AI Retail Agent Team

Multi-Agent Ecosystem for Intelligent Retail

Powered by: Google Gemini 2.0 Flash • Elasticsearch • Google ADK

Technology Stack: Python • FastAPI • Server-Sent Events • Real-time Analytics

Presented by: Yash Kavaiya

October 24, 2025

What is AI Retail Agent Team?

A sophisticated multi-agent coordination system that orchestrates six specialized AI agents to deliver comprehensive retail operations support.

Core Components:

- Retail Coordinator - Main orchestration agent
- Product Search - Visual and text-based product discovery
- Review Analysis - Customer sentiment and feedback insights
- Inventory Management - Real-time stock tracking
- Shopping Cart - Transaction and purchase analytics
- Customer Support - FAQ and issue resolution

Technical Foundation

Backend Stack:

- Google ADK Framework
- Gemini 2.0 Flash Model
- Elasticsearch Database
- FastAPI Web Server
- Server-Sent Events (SSE)
- Python 3.11+

Key Features:

- Real-time streaming
- Multi-agent coordination
- Semantic search
- Vector embeddings
- Session management
- RESTful APIs

The Brain of the System

Core Responsibilities:

- Intelligent Request Routing - Analyzes queries and routes to appropriate specialists
- Multi-Agent Coordination - Orchestrates parallel and sequential workflows
- Context Management - Maintains conversation continuity across agents
- Response Synthesis - Combines outputs into coherent answers
- Escalation Handling - Manages edge cases and fallbacks

Model: Gemini 2.0 Flash for lightning-fast coordination

Visual & Text-Based Product Discovery

Capabilities:

- Text Search - Fuzzy matching on product names and descriptions
- Visual Similarity - ImageBind embeddings (1024-dim vectors)
- Category Browsing - Hierarchical navigation
- Product Comparison - Side-by-side feature analysis
- Similar Products - kNN-based recommendations

Data Index: imagebind-embeddings with rich product metadata

Customer Sentiment Intelligence

Analysis Features:

- Semantic Search - RRF (Reciprocal Rank Fusion) on reviews
- Sentiment Analysis - Positive, negative, neutral classification
- Theme Extraction - Common topics and patterns
- Rating Analytics - Statistical aggregations
- Department Insights - Category-specific feedback

Data Source: Women's Dresses Reviews Dataset with 20+ fields

Real-Time Stock Intelligence

Key Functions:

- Product Inventory Checks - Multi-location availability
- Low Stock Alerts - Critical, high, and medium severity
- Regional Analysis - Store-by-store breakdown
- Demand Forecasting - Predictive restock recommendations
- Seasonal Analytics - Readiness scores for seasonal items

Intelligence: Compares inventory vs. demand forecasts

Purchase Analytics & Behavior Insights

Analytics Capabilities:

- Category Analytics - Spending patterns by product type
- Customer Profiles - Complete purchase history
- Gender Trends - Demographic shopping preferences
- High-Value Transactions - Premium customer identification
- Mall Performance - Location-based metrics
- Payment Analytics - Method usage and preferences

Intelligent FAQ & Issue Resolution

Support Features:

- FAQ Database Search - Full-text semantic search
- Topic-Based Queries - Targeted policy information
- Order Inquiries - Status tracking and updates
- Returns & Refunds - Policy explanation and processing
- Warranty Support - Claims and coverage details

Knowledge Base: Elasticsearch-powered FAQ system

Powering Real-Time Data Access

Indices Used:

1. imagebind-embeddings - Product catalog with visual embeddings
2. retail_store_inventory - Real-time stock levels
3. customer_shopping_data - Transaction history
4. womendressesreviewsdataset - Customer reviews
5. faqs_data - Support documentation

Search Techniques: kNN, semantic search, full-text, aggregations

Real-Time Interactive Interface

Frontend Stack:

- Vanilla JavaScript - No framework dependencies
- Server-Sent Events - Real-time message streaming
- Responsive Design - Mobile-first approach
- Session Management - Persistent conversations
- Local Storage - Conversation history

Server: FastAPI with async SSE endpoints

Intelligent Workflow Orchestration

Common Workflows:

- Product Browsing:
  Search → Inventory Check
- Informed Purchase:
  Search → Review Analysis → Inventory → Cart
- Customer Query:
  Support → FAQ Search → Policy Explanation
- Stock Management:
  Inventory → Demand Forecast → Low Stock Alerts

Technology Stack Overview

AI & ML:

- Google Gemini 2.0 Flash
- ImageBind Embeddings
- Vector Similarity Search
- Semantic Retrieval (RRF)

Infrastructure:

- Elasticsearch Cloud
- FastAPI Web Framework
- Python 3.11+
- Uvicorn ASGI Server

Development: Google ADK (Agent Development Kit)

Practical Applications

1. E-commerce Assistant
   Help customers find products, check reviews, and complete purchases
2. Inventory Management
   Real-time stock monitoring and automated restock recommendations
3. Customer Support
   24/7 automated FAQ responses and policy explanations
4. Business Intelligence
   Shopping trends, customer behavior, and sales analytics
5. Store Operations
   Multi-location coordination and regional performance tracking

Value Proposition

For Customers:

- Intelligent product discovery with visual search
- Real-time stock availability
- Instant FAQ support
- Personalized recommendations

For Business:

- Reduced support costs (24/7 automation)
- Improved inventory management
- Data-driven insights
- Enhanced customer satisfaction

System Capabilities

Performance Metrics:

- Response Time: Sub-second with Gemini 2.0 Flash
- Concurrent Sessions: Unlimited with session management
- Search Performance: Real-time kNN and semantic search
- Data Volume: Handles millions of products and reviews

Scalability: Elasticsearch Cloud + FastAPI async architecture

Upcoming Features

- Voice Integration - Natural language voice commands
- Image Upload Search - Find products from photos
- AR Try-On - Virtual product visualization
- Predictive Analytics - ML-powered demand forecasting
- Multi-Language - Global language support
- Mobile App - Native iOS and Android applications

Agent Structure

```
root_agent = Agent(
    name='retail_coordinator',
    model='gemini-2.0-flash',
    description='Main coordinator...',
    instruction="""
    Coordinate multiple specialized agents
    for retail operations...
    """,
    sub_agents=[
        product_search_agent,
        review_analysis_agent,
        inventory_agent,
        shopping_agent,
        customer_support_agent
    ]
)
```

Thank You!

Follow for More Updates:

@genai-guru

Yash Kavaiya

Gen AI Guru

Easy AI Labs