

2025 TSIA STAR Awards Featured Application



Inbenta Encore: Revolutionizing Enterprise AI with Programmed Intelligence

I. Overview

Inbenta Encore is the enterprise AI orchestration platform that solves the fundamental trade-off between AI convenience and enterprise control. Built on breakthrough **Programmed Intelligence** technology, Encore uses a revolutionary dual-LLM architecture where AI agents prepare content and facilitate natural conversations, while traditional retrieval ensures every response is pre-approved and never deviates. Unlike traditional AI tools that operate in silos or require heavy lifting to scale, Encore turns fragmented automation into intelligent, governed, and repeatable performance. It's not just AI that answers — it's AI that performs exactly as programmed. "Teach it once. Trust it always."

II. What We Did-We observed enterprises facing an impossible choice with AI assistants: choose fast deployment with unpredictable responses (GenAI/RAG) or choose controlled responses with months of manual setup (traditional NLP). Previous attempts to solve these pain points were either unreliable or prohibitively expensive to implement and maintain. To address these limitations, we built Inbenta Encore — the first **Programmed Intelligence** platform that eliminates this false choice through a breakthrough hybrid approach.

The Enterprise AI Dilemma We Solve

Customer Pain Point	How Encore's Programmed Intelligence Solves It
Unpredictable Al Responses GenAl systems generate	Dual-LLM Architecture: LLMs handle conversation and content
inconsistent answers, create compliance risks, and	preparation, but only pre-approved responses are delivered via
produce hallucinations	traditional retrieval — zero content generation in customer interactions
Slow, Expensive Setup Traditional NLP requires	Agent-Powered Content Preparation: LLMs autonomously extract and
months of manual intent creation and costly expert	optimize intents from data sources, reducing setup time from months
configuration	to days
Loss of Enterprise Control GenAl systems can't	Approval-First Responses: Every answer requires explicit approval
guarantee response consistency or maintain audit trails	before deployment and will never deviate without permission
Natural Interaction vs. Precision Trade-off Users	Conversational Precision: Users interact naturally through LLM-
want conversational AI, but enterprises need	powered conversation, but always receive exact, approved responses
guaranteed accuracy	with audit trail for compliance & regulatory requirements
High Maintenance Overhead Traditional systems	Intelligent Maintenance: Agentic AI continuously analyzes
require constant manual updates and intent	performance and recommends improvements, reducing maintenance
management	burden

How It Works: The Encore Intelligence Pipeline

Backend: LLM-Powered Content Preparation

- 1. **Agent-Powered Intent Extraction**: LLMs analyze your data sources and autonomously extract potential intents (question-answer pairs)
- 2. **Intelligent Content Preparation**: Our agents don't just extract they optimize content specifically for retrieval performance in real-time interactions
- 3. Validation & Approval Gateway: Every intent goes through automated validation and human approval before going live

Frontend: LLM-Powered Intent Recognition

- 4. **Conversational Intent Discovery:** When users interact with your assistant, LLMs engage in natural conversation to understand exactly what they're looking for
- 5. **Traditional Retrieval**: Once intent is identified, the system retrieves the exact approved answer using proven matching technology no content generation or modification
- 6. **Continuous Optimization**: Agentic Al monitors performance and recommends improvements to your intent base over time

Ill. B2B Customer Impact

Performance Improvements:

- ==75+% faster== setup and deployment compared to traditional NLP solutions
- =50+% reduction== in overhead costs through precise automation

- ==32% improvement== in customer satisfaction scores industry benchmarks
- ==35+% increase== in first-contact resolution rates of industry benchmarks
- ==Zero hallucinations== or unapproved responses in production

Customer Testimonials: Family Search, Largest genealogy organization in the world.

"With Encore, we finally have AI that feels natural to our customers but gives us complete control over every response. Encore solved our biggest AI challenge: how to scale intelligent automation without losing control." Marc

Results: Reduced abandonment rate, 60-70% productivity enhancement with deep research tools for content building and filtering, ability to manage a large amount of multilingual data with single master RAG and intelligent prompts vs. multiple language-specific databases and 95%+ Accuracy for first contact resolution rates.

IV. Core Differentiators

Dual-LLM Architecture- Revolutionary approach using LLMs as intelligent	Conversational Precision- Users interact
tools, never as content generators:	naturally through conversation, but always
Backend LLMs: Extract and optimize intents from your data sources	receive exact, approved responses:
Frontend LLMs: Converse with users to identify their exact needs	LLM-powered conversation to understand user
Traditional Retrieval: Deliver approved answers without modification or	intent
generation	Traditional retrieval of pre-approved answers
Zero Content Generation: LLMs never create or alter customer-facing	Natural interaction without sacrificing control
responses	or accuracy
responses Self-Setup Experience- Configure Al assistants for any use case without	or accuracy Intelligent Maintenance
<u> </u>	,
Self-Setup Experience- Configure Al assistants for any use case without	Intelligent Maintenance
Self-Setup Experience- Configure Al assistants for any use case without extensive technical implementation:	Intelligent Maintenance Agentic AI continuously analyzes performance
Self-Setup Experience- Configure Al assistants for any use case without extensive technical implementation: Point to your data sources	Intelligent Maintenance Agentic AI continuously analyzes performance and recommends improvements, reducing the

V. Key Innovations

Technical Innovation: First dual-LLM architecture using AI for preparation and recognition while maintaining traditional retrieval for responses

Business Impact: Solves the fundamental trade-off between natural conversation and enterprise control

Market Disruption: Creates a new category between traditional NLP and generative Al

Practical Application: Real-world deployments showing measurable improvements in both setup speed and response reliability

VI. The Programmed Intelligence Advantage

Setup Like GenAl, Perform Like Traditional NLP, Interact Like Modern Al

- Coverage: Extract comprehensive question sets from your data at scale
- Speed: Go from data source to deployed assistant in days, not months
- Control: Every response requires approval and won't deviate without permission
- Natural Interaction: Users converse naturally while getting precise, approved answers
- Reliability: No hallucinations, no surprise responses, no compliance risks
- Efficiency: Automated intent generation with human governance

VII. Why "Encore" Works

The name captures our position perfectly: this isn't our first performance in enterprise AI - we've been delivering reliable solutions for 20 years. Encore represents our next act, where we've learned from both the promises and pitfalls of generative AI to deliver something better. "The perfect blend of convenience and control."

VIII. Next Steps

Inbenta is committed to expanding Programmed Intelligence across every aspect of enterprise AI orchestration. We're enhancing our agentic capabilities for even more intelligent content preparation and performance optimization, while maintaining our core principle: AI that performs exactly as programmed, every time.

Inbenta Encore: Where Enterprise AI Finally Delivers on Both Promise and Programmed Intelligence



For more information, visit the TSIA STAR Awards Winners Page