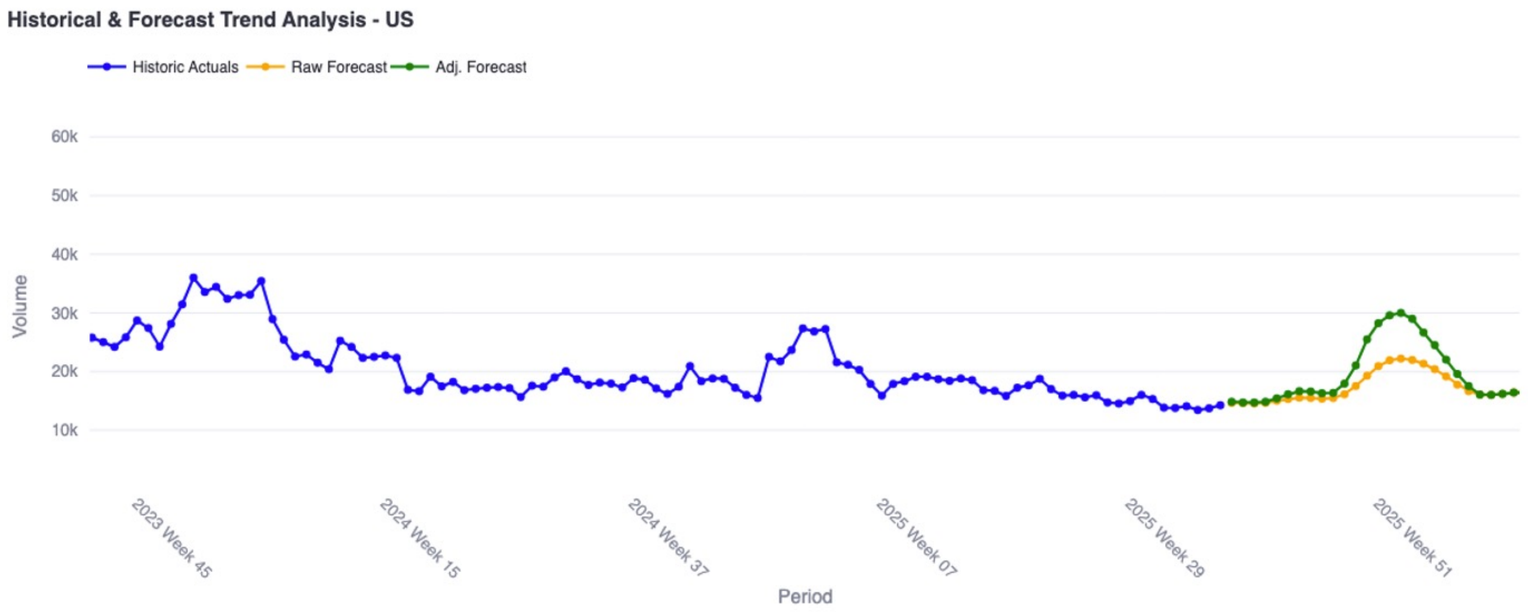
Finny: A Multi-Agent System for Structured Decision-Making with LLMs

ICLR – Logical Reasoning of LLM

Harshitha Ravindra, Utkarsh Bajaj, Madhur Mehta, Puneet Agarwal
Customer & Partner Trust (CPT) Forecasting, Amazon

Forecasting has long depended on analyst judgment that's hard to scale and harder to make consistent. Finny addresses this directly. It is a multi-agent system that **reduced manual analysis time by 70%** within the CPT forecasting team alone. The system combines **RAG-based SOP retrieval with a two-agent design** to bring **domain knowledge** into the forecasting loop automatically



Challenge

At scale, statistical forecasting cannot systematically apply **domain knowledge** codified in SOPs and internal guidelines. Rigid algorithmic approaches **cannot recalibrate when market dynamics shift** without significant manual intervention

Fig 1: *Through pattern analysis guided by knowledgebase, FINNY selectively adjusts forecast only where gap exists. Overlapping orange/green lines indicate adequate forecast, demonstrating intelligent intervention only when warranted*

Approach

- **RAG pipeline** performs semantic retrieval over SOP embeddings, injects statistical features (YoY/QoQ growth, volatility), and synthesizes context-aware adjustments with explicit reasoning chains
- **Production Deployment:** Cloud-based infrastructure with ECS Fargate auto-scaling, exponential backoff for API throttling, powered by Amazon Bedrock and Claude models. Data persistence with AWS tools and logging in S3
- Outputs structured as: **PERIOD | ADJUSTMENT | REASON**
- Two specialized agents separate knowledge-intensive analysis from interactive refinement:
  - **Knowledge Base (KB) Agent -** statistical computation and SOP retrieval across timeseries/granularities
  - **Conversational Agent -** iterative refinement through natural language; forecasters reach acceptable output in **2 prompts on average**

Key Insights

- RAG bridges **unstructured SOPs with structured forecasting** and reasoning simultaneously across quantitative signals and qualitative business rules
- Multi-agent specialization **reduces token consumption** by separating knowledge-intensive analysis from interactive refinement
- **Context management** (granularity filtering, adaptive truncation) scaled the system to 30+ granularities within API constraints
- Structured output format ensured every adjustment came with **a traceable explanation**

Future Work

- Generalize architecture to **adjacent domains**: financial forecasting, capacity planning, inventory optimization
- Automated evaluation framework with **feedback loops** from approved adjustments for continuous improvement

Evaluation & Results

Finny was assessed using a three-layer framework designed to validate both quality and real-world reliability:

- **LLM-as-a-Judge:** the Conversational Agent can validate KB Agent outputs against business constraints and provide alternate recommendations
- **Human-in-the-Loop (HITL):** Expert judges assessed forecast adjustment for each time series
- **Offline Human Evaluation**: Results were compared to previous human adjusted forecasts

HITL results (124 time series, 52 weeks, 5 judges)	
Partial alignment with experts	66.1%
Complete alignment	31.5%
Non-alignment cases	2.4%
Average quality rating	3.93 / 5
Offline Results (16 timeseries, 80 weeks)	
MAPD from expert adjustments	5.89%
Pearson correlation	0.993

Table 1: Results of UAT

Architecture

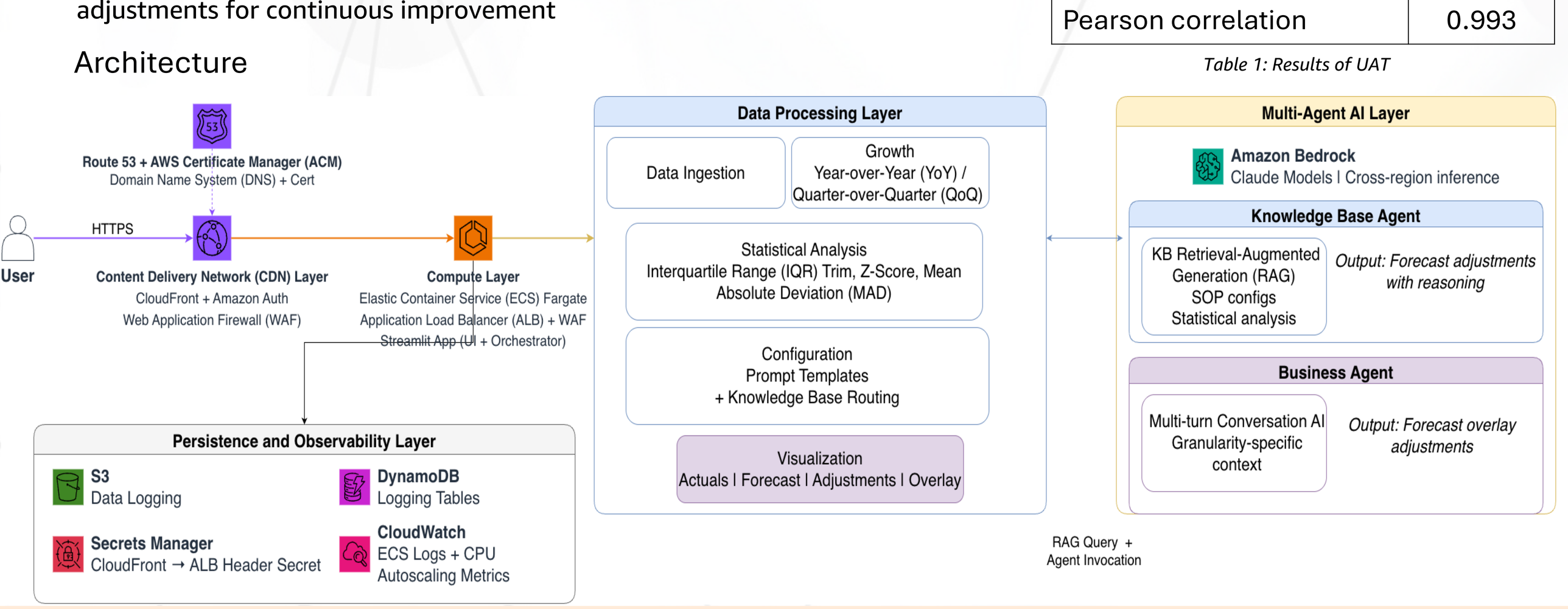


Fig. 2: Finny's hierarchical two-agent architecture, from raw inputs through bulk adjustment generation to conversational refinement and final output