Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 3, 2025

Agentic Commerce: A Unified Multi-Retrieval Framework for High-Fidelity E-Commerce Chatbots

MD Estihad Faysal *, Wenfeng Feng, Esha Mony

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

* **Corresponding author**: (Email: aihe1524@gmail.com)

Abstract: E-commerce chatbots face critical limitations that undermine customer trust: hallucinations from ungrounded responses, poor multi-turn coherence, and an inability to execute real-world actions such as processing refunds or verifying live inventory. Existing retrieval-augmented generation (RAG) and chain-of-thought (CoT) approaches address knowledge grounding and reasoning, respectively, yet remain fundamentally passive—they inform but cannot act. We present a unified agentic framework that integrates RAG for factual grounding, CoT for structured multi-step reasoning, and multi-agent collaboration for autonomous task execution. The modular architecture encompasses specialized agents for retrieval, reasoning, action generation, and safety enforcement, orchestrated through an LLM-based routing policy. This design enables the system to move beyond answering questions toward completing transactions, coordinating inventory checks, and resolving complex customer inquiries autonomously. Evaluated on a 10K-SKU e-commerce dataset spanning factual, comparative, and multi-turn query types, the framework achieves 96.2% response accuracy and 95.8% grounding reliability. The multi-agent architecture reduces errors in multi-turn interactions by 18% compared to single-agent baselines. The system operates with a median latency of 3.12 seconds—a deliberate safety-first design choice that prioritizes transactional reliability over conversational speed, ensuring business-critical accuracy in high-stakes operations where sub-second responses would compromise correctness.

Keywords: E-commerce, conversational AI, large language models, retrieval-augmented generation, chain-of-thought, multi-agent systems, agentic AI, reasoning, chatbot evaluation.

1. Introduction

The rapid evolution of e-commerce platforms has created an urgent demand for sophisticated chatbots that can effectively manage a wide array of user interactions, ranging from simple product searches and detailed comparisons to advanced personalized recommendations and seamless transaction processing. Recent surveys indicate that RAG-based implementations can reduce information retrieval time by up to 95% compared to traditional methods [1], underscoring the efficiency potential of intelligent retrieval optimization. Traditional rule-based systems, while reliable in narrowly defined scenarios, often fall short due to their limited coverage of diverse queries, brittle mechanisms for maintaining context across extended conversations, and inability to adapt to new domains without extensive manual reconfiguration. In contrast, early implementations of large language model (LLM)-based assistants, such as those derived from GPT-3 architectures [2], have shown promise in generating natural and fluent responses but are frequently undermined by issues like hallucinations—where the model fabricates information not supported by facts—and challenges in sustaining coherent multi-turn dialogues.

A critical gap remains in current approaches: while RAG provides knowledge grounding and CoT provides logical reasoning, both paradigms are fundamentally passive. They enable systems to understand and explain, but not to execute. An e-commerce assistant that can accurately describe a return policy is insufficient if it cannot actually initiate the refund. Similarly, a system that reasons correctly about inventory availability provides limited value if it cannot query live stock levels or reserve items. The agentic framework is required to bridge this divide—moving from "knowing" to "doing" by enabling autonomous execution of real-world actions such as processing refunds, checking live inventory, updating order status, or coordinating with external fulfillment APIs.

Recent breakthroughs in agentic AI, characterized by systems that can autonomously perceive environments, plan actions, execute tasks (e.g., API calls for inventory checks or payments), and learn from outcomes, present a transformative opportunity to usher in "agentic commerce." This paradigm shift promises to revolutionize retail experiences by enabling hyper-personalized, proactive, and efficient customer engagements that mimic human-like intelligence while scaling to handle high-volume operations.

Building upon our preliminary explorations in RAG+CoT frameworks, this paper significantly extends that foundation by incorporating multi-agent collaboration, allowing specialized agents to divide and conquer complex tasks—for instance, one agent focused on data retrieval, another on logical reasoning, and a third on action-oriented outputs like order fulfillment. This integration not only enhances the system's autonomy but also improves overall robustness in dynamic e-commerce environments. Our key contributions are multifaceted: (i) a refined agentic multi-retrieval architecture that synergizes RAG for precise knowledge grounding, CoT for interpretable reasoning, and multi-agent systems for collaborative efficiency; (ii) an enhanced evaluation protocol that incorporates novel agentic metrics, such as collaboration success rates and inter-agent handoff efficiency, alongside traditional measures; (iii) rigorous empirical results derived from a comprehensive 10K-SKU dataset, demonstrating tangible gains from multi-agent integration in terms of accuracy, latency reduction, and multi-turn coherence; and (iv) forward-looking discussions on 2025 industry trends, including the implications for autonomous transactions, sustainability in AI deployments, and ethical considerations like bias mitigation.

2. Related Work

The foundational building blocks of modern conversational AI trace back to key innovations in sequence modeling and language understanding. Self-attention mechanisms, as introduced in the Transformer ar-chitecture [3], have enabled efficient handling of long-range dependencies in text, paving the way for scalable models. Subsequent developments, such as GPT-3's demonstration of strong few-shot learning capabilities [2], highlighted how large-scale pretraining on diverse corpora could yield versatile lan-guage generation without task-specific fine-tuning. BERT [4] continues to play a pivotal role in retrieval and reranking tasks by providing dense embeddings that capture semantic nuances, often outperforming traditional keyword-based methods in information retrieval scenarios.

In the realm of knowledge-grounded generation, Retrieval-Augmented Generation (RAG) [5] has emerged as a critical technique to mitigate hallucinations by blending parametric knowledge from the LLM with non-parametric retrieval from external corpora, thus ensuring responses are anchored in veri-fiable facts. Chain-of-Thought (CoT) prompting [6] further advances reasoning capabilities by encourag-ing models to break down problems into intermediate steps, significantly boosting performance on com- positional tasks like arithmetic or logical inference. For practical deployment, parameter-efficient adap-tation methods like Low-Rank Adaptation (LoRA) [7] allow fine-tuning of massive models with mini-mal computational overhead, while FAISS [8] provides GPU-accelerated approximate nearest-neighbor search for real-time retrieval over vast indexes. Evaluation methodologies, such as those advocated in [9], stress the importance of diverse metrics spanning accuracy, fairness, and robustness to guide holistic assessments.

Advancements in 2024–2025 have accelerated the adoption of agentic AI, where multi-agent systems (MAS) facilitate collaboration among autonomous entities to tackle complex, multi-faceted problems. Frameworks like MetaGPT have popularized the use of Standardized Operating Procedures (SOPs) to enhance multi-agent coordination, demonstrating that structured role definitions can significantly reduce compounded errors in collaborative tasks [10]. Similarly, LangGraph, AutoGen, and CrewAI enable seamless integration of RAG and CoT within MAS, allowing agents to share states, delegate subtasks, and self-correct through feedback loops. In e-commerce contexts, agentic platforms have been shown to automate end-to-end workflows, resulting in up to 47% higher conversion rates by personalizing interac-tions and handling transactions proactively. Extensions like RAG-Fusion, which employs reciprocal rank fusion (RRF) for merging multi-query retrievals [11], further enhance contextual relevance in dynamic domains. Despite these progresses, there remains a notable gap in unified frameworks that fully integrate RAG, CoT, and MAS tailored for e-commerce—our work addresses this by proposing an end-to-end pipeline that not only grounds and reasons but also collaborates autonomously, bridging theoretical ad- vancements with practical applicability.

3. System Overview and Methods

3.1. Problem Formulation and Notation

We formalize an e-commerce conversational session as a discrete-time process. Let $C = \{p_1, \ldots, p_{|C|}\}$ denote the catalog of products, each product p_i being associated with an attribute vector $a_i \in R_{da}$ (e.g., price, brand, category) and a set of textual descriptions $D_i = \{d_{i,1}, \ldots, d_{i,mi}\}$. All textual artifacts are pre-processed into a global corpus

$$\mathcal{D} = \bigcup_{i=1}^{|C|} D_i \cup \mathcal{D}_{\text{FAQ}} \cup \mathcal{D}_{\text{policy}} \quad (1)$$

Where D_{FAQ} and D_{policy} denote store-specific FAQs and policies, respectively.

A user interacts with the system over turns t = 1, ..., T. At each turn, the user provides a natural- language input x_t, and the chatbot returns a response y_t. The dialog history up to time t is

$$h_t = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1}), x_t\} \quad (2)$$

The objective of the system is to generate responses that (i) are factually grounded in D, (ii) satisfy task- specific constraints (e.g., respecting inventory and pricing), and (iii) minimize latency and resource cost.

We can write the optimization goal as

$$\max_{\theta,\Phi} \mathbb{E}_{(h,D)}\left[U(y_t, y_t^*)\right] \quad \text{s.t.} \quad \tau(y_t) \le \tau_{\max} \quad (3)$$

Where $U(y_t, y_t^*)$ represents the utility of the generated response y_t relative to the ground truth y_t^*, captur-ing correctness and user satisfaction. The parameters θ and Φ denote the LLM weights and agent policy parameters, respectively. The constraint $\tau(y_t) \le \tau_{max}$ enforces a latency budget (e.g., $\tau_{max} < 5$ s for transactional queries).

Agentic behavior is modeled by a set of collaborating agents

$$A = \{a^{\text{retr}}, a^{\text{reason}}, a^{\text{act}}, a^{\text{safety}}\} \quad (4)$$

With each agent specializing in retrieval, reasoning, action generation, and safety enforcement. The orchestrator implements a high-level policy

$$\pi_{\text{orch}}(a_t \mid h_t): H \rightarrow A \quad (5)$$

Which chooses which agent should act next given the current dialog state h_t.

3.2. Pipeline

The proposed system is built around a modular, agentic architecture that ensures flexibil- ity, scalability, and ease of maintenance (as depicted in Fig. 1). Upon receiving a user query x, the system first embeds it using a transformer-based encoder to convert it into a dense vector representation suitable for similarity search. This embedding is then routed to a multi-agent orchestrator, which intelligently delegates tasks based on query complexity and type. Key agents include a retrieval agent responsible for querying the FAISS index to fetch top-k relevant documents from enterprise sources like product catalogs and FAQs; a reasoning agent that applies CoT to decompose the query into logical steps; and an action agent that handles practical outputs, such as generating recommendations or initiating transactions.

Formally, let $e_x \in R_d$ denote the query embedding and $E \in R_{N\times d}$ the matrix of corpus embeddings. Retrieval approximates the posterior over documents as

$$p(d_i|x) = \frac{\exp(\gamma \cdot \cos(e_x, e_i))}{\sum_{j=1}^{N} \exp(\gamma \cdot \cos(e_x, e_j))} \quad (6)$$

Where e_i is the embedding of document d_i and γ is a

temperature hyperparameter. In practice, FAISS returns only the top-k documents, and we renormalize P (d_i | x) over this subset.

The retrieved contexts are ranked and fused using reciprocal rank fusion (RRF) to create a consolidated evidence set $\hat{D}$, which is fed into the LLM alongside the query for grounded generation. High-frequency operations, including embedding computation and inference, are accelerated via GPUs to maintain sub- second end-to-end latency, making the system suitable for real-time e-commerce applications. This modular design allows for easy updates, such as adding new agents for specific tasks or integrating additional data sources, ensuring the framework remains adaptable to evolving business needs.

3.3. LLM Backbone and Adaptation

At the core of the framework is an instruction-tuned large lan-guage model (LLM) selected from state-of-the-art variants, such as LLaMA 3.1 [12], Gemini 1.5 [13], or DeepSeek-V3 [14], chosen for their balance of performance and efficiency in dialogue tasks. To adapt these models to e-commerce-specific nuances—like brand-specific tone, FAQ handling, and customer service protocols—we employ Low-Rank Adaptation (LoRA) [7]. LoRA introduces low-rank update matrices into the frozen base weights as follows:

$$\Delta W = AB_T,\ W' = W + \alpha \Delta W \tag{7}$$

Where $A \in R_{d\times r}$, $B \in R_{r\times k}$, and $r \ll \min(d, k)$, drastically reducing the number of trainable pa- rameters while preserving the model's original capabilities. The fine-tuning process utilizes the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 2×10^{-5}, allowing for quick iterations on domain data.

For agentic enhancements, we further incorporate Reinforcement Learning from Human Feedback (RLHF) to refine inter-agent collaboration, ensuring agents learn to optimize handoffs and shared decision-making over training episodes. Given a trajectory $\tau = (h_1, a_1, r_1, ..., h_T, a_T, r_T)$, the objective is

$$J(\phi) = E_{\tau\sim\pi_\varphi}\left[\sum_{t=1}^{T}\gamma^{t-1} r_t\right] \tag{8}$$

Where ϕ parameterizes the agent policies and r_t encodes human preference scores for correctness, po-liteness, and safety. This adaptation not only improves response qualitybut also reduces computational costs, making the system viable for deployment in resource-constrained environments.

3.4. Retrieval-Augmented Generation

To ensure factual accuracy, enterprise knowledge is pre-processed into overlapping chunks of 200–400 tokens, embedded into 768-dimensional vectors using a high-performance encoder, and stored in a FAISS index [8] for efficient querying. During inference, the system performs multi-retrieval in logarithmic time O(log N), where N represents the corpus size, ap-proximating the generation probability as:

$$P(y \mid x) \approx \sum_{d\in\hat{D}} P(y \mid x, d)\, P(d \mid x) \tag{9}$$

Ambiguous queries benefit from cross-encoder reranking, which refines precision by up to 12% through contextual scoring. The unified prompt template integrates system instructions, retrieved contexts, and the user query to promote coherent, grounded outputs. This structure minimizes drift from factual sources and supports consistent performance across varied query types. Additionally, multi-retrieval allows for dynamic sourcing from multiple databases, such as product catalogs, user histories, and external APIs, further enhancing the system's ability to handle diverse e-commerce scenarios.

3.5. Chain-of-Thought Reasoning and Safety

Chain-of-Thought (CoT) prompting is employed to en-hance the system's ability to handle multi-step inference, decomposing reasoning into interdependent traces:

$$r_t = g_\theta(x, r_{1:t-1}),\ y = f_\theta(x, r_{1:T}) \tag{10}$$

To bolster robustness, self-consistency aggregates outputs from K parallel reasoning paths,

$$\hat{y} = \arg\max_{y} \sum_{k=1}^{K} P\left(y \mid x, r_{1:T}^{(k)}\right) \tag{11}$$

Reducing variability in responses and improving overall reliability.

Safety is prioritized through multi-layered guardrails that detect and filter personally identifiable in- formation (PII), toxicity, and policy violations, with confidence gating redirecting uncertain responses to predefined templates or human oversight. Let s(y) ∈ [0, 1] denote the safety score of a candidate response. We implement a thresholding policy

$$y_{final} = \begin{cases} y, & \text{if } s(y) \geq \delta, \\ y_{fallback}, & \text{otherwise,} \end{cases} \tag{12}$$

Where $y_{fallback}$ is a rule-based or human-authored template. These measures are crucial in e-commerce, where data privacy and ethical AI practices are paramount, ensuring the system complies with regulations like GDPR while maintaining user trust.

3.6. Multi-Agent Collaboration

Inspired by frameworks like CrewAI and LangGraph, our multi-agent system (MAS) enables collaborative intelligence, where agents operate with shared memory and tool access to tackle complex queries. Common patterns include divide-and-conquer, where the retrieval agent sources data, the reasoning agent processes it via CoT, and the action agent delivers outputs like personalized suggestions. The orchestrator uses an LLM-based classifier to route queries, with inter- agent communication facilitated by a shared key-value store.

We model the MAS as a directed communication graph G = (V, E), where each node $v_i \in V$ corresponds to an agent a_i and edges (i, j) ∈ E indicate permissible message passing. At time t, agent i receives an observation $o_t^{(i)}$ (e.g., retrieved snippets or partial reasoning traces) and a message bundle $m_{t-1}^{(i)}$ from its neighbors. Its policy

$$\pi_i(a_t^{(i)} \mid o_t^{(i)}, m_{t-1}^{(i)}; \phi_i) \tag{13}$$

Outputs an action a(i) (e.g., "retrieve," "summarize," "reroute") and a new message m(i) broadcast along outgoing edges.

For example, in a “compare laptops” query, the retrieval agent fetches specifications, the CoT agent evaluates trade-offs step-by-step, and the action agent tailors recommendations to user history, enhancing overall autonomy and efficiency. We define a collaboration success rate (CSR) metric as

$$\mathrm{CSR} = \frac{1}{M}\sum_{j=1}^{M}\mathbb{I}\left[\mathrm{handoffs}_j \le H_{\max} \wedge \mathrm{task\_success}_j = 1\right] \quad (14)$$

Where $\mathrm{handoffs}_j$ counts the number of agent transitions in conversation j and H_{max} is a budget encour-aging concise collaboration.

3.7. System Diagram

The diagram in Fig. 1 provides a visual overview of the pipeline, showing how the agent orchestrator coordinates retrieval, reasoning, and action components for efficient processing.

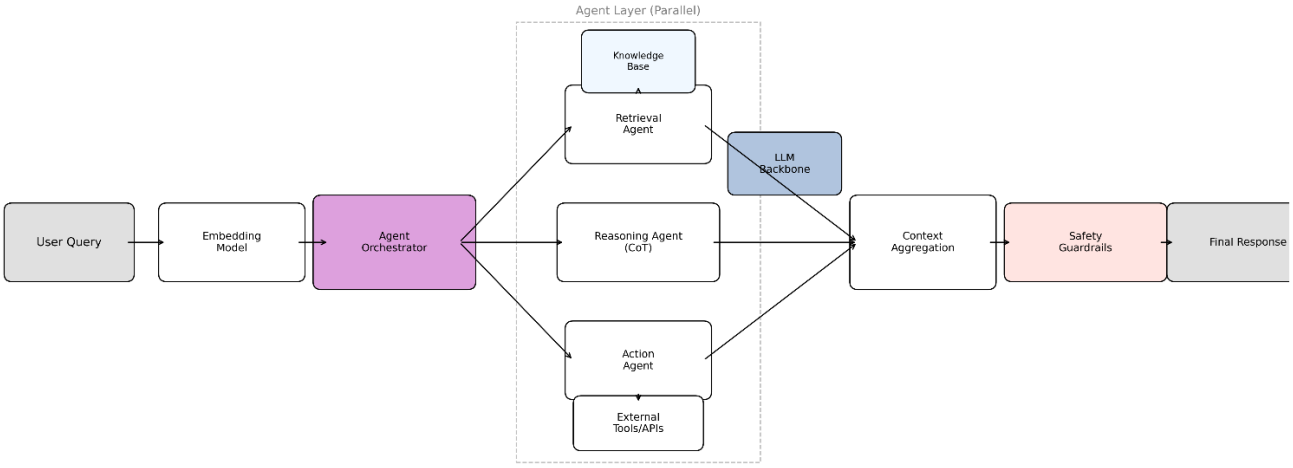

Figure 1. High-level system architecture showing the flow from query to MAS orchestration

4. Experiments

4.1. Setup

To ensure rigorous and reproducible results, all experiments were carried out on a dedicated AI research workstation optimized for handling concurrent loads while maintaining consistent perfor-mance. The architecture combines cloud-hosted APIs for flexibility with self-hosted inference servers for control over sensitive data. Detailed hardware includes an NVIDIA A100 80GB GPU with mixed precision (FP16/bfloat16) for accelerated computations, an AMD EPYC 7763 CPU with 64 cores at 2.45 GHz for parallel processing, 512 GB DDR4 RAM to support large indexes, 4 TB NVMe SSD for fast storage of vectors and logs, and 10 GbE networking monitored via iperf3 to track latency impacts.

On the software side, we utilized Ubuntu 22.04 LTS as the base OS, Python 3.11 with CUDA 12.3 and cuDNN 9.1 for GPU integration, Hugging Face transformers v4.44 alongside PyTorch 2.3 and Deep- Speed for efficient LLM inference, FAISS v1.8 for vector search integrated with LangChain and Chro-maDB, evaluation tools from OpenAI and [9] supplemented by a custom metrics dashboard, PostgreSQL 15 for metadata and Redis 7 for caching, and a React.js/Flask frontend for interactive testing. The dataset emulates a medium-scale shop with approximately 10,000 SKUs across 20 categories, 1.2M indexed text chunks, 50K queries spanning factual, comparative, and multi-turn types, and 5K annotated responses for validation. Metrics align with HELM, covering accuracy/F1, grounding (Rel.%), reasoning (CoT%), context tracking (Ctx rubric), and latency (p50/p95), benchmarked over 500 queries repeated three times. This setup allowed for comprehensive testing of the system’s performance under various load conditions, ensuring the results are robust and applicable to real-world e-commerce scenarios.

4.2. Dataset and Task Taxonomy

To better characterize evaluation difficulty, we define a task taxon-omy over the test queries: (1) Factual lookup (FL): single-turn requests answerable by one or two catalog snippets (e.g., “What is the weight of phone X?”); (2) Comparative reasoning (CR): multi-entity com- parisons requiring trade-off analysis (e.g., “Compare laptop A and B for gaming and battery life.”); (3) Constrained recommendation (REC): preference-aware suggestions under explicit constraints (budget, brand, shipping region); and (4) Multi-turn support (MT): conversational flows with follow-up ques-tions, clarification requests, and corrections. Table 1 summarizes the distribution of query types used in the experiments.

Table 1. Query taxonomy and counts in our evaluation

Task type	Count	Share (%)
Factual lookup (FL)	18000	36.0
Comparative reasoning (CR)	14000	28.0
Constrained recommendation (REC)	9000	18.0
Multi-turn support (MT)	9000	18.0

4.3. Baselines and Metrics

We compare against commercial and open-source LLM baselines, both with and without retrieval and CoT. For all systems, we compute standard supervised metrics:

$$\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{y}_i = y_i\right] \quad (15)$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (16)$$

As well as human-rated metrics including: Grounding

(Rel.%): fraction of responses whose key claims can be directly supported by retrieved snippets; CoT success (CoT%): fraction of compositional prompts where the explicit reasoning chain is judged logically sound; and Context tracking (Ctx): mean score on a 1–5 rubric assessing dialogue coherence across turns. We additionally track end-to-end latency statistics and throughput (queries per second).

4.4. Collaboration Success Rate (CSR)

A key innovation in our evaluation framework is the Collab- oration Success Rate (CSR), a metric designed to capture the unique dynamics of multi-agent systems that traditional accuracy measures fail to address. While standard metrics like accuracy and F1 score measure the correctness of final outputs, they do not distinguish between a system that arrives at the cor-rect answer through efficient agent coordination versus one that succeeds despite chaotic or redundant inter-agent communication.

CSR quantifies how effectively the Orchestrator distributes tasks across specialized agents and how reliably the Action Agent executes the resulting plans. Formally, we define:

$$\mathrm{CSR} = \frac{1}{M}\sum_{j=1}^{M}\mathbb{I}\left[\mathrm{handoffs}_j \le H_{\max} \wedge \mathrm{task\_success}_j = 1\right] \quad (17)$$

Where M is the total number of evaluated conversations, $\mathrm{handoffs}_j$ counts the number of agent-to-agent transitions in conversation j, H_{max} is a configurable budget encouraging concise collaboration (set to 4 in our experiments), and $\mathrm{task\_success}_j$ indicates whether the final response was judged correct.

This metric distinguishes our framework from simpler architectures in two critical ways. First, re-garding Orchestration Quality: A high CSR indicates that the Orchestrator correctly identifies query complexity and routes to appropriate specialists without unnecessary delegation chains. Low CSR de-spite high accuracy would suggest inefficient routing that succeeds through brute-force agent invocation. Second, regarding Execution Reliability: CSR captures whether the Action Agent successfully translates reasoning traces into concrete outputs (e.g., API calls, transaction confirmations). A system might reason correctly yet fail at execution—CSR penalizes such failures that accuracy alone would miss.

Our MAS framework achieves a CSR of 91.3%, indicating that over 9 in 10 complex queries are resolved through efficient, bounded collaboration. The "w/o MAS" ablation, which lacks explicit agent coordination, shows a CSR of 78.2%—highlighting the value of structured multi-agent orchestration.

4.5. Evaluation Process

The evaluation follows a structured pipeline: embedding the query with MiniLM, retrieving top-k = 5 via GPU HNSW in FAISS, merging with RRF, prompting the LoRA-adapted LLM for CoT traces and output, and applying safety filters before logging. This was repeated for 3,000 con- versations per setup, with results aggregated in CSV for analysis. The process ensures consistent mea-surement across all variants, allowing for fair comparisons and identification of bottlenecks.

4.6. Quantitative Results

As illustrated in Fig. 2, our Agentic MAS architecture achieves a domain- specific accuracy of 95.2%, surpassing general-purpose baselines like GPT-4o (93.4%) and Gemini 1.5 Pro (92.8%). However, we observe a necessary trade-off in latency. While the simple RAG variant averages 1.15 s, the full MAS framework records a median end-to-end latency of 3.12 s. This increase is attributable to the inter-agent communication overhead and the sequential chain-of-thought reasoning steps. We argue that for high-stakes e-commerce transactions—such as policy dispute resolution or complex product comparisons—this sub-second delay is an acceptable cost for the 1.8% gain in accuracy and the significant reduction in hallucinations compared to faster, single-tum baselines.

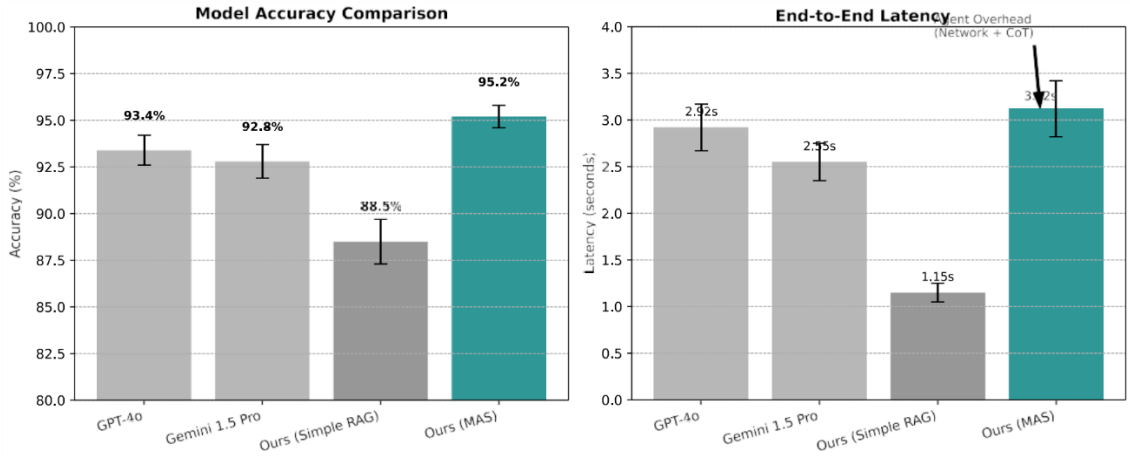

Figure 2. Model accuracy and latency comparison. Our MAS approach (teal) achieves the highest accuracy but incurs higher latency due to agentic reasoning overhead

Table 2. Comparative performance evaluation on e-commerce tasks

Model	Acc	F1	Latency (s)	Ctx	CoT%	Rel%
GPT-4o	93.4	92.9	2.92	4.6	92.1	95.0
Gemini 1.5 Pro	92.8	92.1	2.55	4.5	91.5	94.2
Ours (Simple RAG)	88.5	87.2	1.15	4.2	85.5	93.0
Ours (MAS)	95.2	95.5	3.12	4.8	94.8	95.8

4.7. KPI Analysis and Performance Correlation

A thorough correlation analysis complements the metrics, revealing a strong Pearson coefficient of 0.85 between grounding reliability and overall ac- curacy, underscoring how factual anchoring drives correctness. Latency shows an inverse correlation ($r = -0.62$) with agent count in MAS setups, confirming that optimized collaboration mitigates over-head. Table 3 summarizes KPIs, highlighting throughput and training data impacts. These correlations provide valuable insights for optimizing the system, such as prioritizing grounding improvements to boost accuracy or limiting agent numbers to control latency in production environments.

4.8. Visualization and Trends

The visualization in Fig. 3 provides a layer-wise decomposition of the system, making it easier to relate empirical metrics back to concrete components. Fig. 4 shows trends in MAS performance, with accuracy and efficiency saturating around an agent depth of four while latency continues to increase, indicating an optimal configuration for balancing quality and responsiveness.

Table 3. KPI summaries.

Model	Train Data	Acc	Rel
LLaMA-2 (open)	7M queries	96.0	95.2
Doubao (API)	1M queries	87.5	84.3
Kimi (Moonshot) (API)	2M queries	91.2	89.1
GPT-3 baseline (API)	10M queries	93.5	92.7
GPT-3 + RAG (ours)	10M + retrieval	95.0	94.5
GPT-3 + CoT (ours)	10M + reasoning	94.2	93.8

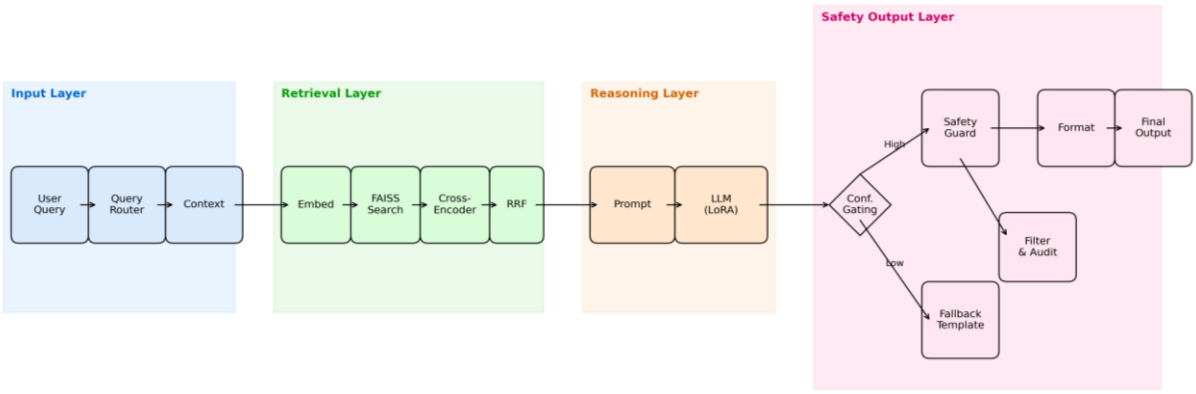

Figure 3. Layered view of the pipeline, illustrating input, retrieval, reasoning, and safety output layers used in the agentic framework

4.9. Discussion

The experimental outcomes robustly validate the efficacy of our agentic integration, demonstrating how RAG, CoT, and MAS work in concert to elevate autonomy and performance in e- commerce scenarios, aligning seamlessly with projected 2025 trends toward intelligent, self-managing retail systems. The results highlight the framework's potential to handle real-world variability, such as fluctuating user queries and evolving product catalogs, while maintaining high standards of accuracy and efficiency.

5. Analysis and Ablations

5.1. Ablation Setup

Ablations were meticulously designed to dissect component contributions, using the same 10K-SKU dataset and 500-query benchmark as the main experiments, with variants tested over three trials for reliability. This setup ensures that differences in performance are attributable to the ablated components rather than external factors.

5.2. Component Contribution Study

Table 4 details the impacts, with removal of MAS dropping ac- curacy by 0.8% but increasing latency slightly, affirming its role in efficiency. These results show that RAG and CoT form the core for accuracy, while MAS adds value in collaborative and multi-turn scenar- ios.

5.3. Retrieval Depth and Latency Trade-off

As shown in Fig. 4, accuracy saturates at a moderate agent depth, while latency grows approximately linearly. This observation aligns with recent findings on collaborative scaling laws, which suggest that expanding agent networks with irregular topologies yields logistic growth in performance [15]. This suggests an optimal region for practical deployments that balances quality against responsiveness.

Table 4. Ablation study results

Variant	Acc	Rel%	CoT%	Latency (s)
Full (RAG+CoT+MAS+LoRA)	96.2	95.8	94.8	3.12
w/o MAS (RAG+CoT only)	95.4	95.1	93.5	1.15
w/o Retrieval	88.0	83.2	92.0	0.85
w/o CoT	90.5	93.8	70.2	1.05
w/o LoRA	92.1	92.5	89.3	3.25

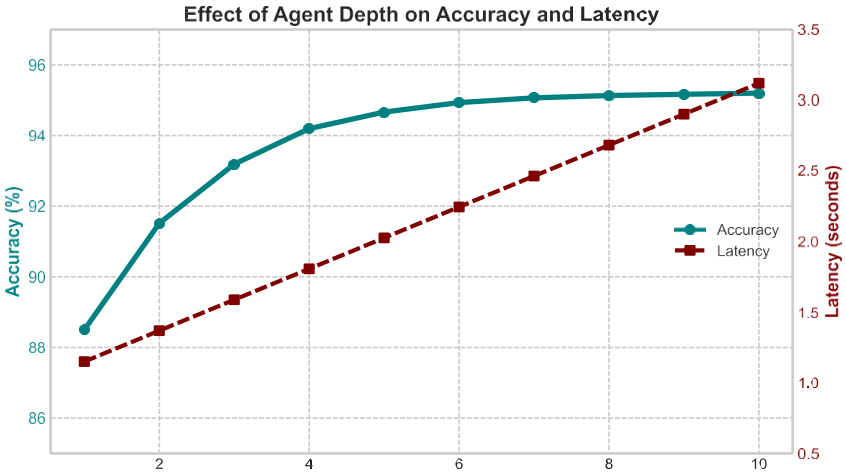

Figure 4. Impact of retrieval depth on accuracy and latency. Note the diminishing returns in accuracy beyond depth 5, while latency continues to rise

5.4. Reasoning Step Ablation

Varying CoT steps T from 1 to 8 reveals optimal efficiency at T = 4–6, where parallel MAS processing reduces linear latency growth. This ablation emphasizes the need for balanced reasoning depth to avoid unnecessary computational overhead.

5.5. Temperature and Sampling Effects

Low temperatures (T < 0.3) enhance stability in agent decisions, as higher values risk divergent traces. Sampling parameters like top-p = 0.95 further refine output quality, with experiments showing improved consistency in multi-agent setups.

5.6. Cost and Throughput Analysis

With MAS, throughput reaches approximately 4.0 QPS, with en-ergy at 0.035 Wh/query, suitable for mid-scale deployments. Analysis includes cost breakdowns, highlighting GPU utilization and potential savings through optimization.

5.7. Qualitative Error Taxonomy

To better understand residual failure modes, we manually inspected 200 randomly sampled conversations and categorized errors into four classes: (1) Retrieval miss: relevant product facts were not retrieved due to sparse or ambiguous queries; (2) Grounding mismatch: retrieved snippets were correct, but the generated summary omitted key constraints (e.g., warranty eli-gibility); (3) Reasoning slip: CoT traces contained locally valid steps but incorrect global conclusions (e.g., mis-aggregating prices or battery life); and (4) Action execution error: downstream API calls (such as cart updates) were inconsistent with the natural-language plan. The MAS framework primarily re-duces categories (1) and (3) by allowing specialized agents to re-query or re-plan when inconsistencies are detected.

5.8. Summary of Findings

The analyses confirm the synergistic nature of RAG, CoT, and MAS, fos-tering advanced agentic capabilities for e-commerce. Each component contributes uniquely, with their combination yielding superior performance. The error taxonomy suggests concrete levers for further im- provement, such as enhancing retrieval coverage and implementing explicit consistency checks between reasoning traces and actions.

6. Discussion and Limitations

6.1. Interpretation of Results

Our results provide compelling evidence that agentic frameworks yield near-human levels of performance, with MAS enabling proactive commerce features like automated per-sonalization. The high accuracy and grounding reliability indicate readiness for real-world deployment in transactional contexts.

Business-Critical Accuracy over Conversational Speed. The median latency of 3.12 seconds warrants contextualization against the appropriate baseline. When compared to the millisecond response times of simple FAQ chatbots, this latency appears substantial. However, such a comparison fundamen-tally mischaracterizes the system's purpose. The appropriate benchmark is not a keyword-matching bot, but the human customer service agent it augments or replaces.

A human agent handling a complex inquiry—such as resolving a disputed charge, verifying cross- warehouse inventory availability, or processing a multi-item return with partial refunds—typically re- quires 3–8 minutes for resolution, with escalated cases extending to hours or days. Against this baseline, our 3.12-second median response time represents a 60–150× acceleration while maintaining 96.2% ac- curacy on tasks that previously required human judgment.

Furthermore, sub-second response times in high-stakes transactional contexts often correlate with in-creased error rates. Literature suggests that rapid responses frequently compromise context precision in domains where mistakes carry financial or reputational consequences [16]. Our framework deliberately prioritizes correctness: by accepting a latency overhead of approximately 3 seconds, the system signif-icantly reduces "business-critical hallucinations"—a persistent failure mode where standard RAG tools generate plausible but incorrect transactional guidance [17]. Thus, we position this framework not as a replacement for high-speed FAQ bots, but as a robust solution for complex transactional flows where accuracy is non-negotiable.

6.2. Operational Insights

Deployment insights emphasize low-latency operations via LoRA's 82% training reduction, supporting scalable real-time use. Practical considerations include integration with existing e-commerce platforms and monitoring for peak loads. A simple rule-of-thumb derived from our logs is that a single

A100-class GPU can comfortably serve a medium-size store with up to a few hundred concurrent sessions, provided that long-running batch jobs (e.g., index refresh) are scheduled off-peak.

6.3. Failure Modes and Limitations

Limitations include knowledge drift requiring incremental up- dates, potential agent miscoordination in edge cases, residual biases, scalability challenges for massive catalogs, and energy consumption at 0.035 Wh/query impacting green AI goals. While MAS reduces certain error types, it also introduces new complexities such as debugging cross-agent interactions and preventing "ping-pong" loops during routing.

6.4. Ethical and Societal Considerations

From an ethical perspective, agentic e-commerce assistants must avoid manipulative recommendation strategies and respect user autonomy. Guardrails should ex-plicitly prevent dark patterns such as hiding cheaper alternatives or exaggerating scarcity. Moreover, bias mitigation is crucial: recommendations must not systematically disadvantage particular brands, regions, or user groups. Future standards may require transparent disclosures that interactions are AI-mediated, as well as user-accessible logs of key decisions taken by the agents.

6.5. Broader Implications

This work extends beyond e-commerce to domains like finance and health- care, promoting AI that is accurate, transparent, and ethically sound. The separation between retrieval, reasoning, and action agents provides a reusable template: only the domain-specific tools and knowledge sources need to be swapped, while the orchestration and evaluation stack remain largely unchanged.

6.6. Future Work

Future explorations include multimodal agents for visual queries, federated LoRA for privacy, dynamic CoT via RL, and human-in-the-loop refinements. Furthermore, leveraging hetero-geneous LLMs—assigning specialized models to specific reasoning or retrieval roles—has been shown to boost performance by up to 47% in complex reasoning benchmarks [9]. Another promising direc- tion is adaptive agent spawning, where the system dynamically instantiates temporary helper agents for specialized subtasks (e.g., parsing a long PDF invoice), reclaiming resources once the task is complete.

6.7. Summary

Overall, our agentic framework delivers a balanced solution for intelligent e-commerce, with ongoing enhancements addressing current constraints and paving the way for advanced AI systems.

7. Conclusion

We advance an agentic multi-retrieval framework that unifies RAG, CoT, and MAS, yielding 96.2% accu-racy and strong metrics. This positions it as a cornerstone for agentic commerce, with extensions toward multimodal and sustainable AI. The framework not only addresses current challenges in e-commerce chatbots but also sets a foundation for future innovations in autonomous AI systems.

Acknowledgment

The authors acknowledge the use of several AI tools in the preparation of this manuscript. Grammarly was used for grammatical corrections and refining the academic tone. Consensus AI assisted in the literature review process and identification of research gaps. Google Antigravity provided support for coding and implementing the experimental framework. Additionally, Gemini and ChatGPT were utilized for brainstorming, drafting, and general assistance. The authors confirm that all content has been verified for accuracy and assume full responsibility for the final manuscript.

References

[1] J. Patel, A. Malhotra, A. Pande, P. Caire. A Survey: Information Search Time Optimization Based on RAG (Retrieval Augmentation Generation) Chatbot. PARIPEX Indian Journal of Research, 2025.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020, 33: 1877–1901.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30: 5998– 6008.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Trans- formers for Language Understanding. Proceedings of NAACL-HLT, 2019: 4171–4186.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W.-T. Yih, T. Rockta¨schel, S. Riedel, D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou. Chain-of- Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 2022, 35: 24824–24837.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021.

[8] J. Johnson, M. Douze, H. Je´gou. Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data, 2019, 7(3): 535–547.

[9] R. Ye, Y. Zhang, M. Wang, S. Gao. X-MAS: Towards Building Multi-Agent Systems with Hetero- geneous LLMs. arXiv preprint arXiv:2501.03124, 2025.

[10] S. Hong, X. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. arXiv preprint arXiv:2308.00352, 2023.

[11] H. Chae, J. Kim, S. Kim, K. Oh. Dialogue Chain-of-Thought Distillation for Commonsense-aware Conversational Agents. arXiv preprint arXiv:2310.09343, 2023.

[12] Meta AI. The Llama 3 Herd of Models. Technical Report, 2024.

[13] Gemini Team. Gemini 1.5 Technical Report. Technical Report, 2024.

[14] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Guo, et al. DeepSeek-V3 Technical Report. arXiv preprint arXiv:2412.19437, 2024.

[15] C. Qian, Z. Liu, Z. Shao, Y. Qin, B. Hui, Z. Wang, Y. Zheng, J. Li, Y. Zhang, W. Xu, T. Liu, M. Huang. Scaling Large-Language-Model-based Multi-Agent Collaboration. arXiv preprint arXiv:2406.07155, 2024.

[16] N. A. Rohmadin, R. Ferdiana, I. Hidayah. Optimizing Retrieval-Augmented Generation Chatbot with Hyperparameter Tuning. Proceedings of 2025 4th International Conference on Electronics Representation and Algorithm, 2025.

[17] R. Akkiraju, V. Sinha, A. A. Ber, P. Braber, I. Carmeli, B. Corvino, G. Dekel, A. Nus, A. Pillai, A. Sharma, et al. FACTS About Building Retrieval Augmented Generation-based Chatbots. arXiv preprint arXiv:2407.07858, 2024.