Addressing security challenges in AI-powered commerce requires treating information sources and operational capabilities with governance rigor traditionally reserved for executable code. Standard software security methodology mandates review processes, validation testing, and access controls before deployment, with change histories maintained through versioning systems and formal approval workflows. Similar rigor must apply to information consumed by AI systems and functions that these systems can invoke. Product descriptions, customer feedback, and supplier information require validation pipelines capable of detecting and eliminating embedded commands before content enters AI processing workflows.

1.9 Separation, Limitation, Validation, and Documentation

Functions and interfaces available to AI agents require permission configurations enforcing minimal access rights, timeout controls preventing resource monopolization, and thorough audit records documenting each invocation with sufficient forensic detail. Separation architectures divide untrusted information from privileged functions through protective boundaries, limiting damage when defensive controls fail. Retrieved information includes metadata indicating source, custody chain, and reliability assessment, allowing AI systems to weight information appropriately and favor verified sources over unattributed submissions. Limitation frameworks restrict permissible agent operations, requiring human authorization for elevated-risk actions such as large refunds or major price changes, while providing preview capabilities showing potential outcomes before execution.

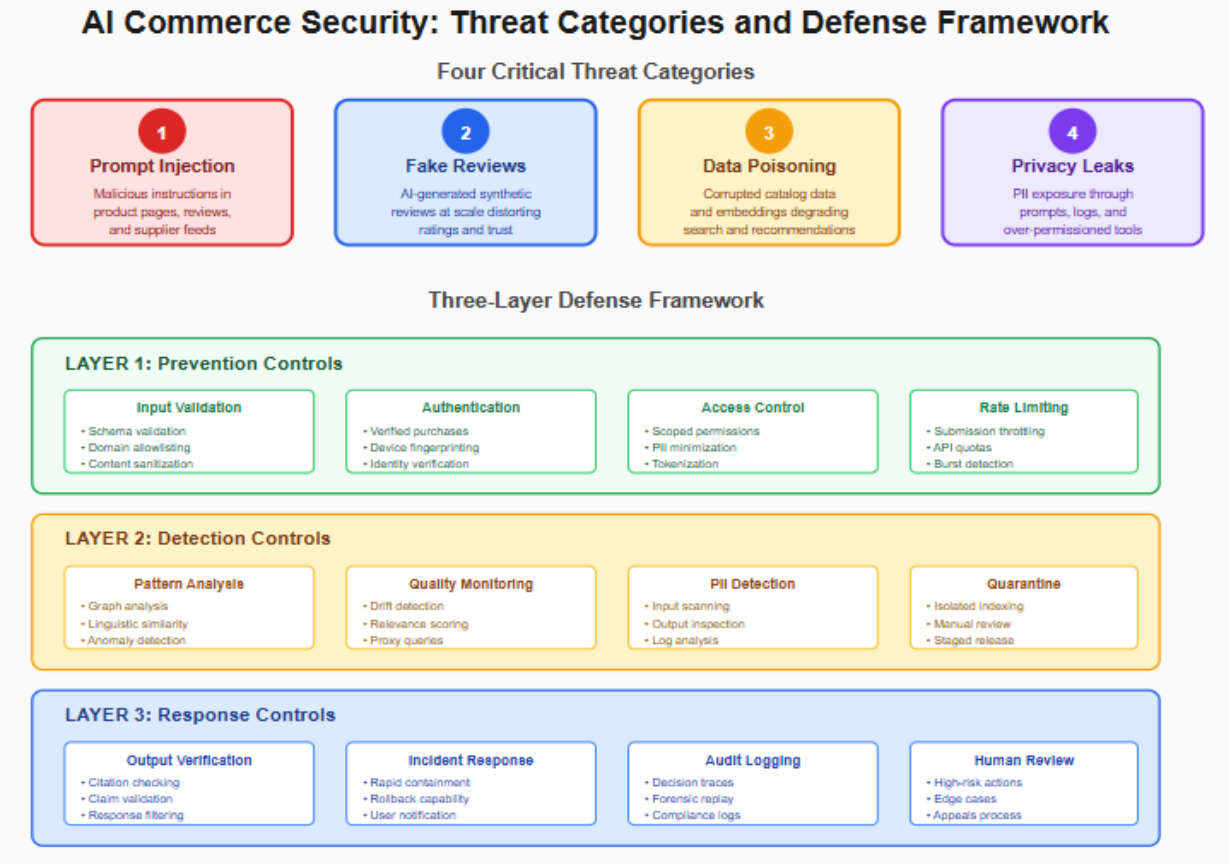
1.10 Transparency and Layered Protection

Validation mechanisms assess AI-produced outputs against trusted reference materials, flag unsubstantiated claims, and block responses displaying unusual characteristics, including unverified links or competitor references. Documentation systems preserve decision lineage connecting information inputs, retrieval results, function executions, and produced responses within searchable repositories, enabling continuous oversight and retrospective investigation. This transparency framework extends past standard application monitoring to capture model behavior characteristics, certainty measurements, logic sequences, and exceptional situations, activating backup procedures. Together, these separation, limitation, validation, and documentation principles form a multi-tier protective strategy recognizing the probabilistic basis of AI technologies while maintaining security guarantees and operational dependability requirements critical for commerce operations.

Table 1: AI Commerce Attack Surface Taxonomy [1][2][3][4]

Attack Category	Entry Points	Threat Mechanism	Affected Components	Primary Impact
Prompt Injection	Product descriptions, reviews, supplier feeds, and help documentation	Embedded malicious instructions override system behavior	AI assistants, chatbots, search systems	Unauthorized discounts, false information, competitor promotion
Fake Reviews	Review submission forms, API endpoints, and bulk import tools	Synthetic content generation at scale	Rating systems, search rankings, recommendation engines	Distorted product scores, eroded trust, biased visibility
Data Poisoning	Supplier uploads, UGC pipelines, web scraping, third-party catalogs	Corrupted attributes contaminate embeddings	Vector databases, product catalogs, recommendation models	Degraded search relevance, inappropriate suggestions
Privacy Leakage	Conversational prompts, tool APIs, logging systems, analytics pipelines	Excessive PII exposure in processing flows	Customer databases, transcript logs, model contexts	Regulatory violations, unauthorized disclosure

Fig. 1: AI Commerce Security Framework - Threat Categories and Defense Mechanisms



2. Prompt Injection Attacks in Commerce Systems

Prompt injection constitutes a significant vulnerability in AI-enhanced retail platforms where adversarial text concealed within information sources manipulates language model operations to generate unintended results. Contrasting with conventional injection exploits that compromise databases or system commands through structured syntax, prompt injection capitalizes on natural language understanding by inserting hostile directives into content that models process as authentic instructions [3][4]. Commercial environments magnify these dangers because e-commerce systems ingest enormous volumes of unverified material from sellers, shoppers, and third-party sources, each potentially harboring concealed commands intended to undermine platform security.

2.1 Commerce-Focused Injection Characteristics

Within e-commerce ecosystems, prompt injection manifests as a specialized threat where hostile text buried in merchandise details, assistance materials, or shopper contributions attempts to supersede platform directives and modify AI operations. The vulnerability exploits core language model design, which handles retrieved material and embedded commands without an innate capacity to differentiate trusted platform instructions from malicious directives hidden in external sources [3]. Retail platforms encounter heightened vulnerability because their intelligent systems must access and evaluate content from inherently unreliable sources to enable practical shopping functionality, establishing numerous compromise pathways absent in restricted applications.

2.2 Contrasts with Conventional Injection Exploits

Established injection vulnerabilities, including SQL injection or shell command injection, exploit structured interfaces where attackers introduce special symbols or syntax to escape data boundaries and trigger unauthorized execution. Prompt injection functions through fundamentally different mechanics by capitalizing on semantic comprehension abilities within language models [4]. Instead of exploiting parser weaknesses or special character handling, prompt injection harnesses the model's linguistic competence to embed directives appearing contextually legitimate while serving hostile purposes. HTML-embedded injection methods illustrate how attackers conceal malicious prompts within apparently harmless web materials that enter the model's processing scope [3]. The adaptive characteristics of these exploits, where directives can be expressed through innumerable semantically identical variations, make signature-based identification approaches largely unsuccessful [4].

2.3 Compromise Pathways in Digital Retail

E-commerce infrastructures expose numerous surfaces through which prompt injection exploits may enter AI workflows. Merchandise pages and assistance documentation accessed by AI shopping advisors to address customer inquiries constitute primary vulnerability points, as merchants or hostile actors may embed concealed directives within descriptions, technical specifications, or guidance materials. These directives could instruct the AI to promote particular merchandise irrespective of shopper requirements, reveal competitor strategy information, or authorize improper discounts. Compromised vendor feeds transmitted as PDF files, HTML documents, or structured datasets provide additional injection channels, especially when external content receives inadequate examination before incorporation into product databases consulted by AI systems [3].

2.4 Shopper Contributions as Compromise Vectors

Customer-generated materials, including merchandise evaluations, inquiry-response sections, and discussion boards, represent particularly challenging vulnerability surfaces because commercial platforms intentionally present this content to AI advisors to deliver genuine shopper insights. Hostile actors can post evaluations or inquiries containing embedded commands engineered to activate when an AI advisor accesses that material during later customer engagements. The adaptive properties of prompt injection allow attackers to disguise malicious directives using diverse linguistic constructions that bypass straightforward pattern recognition [4]. The magnitude of customer-generated content on substantial e-commerce platforms renders manual inspection impractical, while expectations that AI systems incorporate current shopper feedback create urgency to minimize delay between contribution and accessibility.

2.5 Multi-Tier Protection Framework

Successful mitigation of prompt injection in retail systems demands multiple protection tiers functioning at distinct phases of AI processing workflows. No individual safeguard provides adequate protection given the semantic adaptability of natural language exploits, requiring comprehensive approaches that diminish vulnerability surfaces, restrict AI capability scope, and validate results before customer presentation. The framework must reconcile security mandates against functional requirements for AI systems to consult varied content repositories and execute valuable retail functions.

2.6 Content Restriction and Cleaning Procedures

Content restriction establishes initial protective barriers by confining information sources to approved domains, maintaining established reliability relationships, and implementing thorough cleaning to eliminate potential injection mechanisms. Script removal eliminates executable elements from HTML and PDF materials before content enters language model processing [3]. Command detection systems examine retrieved text for characteristics suggesting embedded directives, including expressions attempting to supersede platform operations, atypical formatting indicating concealed text, or content inconsistent with anticipated domain context. Domain approval prevents AI systems from accessing material from arbitrary internet locations, constraining exposure to managed vendor portals and authenticated customer feedback channels.

2.7 Content Sourcing and Origin Documentation

Content sourcing practices guarantee that every text segment handled by AI systems includes metadata specifying its origin, reliability classification, and verification status. Origin documentation associates each accessed passage with particular document markers, timestamps, and custody records, enabling forensic examination when questionable results appear. Verification protocols employ cryptographic signatures or secure checksums to confirm that content originates from stated sources and remains unaltered during transmission or retention. These procedures enable AI systems to preferentially trust information from verified sources over unconfirmed content and impose stricter examination on material from less reliable origins [4].

2.8 Capability Reduction and Authorization Boundaries

Capability reduction decreases potential exploit consequences by constraining the functions AI agents can execute and the resources they can consult. Bounded authorizations ensure each function invocation grants exclusively the minimal privileges required for legitimate task completion, preventing compromised agents from consulting extensive customer repositories or financial infrastructure. Duration limits terminate functions surpassing anticipated completion intervals, countering attempts to employ AI agents for resource depletion or prolonged reconnaissance. Manual approval mandates require human authorization for critical functions, including substantial refunds, major price adjustments, or sensitive customer information access, ensuring that injected directives cannot independently trigger high-impact operations.

2.9 Result Validation and Assertion Verification

Result validation procedures examine AI-produced responses before customer delivery to identify indicators of successful prompt injection. Reference mandates require factual assertions to cite particular trusted sources, enabling confirmation that statements correspond with authoritative documentation rather than injected falsehoods. Assertion verification cross-examines AI declarations against product catalog information, pricing repositories, and policy databases to detect inconsistencies suggesting manipulation. Deviation identification flags responses containing unexpected components such as competitor references, unvetted hyperlinks, or policy breaches that may signal successful injection exploits. These validation tiers function with negligible latency consequences through concurrent processing and caching of validation information.

2.10 Oversight Infrastructure and Performance Indicators

Extensive oversight infrastructure monitors signals revealing both successful exploits and protective mechanisms. Performance indicators include tallies of thwarted injection attempts identified through content examination, capability misuse prevention incidents where authorization boundaries prevented unauthorized functions, platform latency measurements confirming security mechanisms satisfy performance expectations, and erroneous blocking rates quantifying unnecessary restriction of legitimate

content. Pattern examination identifies evolving exploit approaches requiring defensive modification, while deviation identification surfaces unusual agent operations warranting scrutiny.

2.11 Traceability Systems and Event Reconstruction

Traceability systems maintain comprehensive decision records linking accessed content, platform directives, function invocations, and produced results in tamper-resistant logs supporting forensic inquiry and compliance demonstration. Event reconstruction capability allows security personnel to reconstruct the precise information conditions and logic sequences that yielded particular results, enabling root cause determination when injection exploits succeed. Preservation strategies balance investigative requirements against storage expenses and privacy obligations, typically maintaining granular logs for recent engagements while archiving condensed information for historical examination. This transparency infrastructure fulfills dual objectives of incident remediation and continuous enhancement of protective mechanisms through examination of exploit attempts and safeguard failures.

Table 2: Prompt Injection Defense Mechanisms [3][4]

Defense Layer	Control Mechanism	Implementation Method	Detection Capability	Limitation Addressed
Input Isolation	Domain allowlisting, script stripping	Whitelist approved content sources, remove executable code	Identifies unauthorized domains, suspicious formatting	Prevents injection from untrusted sources
Retrieval Hygiene	Source tagging, cryptographic attestation	Metadata association, digital signatures	Traces content provenance, verifies authenticity	Enables trust-based content weighting
Tool Minimization	Scoped permissions, timeout controls	Least-privilege access, execution time limits	Blocks unauthorized operations, resource exhaustion	Constrains the impact of successful injection
Output Verification	Citation requirements, claim validation	Cross-reference trusted sources, anomaly detection	Identifies unsupported claims, unexpected content	Prevents the dissemination of injected misinformation
Auditability	Decision logging, replay capability	Comprehensive trace capture, forensic reconstruction	Reveals injection success patterns, control failures	Enables incident investigation and improvement

3. Fake Review Detection and Review Integrity Management

Customer feedback mechanisms serve as foundational trust elements within digital retail environments, shaping purchase choices, search algorithm outputs, and platform reputation. The advancement of generative AI has amplified complications surrounding fraudulent customer assessments, permitting bad actors to generate persuasive synthetic feedback at remarkable velocity and complexity [5][6]. These manufactured evaluations corrupt product scoring systems, distort algorithmic visibility mechanisms, and undermine shoppers' faith in marketplace infrastructures. Maintaining review authenticity demands extensive technical safeguards, open governance structures, and persistent oversight to protect genuine customer input while sustaining operational capacity in large-volume retail contexts.

3.1 Machine-Generated Feedback Complications

Generative AI advancement has radically transformed both the cost structure and identification difficulty of fraudulent feedback operations. Contemporary language models generate contextually suitable,

grammatically accurate, and emotionally convincing review content replicating genuine customer experiences with limited human participation [5]. The production capacity for synthetic feedback has grown dramatically, permitting orchestrated operations to saturate product listings with numerous manufactured assessments within condensed periods. This quantity overwhelms conventional manual oversight methods and strains automated identification systems constructed to recognize less sophisticated fraud signatures.

3.2 Advanced Deception Methodologies

The complexity of machine-generated feedback transcends basic text production to include tactical manipulation of assessment attributes that historically indicated genuineness. Machine learning approaches allow fraudsters to examine authentic feedback structures and duplicate stylistic components, emotional patterns, and compositional characteristics that formerly differentiated real customer input [6]. Synthetic assessments can be customized for particular merchandise categories, integrate appropriate specialized vocabulary, and modify linguistic structures to escape grouping algorithms intended to identify repetitive operations. This flexible capacity demands an identification infrastructure to utilize progressively sophisticated examination techniques that analyze deeper semantic and behavioral indicators rather than superficial textual characteristics.

3.3 Ramifications for Digital Marketplaces

Fraudulent feedback expansion creates compounding adverse impacts across various aspects of digital retail functions. Merchandise scores distorted by synthetic favorable assessments mislead shoppers toward purchasing substandard items, diminishing customer contentment and elevating merchandise return frequencies. Alternatively, orchestrated unfavorable assessment campaigns harm authentic vendors' standings and inhibit transactions for quality merchandise. Search ranking algorithms incorporating assessment indicators as positioning elements transmit the consequences of feedback fraud, systematically skewing product discoverability toward fraudulent entries [5]. Shopper confidence declines as purchasers experience disparity between assessment assertions and actual merchandise quality, diminishing overall marketplace participation and transaction finalization frequencies.

3.4 Multi-Stage Authenticity Pipeline

Constructing dependable review authenticity demands sequential pipelines that confirm reviewer legitimacy, document content origins, identify fraudulent signatures, and sustain transparent governance protocols. No isolated safeguard provides adequate defense against sophisticated fraud enterprises, requiring stratified methods that merge preventive confirmation, analytical identification, and corrective intervention. The pipeline construction must reconcile fraud deterrence goals against user experience factors, reducing obstacles for authentic reviewers while establishing barriers to fraudulent contribution.

3.5 Identity Confirmation and Transaction Validation

Identity confirmation procedures establish the foundational stratum of review authenticity by validating that reviewers maintain genuine commercial relationships with assessed merchandise. Transaction validation restricts feedback privileges to shoppers who have finalized authenticated purchases for particular items, removing the capacity of fraudsters lacking purchase records to contribute manufactured assessments [6]. Customer program integration broadens confirmation by utilizing accumulated behavioral information and account records to evaluate reviewer reliability, with mature accounts receiving elevated trust designations than recently established profiles. Additional authentication obligations for premium merchandise classifications impose supplementary obstacles to automated fraud operations while maintaining availability for genuine shoppers.

3.6 Origin Documentation and Attribute Recording

Thorough attribute recording documents the comprehensive origin sequence for each contributed assessment, capturing source platform, hardware attributes, network markers, contribution timestamps, and behavioral indicators that facilitate forensic examination and signature identification. Hardware identification methods recognize questionable signatures such as numerous assessments originating from matching equipment arrangements or network locations [5]. Timestamp examination exposes temporal concentration that may signal orchestrated fraud operations. Machine-generated markers designate contributions displaying linguistic attributes aligned with automated text creation, activating enhanced

examination without automatically dismissing potentially authentic content. This attribute infrastructure enables both immediate fraud identification and retrospective inquiry following recognition of fraud networks.

3.7 Signature Recognition and Connection Examination

Sophisticated examination techniques inspect assessment populations for signatures suggesting orchestrated fraud enterprises. Connection examination builds network models of reviewers, merchandise, and contribution associations, recognizing questionable groups where numerous accounts display correlated operations such as assessing matching merchandise collections within narrow periods [6]. Linguistic resemblance identification utilizes natural language processing to recognize assessments sharing exceptional textual correspondence, indicating template-driven generation or duplication functions. Identity grouping algorithms consolidate behavioral indicators to recognize accounts controlled by shared operators despite superficial variation. Surge identification observes the contribution pace to designate abrupt increases of assessments for particular merchandise, a characteristic signature of compensated fraud operations [5].

3.8 Oversight Structure and Regulation Openness

Open oversight structures establish explicit regulations controlling assessment acceptance, transparently convey these benchmarks to marketplace contributors, and supply organized procedures for disputing moderation choices. Public regulation revelation records the standards utilized to assess assessment genuineness, forbidden assessment activities, and ramifications for regulation breaches, creating mutual comprehension of permissible conduct. Published regulations diminish the uncertainty that fraudsters might leverage and supply an authentic understanding of platform requirements, reducing unintentional regulation breaches.

3.9 Dispute Protocols and Manual Examination

Dispute protocols supply organized pathways through which reviewers can challenge deletion choices, contribute supplementary authentication proof, or explain potentially unclear content. Manual examination for boundary situations guarantees that automated identification infrastructure functioning with incomplete precision does not inappropriately suppress authentic assessments displaying atypical but genuine attributes [6]. Qualified moderators inspect flagged content, considering situational elements that automated infrastructure may inadequately recognize, reconciling fraud deterrence against maintenance of genuine customer expression. Dispute information returns into the identification infrastructure enhancement, recognizing signatures of incorrect positives demanding algorithmic modification.

3.10 Organized Information Administration and Search Coordination

Organized information administration guarantees that assessment summaries presented to search platforms precisely represent current platform conditions following fraud elimination. Schema.org adherence arranges assessment attributes according to normalized vocabularies that search platforms process when building merchandise knowledge structures and establishing search positions [5]. Prompt modifications to organized information following eliminations prevent fraudulent assessments from persistently affecting search prominence even after platform deletion. Version management for summary statistics permits platforms to exhibit authenticity to both shoppers and search platforms, sustaining marketplace reputation.

3.11 Effectiveness Quantification and Consequence Evaluation

Thorough effectiveness quantification monitors both functional measurements reflecting identification infrastructure capability and commercial results exhibiting authenticity program worth. Transaction-validated assessment proportions measure the segment of total assessments linked with confirmed purchases, supplying a fundamental marker of baseline authenticity. Moderation velocity gauges the rapidity at which flagged content obtains manual inspection, reconciling fraud reaction urgency against resource limitations. Elimination frequencies monitor deleted content quantities, while dispute frequencies signal potential excessive enforcement demanding calibration modifications [6].

3.12 Commercial Result Observation

Commercial consequence measurements connect authenticity programs to fundamental business goals, exhibiting program worth to organizational participants. Engagement frequency examination inspects

whether enhanced assessment quality increases customer interaction with merchandise entries. Transaction finalization frequency monitoring evaluates whether authentic assessment populations stimulate elevated purchase completion frequencies compared to fraud-contaminated reference points [5]. Customer contentment surveys and merchandise return frequency observation supply direct input on whether assessment authenticity enhancements convert to superior purchase choice quality. These commercial measurements validate persistent investment in fraud deterrence infrastructure and direct resource distribution toward maximum-consequence authenticity programs.

Table 3: Review Integrity Pipeline Components [5][6]

Pipeline Stage	Control Type	Technical Implementation	Verification Method	Fraud Signal Detected
Authentication	Preventive	Verified purchase linking, loyalty program integration	Transaction record validation	Reviewers without purchase history
Provenance Tracking	Detective	Device fingerprinting, timestamp capture, and AI detection flags	Metadata analysis	Multiple reviews from identical sources
Pattern Detection	Detective	Graph analysis, linguistic similarity, burst detection	Network clustering, NLP comparison	Coordinated campaigns, template reuse
Human Oversight	Corrective	Manual review of edge cases, appeals processing	Expert judgment, contextual evaluation	False positives, ambiguous content
Structured Data	Corrective	Schema.org compliance, aggregate updates	Search engine validation	Persistent fraud influences post-removal

4. Data Poisoning and Product Brain Integrity

The product brain constitutes the aggregate intelligence framework driving contemporary e-commerce operations, incorporating catalog repositories, vector representations, suggestion algorithms, and retrieval mechanisms that convert raw merchandise details into customized shopper experiences. Data contamination exploits target this framework by introducing hostile or corrupted details at multiple entry locations, methodically undermining system capability and distorting customer-facing results [7][8]. Contrasting with discrete security violations affecting singular transactions, contamination exploits corrupt fundamental data structures that shape countless engagements across prolonged durations. Safeguarding product brain authenticity demands thorough safeguards covering data entry, retention, handling, and result production, paired with persistent observation to identify and correct contamination before extensive distribution.

4.1 Intelligence Framework Components

The intelligence framework design includes linked elements that convert varied information origins into operational knowledge controlling customer engagements. Catalog infrastructures preserve organized merchandise details, incorporating specifications, costs, stock levels, and hierarchical classification that constitute the definitive record for product selections. Vector representations convert textual merchandise narratives, visual materials, and behavioral indicators into multidimensional mathematical forms, permitting semantic retrieval and resemblance calculation [7]. Suggestion mechanisms process these representations alongside patron engagement records to produce tailored merchandise proposals, while

retrieval infrastructures employ both organized catalog information and vector forms to understand inquiries and position outcomes.

4.2 Information Movement Design

Information travels through numerous conversion phases from initial entry to customer-facing display, with each phase introducing potential susceptibility to contamination exploits. Raw merchandise details are entered through vendor channels, manual information entry, web collection functions, and customer submissions. Entry pipelines analyze, authenticate, and standardize this heterogeneous input into uniform schemas appropriate for catalog retention. Representation creation procedures convert textual and visual material into vector forms retained in specialized repositories optimized for resemblance retrieval [8]. Suggestion and retrieval algorithms query these repositories immediately, merging vector resemblance measurements with business regulations and customization indicators to produce positioned results. Each conversion phase magnifies the consequence of upstream contamination, as corrupted information is distributed through dependent infrastructures and shapes progressively extensive customer groups.

4.3 Vendor Information Distortion

Vendor-supplied merchandise details constitute a substantial contamination pathway because commercial operations must absorb information from countless external origins to preserve thorough merchandise catalogs. Hostile vendors can introduce deceptive characteristics, improper classifications, or keyword manipulation designed to distort retrieval positions and suggestion arrangement [7]. Subtle characteristic distortion, such as exaggerated capability specifications or minimized merchandise measurements, may bypass elementary authentication while methodically skewing customer choices. Orchestrated exploits across numerous vendor profiles can saturate particular merchandise classifications with contaminated information, overwhelming manual inspection capacity, and creating corrupted reference points that later authentication considers typical.

4.4 Uncontrolled Patron Submissions

Patron-created material incorporating evaluations, inquiries, responses, and merchandise visuals flows immediately into infrastructures that produce representations and educate suggestion frameworks without the thorough examination applied to vendor information. Hostile actors can contribute material containing contaminated details concealed as genuine customer interactions, intentionally introducing skewed merchandise connections into representation domains [8]. Large-quantity contributions of synthetic patron materials can alter semantic associations within vector repositories, prompting suggestion infrastructures to improperly connect merchandise or promote substandard items. The requirement that operations incorporate current patron submissions with limited postponement establishes temporal urgency that diminishes examination and enables contamination distribution.

4.5 External Catalog Susceptibilities

Coordination with external merchandise repositories and classification structures introduces contamination dangers when outside origins lack uniform quality safeguards or experience hostile manipulation. Schema disparities between outside catalogs and internal infrastructures establish mapping uncertainties that hostile actors leverage to introduce malformed information that circumvents authentication procedures constructed for primary information origins [7]. Reliance on outside classifications for merchandise categorization permits hostile actors who breach these infrastructures to methodically miscategorize merchandise across dependent operations. Confidence associations with recognized information suppliers can establish hazardous presumptions about information authenticity, diminishing alertness for progressive quality decline or focused contamination operations.

4.6 Polluted Education Information

Suggestion infrastructures and representation frameworks educated on web-collected information or combined operation engagements inherit contamination existing in education collections. Public web material incorporates considerable quantities of distorted merchandise details, manufactured evaluations, and optimization-focused interference that, when integrated into education repositories, instruct frameworks to duplicate contaminated signatures [8]. Historical engagement information utilized for cooperative screening can mirror previous contamination operations, prompting suggestion infrastructures to maintain rather than rectify earlier pollution. The computational cost of re-educating substantial

frameworks establishes urgency to optimize information employment rather than forcefully screen potentially contaminated specimens, exchanging education productivity for information quality.

4.7 Stratified Protection Approach

Successful safeguarding against information contamination demands multi-tier designs executing safeguards at every phase of information entry, conversion, retention, and employment. No isolated protective procedure supplies thorough protection because hostile actors persistently modify to leverage spaces between safeguard strata. Stratified protections establish numerous independent chances to identify and obstruct contamination efforts, diminishing the likelihood that sophisticated exploits successfully breach core product brain elements. The approach reconciles deterrence of contamination entry against identification and correction of pollution that circumvents initial safeguards.

4.8 Entry Safeguards and Authentication

Schema authentication imposes rigorous structural and semantic limitations on arriving information, dismissing contributions that breach anticipated arrangements, information categories, or referential consistency obligations. Category imposition prevents classification confusion exploits where adversaries contribute numeric information in text locations or Boolean quantities in numeric spans to leverage downstream handling presumptions [7]. Span inspection authenticates that quantitative characteristics fall within reasonable boundaries for their merchandise classifications, obstructing efforts to assert impossible capability attributes or illogical measurements. List restrictions confine categorical characteristics to predetermined authentic quantities, preventing introduction of hostile-managed classifications that might distort retrieval grouping or suggestion procedures.

4.9 Isolation Framework

Isolation infrastructures direct information displaying questionable attributes to separate repositories, where material experiences are enhanced for examination before coordination into production infrastructures. Questionable markers incorporate rapid contribution quantities from recent profiles, characteristic patterns substantially divergent from historical classification standards, or material flagged by automated quality evaluation instruments [8]. Isolation separation prevents potentially contaminated information from promptly shaping customer-facing results while permitting exhaustive inspection without obstructing authentic contributions. Progressive release protocols incrementally present isolated material to restricted patron groups, observing for irregular behavioral indicators before complete production implementation. Preservation protocols reconcile isolation capacity limitations against the requirement to preserve questionable material for forensic examination following verified contamination episodes.

4.10 Origin and Duration Administration

Thorough origin documentation connects every information component with attributes recording its origin, entry timestamp, conversion record, and authentication condition. Origin markers permit rapid recognition and elimination of all information beginning from breached vendors or profiles when contamination is identified [7]. Duration protocols automatically terminate information components after predetermined intervals, compelling revalidation against current origins and constraining the temporal consequence of historical contamination. Conversion record documentation maintains comprehensive accounts of handling phases applied to each information component, enabling forensic rebuilding of how contaminated information was distributed through the infrastructure and recognizing which downstream elements require correction.

4.11 Quality Observation and Deviation Identification

Representative query infrastructures persistently perform characteristic retrieval and suggestion solicitations, contrasting current results against anticipated reference points to identify relevance decline, indicating contamination or quality deviation. Automated relevance evaluation inspects whether top-positioned outcomes suitably correspond query purpose utilizing both algorithmic measurements and synthetic patron input [8]. Deviation identification algorithms observe statistical patterns of representation vectors, catalog characteristics, and suggest signatures, warning when measurements diverge from historical standards in manners indicating methodical contamination rather than organic progression.

Irregularity recognition flags abrupt increases in particular merchandise prominence, atypical classification connections, or correlation signatures inconsistent with genuine patron conduct.

4.12 Restoration and Reversal Competencies

Versioned retention preserves historical records of representations, catalog conditions, and suggestion framework settings, permitting rapid restoration to last-authenticated arrangements when contamination is identified. Reversal protocols must thoroughly evaluate temporal reach to prevent canceling authentic modifications while removing contaminated information, typically demanding forensic examination to recognize the contamination introduction [7]. Progressive correction approaches selectively eliminate or substitute verified contaminated components rather than wholesale restoration, reducing interference to authentic current modifications. Concurrent handling competencies permit production and authentication of purified representations and repositories without interrupting production infrastructures, permitting exhaustive quality confirmation before transition to corrected editions.

4.13 Functional Measurements

Thorough functional measurements supply prominence into both contamination exploitation occurrence and protective safeguard capability. Invalid characteristic frequencies measure the segment of absorbed information dismissed by authentication safeguards, creating reference points for typical dismissal concentrations and emphasizing irregular increases indicating orchestrated contamination operations [8]. Origin documentation quantifies the proportion of catalog components with comprehensive origin records, recognizing spaces in the documentation framework that could hide contamination origins. Isolation duration monitors the interval between questionable material recognition and disposition choice, reconciling exhaustive inspection against customer interaction consequences from postponed material accessibility. Deviation warning resolution duration gauges how rapidly identified irregularities obtain inquiry and correction, immediately associated with the vulnerability period during which contaminated information shapes customer engagements.

5. Privacy Protection and PII Management in Commerce Assistants

Privacy safeguarding represents a critical obligation for AI-enhanced retail advisors that handle confidential shopper details during dialogue engagements and independent transaction completion. Personally identifiable details travel through numerous infrastructure elements incorporating dialogue prompts, instrument activations, choice recordings, and evaluation conduits, establishing countless disclosure channels where insufficient safeguards can produce unapproved revelation or preservation breaches [9][10]. Regulatory structures incorporating GDPR, CCPA, and domain-specific privacy legislation enforce rigorous duties on personal data treatment, obligating organizations to deploy technical and administrative procedures guaranteeing secrecy, authenticity, and patron authority. Successful PII administration reconciles operational obligations for tailored guidance against privacy mandates through detail reduction, entry limitations, duration oversight, and open patron safeguards.

5.1 Privacy Danger Environment

AI retail advisors face privacy dangers across numerous functional aspects where personal data handling happens. The dialogue character of these infrastructures establishes urgency to supply extensive situational details to language frameworks to permit precise comprehension and tailored replies, potentially revealing more shopper information than absolutely required for assignment fulfillment [9]. Instrument connections that link AI advisors to shopper repositories, purchase records, and settlement infrastructures create channels through which surplus authorizations can provide wider information entry than singular functions demand. Recording framework capturing dialogue interactions and infrastructure choices for troubleshooting, examination, and framework enhancement may preserve personal details past authentic preservation intervals or reveal them to evaluation personnel lacking suitable entry permission [10].

5.2 Situational Prompt Revelation

Situational prompts built to supply AI advisors with shopper background details constitute a principal PII revelation procedure. Infrastructure architects encounter motivations to incorporate thorough shopper descriptions, transaction records, navigation signatures, and demographic characteristics in prompts to

optimize tailoring excellence and reply precision [9]. Nevertheless, this habit broadens the vulnerability domain by positioning confidential details in framework context spaces where prompt introduction exploits, recording habits, or framework education information gathering might reveal it. External framework APIs that handle prompts outside organizational framework boundaries present supplementary revelation dangers when shopper information is transmitted to outside suppliers. The difficulty magnifies as dialogue engagements gather context across numerous interactions, steadily broadening the quantity of personal details functioning in the infrastructure.

5.3 Surplus Authorization Entry

AI advisors linked to backend infrastructures through API connections commonly obtain wider authorizations than singular functions require, breaching minimal-authorization concepts fundamental to privacy safeguarding. An advisor accessing purchase conditions for shopper inquiry may obtain credentials providing entry to complete purchase repositories rather than limited authorizations confined to the particular shopper's accounts [10]. Settlement handling instruments might reveal complete credit identification quantities when exclusively final-segment confirmation is operationally demanded. Shopper assistance connections could allow adjustment of profile configurations when read-exclusive entry would satisfy inquiry settlement. These surplus authorizations establish privacy dangers both through potential AI misconduct stemming from prompt introduction or framework mistakes and through broadened vulnerability domains for credential robbery or infrastructure breach.

5.4 Unshielded Evaluation and Recording

Dialogue recordings and evaluation conduits constructed to enhance AI advisor capability commonly capture personal details without suitable shielding or entry safeguards. Development personnel examining dialogue signatures to recognize enhancement chances may face shopper identities, locations, purchase particulars, and monetary details incorporated in recorded interactions [9]. Evaluation infrastructures combining usage signatures across shopper groups might preserve detailed particulars, permitting singular re-recognition despite the combination purpose. Extended preservation intervals for troubleshooting and framework education objectives can breach regulatory obligations constraining personal information retention duration. The division between functional infrastructures with rigorous entry safeguards and evaluation surroundings with wider personnel entry establishes potential privacy spaces where personal details travel to persons lacking an authentic requirement.

5.5 Privacy-Through-Construction Safeguard Structure

Successful privacy safeguarding demands privacy-through-construction designs where technical safeguards impose detail reduction, entry limitations, and duration administration throughout infrastructure construction rather than depending on procedural adherence. These safeguards must function openly to patrons, supply detailed arrangement choices honoring singular privacy inclinations, and preserve comprehensive examination records exhibiting regulatory adherence [10]. The structure includes both preventive procedures constraining initial PII revelation and investigative safeguards recognizing privacy breaches demanding correction.

5.6 Detail Reduction and Attribute-Tier Entry

Detail reduction concepts limit information movements to the smallest required for particular operational objectives, deployed through attribute-tier entry safeguards that prevent infrastructures from consulting characteristics beyond functional obligations. AI advisors accessing shopper details obtain exclusively attributes applicable to present assignments rather than complete shopper descriptions [9]. Substitution replaces confidential markers like profile quantities or electronic correspondence locations with non-confidential surrogates in situations where immediate quantities are unnecessary. Prompt building procedures apply algorithmic screening to eliminate personal detail classifications immaterial to present dialogue situations before transmitting information to language frameworks. These reduction safeguards diminish both unapproved disclosure risks and regulatory compliance risks by constraining the classifications and quantities of personal data stored in infrastructure.

5.7 Entry Administration and Duration Limitations

Limited entry credentials impose minimal authorization concepts by providing AI advisors restricted authorizations confined to particular resources, functions, and duration spaces. Competency limitations

prevent advisors from executing functions past established operational boundaries, such as obstructing adjustment functions when exclusive read entry is demanded [10]. Resource limiting confines repository inquiries to particular shopper accounts rather than allowing table-wide entry. Duration limitations deploy automatic authorization termination, demanding re-confirmation for prolonged sessions and preventing abandoned or breached credentials from preserving unlimited entry. These entry administration safeguards establish multi-tier strata where numerous independent permissions must be accomplished for confidential functions, diminishing the consequences from singular safeguard breakdowns.

5.8 Duration Oversight and Preservation Administration

Duration oversight creates preservation regulations and technical implementation procedures guaranteeing personal detail elimination or anonymization when authentic handling objectives terminate. Dialogue transcript preservation deploys duration-restricted regulations automatically removing comprehensive recordings after intervals adequate for instant functional requirements like conflict settlement [9]. Shielded evaluation repositories maintain combined signatures and anonymized specimens, enabling framework enhancement while eliminating personal markers that would prolong preservation duties. Progressive preservation designs preserve elevated-detail current information, enabling functional obligations while steadily diminishing detail concentrations for historical details, reconciling evaluation worth against privacy reduction. Automated elimination procedures remove manual participation obligations that present postponement and disparity in preservation regulation implementation.

5.9 Patron Rights and Openness

Open privacy habits and reachable patron safeguards create confidence and permit persons to practice regulatory rights over their personal details. Direct-language privacy announcements convey information gathering objectives, handling functions, preservation intervals, and external distribution in vocabulary comprehensible to non-technical populations [10]. These announcements prevent legal terminology and surplus length that discourage reading, concentrating on details most applicable to patron choice-making about assistance participation.

5.10 Individual Entry and Elimination Procedures

Individual entry procedures permit shoppers to access duplicates of personal details organizations preserve about them, satisfying regulatory obligations under GDPR, CCPA, and comparable legislation. Automated access infrastructures gather information across distributed retention locations, incorporating transactional repositories, dialogue recordings, evaluation databases, and reserve infrastructures [9]. Confirmation protocols validate solicitor legitimacy while reducing supplementary personal detail gathering exclusively for entry solicitation handling. Elimination procedures deploy thorough deletion across all retention locations, incorporating reserve infrastructures and evaluation derivatives, with technical confirmation validating complete removal. These procedures must finalize within regulatory periods while preserving examination records exhibiting adherence with elimination solicitations.

5.11 Permission Administration Coordination

Permission administration infrastructures capture, retain, and impose patron privacy inclinations across AI advisor engagements. Detailed permission choices allow patrons to independently permit distinct handling functions such as tailoring, evaluation participation, and external distribution [10]. Permission capture interfaces display selections transparently without deceptive signatures that distort patrons toward privacy-reducing choices. Implementation procedures inspect the permission condition before commencing handling functions, obstructing functions lacking suitable permission. Permission withdrawal competencies permit patrons to cancel formerly provided authorizations, activating information elimination procedures for details gathered under the cancelled permission. Coordination with legitimacy administration infrastructures guarantees permission inclinations continue across sessions and apparatus, supplying uniform privacy safeguard interactions.

5.12 Identification and Reaction

Privacy episode identification infrastructures observe AI advisor functions for PII revelation breaches, demanding instant reaction. Automated PII identification examines both inputs obtained from patrons and results produced by advisors, recognizing personal detail classifications appearing in situations breaching

privacy regulations [9]. Signature recognition recognizes surplus information entry efforts, atypical inquiry signatures against shopper repositories, or irregular detail movements to evaluation infrastructures. Immediate obstruction prevents privacy breaches from contacting patrons or outside infrastructures when identification happens before result transmission. Warning elevation informs security and privacy personnel of potential episodes demanding inquiry, with severity classification permitting suitable reaction emphasis.

5.13 Episode Reaction Protocols

Organized episode reaction protocols establish positions, communication arrangements, and correction phases for privacy breaches. Initial evaluation establishes breach reach, incorporating affected persons, revealed information classifications, and outside revelation scope. Limitation procedures stop continuing breaches through entry cancellation, infrastructure separation, or assistance suspension, depending on episode severity [10]. Shopper announcement satisfies regulatory duties and preserves confidence through open communication about breach situations and correction functions. Regulatory reporting adheres to the violation announcement obligations under relevant privacy legislation. Following-episode examination recognizes fundamental origins and deploys corrective procedures preventing repetition, with discoveries informing persistent enhancement of privacy safeguards.

5.14 Adherence Measurements and Observation

Thorough adherence measurements supply prominence to privacy safeguard capability and regulatory compliance. PII shielding documentation gauges the segment of recorded information successfully shielded or substituted before evaluation handling, creating reference protection concentrations and recognizing spaces demanding correction [9]. Obstructed surplus-authorization efforts measure occasions where entry safeguards prevented instruments from consulting information beyond limited permissions, exhibiting minimal-authorization implementation capability. Preservation compliance measurements monitor information elimination promptness against regulation obligations, recognizing procedure obstacles producing adherence postponements. Patron contentment gauges evaluate whether privacy safeguards and openness procedures satisfy shopper requirements, reconciling security obligations against patron interaction [10]. These measurements enable both functional observation and regulatory adherence exhibitions, supplying proof of authentic-faith privacy safeguarding attempts.

Table 4: Privacy Control Framework [9][10]

Control Domain	Privacy Principle	Technical Implementation	Enforcement Mechanism	Compliance Metric
Data Minimization	Collect only necessary information	Field-level access control, tokenization	Programmatic filtering, surrogate substitution	PII redaction coverage percentage
Access Management	Least-privilege authorization	Scoped tokens, capability restrictions, and temporal constraints	Permission verification at invocation	Blocked over-permission attempts
Lifecycle Governance	Time-limited retention	Transcript TTLs, masked analytics lakes	Automated deletion workflows	Retention adherence rate
User Rights	Individual control and transparency	Plain-language notices, subject access workflows	Self-service portals, verification procedures	Request fulfillment timeliness
Incident Detection	Real-time violation identification	PII detection in inputs/outputs, anomaly monitoring	Automated blocking, alert escalation	Detected exposure incidents

Conclusion

The integration of generative artificial intelligence into e-commerce platforms has created unprecedented opportunities for personalized customer experiences while simultaneously introducing complex security vulnerabilities requiring comprehensive defensive strategies. Prompt injection attacks, synthetic review proliferation, data contamination operations, and privacy exposure incidents represent interconnected threat categories that demand coordinated technical controls, transparent governance frameworks, and persistent monitoring infrastructure. Organizations deploying AI-enhanced retail systems must implement layered defenses spanning input validation, access restriction, provenance documentation, and output verification to protect against adversarial manipulation. Authentication mechanisms preserving review integrity, quarantine systems isolating suspicious content, and privacy-by-design architectures minimizing personal information exposure constitute essential components of secure AI commerce operations. Regulatory compliance obligations under consumer protection and privacy statutes necessitate not merely reactive incident response but proactive control implementation demonstrating good-faith security efforts. The dynamic nature of AI security threats requires continuous adaptation of defensive measures as attack techniques evolve and new vulnerabilities emerge. Success demands cross-functional collaboration among engineering, security, legal, and customer experience teams, supported by executive commitment to balancing innovation velocity against risk management imperatives. Ultimately, sustainable competitive advantage in AI-powered commerce depends upon establishing customer trust through demonstrated security capabilities and transparent privacy practices.

References

[1] Shervin Ghaffari, et al., "Generative-AI in E-Commerce: Use-Cases and Implementations," in 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), 25 March 2024. https://ieeexplore.ieee.org/document/10475266

[2] Nir Kshetri, "Generative Artificial Intelligence and E-Commerce," IEEE Access, vol. 12, pp. 15776-15800, 31 January 2024. https://ieeexplore.ieee.org/document/10417762

[3] Ionuţ-Vlăduţ Dinu, et al., "Disrupting Large Language Models with Hidden Prompt Injection Attacks Embedded in HTML Pages," in 2025 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2025 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), 15 July 2025. https://ieeexplore.ieee.org/document/11075247

[4] Zhilong Wang, "To Protect the LLM Agent Against the Prompt Injection Attack with Polymorphic Prompt," in 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S), 09 July 2025. https://ieeexplore.ieee.org/document/11068353

[5] Pasupathi Pandi. T and N. Siva Kumar, "Fake Review Detection in E-Commerce Using Machine Learning and NLP Technique," in 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), 15 July 2025. https://ieeexplore.ieee.org/document/11069636

[6] Ansh Vashist, et al., "Detecting Fake Reviews on E-commerce Platforms Using Machine Learning," in 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), 17 January 2025. https://ieeexplore.ieee.org/document/10837502

[7] Chenwang Wu, et al., "Influence-Driven Data Poisoning for Robust Recommender Systems," IEEE Transactions on Knowledge and Data Engineering, 10 May 2023. https://ieeexplore.ieee.org/document/10122715

[8] Zhiye Wang, et al., "Revisiting Data Poisoning Attacks on Deep Learning Based Recommender Systems," in 2023 IEEE Symposium on Computers and Communications (ISCC), 28 August 2023. https://ieeexplore.ieee.org/document/10218302

[9] Dorababu Nadella, "Securing Data at Rest: Using ML-Driven Personally Identifiable Information(PII) Detection and Privacy-Preserving Techniques," in 2024 International Conference on Engineering and Emerging Technologies (ICEET), 12 March 2025. https://ieeexplore.ieee.org/document/10913953

[10] Victor Morel, et al., "AI-Driven Personalized Privacy Assistants: A Systematic Literature Review," in 2024 IEEE International Conference on Trust, Privacy and Security in Intelligent Systems (TPS), 11 September 2025. https://ieeexplore.ieee.org/document/11159499