Finny: A Multi-Agent System for Structured Decision-Making with LLMs

Harshitha Ravindra **Utkarsh Bajaj** **Madhur Mehta**
Amazon
{harravin, ubajaj, msmehta}@amazon.com

Abstract

Finny is a multi-agent system that demonstrates how large language models can perform structured decision-making by applying domain-specific rules to multiple related scenarios. Leveraging foundation models with Retrieval-Augmented Generation (RAG), the system applies Standard Operating Procedures (SOPs) for intelligent forecast refinement at scale. Finny employs a two-stage architecture: a knowledge base agent that retrieves and applies domain rules while analyzing historical patterns, and a conversational agent enabling interactive refinement. In user acceptance testing (UAT), the system achieved 97.6% alignment with expert judgment across 124 evaluations (31.5% complete, 66.1% partial), with quantitative validation showing 5.89% mean deviation and 0.993 correlation against human decisions across 1,280 data points. This production-deployed system reduces manual analysis time by 70%, translating to 2,400 annual hours savings in the piloted teams.

1 Introduction

Large language models have shown remarkable capabilities in natural language understanding, yet their application to structured decision-making tasks like forecast adjustment or finetuning requiring consistency across multiple related scenarios remains challenging. Real-world business applications often require applying explicit domain rules to dozens of related cases while maintaining logical coherence and providing explainable reasoning.

We present Finny, a multi-agent system for forecast adjustment that addresses these challenges. The system must apply domain-specific Standard Operating Procedures (SOPs) - essentially symbolic rules - to dozens of related forecast granularities while maintaining consistency and providing natural language explanations for each decision. This demonstrates practical LLM reasoning in production environments where decisions must be both accurate and explainable.

Challenges Addressed: (1) Synthesizing unstructured domain knowledge (SOPs) with structured statistical patterns; (2) Making logical adjustments and maintaining consistency across multiple related forecast granularities; (3) Managing token constraints while processing multiple related scenarios; (4) Ensuring production reliability and explainability.

Key Contributions: (1) RAG-enhanced decision-making synthesizing domain knowledge with statistical patterns; (2) Hierarchical multi-agent architecture achieving 1.5-2x token reduction while maintaining consistency; (3) Context management enabling typical processing of 30+ granularities within API constraints, with tested scalability to 100+ granularities; (4) Comprehensive evaluation across 124 production assessments demonstrating 97.6% alignment with expert judgment.

2 Related Work

Retrieval-Augmented Generation: RAG (1) has proven effective for knowledge-intensive NLP tasks by enabling LLMs to retrieve relevant information from external knowl-

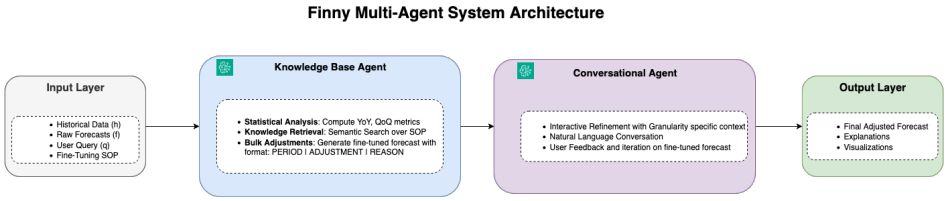

Figure 1: Finny's two-stage multi-agent architecture. The Knowledge Base Agent performs bulk analysis using RAG to retrieve SOPs and compute statistical features, generating comprehensive decisions with explicit reasoning. The Conversational Agent enables interactive refinement with granularity-specific context, achieving significant token reduction.

edge bases before generating responses. While most RAG applications focus on question answering and text generation, we demonstrate its application to structured decision-making by retrieving domain-specific rules (SOPs) and synthesizing them with statistical evidence to generate consistent decisions across multiple related scenarios.

Multi-Agent Systems: Multi-agent architectures (2) have been proposed to decompose complex reasoning tasks, with agents specializing in different aspects of problem-solving. Our work extends this by demonstrating how hierarchical agent specialization can maintain consistency across multiple related decisions while managing computational constraints in production environments. The separation of knowledge-intensive retrieval from interactive refinement achieves significant token reduction while preserving decision quality.

Time Series Forecasting: Traditional forecasting approaches include deep learning models such as Temporal Fusion Transformers (3) and DeepAR (5), which generate predictions from historical data. Recent work has explored foundation models for forecasting (4). These approaches focus on end-to-end prediction from historical patterns.

LLMs for Structured Tasks: While LLMs have shown capabilities in various structured domains, most applications focus on end-to-end generation rather than augmenting existing outputs with domain expertise. Our work demonstrates practical application of LLMs in a production system that augments statistical forecasts by applying domain-specific rules, requiring consistent rule application across multiple related scenarios with explicit reasoning chains enabling expert verification and trust.

3 Methodology

3.1 Problem Formulation

Given historical time series data (actuals $\mathbf{h} = \{h_1, ..., h_n\}$ and raw forecasts $\mathbf{f} = \{f_{n+1}, ..., f_{n+m}\}$) across multiple granularities, we generate adjusted forecasts $\mathbf{f}' = \{f'_{n+1}, ..., f'_{n+m}\}$ by applying domain knowledge:

$$f'_i = f_i \cdot (1 + \alpha_i), \quad \alpha_i \in [-1, 1] \tag{1}$$

where α_i represents percentage adjustment determined by the LLM based on retrieved SOP guidelines and historical patterns. The system must maintain consistency across related granularities while providing explainable reasoning for each decision.

3.2 Multi-Agent Architecture

Finny employs hierarchical two-stage agents (Figure 1). The *Knowledge Base Agent* performs comprehensive analysis using Retrieve-and-Generate API, computing year-over-year(YoY)/ Quarter-over-Quarter(QoQ) growth rates, retrieving relevant SOP instructions, and generating bulk adjustments with structured format (PERIOD | ADJUSTMENT |

REASON). This structured output enforces explicit reasoning chains, enabling verification of decision logic.

The *Conversational Agent* enables interactive refinement through natural language dialogue, operating on granularity-specific context and conversation history. This separation achieves 1.5-2x token reduction by avoiding repeated SOP retrieval while maintaining consistency through shared knowledge base access.

3.3 RAG Pipeline

Three-stage process bridges unstructured domain expertise with structured decision-making:

Stage 1 - Statistical Analysis: Compute growth metrics, trend indicators, and volatility measures from historical data. This provides empirical evidence for decision-making.

Stage 2 - Knowledge Retrieval: Perform semantic search over SOP knowledge base using forecast context (product category, region, seasonality). Retrieved rules serve as guidelines for decisions.

Stage 3 - Synthesis: The LLM combines statistical evidence with retrieved rules to generate decisions. Example: "YoY growth is 15% (statistical evidence) AND SOP recommends conservative adjustments for seasonal products (domain rule) THEREFORE apply 12% adjustment."

3.4 Context Management

Three complementary strategies enable processing within token limits: (1) Granularity-specific filtering extracting only relevant sections, reducing context from $O(N \cdot M)$ to $O(M)$ tokens (for processing N granularities with M tokens per granularity); (2) Conversation history truncation maintaining 4,000 token sliding window; (3) Adaptive report truncation compressing prompts exceeding 25,000 tokens while preserving critical statistics.

4 Production Deployment and Evaluation

Deployment Infrastructure: Finny is deployed on cloud infrastructure with auto-scaling for concurrent user sessions. The system implements exponential backoff with jitter for API throttling management and lazy-loading for memory optimization. The Streamlit-based interface provides interactive visualizations and real-time chat with structured output parsing.

Quantitative Validation: We conducted quantitative comparison between Finny's automated adjustments and human expert adjustments across 1,280 forecast periods spanning 16 granularities and 80 weeks. Mean Absolute Percentage Deviation (MAPD) of 5.89% indicates automated adjustments deviate from human expert adjustments by less than 6% on average. Pearson correlation of 0.993 demonstrates near-perfect linear relationship between Finny's recommendations and human expert patterns. Additional metrics: RMSE 870.32, MAE 463.00. Mean forecast values were highly similar: human adjustments averaged 8,169.49 while Finny's outputs averaged 8,207.64 (less than 0.5% difference).

UAT results: We evaluated Finny across 124 production assessments from 5 expert judges evaluating 109 granularities spanning diverse product categories and geographic regions. Experts indicated Complete Alignment (CA), Partial Alignment (PA), or No Alignment (DNA). Results: 31.5% CA (39/124), 66.1% PA (82/124), 2.4% DNA (3/124), yielding 97.6% alignment with agreement score 0.645 and average rating 3.93/5.

Interaction Efficiency: For cases where judges indicated PA or DNA with the Knowledge Base Agent's initial output, the Conversational Agent required an average of only two prompts to achieve acceptable refinements.

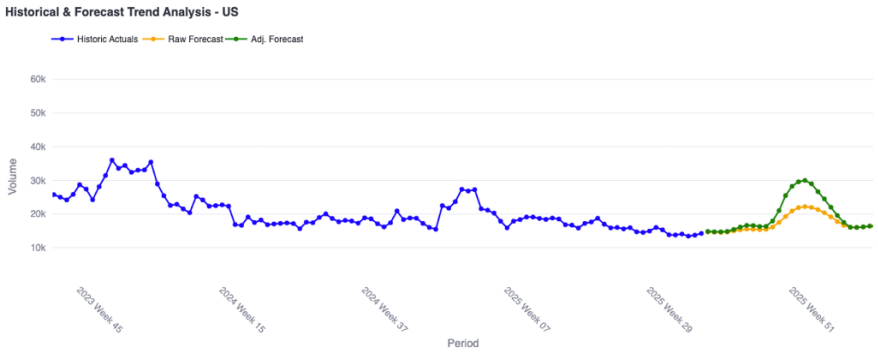

Figure 2: Finny selectively applies adjustments only where analysis indicates necessity. Overlapping orange/green lines indicate periods where no adjustment was needed, demonstrating the system's precision in identifying when intervention is warranted versus when the baseline is adequate.

5 Conclusion

Finny demonstrates practical application of foundation models to structured decision-making requiring consistency across multiple related scenarios. The two-stage architecture provides a scalable framework for applying domain expertise through LLMs. Token limits emerged as the primary constraint, addressed through multi-agent separation of knowledge retrieval from interactive refinement. The system's explainability through natural language reasoning and structured output format (PERIOD | ADJUSTMENT | REASON) enables trust and adoption in business-critical environments.

With 97.6% expert alignment and 70% time savings (2,400 annual hours), Finny validates foundation model value in business-critical workflows requiring consistent rule application across multiple forecasting periods or related decisions. Future work includes developing automated evaluation of consistency across scenarios, learning new rules from expert feedback on approved decisions, and scaling to thousands of related decisions through distributed processing.

References

[1] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 33, 9459-9474.

[2] Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. *UIST*, 1-22.

[3] Lim, B., et al. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecasting*, 37(4), 1748-1764.

[4] Garza, A., & Mergenthaler-Canseco, M. (2023). TimeGPT-1. *arXiv:2310.03589*.

[5] Salinas, D., et al. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecasting*, 36(3), 1181-1191.

A Appendix

A.1 Detailed Architecture Diagram

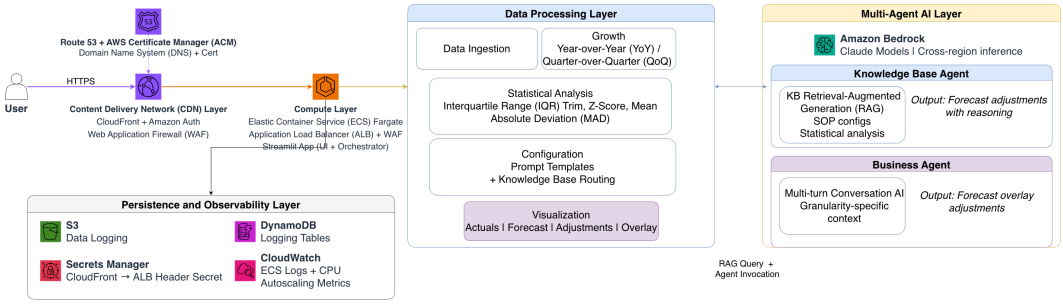

Figure 3: Detailed architecture of Finny's multi-agent system.

A.2 Evaluation Metrics Summary

Quantitative Metrics (1,280 forecast periods):

- Mean Absolute Percentage Deviation (MAPD): 5.89%
- Pearson Correlation: 0.993
- Root Mean Square Error (RMSE): 870.32
- Mean Absolute Error (MAE): 463.00
- Mean difference: <0.5% (8,169.49 vs 8,207.64)

Human Expert Evaluation (124 assessments):

- Complete Alignment: 31.5% (39/124)
- Partial Alignment: 66.1% (82/124)
- No Alignment: 2.4% (3/124)
- Agreement Score: 0.645
- Average Quality Rating: 3.93/5

Operational Impact:

- Manual analysis time reduction: 70%
- Annual hours saved: 2,400 hours (piloted teams)
- Average prompts for refinement: 2 (for PA/DNA cases)
- Typical production scale: 30+ granularities
- Demonstrated scalability: 100+ granularities
- Token reduction: 1.5-2x through multi-agent architecture

A.3 Acknowledgements

We extend our sincere gratitude to our leaders - Puneet Agarwal, Vishnu Chillara, and Pinakini Mohanty for their unwavering support and guidance throughout this research. Our appreciation extends to the entire Demand Planning and Workforce Intelligence team for their continuous support and collaboration.

A.4 Use of Large Language Models

Large language models were used to assist with code generation, debugging, and editing during the preparation of this paper. The authors remain fully responsible for all technical content, experimental results, and claims made in this paper.