

tsia

STAR Awards

2026

2026 TSIA STAR Awards
Featured Application

Fewer Tokens, More Resolutions: Inbenta Encore is Low-Compute by Design

I. Overview

CX buyers shouldn't have to optimize AI costs by rationing tokens or doing less. They should be able to optimize by needing fewer tokens — and doing more with them. Most of the market offers one of the first two: negotiate the rate or downgrade the model, cap the volume, and narrow the scope. Inbenta Encore takes the third path: **low-compute by design**. Encore uses proprietary **Knowledge Engineering** to structure enterprise content into approved intents once, at ingestion. At runtime, Encore delivers the answer through deterministic retrieval, not generation. GenAI is used to surface intent gaps and build on the knowledge base, not to author live answers each time a customer asks. As a result, Encore uses fewer tokens per resolution, with no ceiling on how many resolutions the business goes after, and a unit price it can plan on. **Less compute, more results.**

II. What We Did

Enterprise CX teams have had two bad options: generative AI that ships fast but bills more with each new adoption, or hand-built knowledge that's forecastable but takes quarters and headcount to stand up. Either way, a CFO eventually stops underwriting the next phase, and the channel the business funded to grow gets quietly scaled back. Gartner expects over 40% of agentic AI projects to be cancelled by 2027 on cost and unclear value. The standard response is to manage consumption: dashboards, rate negotiations, model downgrades, spend caps — in other words, buying tokens cheaper or doing less with them. **Encore removes this tradeoff. It engineers value at the knowledge layer, so higher volume doesn't mean higher cost**.

The Cost Problem We Solve

Customer Pain Point	How Encore's Knowledge Engineering Solves It
A bill you can't put in a plan. Consumption depends on how badly retrieval performed and how often the customer rephrased the question.	**Governed, repeatable consumption.** The same question matches the same intents and consumes the same work for a consistent unit cost the business can plan around.
Spend that grows with success, at a price you don't set. Each conversation re-pays for retrieval and reasoning, so the only levers left are a cheaper rate or less usage.	**Engineering at ingestion.** Semantic work happens once per source rather than once per query, and a verified intent match needs no generative call, so volume growth doesn't cost more.
Maintenance that scales with the knowledge base. Full reprocessing costs so much that refreshes get skipped, and stale knowledge is what produces escalations.	**Differential maintenance.** Encore tracks which source each intent came from. When a source changes, Encore compares old and new versions, reprocesses only the intents affected by that change, and flags material changes for expert review. Maintenance work drops by more than half.
Setup priced in quarters and headcount. Hand-built knowledge graphs need linguists working at six to seven intents per page.	**Automated intent extraction.** One head-to-head build on identical content took four hours and about $300 in tokens, against four FTEs for two months.

III. Where Generation Belongs

The costliest habit in contact center AI is paying a model to compose what the business already knows. It persists because the market treats an LLM and generative AI as one thing. A large language model is a capability — it reads meaning, classifies, extracts, summarizes, translates and rewrites. Generation is one way to use that capability: asking the model to compose new content in the moment, out of whatever context it was handed. That mode carries the highest price and the least-verifiable output, yet the industry has unfortunately defaulted to it. **Encore doesn't**. It reserves generation for two things: building knowledge once and checking it before it's reused. That's why it can scale affordably— adding a new question, channel, or language means adding more structured knowledge, not more generative calls at runtime.

Where generation earns its keep

- **Building the knowledge, once.** A large-context model reads connected sources and extracts question-and-answer intents — each a small graph of phrasings and embeddings — tagged to its source. A smaller model validates each intent, and a subject matter expert approves it. Roughly 30 intents per page (vs. six to seven built by hand).
- **Reading the question.** The number of real questions a business gets is finite; the ways customers phrase them are not. A verbose question is reduced to its semantic core before matching, so surrounding words don't pull wrong answers.

- **Composing across intents.** A compound question is answered by assembling several approved intents rather than by widening the retrieval window. Where one part has no answer, Encore resolves what it can and escalates only the remainder.
- **Phrasing, and finding what's missing.** The retrieved answer is rendered conversationally in the customer's language, while Encore's self-healing agent, Elevate, scans live conversations for specific, closeable gaps.

Where it doesn't belong

- **Authoring the customer-facing answer live.** Businesses pay full price to re-derive what the business already knows. Where nothing matches, Encore says so and routes the gap to an expert.
- **Adjudicating contradictions, or running steps with one correct path.** Conflicting sources surface at ingestion and go to the person who knows which is right, resolved once. A deterministic route costs no inference and doesn't vary between contacts.

IV. The Value Encore Delivers

Cost to serve

- **~40% lower cost per interaction** than the RAG-first architectures we displace, head-to-head on the same content.
- **Model usage about one-third** that of a comparable generative exchange.
- **Over 50% less** ongoing knowledge maintenance, and roughly **90% less** knowledge engineering workload.
- Ingestion in **30 to 60 minutes across 800+ integrations**, with no content migration. Case study: 27 sources produced over 4,500 approved intents on a Tier-1 bank's content, against a market norm of three to six months

Service quality

- **98% answer accuracy, 35% better first-contact resolution, 30% better CSAT.**
- Full traceability on any answer: which intents matched, and which sources they were built from.
- Client example: One airline customer runs over **10 million queries a year** on this architecture.

Case study: **A global genealogy organization** came to us with first-contact resolution near 65%, overflow into email and phone, and significant abandonment — having already paid for an expert-built knowledge graph. We structured their content and outperformed that graph on the questions their users actually ask. Service improved, and one multilingual knowledge base replaced their language-specific databases.

Testimonials

"We deployed Encore earlier this year, and it has transformed Mission Regan. We've more than doubled the number of people we're able to serve. What amazed me most is what it didn't cost. We got enterprise-grade capability at a fraction of the price, deployed in a matter of 3 to 4 weeks, and we haven't had to backfill several positions because so much of our customer experience is now automated." — Josh Willis, Founder/Director, **Mission Regan**

"Inbenta's chatbot now handles on average 91% of chat inquiries on its own, with less than 10% requiring escalation to a live agent. It's been a reliable, cost-effective solution that has let us scale guest support without scaling headcounts." — Vanessa Day, Manager Digital Content, **Alterra Mountain Company**

V. The Commercial Model

- **Priced per interaction, not per token.** The customer buys a unit cost they can forecast, insulated from token-price movements, model-tier changes and their own traffic mix
- **Priced on resolution, not deflection.** On voice deployments, the charge attaches to contacts that end resolved, not ones that are merely deflected and end up costing not just tokens but human hours when a query is escalated.
- **Model-agnostic by design.** Encore routes to whichever commercial model clears the accuracy bar, re-benchmarked as prices and capabilities move — and the structured knowledge belongs to the customer if they leave.

VI. Execution and Scale

Most AI business cases quietly omit the specialist team required to keep the thing running, which is where the projected savings go. Encore's administrators are subject matter experts — the people who already own the content — not prompt engineers or developers. Reviewing and approving intents is one person's work, not a department's. Adding a language, channel or business unit carries near-zero marginal cost and no services engagement. New business units can stand up their own assistants without IT. **Optimize by needing fewer tokens — then do more with them.**


tsia
STAR Awards
2026