Customer Testimonial

Agent AI: From Concept to Production in 1 month.

A STEP-BY-STEP CASE STUDY

Are you developing your own AI model or agent? Need a partner to help validate, monitor, and remediate issues?

About Us

STARSEER: VALIDATE, MONITOR, & REMEDIATE AI

Starseer helps teams build, validate, monitor, and remediate AI models and agents pre and post-production.

The platform provides deep runtime insight into how AI systems behave in real environments—capturing performance, drift, decision paths, and failure signals across cloud and edge deployments. Starseer enables teams to validate AI before release, continuously monitor behavior at runtime, and respond quickly when incidents occur.

With built-in assurance and remediation workflows, Starseer gives engineering teams the control and confidence needed to deploy AI systems safely, reliably, and at scale.

CUSTOMER CASE STUDY

Agentic AI: From Concept to Production in 1 Month

How Starseer delivered an offline-capable, enterprise-grade AI agent that replaced frontier model dependencies with a deployable solution that fits on a laptop.

1	100%	90%+	100%
Month to Production	**Offline Capable**	**Cost Reduction**	**Control & Validated**

Executive Summary

Businesses are being forced to innovate faster than ever in the AI age. Whether it is a meeting note taker, Gemini drafting emails, or integrating Claude Code-like assistants for software development, every team across the business is expected to figure out how AI workflows can increase productivity and efficiency.

That charge hits significant obstacles when businesses look to integrate GenAI within flagship products that generate millions of dollars in revenue. The news is filled with headlines highlighting the challenges of product teams that rushed GenAI into products too early, resulting in lost trust and revenue. Product and security teams that process or handle sensitive data that can’t, won't, or shouldn't be shared with frontier APIs may see innovation as out of reach.

This is the story of how Starseer scoped, built, and shipped a custom private agentic solution that moved a customer from frontier model providers to an upgradable agent that fits on a laptop in 1 month.

Rapidly developing, validating, and deploying AI agents - safely.

1. THE CHALLENGE

From the initial engagement, the customer's product team emphasized the importance of speed: We need this new feature set to be demonstrable within 1 month. Starseer's team was ready to meet the challenge.

Like many goals of agentic workflows, the task is paradoxically simple: We want our customers to be able to describe goals in natural language. These natural language goals are provided to an agent that leverages a database of curated knowledge to generate tasks and plans grounded against real data. The end result is increased customer satisfaction as less time is spent coming up with details and more time reviewing and executing them.

2. CRITICAL CONSTRAINTS

- **Offline Operation:** The agentic solution must work completely offline with no external API dependencies.
- **Geographic Restrictions:** Model selection limited to specific geographic origins due to regulatory and security requirements.
- **Enterprise Deployment:** Solution must be containerized and deployable with documented cluster requirements.

These types of restrictions are a natural fit for Starseer's platform, which has maintained full offline capability from the very beginning. Our team is deeply familiar with air-gapped and restricted operating environments.

3. THE SOLUTION

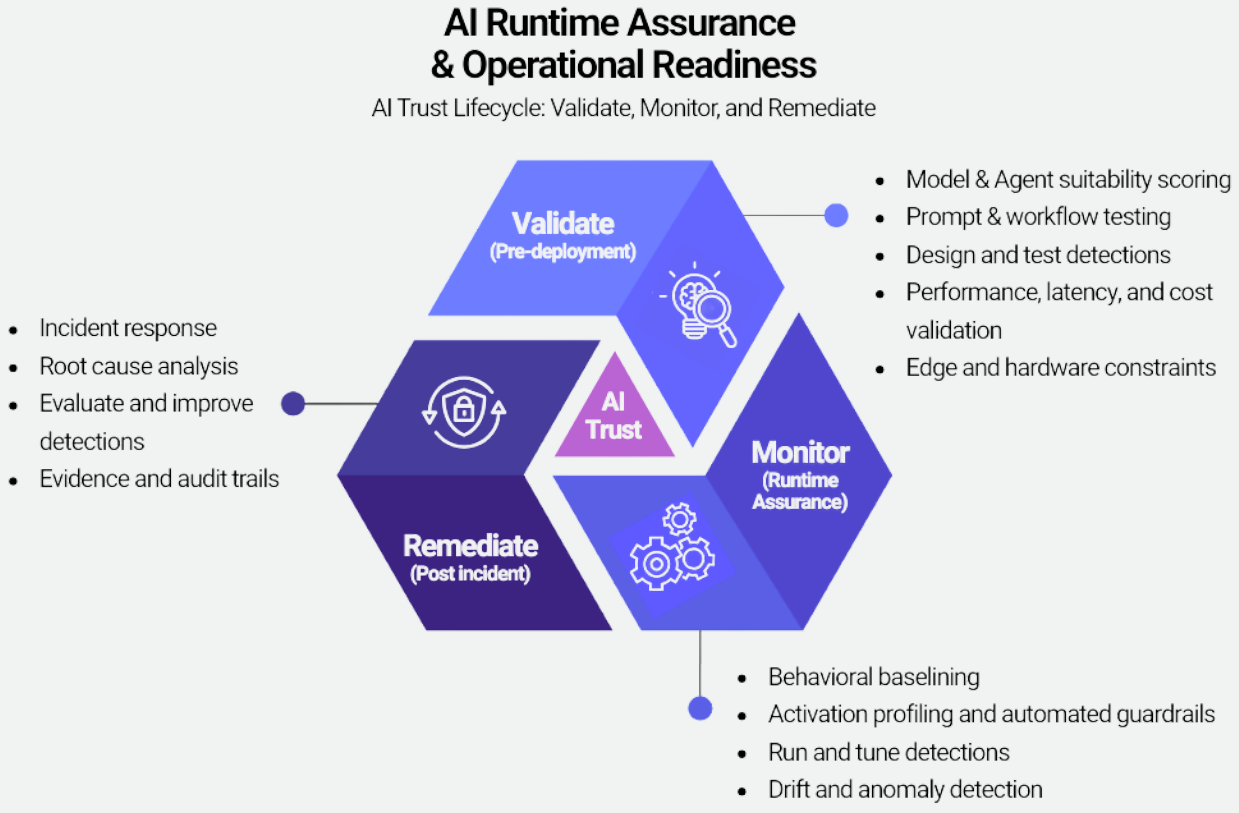
Phase 1: Scoping and Model Selection
Given the immediate timeline, Starseer's team engaged in focused discussions to ensure expectations were aligned on the first deliverable. An end-to-end demonstration of the agentic workflow was the priority, with enhanced reliability and edge case handling to be addressed in subsequent phases.

The customer had already attempted initial efforts with popular frontier models that weren't delivering results. Starseer's team analyzed the failure modes and recommended an alternative approach: IBM's Granite model family, an under-the-radar but high-performing option that met the geographic requirements while delivering superior results for this use case.

Phase 2: RAG Pipeline Architecture

Starseer designed and implemented a Retrieval-Augmented Generation (RAG) pipeline that enables the AI agent to anchor its outputs against the customer's proprietary knowledge base. This architecture ensures that generated outputs are contextually relevant, accurate, and aligned with existing organizational standards.

Component	Purpose
Knowledge Base	Structured indexing of customer's proprietary content library with semantic search capabilities
Vector Database	High-performance vector storage for similarity search and retrieval
Embedding Model	Compact, high-performance embedding model for semantic representations
Retrieval Stage	Semantic search across indexed content to surface relevant context



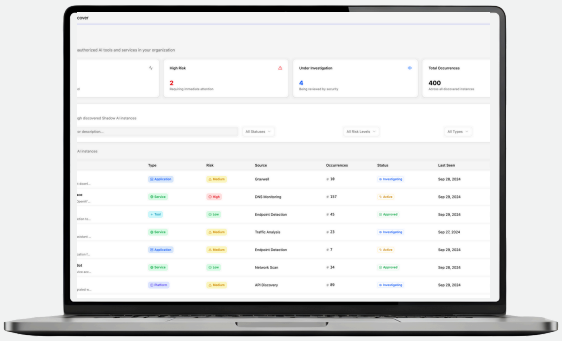
Phase 3: Agentic Workflow Pipeline
The core of the solution is a four-stage agentic pipeline that transforms natural language objectives into validated, production-ready outputs:

Stage 1	**Retrieval:** Query the vector database for relevant context from the knowledge base. Embeds the natural language objective and performs semantic search to retrieve similar examples, phase templates, and implementation patterns.
Stage 2	**Plan:** Generate a structured execution plan using the LLM with RAG context. Constructs a prompt with the objective, target parameters, and retrieved context to produce a structured execution plan.
Stage 3	**Build:** Convert the execution plan to validated, schema-compliant output. Deterministic transformation that maps the plan structure to the required schema with proper compliance.
Stage 4	**Validate:** Ensure generated output meets schema requirements. Includes schema compliance, syntax validation, and automatic retry logic with error feedback to the planner.

The validation stage includes intelligent retry logic: if validation fails, the workflow returns to the Plan stage with error context, allowing the agent to self-correct. This approach dramatically improves reliability without human intervention.

4. THE DELIVERABLES

Starseer delivered a complete, production-ready solution within the 30-day timeline:



Containerized Agentic Workflow
Natural language objective input
Foundational RAG system referenced for output generation
Validated structured output returned via documented API

Enterprise Deployment Package
Deployment architecture for containerized or microservice environments
Documented cluster requirements and configuration

Documentation and Future-Proofing
Complete API documentation for the agentic workflow
Model evaluation rubric for future upgrades as new models become available
Full operational control: upgrades happen on the customer's schedule, not dictated by frontier model vendor changes or deprecations

5. THE RESULTS

90%+ Cost Reduction Eliminated recurring frontier API costs. The solution runs entirely on local infrastructure with no per-token charges.	**100% Offline Capable** Complete air-gap compatibility. No external network dependencies required for operation.
100% Control Upgrades on the customer's schedule, not dictated by frontier model vendor changes or deprecations.	**Future-Proof Architecture** Model evaluation rubric enables systematic upgrades as the rapidly evolving AI landscape produces better options.

Why Starseer

Starseer brings a unique combination of capabilities to enterprise AI deployments:

- **Depth of Visibility:** Starseer's platform technology opens the black box, delivering deeper visibility into AI models and agents than previously available. By providing granular insights into model behavior, we enable rapid root cause analysis and fine-tuning—whether delivered by our team or empowered for yours.
- **Deep Model Expertise:** Our team understands not just how to use AI models, but how they work internally. This enables us to select the right model for each use case and optimize performance beyond what generic implementations achieve.
- **Offline-First Architecture:** Starseer's platform was built from day one to operate in restricted environments. We don't retrofit solutions for air-gapped deployments; it's our native operating mode.
- **Rapid Delivery:** We scope aggressively and deliver incrementally. One month to a working proof of concept is our standard, not an exception.
- **Security-First Mindset:** With roots in FFRDC and national security environments, Starseer brings deep expertise in deploying and securing AI systems for the most critical commercial and public sector enterprises. We understand the security and compliance requirements that enterprise customers face—and we build solutions that work within those constraints, not around them.

Ready to accelerate your AI initiative? Contact Starseer to discuss how we can deliver production-ready AI solutions for your organization