$172,000

ANNUAL OPEX RECOVERED

INDUSTRY	MODELS
ENTERPRISE E-COMMERCE	CLAUDE 3.5 SONNET + PINECONE
TIMELINE	STATUS
22 DAYS	OPERATIONAL — PHASE II SCALING

Autonomous Customer Support Routing & Resolution

Deployed an autonomous RAG support agent for a global e-commerce brand. Replaced three full-time tier-one manual support roles with a deterministic LangGraph pipeline.

The Baseline Inefficiency

A global e-commerce brand processing 3,400 inbound support tickets monthly was operating with an 8.4-hour mean resolution time. Tier-one agents were spending 70% of their billable hours manually retrieving shipping status and return policies from disjointed internal databases. The projected Q3 hiring plan required an additional $172,000 in support headcount just to maintain current SLAs.

The Architectural Solution

We mapped their Zendesk taxonomy and deployed a customized LangGraph orchestration pipeline. Support queries are routed through a Pinecone serverless vector database containing 4,000+ internal operational documents. Anthropic's Claude 3.5 Sonnet generates the response, locked by strict constitutional guardrails to prevent hallucination. Complex anomalies trigger a deterministic fallback, routing the parsed data to a human escalation node via Make.com.

The Fiscal Outcome

The architecture was moved to production in 22 days. The automated pipeline achieved an 81% autonomous resolution rate within the first month. The 8.4-hour manual resolution latency dropped to 22 seconds. The client canceled the $172,000 Q3 headcount expansion and moved to our Phase II operational retainer.

Quantifiable Outcomes

ESCALATION DROP

-81%

Total reduction in tickets requiring human intervention.

OPEX RECOVERY

$172,000

Annualized physician hours recovered via automated triage processing.