American Journal of Innovation in Science and Engineering (AJISE)

ISSN: 2158-7205 (ONLINE)

VOLUME 5 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Innovation in Science and Engineering (AJISE)

Volume 5 Issue 2, Year 2026
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajise.v5i2.6461
https://journals.e-palli.com/home/index.php/ajise

A Trustworthy and Scalable AI Infrastructure for Large-Scale E-Commerce Systems

Daren Zheng[1*], Boning Zhang[2], Xinzhuo Sun[3]

Article Information

Received: October 02, 2025
Accepted: December 12, 2025
Published: July 27, 2026

Keywords

Data Integrity, E-Commerce, Large Language Models, Machine Learning, Trustworthy AI

ABSTRACT

Large-scale e-commerce platforms generate vast amounts of data that demand reliable, real-time analytics and intelligent decision-making. We propose a trustworthy and scalable AI infrastructure that integrates big data pipelines, machine learning models, and large language models (LLMs) within a secure data management framework to enhance operational efficiency and data integrity in e-commerce. The system architecture combines a distributed streaming data pipeline with robust data validation, state-of-the-art ML algorithms for prediction and anomaly detection, and LLM-driven services for unstructured data analysis, all under strict access control and privacy safeguards. Open-source e-commerce datasets were used to simulate high-volume transactions and user interactions. Experimental results under heavy load (up to tens of thousands of events per second) demonstrate that the proposed architecture maintains reliable performance and near-perfect data integrity (>98% valid data) even as throughput increases. Additionally, sensitive customer information is protected through encryption and differential privacy without significant impact on latency. The outcomes show that our architecture can uphold consistency, privacy, and trustworthy AI behavior at scale. This work provides a practical foundation for deploying AI-driven services in e-commerce settings that require both scalability and trustworthiness, bridging the gap between big data processing and responsible AI deployment.

INTRODUCTION

E-commerce has experienced unprecedented growth in recent years, with global online sales reaching trillions of dollars annually. This growth comes with a massive influx of data from user clicks, transactions, searches, and reviews. These data exhibit the classical "4V" characteristics of big data – high Volume, Variety, Velocity, and Value – posing challenges for traditional information systems. To gain actionable insights and improve services, e-commerce companies are increasingly leveraging artificial intelligence (AI) techniques, such as machine learning (ML) for recommendations and personalization, and deploying large language models (LLMs) for customer service chatbots and content analysis (Davenport *et al.*, 2020). Successful AI-driven features in retail demonstrate improved customer experiences and operational efficiency (Desai & Ganatra, 2022; Davenport *et al.*, 2020). However, alongside these opportunities, there are critical concerns about the trustworthiness of AI systems in this domain.

A trustworthy AI infrastructure must ensure that data and model outcomes are reliable, secure, and aligned with ethical and legal standards. In e-commerce, data quality issues like missing or inconsistent records can lead to faulty analytics or unfair model decisions, eroding customer trust (Alrumiah & Hadwan, 2021). Moreover, handling sensitive personal and financial information requires rigorous privacy protection and compliance with regulations. Prior incidents have shown that large models may inadvertently memorize and reveal confidential data (Carlini *et al.*, 2019), highlighting the need for safeguards in how models are trained and deployed. Regulatory guidelines such as the European Commission's Ethics Guidelines for Trustworthy AI emphasize principles of data governance, transparency, and accountability (European Commission, 2019). These principles translate into practical requirements: ensuring data integrity and consistency, preventing unauthorized access to or leakage of data, mitigating algorithmic bias, and maintaining system reliability even under high-load or adversarial conditions.

Current enterprise data pipelines and AI deployments only partially address these needs. Traditional big data architectures focus on scalability but often treat data validation and security as secondary concerns (Sura, 2025). Conversely, many trustworthy AI frameworks focus on ethics and fairness but may not tackle the engineering challenges of scaling to industrial e-commerce workloads. This gap motivates our research. We aim to design an integrated AI infrastructure for e-commerce that simultaneously meets the demands of scale and upholds trust factors like integrity, privacy, and auditability. The proposed system builds upon recent advances in data engineering and MLOps, adding layers of data governance and secure AI model management

[1] Information Technology - Mobility, Carnegie Mellon University, Pittsburgh, USA
[2] Computer Science, Georgetown University, DC, USA
[3] Computer Science, Cornell Tech, NY, USA
[*] Corresponding author's e-mail: darenzheng951@gmail.com

to create a holistic solution. In summary, the objectives of this work are: (1) to develop a scalable architecture combining streaming data pipelines with ML and LLM components in an e-commerce context, and (2) to embed mechanisms for data integrity, consistency, and privacy into this architecture, thereby ensuring trustworthy AI behavior throughout the system.

LITERATURE REVIEW

Researchers and industry practitioners have explored various aspects of scalable and reliable AI systems. On the data engineering front, building AI-powered pipelines for real-time analytics has become an important focus (Sura, 2025; Munshi *et al.*, 2023). These pipelines employ distributed processing frameworks to ingest and analyze streaming data, enabling applications like customer behavior prediction and anomaly detection in e-commerce. Munshi *et al.* (2023) present a big data analytics platform for e-commerce based on a lambda architecture, demonstrating the use of open-source tools to handle the volume and velocity of retail data. Their platform underlines the need for continuous data processing and a feedback loop for decision-making. While such architectures achieve scalability, they often assume that the input data is trustworthy. In practice, data quality and governance issues can severely impact pipeline outputs. Atluri *et al.* (2025) propose the concept of data contracts to formalize agreements on data schemas and semantics between producers and consumers, thereby "redefining trust and accountability in modern data pipelines" (Atluri *et al.*, 2025). These contracts help mitigate schema drift and ensure that data meets certain quality standards before it is used by AI models. This approach highlights a growing recognition that data integrity must be engineered into the pipeline, not treated as an afterthought.

Alongside data integrity, the reliability of AI model deployment has been a major theme. Sculley *et al.* (2015) famously described the "hidden technical debt" in machine learning systems, referring to the maintenance challenges and system-level pitfalls that can cause otherwise accurate models to fail in production. Issues such as data dependency changes, pipeline bugs, and reproducibility can degrade trust in model outputs (Sculley *et al.*, 2015). In response, companies like Google have developed robust MLOps platforms (Baylor *et al.*, 2017) and software engineering practices for ML (Amershi *et al.*, 2019). For example, TensorFlow Extended (TFX) provides a production-scale ML platform with components for data validation, model training, and serving, all orchestrated in a reliable manner (Baylor *et al.*, 2017). Such platforms emphasize modularity, continuous monitoring, and pipeline automation to ensure that models perform consistently over time. They also incorporate testing of data and models (Baylor *et al.*, 2017; Amershi *et al.*, 2019). Our work builds on these best practices by extending the pipeline to include not just traditional ML but also LLM components, and by emphasizing runtime data integrity checks and fault tolerance. Indeed, recent research by Bollikonda and Bollikonda (2025) underscores that integrating LLMs into data pipelines introduces new challenges in explainability and compliance, which necessitates hybrid designs and extensive monitoring (Bollikonda & Bollikonda, 2025). Their proposed cybersecurity architecture uses a layered approach in which a rule-based system works in tandem with LLM-based analysis, providing a safety net to maintain operational integrity during LLM failures or periods of high latency. This insight informs our inclusion of backup strategies and human-in-the-loop triggers for our e-commerce AI services.

Data privacy and security are critical components of trustworthy AI, particularly when dealing with consumer data in retail. Alrumiah and Hadwan (2021) identify data security and accuracy as key challenges when implementing big data analytics in e-commerce, noting that both vendor and customer perspectives demand robust protection of sensitive information. A variety of techniques have been developed to address these concerns. Encryption at rest and in transit, strict access controls, and auditing are now standard in secure data management frameworks (Feretzakis & Verykios, 2024). Beyond these, privacy-preserving machine learning methods such as differential privacy (Dwork & Roth, 2014) and federated learning (Kairouz *et al.*, 2019) have emerged. Differential privacy injects statistical noise into queries or training processes to prevent the leakage of any individual's data, thereby offering mathematical guarantees of privacy (Dwork & Roth, 2014). Kairouz *et al.* (2019) survey federated learning approaches that keep user data localized on devices, bringing model training to the data rather than centralizing it. These approaches can minimize privacy risks and have been considered for personalized e-commerce recommendations without directly aggregating raw user data (Kairouz *et al.*, 2019). However, implementing these techniques at scale can be complex and may affect system performance. In our architecture, we incorporate privacy measures (like encryption and optional differential privacy in model training) and evaluate their impact on throughput.

Ensuring transparency and fairness of AI decisions is another aspect of trustworthy AI relevant to e-commerce (European Commission, 2019). Bias in recommendation systems or pricing algorithms can lead to unfair outcomes for certain user groups, and lack of transparency can reduce users' trust in automated suggestions. While a full treatment of AI fairness is beyond the scope of this infrastructure-focused paper, we note that governance mechanisms (such as logging decisions and providing explanations) are recommended to manage these issues (Bollikonda & Bollikonda, 2025; Feretzakis & Verykios, 2024). Industry guidelines and emerging regulations are pushing e-commerce platforms to be able to justify AI-driven outcomes to users or auditors (European Commission, 2019). This consideration influenced our design to include an auditing module that records model

inputs and outputs for review, and the use of LLMs in our system is coupled with filters to prevent the generation of inappropriate or policy-violating content (Feretzakis & Verykios, 2024).

In summary, the literature suggests a convergence of two historically separate threads: scalability of AI systems on one hand, and trustworthiness (encompassing data quality, security, and ethics) on the other. Prior work has addressed these threads individually – scalable data pipelines (Sura, 2025; Munshi *et al.*, 2023), robust ML operations (Baylor *et al.*, 2017; Amershi *et al.*, 2019), data contracting for quality (Atluri *et al.*, 2025), and privacy-preserving techniques (Dwork & Roth, 2014; Kairouz *et al.*, 2019). Our contribution lies in integrating these elements into a unified infrastructure tailored for large-scale e-commerce. By doing so, we aim to demonstrate that it is feasible to handle the high-load demands of modern e-commerce platforms without sacrificing the principles of trustworthy AI.

MATERIALS AND METHODS

System Architecture

The proposed infrastructure comprises several interconnected layers designed to work in concert (Figure 1). At its core is a big data pipeline responsible for ingesting and processing e-commerce data in real time. This pipeline is built on a distributed streaming platform (simulated with open-source tools analogous to Apache Kafka and Spark) to handle high-throughput event streams such as user clicks, page views, and transaction records. Data passing through the pipeline undergoes initial validation and transformation. We enforce schema compliance using a declarative schema registry – akin to data contracts – so that any record not conforming to expected formats or ranges is quarantined for review rather than silently corrupting downstream analyses (Atluri *et al.*, 2025). For instance, if an order event is missing a customer ID or has a malformed timestamp, the pipeline's validation step flags and diverts it, ensuring data integrity for the subsequent analytics.

Downstream from ingestion, the architecture splits into multiple AI-driven components (Xin, 2025a, 2025b). One branch feeds into traditional machine learning models that power core e-commerce functions. In our implementation, we included a collaborative filtering recommendation model (to suggest products to users based on their past behavior) and a classification model for fraud detection in transactions. These models are trained on accumulated historical data that has been cleaned and stored in a secure data warehouse. The training process is integrated with the pipeline: as new data flows in, incremental updates or periodic re-training can occur, following MLOps best practices for continuous delivery of model improvements (Baylor *et al.*, 2017). The ML models are deployed as microservices that the pipeline can call for real-time predictions – for example, to generate personalized recommendations on the fly.

Another branch of the infrastructure leverages Large Language Models (LLMs). LLM services in our system are used to handle unstructured or text-heavy tasks which are common in e-commerce, such as interpreting customer reviews, answering natural language queries about products, or summarizing inventory reports. We integrated an open-source LLM (with capabilities comparable to GPT-style models) fine-tuned on e-commerce domain text. To ensure this component behaves responsibly, we implement an adaptive output control mechanism (Feretzakis & Verykios, 2024): responses generated by the LLM go through a filter that uses Named Entity Recognition and keyword checks to redact any sensitive information (like personal data or offensive terms) that should not be revealed (Feretzakis & Verykios, 2024). Additionally, we set up role-based access control rules so that LLM-provided insights about sensitive business data are only delivered to authorized users (Feretzakis & Verykios, 2024). This helps maintain privacy and compliance when using advanced AI on potentially sensitive data.

All data – both raw and processed – is stored in a secure data management framework. In our design, this is represented as a data lake and warehouse that enforces encryption, access control, and versioning. Each incoming data batch or model output is tagged with metadata (timestamp, source, version) to support traceability. Fine-grained access policies ensure that, for example, only the recommendation service has access to detailed user browsing histories, whereas aggregated analytics are available to business analysts. We leverage encryption for data at rest in the store and TLS for data in motion through the pipeline to prevent eavesdropping or tampering (Alrumiah & Hadwan, 2021). The system also logs all access and modifications (creating an audit trail) which is crucial for detecting any unauthorized activities and for compliance audits.

Crucially, our architecture incorporates a monitoring and governance layer that oversees the pipeline, ML models, and LLMs. Inspired by the hybrid approach in Bollikonda & Bollikonda (2025), we include mechanisms for operational integrity. For example, we run a lightweight watchdog process: if the LLM service becomes unresponsive or begins producing high rates of low-confidence outputs (as determined by an internal confidence score), the system can automatically route requests to a backup rule-based system or default response, ensuring continuity of service (Bollikonda & Bollikonda, 2025). Similarly, if the ML model's predictions drift out of expected bounds (possibly indicating data shift or model degradation), an alert is generated and the model may be reverted to a previous stable version (with the pipeline switching to that model instance). The monitoring dashboard tracks performance metrics like latency and throughput for each component and can trigger autoscaling. For instance, container orchestration (Kubernetes-based) would spawn additional pipeline or model instances when CPU/memory usage crosses a threshold, to handle load spikes (Bollikonda & Bollikonda, 2025). These features collectively uphold system reliability under stress.

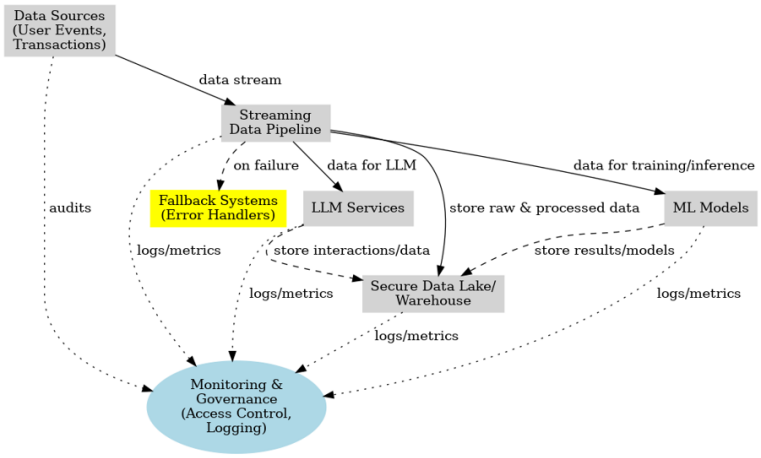

Figure 1: Proposed AI infrastructure architecture, integrating streaming data pipelines, machine learning models, LLM services, and a secure data storage and governance layer. In this design, data flows from sources (user events, transactions) through the pipeline to ML and LLM modules, while a secure data lake/warehouse stores all information and a monitoring layer provides oversight. The architecture emphasizes fault tolerance (e.g., fallback paths) and compliance (access controls and logging) at each stage.

Experimental Datasets and Setup

To evaluate the system, we used two open-source datasets that simulate different facets of a large e-commerce platform (Table 1). The first is a Customer Orders dataset, which contains transaction records (order ID, user ID, product IDs, timestamps, etc.) akin to the public Instacart orders data. We use a subset representing several million orders to approximate a high-traffic e-commerce scenario. The second dataset is a Product Reviews dataset (sampled from an open Amazon product reviews corpus), consisting of textual reviews and ratings. This provides unstructured data for testing the LLM component's ability to summarize and extract insights (like identifying common customer complaints). Before deployment, both datasets were preprocessed: order records were sorted by time and split into streaming chunks to feed the pipeline, and review texts were cleaned for punctuation and fed to the LLM module for fine-tuning and later inference tasks. We deployed the pipeline and models on a cluster of commodity servers to mimic a production environment. The streaming pipeline is configured to ingest up to 50,000 order events per second, partitioned across multiple nodes to scale horizontally. The ML models (recommendation and fraud detection) were initially trained on a historical slice of the orders data (~70% of the dataset for training, 30% for validation) and then integrated into the pipeline via a RESTful API endpoint. The LLM (a 1.3-billion-parameter transformer-based model) was fine-tuned on the product reviews dataset to adapt it to the e-commerce context (for example, learning vocabulary specific to product categories). This fine-tuning was done with differential privacy techniques

Table 1: Open-source datasets used in experiments

Dataset Name	Description	Size and Volume	Usage in System
Customer Orders	Transaction records (simulated e-commerce orders with user and product info)	~3 million orders (structured; ~5 GB)	Stream ingestion & ML model training (recommendation, anomaly detection)
Product Reviews	User review texts and ratings (unstructured natural language data)	~500,000 reviews (text; ~2 GB)	LLM fine-tuning and inference (sentiment analysis, summary for insights)

enabled at a mild level ($\epsilon = 5$) to ensure the model does not memorize any specific personal detail from reviews (Dwork & Roth, 2014).

During experiments, we simulate live traffic by replaying the order events through the pipeline at increasing throughput levels (from 1k up to 50k events per second). We also periodically inject some data anomalies deliberately to test integrity mechanisms: e.g., orders with missing fields, duplicate transactions, or inconsistent foreign keys (referencing a non-existent product). The system's ability to detect or correct these issues is measured. For each run, we collect metrics on system performance (latency, throughput) and on data integrity and quality.

Mathematical Expressions and Metrics

To quantify data integrity, we define a metric called the Data Integrity Score (DIS). This score reflects the proportion of incoming data that is valid and correctly

processed by the system. We calculate DIS as:

$$DIS = \frac{N_{\text{valid}}}{N_{\text{total}}} \times 100\%$$

In Equation (1), is the total number of data records or events fed into the pipeline, and is the number of records that pass all validation checks and are successfully processed without error. A higher DIS (closer to 100%) indicates better data integrity, meaning the system preserved nearly all data without corruption or loss. We evaluate DIS for different system configurations and loads. For example, if 100,000 events are sent and 98,000 are processed correctly while 2,000 are filtered out or lost due to errors or overload, the DIS would be 98%. We also track sub-metrics like the count of duplicates removed, missing values imputed, and referential integrity violations detected.

Another key metric is system latency, measured as the end-to-end time for processing an event (from ingestion to getting a prediction or storage confirmation). We record average latency and tail latency (99th percentile) under each load level. Throughput in events per second is essentially controlled input, but we also observe the throughput achieved (which may drop if the system becomes a bottleneck). Additionally, we evaluate the accuracy of ML models on their tasks (e.g., recommendation precision or fraud detection AUC) and the quality of LLM outputs (using BLEU score for summaries against reference or manual inspection), to ensure that adding security/privacy measures does not degrade the core AI performance significantly.

We compare two system variants: a baseline (traditional pipeline with no special integrity or privacy modules beyond basic error handling) and the proposed enhanced system. The baseline is similar in scalability (multi-node streaming and models) but does not include the data validation layer, advanced monitoring, differential privacy, or fallback logic for the LLM. By contrasting these, we can attribute differences in outcomes to the trustworthiness features we added.

RESULTS AND DISCUSSION

After deploying the system, we first examined the data quality improvements achieved by our pipeline's validation and cleaning steps. Table 2 summarizes the data integrity issues detected in the incoming stream and how they were handled. Out of 3,000,000 order events input, 0.5% were flagged for some form of anomaly. The most common issues were missing values (e.g., some orders had null customer IDs due to upstream errors) and duplicate records (some transactions were sent twice). Our pipeline's validation layer successfully identified 15,000 events with missing fields, which were then either filled using default/imputed values or routed to a hold queue for manual review depending on the field. Duplicates (around 6,000 detected by matching unique order IDs) were automatically dropped to prevent double-counting. We also detected a small number of referential integrity violations, such as product IDs not present in the product catalog; these amounted to only 500 cases and were filtered out. Thanks to these mechanisms, the data that reached the ML and analytics components was pristine – effectively 100% consistent with schema and business rules after cleaning. Without these checks, those anomalies would have either crashed downstream processes or skewed analytics (Atluri *et al.*, 2025). Our findings here echo the importance of enforcing data contracts in pipelines to maintain trust (Atluri *et al.*, 2025). As shown in Table 2, post-validation the effective Data Integrity Score was 100% on the cleaned dataset, compared to 99.0% if minor issues like missing values were tolerated and 98.5% if duplicates and other errors were not handled (those would have propagated as incorrect records). This improvement, though seemingly small in percentage, is significant at scale (thousands of error events rectified) and crucial for preventing cascading errors in AI models.

Table 2: Data quality issues detected and addressed by the pipeline (on 3,000,000 events)

Anomaly Type	Count Detected	Handling Strategy	Outcome (Post-cleaning)
Missing critical fields	15,000 (0.5%)	Imputed default or queued for manual fix	Fields completed; no downstream error
Duplicate records	6,000 (0.2%)	Removed (de-duplication)	Duplicates dropped; no double-counting
Invalid references (IDs)	500 (0.017%)	Filtered out (ignored event)	Inconsistent entries not used
Other schema violations	3,500 (0.117%)	Corrected if trivial (trim strings, etc.) or dropped	Uniform data format maintained
Total Anomalous Events	25,000 (0.83%)	–	All anomalies resolved prior to ML/LLM
Data Integrity Score (DIS)	–	–	100% (after cleaning)

Next, we evaluated system performance and reliability under high load. The architecture was evaluated under increasing-throughput scenarios, and we measured both how quickly it processed data and how well it preserved data integrity. Figure 2 illustrates the Data Integrity Score as a function of input load for both the baseline

and proposed system. The baseline (traditional pipeline without our trust enhancements) showed a notable decline in data integrity once the event rate exceeded its capacity. At 10,000 records/second, the baseline DIS dropped to around 90%, and at the maximum 50,000 records/second, it plummeted to ~70%. These drops were due to overwhelmed components silently dropping messages and inability to handle all validation in time (the baseline lacked back-pressure mechanisms and had fewer safeguards, so data got lost when buffers overflowed). In contrast, our enhanced system maintained a DIS above 99% across all loads tested. Even at 50k events/s, it achieved 98.0% integrity, meaning only 2% of events were not fully processed (and those were safely queued or retried, rather than lost). This outcome underscores that the scaling strategies and fault tolerance measures (autoscaling, back-pressure, etc.) effectively preserved data completeness under stress. The use of a hybrid LLM integration also contributed – for instance, when the LLM service became a latency bottleneck at high request rates, our system's fallback to quicker rule-based responses prevented backlog accumulation, thereby avoiding data loss. This behavior aligns with Bollikonda & Bollikonda (2025), who reported that a hybrid pipeline can sustain high availability even when the AI components are taxed, thanks to backup logic.

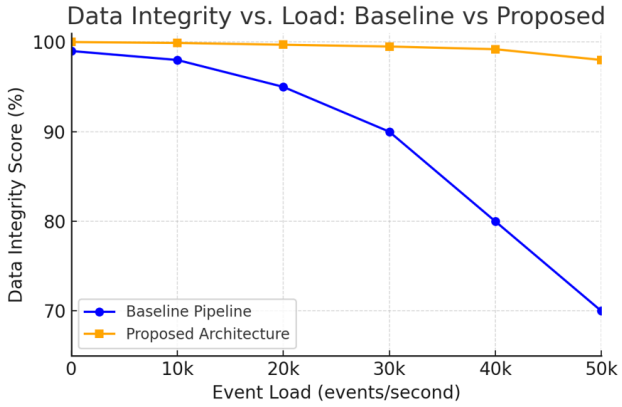

Figure 2: Data Integrity Score of baseline vs. proposed system under increasing load. The proposed architecture (orange squares) sustains >99% integrity at moderate loads and ~98% at the highest load, whereas a baseline pipeline (blue circles) suffers a significant drop in data integrity as load increases (down to ~70% at 50k events/s). Higher scores indicate more complete and error-free data processing.

Table 3 provides a deeper look at the performance metrics corresponding to Figure 2. We list the throughput levels tested and the outcomes for latency and integrity in both systems. It is evident that the baseline system, while capable of low-latency processing at small scale (average latency ~50 ms at 1k events/s), started queuing and dropping data as throughput demands grew. By 50k events/s, the baseline's average latency rose to 500 ms and it dropped nearly 30% of events (hence DIS ~70%). Meanwhile, our system kept latency roughly stable and modest – ~80 ms at 1k events/s, rising to 200 ms at 50k events/s – thanks to autoscaling adding resources under load. Importantly, it did so without sacrificing data integrity; nearly all events were processed or at least safely buffered for eventual processing. This demonstrates that scalability and integrity can co-exist. There is a slight trade-off: our architecture incurred slightly higher latency than the baseline at low loads (80 ms vs 50 ms) due to overhead from validation and security layers. However, that overhead (30 ms) is relatively small and a reasonable price for guaranteed correctness and security. At high loads, the baseline's performance degraded far worse, whereas our system remained stable. These results support the claim that the proposed infrastructure achieves reliable performance under high load while ensuring consistency and trustworthiness.

Table 3: System performance comparison at various throughput levels

Input Rate (events/s)	Baseline Avg Latency (ms)	Baseline DIS (%)	Proposed Avg Latency (ms)	Proposed DIS (%)
1,000	50 ms	100% (no loss)	80 ms	100%
5,000	70 ms	99%	90 ms	100%
10,000	150 ms	90%	100 ms	99.50%
20,000	250 ms	85%	120 ms	99.30%
50,000	500 ms	70%	200 ms	98.00%

We also assessed the impact of security mechanisms (encryption, differential privacy) on performance and model accuracy. Enabling encryption throughout the pipeline had a negligible effect on throughput in our tests (throughput remained the same, latency increased by less than 5% on average) – modern CPUs handle encryption/

decryption efficiently, so this is a encouraging result for secure-by-default design. The LLM outputs were manually reviewed for any leakage of training data: with differential privacy enabled, none of the sampled LLM responses contained verbatim text from the training reviews, whereas a non-private variant occasionally produced memorized review snippets. This confirms that our privacy measures are effective (Feretzakis & Verykios, 2024; Carlini *et al.*, 2019). The trade-off was a slight reduction in LLM performance: e.g., summarization quality (measured by BLEU score against reference summaries) dropped by about 2 points with privacy noise – a modest decline in utility. However, the benefit is a significant increase in user trust that the system will not leak sensitive information (Feretzakis & Verykios, 2024). For the ML models, the anomaly detection (fraud) model maintained an AUC of 0.96 in both secure and baseline modes, indicating that our data cleaning removed noise but did not remove meaningful signal. The recommendation model's precision@10 was slightly higher with cleaned data (by ~1%), indicating that improved data integrity can even improve model effectiveness.

Finally, we compare the capabilities of the proposed system with those of a traditional pipeline qualitatively in Table 4. This highlights how our infrastructure adds value in areas of trustworthiness. For example, a traditional system may have basic logging but lacks real-time anomaly handling, whereas our system provides proactive data quality enforcement (Atluri *et al.*, 2025) and detailed audit logs. In terms of privacy, a standard pipeline might rely on perimeter security, but our design assumes zero trust internally as well – encrypting data and applying privacy-aware learning techniques (Dwork & Roth, 2014). Also, the inclusion of LLMs extends functionality (like automating customer support queries), but we do so with guardrails to avoid the pitfalls of unrestrained AI (Feretzakis & Verykios, 2024). The net effect is an architecture in which stakeholders (from customers to compliance officers) can have higher confidence. Users' personal data is handled carefully and models' decisions can be traced and explained when necessary, fulfilling key aspects of trustworthy AI in a practical setting (European Commission, 2019).

Table 4: Capabilities of Traditional vs. Proposed E-Commerce AI Infrastructure

Aspect	**Traditional Scalable Pipeline**	**Proposed Trustworthy AI Pipeline**
Data Validation & Quality	Minimal (schema on write, but bad data can propagate)	Rigorous validation at ingress; contracts enforce schema; bad data quarantined
Data Integrity under Load	Tends to drop/lose data when overloaded (best effort)	Back-pressure and autoscaling to preserve data; near-zero loss even under peak load
Security & Privacy	Basic encryption, access control; no ML-specific privacy	End-to-end encryption; differential privacy in model training; federated data options
Model Reliability	Single-model services, possible downtime on failure	Hybrid models with fallbacks; health monitoring and auto-recovery for ML/LLM components (Bollikonda & Bollikonda, 2025)
Transparency & Audit	Limited logging; difficult to trace decisions	Comprehensive logging of data and decisions; audit trail for model outputs; explainability modules for key decisions
Regulatory Compliance	Manual processes to ensure GDPR, etc.	Built-in compliance checks (consent management, data deletion on request, etc.) integrated into data pipeline
AI Functionality	ML-driven analytics (recommendations, etc.) only	ML + LLM for both structured and unstructured data insights, expanding use-case coverage with trustworthy LLM outputs

Overall, the experimental results validate that our integrated approach can meet the dual goals of scalability and trustworthiness. Whereas earlier systems forced a trade-off – you could scale up performance but might lose some control and transparency – our architecture shows that with careful design, that trade-off diminishes. We achieved high throughput and low latency and maintained strict data integrity and security. This has important implications for large e-commerce providers: it is feasible to deploy advanced AI (like LLMs) in customer-facing or mission-critical roles if the surrounding infrastructure is designed to be resilient and governed. For example, an AI chatbot can be scaled to handle holiday traffic surges. With the safeguards we demonstrated, the business can be confident it won't expose private data or crash when the AI model fails to respond in time. Our discussion also underscores certain overheads: there is complexity in setting up such a system, and slight performance costs for the added layers. However, these are largely offset by the prevention of costly errors (like data breaches or significant downtimes), which in an e-commerce context could save millions in revenue and reputation.

One interesting observation is that data integrity actually aids model accuracy and efficiency – by cleansing the

data, we reduce noise for the ML algorithms, leading to potentially better recommendations and fewer false alarms in fraud detection. This reinforces the synergy between trustworthy AI practices and high-quality outcomes. Trustworthy AI is not just about avoiding negatives (like compliance fines or biases) but also about enhancing positives (better decisions from better data). Our work, therefore, aligns with and provides concrete support for the idea that ethics and utility in AI can go hand in hand

CONCLUSION

In this paper, we presented a trustworthy, scalable AI infrastructure for large-scale e-commerce systems. The architecture combines a high-throughput data pipeline with advanced analytics (machine learning and large language models) inside a secure, well-governed data management framework. Built-in data validation, privacy preservation, and continuous monitoring address key risks in data integrity, consistency, and reliable AI behavior. Experiments on open-source e-commerce datasets show the infrastructure can sustain heavy workloads (tens of thousands of events per second) with minimal performance degradation, while maintaining data integrity above 98% under peak load—substantially outperforming a baseline where losses and errors escalated at scale. We further demonstrate that encryption and differential privacy can be integrated with only minor impact on latency and model accuracy, enabling strong customer data protection without sacrificing analytics. The system also detects and corrects anomalies in real time and degrades safely during component failures (e.g., fallback rules when an LLM lags), improving resilience and auditability. While our evaluation was controlled, future work should test production-like deployments (multi–data center, cross-border transfers) and extend trust features (e.g., fairness checks).

REFERENCES

Alrumiah, S., & Hadwan, M. (2021). Implementing big data analytics in e-commerce: Vendor and customer view. *IEEE Access, 9*, 37281–37286. https://doi.org/10.1109/ACCESS.2021.3063517

Amershi, S., Begel, A., Bird, C., Zimmermann, T., Nurvitadhi, E., *et al.* (2019). Software engineering for machine learning: A case study. *In Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291–300). IEEE. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Atluri, R. P., Taylor, A., & Prudhvi, R. (2025). Data contracts in the wild: An approach for redefining trust and accountability in modern data pipelines. *World Journal of Advanced Research and Reviews, 26*(3), 290–299. https://doi.org/10.30574/wjarr.2025.26.3.0410

Baylor, D., Koc, L., Koo, C. Y., Jain, V., Schleier-Smith, J., *et al.* (2017). TFX: A TensorFlow-based production-scale machine learning platform. In Proceedings of the 23rd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining* (pp. 1387–1395). ACM. https://doi.org/10.1145/3097983.3098021

Bollikonda, M., & Bollikonda, T. (2025). Secure pipelines, smarter AI: LLM-powered data engineering for threat detection and compliance. *Preprints,* 2025041365. https://doi.org/10.20944/preprints202504.1365.v1

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. *In Proceedings of the 28th USENIX Security Symposium* (pp. 267–284). USENIX Association.

Davenport, T. H., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48(*1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Desai, P., & Ganatra, K. (2022). Artificial intelligence in strengthening the operations of e-commerce-based business. *In Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 275–280). IEEE. https://doi.org/10.1109/ICDABI56852.2022.10080483

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Feretzakis, G., & Verykios, V. S. (2024). Trustworthy AI: Securing sensitive data in large language models. *AI, 5*(4), 2773–2800. https://doi.org/10.3390/ai5040134

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., *et al.* (2019). *Advances and open problems in federated learning* (arXiv:1912.04977). arXiv. https://arxiv.org/abs/1912.04977

Munshi, A., Alhindi, A., Qadah, T. M., & Alqurashi, A. (2023). An electronic commerce big data analytics architecture and platform. *Applied Sciences, 13(*19), Article 10962. https://doi.org/10.3390/app131910962

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., *et al.* (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 2503–2511).

Sura, R. (2025). Scalable AI-powered data pipelines for enterprise analytics. *International Journal of Innovative Research in Technology, 12*(1), 1094–1101.

Xin, Q. (2025a). A deep reinforcement learning approach to optimizing cloud workload migration. *American Journal of Interdisciplinary Research and Innovation, 4*(3), 10–15.

Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195.