

MULTI-AGENT DECISION SUPPORT SYSTEM FOR RETAIL OPERATIONS

Presented by:

Victor Dumaslan
Dongmei Han
Niranjan Rao Saraf Srinivas
Thi Thao Thien Tien

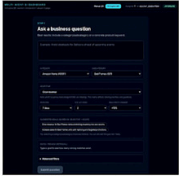
Project Advisors:

Dr. Simon Shim

Introduction

Retail operators face significant cognitive load when making decisions around pricing, inventory, and demand. Most existing AI tools present surface-level information without actionable guidance - working for humans rather than with them. This results in decision fatigue, missed opportunities, and suboptimal outcomes in fast-moving retail environments.

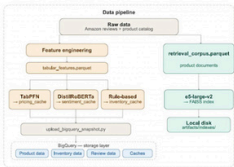
- Operators need AI that reduces cognitive load while preserving human authority over final decisions
- Current tools offer fragmented analytics with no mechanism for structured multi-perspective reasoning
- Our system targets three core areas: pricing optimization, inventory positioning, and demand signal interpretation



Methodology

Data Collection & Pipeline

We sourced data from the Amazon Reviews 2023 dataset (McAuley Lab), focusing on the Home & Kitchen category. Raw JSONL files were cleaned and validated using quality gates - requiring valid product identifiers, ratings between 1.0-5.0, minimum 30-character reviews, and at least 15 reviews per product. Price outliers were handled via 1st-99th percentile winsorization. The final curated dataset contains 50,000 products and 5.3 million reviews, split 70/15/15 into train, validation, and test sets



- Raw JSONL data ingested and validated before storage in Google BigQuery
- Reviews and metadata transformed into Parquet format for downstream agent use
- Synthetic inventory data generated to simulate stock and sales signals for the Inventory Agent

Analysis and Results

To support fast runtime performance, specialist agent outputs - including pricing, sentiment, and inventory signals - were precomputed and stored as caches in BigQuery. The retrieval index and embedding model are loaded into memory at startup, eliminating per-request re-encoding

Model	Recall@10	nDCG@10	MRR
BM25	0.9500	0.9503	0.9307
bge-large-en-v1.5	0.9577	0.9775	0.9707
e5-large-v2	0.9585	0.9786	0.9720

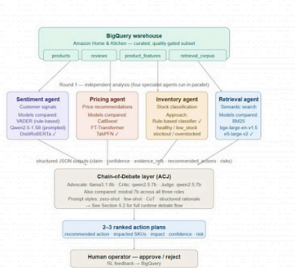
Agent Design & Chain-of-Debate Framework

Specialist agents analyze each retrieved product in parallel before passing structured outputs to the Advocate-Critic-Judge (ACJ) debate layer. Each agent underwent model comparison to identify the best performing approach.

- Retrieval Agent:** Compared sparse vs dense retrieval
- Sentiment Agent:** Compared rule-based, prompted LLM, and fine-tuned encoder
- Pricing Agent:** Compared gradient-boosted, deep learning, and in-context tabular inference
- ACJ Debate Layer:** Evaluated 27 combinations of LLMs across Advocate, Critic, and Judge roles

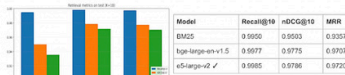
Agent	Models Compared	Evaluation Metrics
Retrieval	BM25, bge-large-en-v1.5, e5-large-v2	Recall@10, nDCG@10, MRR
Sentiment	VADER, Qwen2.5-7B, DistilRoBERTa	Macro F1, MCC, Accuracy
Pricing	CatBoost, FT-Transformer, TabPFN	RMSE, MAE, Violation Rate
ACJ Layer	Llama3.1-8B, Qwen2.5-7B, Mistral-7B	Conflict, Justification, Constraint, Latency

In the ACJ debate layer, the Advocate proposes candidate actions, the Critic identifies risks and conflicts, and the Judge synthesizes the final 2-3 ranked action plans. Operators retain final approval authority, with all decisions logged to BigQuery for reinforcement learning feedback.



Retrieval Results

The dense retrieval model e5-large-v2 outperformed both the BM25 baseline and bge-large-en-v1.5 across all metrics, confirming that semantic retrieval provides stronger candidate ranking than keyword-based search. High retrieval quality is critical for the system since all downstream agents depend on the retrieved candidate set.

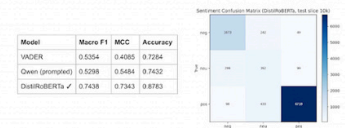


Pricing & Sentiment Results

TabPFN outperformed both CatBoost and FT-Transformer on the pricing task within the $\pm 10\%$ policy bound, demonstrating that in-context tabular inference generalizes better than traditional tree ensembles on this dataset



The fine-tuned DistilRoBERTa outperformed both VADER and prompted Qwen across all classification metrics, performing strongly on positive and negative classes. However, the neutral class showed lower precision, which is a known challenge in sentiment classification due to class ambiguity.



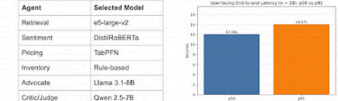
ACJ Results

The selected ACJ configuration (Llama 3.1-8B as Advocate, Qwen 2.5-7B as Critic and Judge) achieved the best performance across all rubric metrics while maintaining the lowest latency among all tested combinations. All outputs were valid and correctly structured with an ok_rate of 1.00. A prompt-style ablation further revealed that few-shot prompting improved constraint respect at the cost of conflict detection.

Advocate	Critic	Judge	Conflict	Justification	Constraint	Latency
llama3.1.8b	qwen2.5.7b	qwen2.5.7b	0.90	0.97	0.85	7.63s
mistral.7b	qwen2.5.7b	mistral.7b	0.90	0.83	0.78	8.98s
mistral.7b	qwen2.5.7b	qwen2.5.7b	0.80	1.00	0.73	8.60s
llama3.1.8b	qwen2.5.7b	mistral.7b	0.80	0.80	0.92	8.30s

System Performance

The end-to-end system was evaluated across 20 test queries, achieving a median latency (p50) of 12.08 seconds and p95 of 14.07 seconds. An ok_rate of 1.00 was maintained across all queries, confirming consistent and constraint-compliant outputs.



Summary/Conclusions

This project implemented a multi-agent decision support system for retail operations using a Chain-of-Debate (CoD) framework, combining semantic retrieval, sentiment analysis, pricing, and inventory reasoning while preserving human control over final decisions.

Model comparisons confirmed e5-large-v2, DistilRoBERTa, and TabPFN as the best performing models. The ACJ debate layer achieved an ok_rate of 1.00 across all queries. Key limitations include hardware-dependent latency and reliance on synthetic inventory data.

Key References

- [1] Liang et al. (2023) - Encouraging Divergent Thinking in Large Language Models Through Multi-Agent Debate
- [2] Estornell & Liu (2024) - Multi-LLM Debate: Framework, Principals, and Interventions
- [3] Ling et al. (2024) - MeMAD: Structured Memory of Debates for Enhanced Multi-Agent Reasoning
- [4] Zhang et al. (2025) - If Multi-Agent Debate is the Answer, What is the Question
- [5] Zhai et al. (2025) - Beyond Retrieval-Ranking: A Multi-Agent Cognitive Decision Framework for E-Commerce Search
- [6] McGrath et al. (2025) - Collaborative Human-AI Trust (CHAI-T): A Process Framework for Active Management of Trust in Human-AI Collaboration

Acknowledgements

We would like to thank Professor Simon Shim for his guidance and support throughout this project.

We also thank the Applied Data Science Department and the university GPU lab for providing the computational resources necessary for model training and evaluation.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.