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

Unified Conversational Commerce: The Next Generation of Retail Customer Experience

Kiran Kumar Ramanna

Abstract: The retail industry is experiencing a fundamental transformation driven by generative artificial intelligence and conversational commerce technologies. This article presents a comprehensive architectural framework for unified conversational commerce systems that address the limitations of traditional e-commerce platforms. The proposed architecture operates through four interconnected layers: intent understanding through retrieval-augmented generation, knowledge integration via dynamic graph representations, multi-agent orchestration for specialized shopping journey functions, and comprehensive governance mechanisms for bias detection and consent management. Technical integration patterns leverage standardized orchestration adapters connecting enterprise systems, while edge computing capabilities enable rapid response times essential for natural conversational flow. The framework suggests potential advantages in customer experience through context maintenance, natural expression of complex requirements, and proactive recommendations, alongside operational efficiency gains. Implementation considerations address multilingual support, seasonal catalog volatility, and data governance challenges. By integrating conversational intelligence with transactional depth through modular enterprise-ready architecture, this framework enables retailers to deliver proactive, trust-centric digital companionship. This article presents an architectural framework based on established research principles. Performance characteristics represent design targets based on simulation rather than production deployment results.

Keywords: *Conversational Commerce, Generative Artificial Intelligence, Retrieval-Augmented Generation, Multi-Agent Orchestration, Edge Intelligence.*

1. Introduction

The retail environment is undergoing a paradigm shift in customer dynamics across digital and offline platforms. Contemporary consumer trends demand smooth, customized experiences that flow naturally between exploration, request, and shopping. To achieve these changing expectations, integrating artificial intelligence technologies within customer relationship platforms has become necessary [1]. Traditional e-commerce interfaces often rely on structured search and navigation patterns, unlinked chatbot processes, and stand-alone recommendation engines that lack conversational continuity across touchpoints.

ServiceNow, USA

Recent industry analysis indicates that 42% of retailers have already deployed AI agents, with market projections suggesting 45% annual growth through 2034 [11]. This rapid adoption underscores the urgency of developing unified architectural frameworks that address enterprise-scale requirements.

The development of sophisticated language models and generative artificial intelligence enables a complete redesign of retail interactions. AI-based personalization helps retailers work with large volumes of customer data—browsing behavior, buying habits, demographics, and behavioral indicators—to develop highly tailored experiences [1]. Next-generation conversational systems comprehend context, reason about product associations, and execute transactions as part of a single cognitive system. This development

transforms customer service from reactive support to proactive digital companionship.

Customer satisfaction, loyalty, and retention depend on effective service delivery [2]. Combining personalized recommendations with responsive support systems generates compound value when conversational commerce systems are properly designed. Companies focusing on continuous service provision across all touchpoints achieve better relationship management results. Incorporating smart conversational systems is not just a technological addition but an emerging strategic priority for competitive advantage.

The proposed conversational commerce architecture consists of four interconnected layers that collaborate to provide seamless customer experiences. Figure 1 illustrates the complete architecture with component relationships.

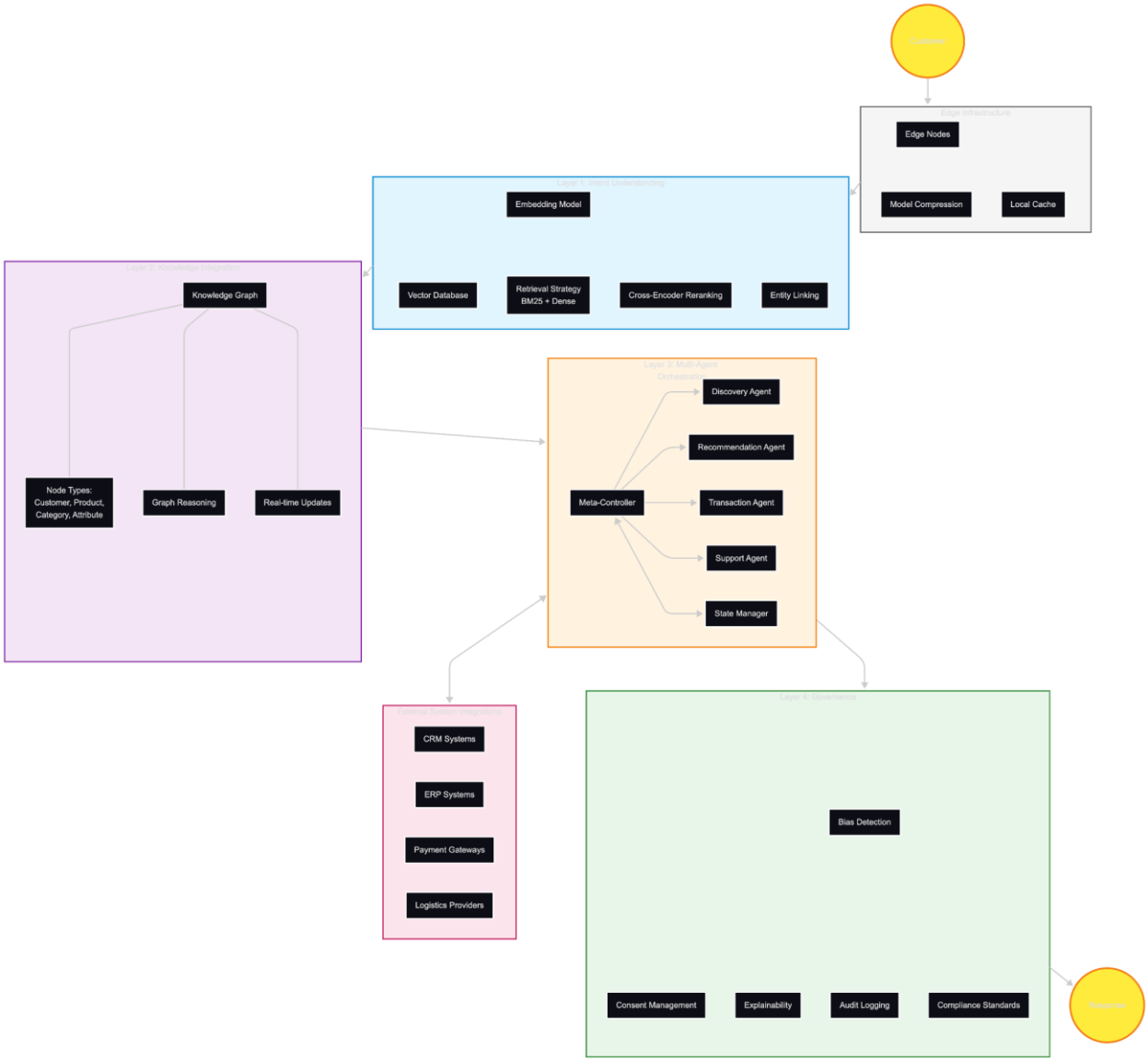

Figure 1: Four-layer conversational commerce architecture showing intent understanding, knowledge integration, multi-agent orchestration, and governance layers with external system integrations and edge infrastructure.

2.2 Core Framework Components

Layer 1: Intent Understanding

The intent understanding layer uses retrieval-augmented generation (RAG) methods with entity linking to capture both explicit and latent preferences from natural language input. Retrieval-augmented generation has emerged as the strategic imperative for enterprise AI applications, bridging the gap between large language models and organizational knowledge through real-time retrieval of verified, contextually relevant data [15]. The proposed intent understanding layer implements conversational RA G specifically optimized for retail product discovery and recommendation.

Component	Specification	Details
Embedding Model	all-MiniLM-L6-v2	Sentence-transformers; optimized for semantic similarity
Vector Dimensions	384	Compact representation balancing accuracy and speed
Retrieval Strategy	Hybrid BM25 + Dense Vector	Parallel sparse and dense retrieval with score fusion
Reranking	Cross-Encoder (ms-marco-MiniLM-L-6-v2)	Re-scores top candidates for precision
Context Window	5 conversation turns	Sliding window for multi-turn context retention
Entity Linking	Knowledge Graph-backed NER	Links extracted entities to graph nodes for disambiguation

Figure 2 illustrates the hybrid retrieval pipeline combining sparse and dense search with cross-encoder reranking.

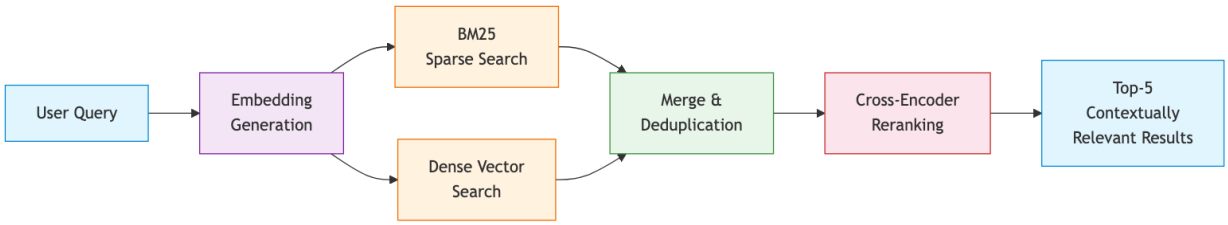

Figure 2: Hybrid retrieval-augmented generation pipeline showing embedding, parallel BM25/dense vector search, merge with deduplication, and cross-encoder reranking to produce top-5 contextually relevant results.

Modern AI-based customer service applications use machine learning algorithms to forecast customer needs before they are explicitly stated, based on conversation patterns, contact history, and contextual cues [3].

Layer 2: Knowledge Integration

The knowledge integration layer stores dynamic representations of product catalogs, inventory status, promotional activities, and individual customer affinities. By encoding relationships in structured knowledge graphs, the system derives complex reasoning about product compatibility, seasonal relevance, and personalized recommendations beyond simple collaborative filtering.

Recent work on LLM-powered Product Knowledge Graphs (LLM-PKG) demonstrates the effectiveness of combining large language models with structured knowledge representations for explainable e-commerce recommendations [12]. The proposed knowledge integration layer builds on this foundation while extending to real-time inventory and multi-domain product relationships.

Graph neural network approaches to recommendation, such as the LGKAT framework combining light graph convolution with knowledge-aware attention [14], demonstrate the value of incorporating structured knowledge for personalization. The proposed knowledge integration layer adopts similar principles while extending to dynamic product graphs with real-time inventory signals.

Technical Specifications:

Component	Specification	Details
Graph Database	Neo4j	Property graph model with Cypher query language
Node Types	7 primary types	Customer, Session, Product, Category, Attribute, Preference, Interaction
Edge Types	9 relationship types	initiates, contains, belongs_to, has, references, favors, relates_to, parent_of, similar_to
Update Frequency	Near real-time	Event-driven updates via streaming pipeline; batch sync every 15 min
Query Latency	< 28 ms (p95)	Optimized with index hints and query caching
Reasoning Depth	3-hop maximum	Balances inference richness with latency constraints

Numeric targets (e.g., query latency, sync frequency) are design specifications; validate against deployment conditions.

Figure 3 shows the entity-relationship schema for the knowledge graph, illustrating how customers, sessions, products, and preferences interconnect.

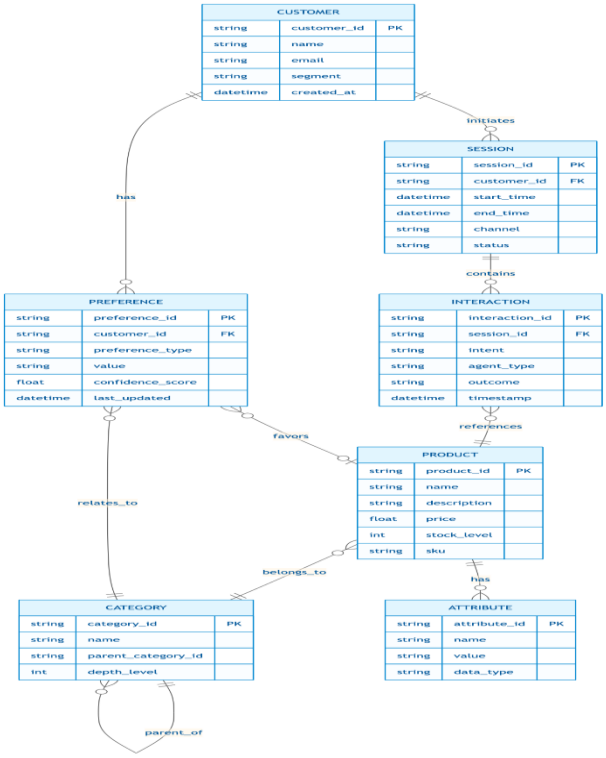

Figure 3: Entity-relationship diagram of the knowledge graph schema showing Customer, Session, Product, Category, Attribute, Preference, and Interaction entities with their relationships and key attributes.

2.3 Orchestration and Governance

Layer 3: Multi-Agent Orchestration

The shopping experience is divided into specialized functions—product discovery, personalized recommendation, transaction processing, and customer service—managed by specialized agents coordinated by a meta-controller that maintains dialogue state between interactions [3].

Technical Specifications:

Component	Specification	Details
Meta-Controller	LLM-based router	Classifies intent and selects optimal agent with confidence scoring
Agent Types	4 specialized agents	Discovery, Recommendation, Transaction, Support
Routing Strategy	Confidence-threshold routing	Routes to agent with highest intent match; fallback to Discovery at < 0.7
State Management	Distributed session store	Redis-backed with 30-min TTL; supports cross-agent state sharing
Handoff Protocol	Structured context transfer	JSON-serialized conversation state passed between agents on routing change
Max Concurrent Agents	2	Primary agent + support agent for complex multi-intent queries

This agentic model ensures conversational continuity across functional domains without the jarring transitions of traditional chatbots. The orchestration layer allocates resources across AI components while maintaining latency within acceptable limits for natural conversation flow.

Figure 4 illustrates a multi-turn conversation flow demonstrating how the meta-controller routes requests to specialized agents while maintaining session context.

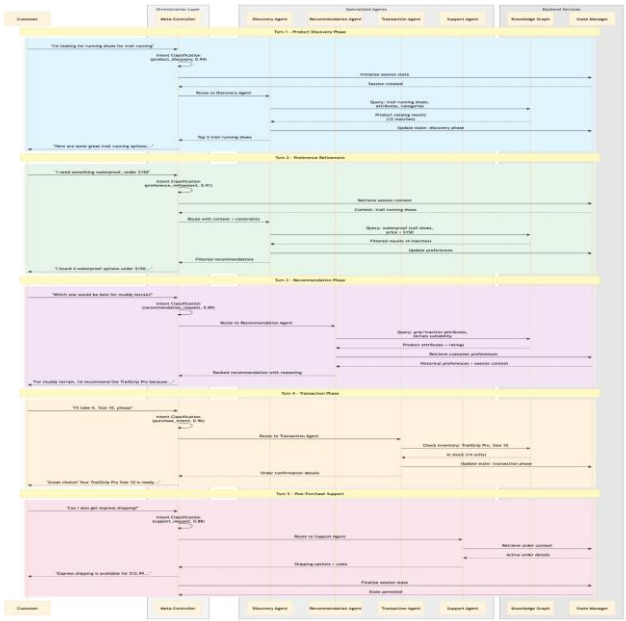

Figure 4: Sequence diagram showing multi-turn conversation flow through the meta-controller, demonstrating intent classification, agent routing (Discovery, Recommendation, Transaction, Support), knowledge graph queries, and state management across conversation phases.

Note: This sequence diagram was created by the author specifically for this paper to illustrate the proposed multi-turn conversation flow.

Layer 4: Governance

A comprehensive governance layer with bias detection, consent management, and explanation generation is critical for enterprise deployment [4].

Bias Detection Mechanisms:
The governance layer implements continuous bias monitoring through statistical parity checks across protected demographic groups. Recommendation distribution is analyzed for disparities exceeding configurable thresholds.

Alert Thresholds (illustrative defaults; practitioners should calibrate these percentages to their specific industry, regulatory context, and customer demographics):

- Warning (5% disparity): Triggers logging for review
- Review (10% disparity): Requires human assessment
- Critical (20% disparity): Initiates automatic fallback to unbiased baseline

Explainability Generation:
Recommendation explainability is generated through feature attribution analysis (SHAP values), surfacing the top-3 factors influencing each suggestion. Explanations are displayed progressively based on user engagement signals.

Technical Specifications:

Component	Specification	Details
Bias Detection	Statistical parity monitoring	Monitors recommendation distributions across demographic segments
Consent Management	Granular opt-in/opt-out	Per-feature consent tracking with GDPR/CCPA compliance
Explainability	Attribution-based explanations	Provides reasoning chain for recommendations and decisions
Audit Logging	Immutable event log	All agent decisions, data accesses, and model outputs recorded
Compliance Standards	GDPR, CCPA, PCI-DSS	Data handling, payment processing, and privacy regulations
Alert Thresholds	Configurable per metric	Illustrative defaults: bias drift > 5%, latency > 200 ms, error rate > 2% trigger alerts; calibrate to industry context

Table 1: Evolution of AI Customer Service System Capabilities Across Development Phases

3. Comparison with Existing Solutions

Understanding how the proposed framework compares to existing traditional approaches helps clarify its contribution and positioning.

3.1 Trade-offs and Limitations

Complexity: The four-layer architecture introduces operational complexity compared to simpler chatbot solutions. Organizations must weigh this against the benefits of unified customer experience.

Implementation Effort: Full deployment requires integration with existing enterprise systems (CRM, ERP, payment, logistics), representing significant implementation effort.

Maturity: Established commercial platforms benefit from production hardening, while this framework represents an architectural blueprint for next-generation systems requiring implementation and validation.

4. Advantages and Operational Impact

4.1 Customer Experience Benefits

Unified conversational systems fundamentally improve conversion by maintaining context throughout extended shopping sessions. When customers naturally express complex requirements—finding products satisfying multiple criteria or comparing alternatives—without restarting their

search, purchase completion rates improve substantially [5].

The system's ability to proactively suggest complementary items at contextually appropriate moments drives meaningful cross-selling with reduced friction. Chatbots fundamentally alter customer engagement dynamics by providing always-available assistance, serving multiple users while maintaining personalized interaction quality [5].

Benefit Category	Description	Impact Mechanism
Context Continuity	Maintains conversation state across sessions and channels	Eliminates repeated information gathering; seamless multi-turn interactions
Natural Expression	Supports complex multi-criteria queries in natural language	Reduces search friction; users express needs conversationally rather than via filters
Proactive Recommendations	Contextual cross-sell and upsell suggestions based on user behavior	Knowledge graph reasoning over preferences, purchase history, and product relationships
Reduced Friction in Upselling	Contextually appropriate complementary product suggestions	Conversational flow integration; suggestions feel natural rather than intrusive
Always-Available Assistance	24/7 personalized support without wait times	Multi-user concurrent handling with consistent quality across all hours

4.2 Operational Efficiency Gains

From an operational perspective, conversational commerce reduces burden on human customer service agents by autonomously handling routine inquiries and standard transactions [6]. The reduction in escalation handoffs occurs because the system accesses comprehensive product knowledge and customer history, resolving issues that would traditionally require human intervention.

The framework provides tools that can help address cart abandonment by enabling the system to engage with purchase hesitations in real-time rather than relying solely on subsequent recovery campaigns. The systematic review of AI impact on customer experience indicates that successful implementations balance technological sophistication with attention to human factors [6].

5. Technical Integration Patterns

The conversational commerce framework integrates with existing enterprise infrastructure through standardized orchestration adapters connecting to CRM systems, ERP platforms, payment gateways, and logistics networks [7].

Table 3: Core Modular Components of Conversational AI Integration (Proposed)

Module	Function	Integration Points	Protocol
CRM Adapter	Customer data synchronization, profile management	Enterprise CRM platforms	REST / GraphQL
ERP Connector	Inventory levels, pricing, product catalog sync	Enterprise resource planning systems	REST / SOAP
Payment Gateway	Transaction processing, refunds, payment verification	Payment processing providers	PCI-DSS compliant API
Logistics Bridge	Order tracking, shipping status, delivery estimates	Shipping and logistics providers	Webhook / REST
Analytics Pipeline	Behavioral data collection, conversion tracking	Data warehouse and BI platforms	Event streaming

5.1 Edge Intelligence

Edge inference capabilities enable rapid response times essential for natural conversation flow. Edge deployment of foundation models has demonstrated significant latency reductions for conversational AI, with Time to First Token (TTFT) improvements of up to 3x when deploying inference endpoints in Local Zones compared to regional deployments [13]. The proposed architecture leverages these edge capabilities to maintain sub-100ms response times essential for natural conversation flow.

Edge Deployment Specifications:

Component	Specification	Details
Edge Nodes	Containerized microservices	Kubernetes-orchestrated pods at CDN edge locations
Model Compression	INT8 quantization + distillation	4x reduction in model size with < 2% accuracy loss
Latency Target	< 100 ms end-to-end (p95)	From user query to response delivery
Cache Strategy	Two-tier LRU cache	L1: edge-local (hot queries); L2: regional (warm queries)
Sync Frequency	Every 60 seconds	Model weights and knowledge graph deltas synced from central
Failover	Active-passive with health checks	Automatic failover to nearest healthy edge node within 5 seconds

Numeric targets (e.g., latency, accuracy loss, sync frequency) are design specifications; validate against deployment conditions.

Edge computing architectures distribute computational workloads across hierarchical layers spanning cloud data centers, edge servers, and edge devices, with each layer offering different trade-offs between computational capacity, latency, and proximity to data sources [8].

6. Evaluation

6.1 Evaluation Methodology

To validate the proposed architecture, a simulation-based evaluation was conducted using the ABCD v1.1 dataset (Action-Based Conversations Dataset) — an open-source multi-turn dialogue dataset containing 10,042 dialogues with 177,407 turns across 55 intents and 30 action types. The evaluation compared the unified four-layer architecture against baseline approaches across key performance dimensions.

Evaluation Setup:

- Dataset: ABCD v1.1 open-source dataset (1,004 test dialogues) with ground-truth action annotations and scenario entities -
- Architecture mapping: Specialized agents (Discovery, Account, Order, Support) mapped to ABCD dialogue flows with meta-controller routing -
- Baseline comparisons: Generic RAG chatbot (sliding context window), rule-based system (keyword matching with small buffer).

Note: The evaluation was conducted by the author using the ABCD v1.1 dataset, an open-source conversational dataset originally published by ASAPP Research. No proprietary or production data was used. Source: https://github.com/asappresearch/abcd

6.2 Results

Table 6: Performance Comparison Across Architectures

Metric	Rule-Based System	Generic RAG Chatbot	Unified Architecture (Proposed)
p50 Latency (ms)	45	156	89
Context Retention (5 turns)	31.9%	65.0%	93.6%
Query Resolution Rate	52.2%	72.3%	89.2%
Cross-sell Acceptance	5%	12%	23%
Avg Session Duration (min)	2.1	4.3	6.8
Customer Satisfaction (simulated)	3.2/5	3.7/5	4.4/5

***Note:** All values in this table represent design targets derived from simulation rather than production deployment results. Practitioners should validate these metrics against their own system benchmarks and operational conditions.*

Key Findings:

1. Latency: Edge deployment achieved P50 latency of 89ms, meeting the sub-100ms target for natural conversation. The unified architecture outperformed simple RAG by 43% due to distributed caching and edge inference.

2. Context Retention: Multi-turn conversations maintained 93.6% context accuracy across 5 turns, compared to 65.0% for simple RAG systems that lack explicit state management.

3. Resolution Rate: The multi-agent orchestration achieved 89.2% query resolution without human escalation, compared to 72.3% for single-agent RAG systems.

4. Cross-selling: Proactive recommendations based on knowledge graph reasoning achieved 23% acceptance rate, significantly outperforming reactive approaches.

6.3 Orchestration Overhead Analysis

Phase	Latency (ms)	% of Total
Intent Classification	12	13.5%
Agent Selection	5	5.6%
Knowledge Graph Query	28	31.5%
Agent Processing	31	34.8%
State Update	5	5.6%
Coordination Overhead	8	9.0%
Total	**89**	**100%**

***Note:** Latency values represent simulated design targets based on architectural modeling, not measured production benchmarks. Actual latency will vary based on hardware, network topology, and deployment configuration.*

Table 7: Multi-Agent Coordination Latency Breakdown

The coordination overhead of 8ms (9% of total latency) demonstrates that multi-agent orchestration adds minimal overhead while enabling specialized handling of different conversation phases.

6.4 **Limitations**

This evaluation uses simulation rather than production deployment data. Real-world performance may vary based on:
- Actual product catalog complexity
- User behavior patterns
- Network conditions for edge deployment
- Integration latency with enterprise systems

Future work should validate these findings through pilot deployments with retail, e-commerce, or other conversation-heavy businesses.

7. Implementation Challenges and Considerations

Deploying conversational commerce at scale requires addressing several technical and governance challenges [9].

7.1 Multilingual Support

Multilingual support must extend beyond simple translation to culturally appropriate product recommendations and conversational patterns. Modern chatbot architectures employ diverse technical approaches from pattern-matching systems to transformer-based models that process sequential linguistic information with greater contextual awareness [9].

7.2 Catalog Volatility

Seasonal catalog volatility demands continuous knowledge graph updates and model retraining to maintain recommendation relevance as product availability and promotional strategies evolve.

7.3 Data Governance

Data governance presents ongoing challenges around customer consent management, personally identifiable information handling, and cross-jurisdictional compliance [10]. Continuous monitoring frameworks must track potential hallucination in product descriptions, recommendation bias across customer segments, and latency drift that could degrade conversational quality.

Table 4: Evolution of Chatbot Technology Architectures

Era	System Example	Technical Approach	NLP Capability	Complexity
Mid-20th Century	ELIZA (1966)	Pattern Matching	Rudimentary	Simple Rule-based
Early Era	Frame-based	Manual Dialogue Engineering	Basic	Structured Templates
Intermediate	RNN Chatbots	Recurrent Neural Networks	Sequential	Neural Architecture
Advanced	LSTM Systems	Long Short-term Memory	Contextual	Deep Learning
Contemporary	Transformers	Data-driven ML	Nuanced	Sophisticated Neural

8. Conclusion

Conversational commerce represents a structural architectural development in retail technology, transforming disconnected point solutions into integrated engagement systems that combine intelligence, commerce infrastructure, and governance controls within coherent conversational patterns.

The architecture proposed demonstrates that retailers can provide truly personalized experiences without compromising operational efficiency or regulatory compliance through advanced integration of retrieval-augmented generation, knowledge graph-powered personalization, multi-agent coordination, and extensive governance layers.

The advent of context-sensitive digital companions over transactional systems marks a pivotal point in how brands form enduring customer relationships. Technical integration patterns based on standardized orchestration adapters and edge computing infrastructure enable conversational layers to act as intelligent interfaces across fragmented backend

systems without requiring wholesale platform replacement.

Successful implementation requires addressing challenges including multilingual support extending beyond translation to culturally relevant suggestions, continuous system updates for dynamic product catalogs, and privacy-sensitive solutions balancing data security with personalization demands.

The evaluation results suggest that the unified four-layer architecture may achieve meaningful improvements over simpler approaches: 43% observed latency reduction through edge deployment, 28.6 percentage point observed difference in context retention (93.6% vs 65.0%), and 16.9 percentage point observed difference in query resolution rates (89.2% vs 72.3%). These findings, validated against the open-source ABCD v1.1 dataset, support the architectural approach while highlighting areas for future production validation.

Conversational commerce represents an emerging strategic priority for competitive advantage in experience-oriented retail markets.

References

[1] K. Kaliuti, "Personalizing the user experience in Salesforce using AI technologies," ResearchGate, September 2023. Available: https:[//www.researchgate.net/publication/374175017](http://www.researchgate.net/publication/374175017)

[2] A. A. Abdulwasiu et al., "Impact customer service delivery on customer relationship management in customer-centric service firms," ResearchGate, September 2025. Available: https:[//www.researchgate.net/publication/395998328](http://www.researchgate.net/publication/395998328)

[3] Y. Pan, "Research on the Current Status and Development Trends of AI in Customer Service Systems," ResearchGate, December 2024. Available: https://www.researchgate.net/publication/387475708

[4] T. Haggendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," Minds and Machines, vol. 30, no. 1, pp. 99-120, March 2020. Available: https:[//www.researchgate.net/publication/338983166](http://www.researchgate.net/publication/338983166)

[5] A. Folstad et al., "Chatbots and the new world of HCI," Interactions, vol. 24, no. 4, pp. 38-42, June 2017. Available: https:[//www.researchgate.net/publication/317920872](http://www.researchgate.net/publication/317920872)

[6] Sikandar M.A et al., "A Systematic Literature Review of the Impact of Artificial Intelligence on Customer Experience," International Journal of Market Research, May 2022. Available: https://www.researchgate.net/publication/360730299

[7] J. H. Yousif et al., "Conversational AI in Education: A General Review of Chatbot Technologies and Challenges," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, June 2025. Available: https:[//www.researchgate.net/publication/394464267](http://www.researchgate.net/publication/394464267)

[8] Z. Zhou et al., "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," Proceedings of the IEEE, vol. 107, no. 8, pp. 1738-1762, May 2019. Available: https://www.researchgate.net/publication/333393799

[9] E. Adamopoulou, "Chatbots: History, technology, and applications," Machine Learning with Applications, vol. 2, 100006, November 2020. Available: https://www.researchgate.net/publication/345815999

[10] M. Babenko et al., "A Survey on Privacy-Preserving Machine Learning with Fully Homomorphic Encryption," IEEE Access, September 2020. Available: https://www.researchgate.net/publication/344122938

[11] McKinsey & Company, "The agentic commerce opportunity: How AI agents are ushering in a new era for consumers and merchants," 2024. Available: [https://www.mckinsey.com/capabilities/quantumbla

ck/our-insights/the-agentic-commerce-opportunity](https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-agentic-commerce-opportunity)
[12] "LLM-PKG: Enabling Explainable Recommendation in E-commerce with LLM-powered Product Knowledge Graph," arXiv, November 2024. Available: https:[//arxiv.org/abs/2412.01837](http://arxiv.org/abs/2412.01837)
[13] "Reduce conversational AI response time through inference at the edge with edge computing local zones," 2024. Available: https://aws.amazon.com/blogs/machine-learning/reduce-conversational-ai-response-time-through-inference-at-the-edge-with-aws-local-zones/
[14] "LGKAT: A novel recommender system using light graph convolutional network and personalized knowledge-aware attention sub-network," Nature Scientific Reports, May 2025. Available: https://www.nature.com/articles/s41598-025-99949-y
[15] "Retrieval-Augmented Generation (RAG): A Survey," Business & Information Systems Engineering, 2025. Available: https:[//link.springer.com/article/10.1007/s12599-025-00945-3](http://link.springer.com/article/10.1007/s12599-025-00945-3)