C A S E S T U D Y

How a Major Health Insurance Provider Cut Costs by 85% with AI Agent Orchestration

299,290 real interactions. 18 months in production. Not a pilot.

85.3%	37x	~6x	299K
Cost Reduction	Faster Response	ROI Year 1	Interactions

Powered by **CLU** | cluagents.com
Industry: Health Insurance | Period: August 2024 – January 2026

Executive Summary

A leading health insurance provider serving millions of members faced a critical operational challenge: their customer service infrastructure was consuming up to 85% of operating costs while delivering slow, fragmented experiences. Manual processes across call centers, CRM, ERP, and email created resolution times averaging 20 minutes per interaction.

Using CLU's AI agent orchestration platform, the organization deployed **Martina**—an autonomous AI agent designed to handle customer inquiries end-to-end. Over 18 months of sustained production (not a pilot), Martina processed 299,290 real customer interactions, delivering measurable results that exceeded initial projections.

KEY RESULTS

85.3% reduction in cost per interaction (USD 1.16 → USD 0.17)
37x faster resolution (32 seconds median vs. 20 minutes typical call center)
USD 296,299 in accumulated savings over 18 months of production
8.9% negative sentiment rate — operationally acceptable for digital self-service

The Challenge

The organization's customer service operations were built on a legacy model: human agents handling inquiries across multiple disconnected systems. This created compounding inefficiencies that impacted both cost structure and customer experience.

Operational Reality

- **Cost-to-revenue ratio above 85%** driven by manual handling of routine inquiries
- 20-minute average resolution time as agents navigated between CRM, ERP, spreadsheets, and email
- Churn rates approaching 23% partly attributed to slow, fragmented service experiences
- 72% of customers expected digital-first experiences that the organization could not deliver

The core problem was not a lack of technology—the organization had invested in CRM, ERP, and communication tools. The problem was that these systems were siloed, requiring human agents to act as the orchestration layer between them.

The Solution

CLU deployed an autonomous AI agent (**Martina**) designed to orchestrate customer interactions end-to-end. Unlike a simple chatbot, Martina operates as a coordinated multi-tool agent within CLU's orchestration framework.

Architecture

Martina was built on CLU's GRID framework, which provides the foundational infrastructure for agent orchestration. The CLU Orchestrator coordinates Martina's tools, data sources, and decision logic in real time, enabling her to:

- Capture customer requests across any channel
- Fetch data from CRM, ERP, and internal databases
- Apply business rules and approval workflows automatically
- Escalate to human agents only when necessary, with full context

The deployment moved from integration to production in under 30 days, with the agent handling real customer traffic from day one.

Results in Detail

All metrics below are from 18 months of **sustained production operations** (August 2024 – January 2026), measured against the organization's existing call center baselines. These are not pilot results or projections.

1. Cost Efficiency: 85.3% Reduction

Metric	Before	After
Cost per interaction	USD 1.16	USD 0.17
Estimated total cost (traditional model)	USD 347,178	—
Actual operational cost	—	USD 50,879
Accumulated savings		**USD 296,299**

The agent is approximately 6.8x more cost-efficient than human-handled interactions on a per-unit basis.

2. Speed: 37x Faster Resolution

Metric	Call Center	AI Agent
Median resolution time	20 minutes	32 seconds
Speed improvement	—	**37x faster**

Resolution time measured from first message to last message per conversation. Median (P50) used instead of average due to long tail of asynchronous conversations.

3. Scale: 299,290 Production Interactions

Over the 18-month measurement period, the agent handled 299,290 customer interactions with sustained, stable throughput. Volume grew consistently from approximately 250 daily interactions at launch to over 300 daily interactions by January 2026, demonstrating both reliability and organic adoption.

4. Quality: Low Negative Sentiment

Sentiment analysis across all interactions showed a negative sentiment rate of just **8.9%**—operationally acceptable for digital self-service and comparable to well-run human operations. The current 61% resolution rate reflects the agent's informational scope; with transactional capabilities (account queries, balance management, certificate generation), the projected resolution rate reaches approximately 90%.

What Made the Difference

This was not a chatbot deployment. Three architectural decisions drove the results:

Orchestration, Not Just Conversation

Martina does not simply answer questions—she orchestrates processes. When a customer asks about a claim, the agent coordinates across databases, applies business rules, and returns a resolution. The CLU platform handles the complexity of multi-system coordination so the agent can focus on the customer interaction.

Human-in-the-Loop by Design

The agent escalates to human agents when it encounters edge cases or high-stakes decisions, passing full conversation context and system state. This preserves quality while keeping routine interactions automated. The savings are not achieved by degrading the customer experience.

Production-First Approach

CLU deployed Martina into production handling real customer traffic, not in a sandboxed pilot. Every metric in this case study reflects actual operational performance with real customers, real systems, and real business processes.

Looking Ahead

The next phase focuses on enabling transactional capabilities—allowing the agent to execute actions (balance queries, account management, certificate generation) rather than only providing information. Modeling indicates this will increase the resolution rate from 61% to approximately 90%, further amplifying cost savings and customer satisfaction.

This represents a shift from optimization to a fundamentally different service model: autonomous, real-time, and continuously improving.

Ready to orchestrate your AI agents?

CLU helps enterprises deploy autonomous AI agents that deliver measurable results in production—not just in demos.

Book a live demo at cluagents.com

Disclaimer: *Results are based on actual production data from a specific deployment (August 2024 – January 2026). Individual outcomes may vary depending on industry, operational complexity, data quality, and integration scope. Cost comparisons use the client's pre-deployment call center unit economics as the baseline. Client identity anonymized at their request.*