Case study · IT services · KeyDelta engagement, AI delivery lead and operator advisor · 6 months

MULTI-LOCATION IT SERVICES FIRM · THREE ACQUISITIONS IN 18 MONTHS · NAME WITHHELD

Resolution time from 4.2 hours to 1.6 hours in 6 months.

A post-acquisition IT services firm had tickets bouncing between teams with zero shared context. KeyDelta built an AI platform that pre-solves tickets before a human touches them, cutting resolution time 62% and escalations 44% in 6 months. The operating model came first: defined ownership, a daily quality cadence, and model abstraction that survived an acquisition without a rebuild. This is Operator-Built AI: built on a fixed operating model, not bolted onto chaos.

Situation

A multi-location IT services company, formed through three acquisitions in 18 months, was drowning in service tickets. Each legacy entity had its own ticketing system, its own triage process, and its own tribal knowledge. Reps were spending more time figuring out where a ticket should go than solving it.

- Three acquired teams with three ticketing systems, three triage processes, and zero shared knowledge
- Reps spending 40%+ of their time on ticket routing and context gathering, not problem-solving
- Escalation rates climbing: tickets bounced between teams because nobody had full context
- No model flexibility: locked into a single AI vendor with rising costs and inconsistent quality
- ServiceNow and Jira running in parallel with no integration and duplicate work everywhere

Approach

Before any model shipped, KeyDelta fixed the operating model the AI would run on: who owns hydration quality, a daily ticket-quality cadence, and a defined scope for every agent in the swarm. Only then did we build the AI agent platform that intercepts every incoming ticket, assembles context from multiple systems, and proposes a solution before routing to the right human.

01 **Platform architecture**

Built on Python Django with a LiteLLM proxy for model abstraction. The platform hot-swaps between GPT-4, Claude, and other models without code changes, optimizing for cost, speed, or quality depending on ticket complexity.

02 **System integration layer**

MCP (Model Context Protocol) integration connected the agent swarm to internal knowledge bases, customer records, and configuration data. ServiceNow and Jira integrations unified the fragmented ticketing landscape into a single workflow.

03 **Agent swarm and hydration engine**

Specialized agents executed in a parallel fan-out pattern: context agent, knowledge agent, and similarity agent ran concurrently, then a resolver agent synthesized their outputs into a proposed solution. A confidence score determined auto-route vs. human review.

04 **Scaling and optimization**

AWS Lambda functions handled compute-intensive analysis tasks serverlessly. Continuous learning from every resolution improved hydration accuracy over time. The platform got measurably smarter each month.

Vision	Every ticket arrives at a human already diagnosed, contextualized, and half-solved. Reps close issues instead of chasing them.
Outcomes	Resolution time, escalation rate, and hydration accuracy measured from day one. Model costs tracked per ticket to prove ROI in real time.
Ownership	Service delivery owned hydration quality. Each agent in the swarm had a defined scope: no overlap, no gaps, no committees.
Cadence	Daily ticket-quality reviews and weekly model performance analysis. Bad hydrations were caught in hours, not weeks.
Systems	New agents deploy in days. Model abstraction means no vendor lock-in. The platform survived an acquisition without re-architecting.

Stack

Python Django · LiteLLM proxy · MCP (Model Context Protocol) · ServiceNow · Jira

AWS Lambda

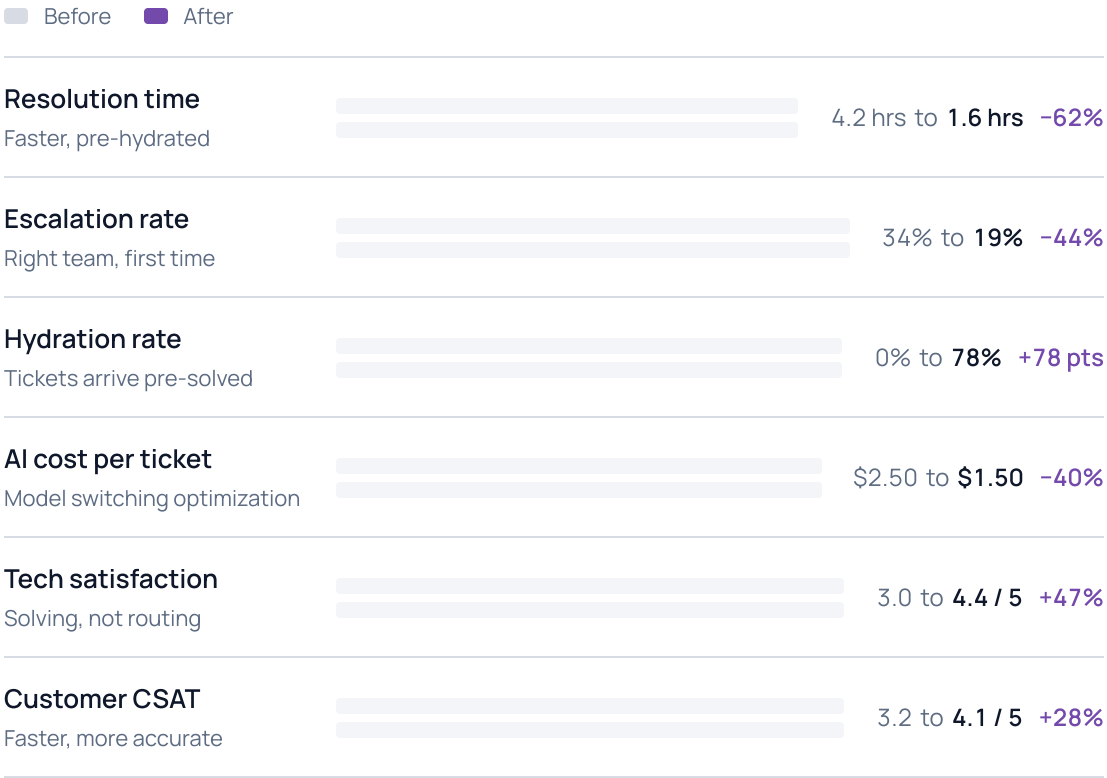
Outcome

Over 6 months, resolution time fell from 4.2 hours to 1.6 and the escalation rate from 34% to 19%, with 78% of tickets arriving pre-solved. AI cost per ticket dropped from $2.50 to $1.50 and customer CSAT rose from 3.2 to 4.1. Builds like this now ship under the KeyDelta Method and stay current under AAMP.

−62%
Resolution time
From 4.2 hours to 1.6 hours over 6 months

−44%
Escalation rate
From 34% to 19% over 6 months

78%
Tickets arrive pre-solved
From 0% hydration, measured at month 6



Each row scales to its own range. Grey is the starting point, Violet is where it landed. Exact values in the table below.

MEASURE	BEFORE	AFTER	CHANGE
Resolution time. Faster, pre-hydrated	4.2 hrs	1.6 hrs	−62%
Escalation rate. Right team, first time	34%	19%	−44%
Hydration rate. Tickets arrive pre-solved	0%	78%	+78 pts
AI cost per ticket. Model switching optimization	$2.50	$1.50	−40%
Tech satisfaction. Solving, not routing	3.0	4.4 / 5	+47%
Customer CSAT. Faster, more accurate	3.2	4.1 / 5	+28%

"We turned our service desk from a routing nightmare into an AI-powered resolution engine. Tickets that used to bounce between three teams now arrive pre-solved. Customer satisfaction jumped from 3.2 to 4.1. Our techs went from frustrated investigators to confident closers. And AI cost per ticket dropped 40% because we could swap models based on complexity instead of paying premium prices for password resets."

KeyDelta operating lead

$125M+ revenue impact, delivered to clients across 30+ transformations. Measurable results by day 30, typical across embedded engagements. 3.8 to 5.1x AI ROI in 6 to 9 months, delivered to clients. Client anonymized; details shared with written approval.

Book a 30-Minute Call

Builds like this ship through our AI agent development and stay current under our managed AI services.