How Aible Flips the Script on Enterprise GenAI

AIBLE

GenAI Estimate Vs. Aible at Fortune 50s

	Others	Aible
TIME TO DEPLOY	8.2 Mths	< 30 Days
COST PER USE CASE	$2M-16M	< $500K
SUCCESS RATE	10% Pilots to Production	Guaranteed

Source: Leading Research Analyst Firm

Top 3 Reasons Why GenAI Projects Fail

Misalignment with Business

- Data Science evaluations of GenAI solutions ≠ business user preferences
- A lot of work is done by data science and IT before business users ever get to confirm whether the AI is useful to them

Speed & Cost of Building Enterprise Solutions

- A lot of effort is spent on failed projects (90% of AI projects fail as per Gartner)
- In GenAI, by the time a project is completed (8 months per Gartner), the tech is out of date

Enterprise Guardrails

- Can't lose control over data (LLMs require data egress)
- Can't make the wrong decisions due to hallucinations
- Can't 'leak' information to unauthorized users through model fine-tuning

Traditional Data Science Process

WITHOUT AIBLE

1 USE CASE IDENTIFICATION

First, **Data Scientists** work with **Business Users** to understand / define the use case. Often business users are asked to provide sample questions and 'golden answers' that Data Scientists can use to evaluate the project - before business users even have a full understanding of the proposed solution.

2 USE CASE

Second, **Data Scientists** look for the data to enable the use case. They look at the sensitivity of the data to make a case for AI governance.

3 GOVERNANCE

Third, **Data Scientists** work with AI governance to figure out enterprise controls on a case-by-case basis even as the underlying technology is rapidly changing.

4 CONFIGURATION

Fourth, **Data Scientists** try different models, model settings, VectorDB settings, prompt augmentation, few-shot learning, even fine-tuning of Specialized/Small Language Models (SLM), etc. They use the sample questions and answers from step 1 to evaluate the different configurations. Unfortunately, they can only check for how well a response matched the words in the 'golden answer' and not whether the answer was useful or accurate.

5 PROTOTYPE

Fifth, **Data Scientists** create one or two prototypes at best and show the solutions to Business Users.

6 DEPLOY

Sixth, **Data Scientists** work with IT teams to get the GenAI solutions deployed at scale.

7 RE-DO /REPEAT

Finally, as the underlying technology changes, **Data Scientists** either put in significant effort to re-do the work or just don't make any changes - thus falling further and further behind the state of the art.

90% OF PROJECTS FAIL

Fundamentally, **this process places humans at the beginning, middle and end** of the process. This would be appropriate in a world where there are very few opportunities to use AI to improve business performance. But you can't solve an exponential opportunity / problem by using a linear solution like people.

TIME TO IMPACT: 7 Days

> *That iterative process was maybe a one-week process, where we tried our foundation models, we tried some prompt augmentation in a couple of days. Is this going to work, no we need more accuracy, ok let me fine tune a Mistral or a smaller model and get more accuracy there…"*

verizon✓

– Michael Hanson, Sr Director, Business Transformation
https://qrco.de/bfOuq8 Gartner Data & Analytics Summit 2024

AI-First Scalable Approach

WITH AIBLE

Aible believes that there are many opportunities to deliver business impact from our data using AI and you need an AI-first scalable approach to extract value from your data. You need to fully automate the first few steps to sufficiently derisk the project before spending peoples' time on it. You need to engage all relevant stakeholders earlier in the process to ensure success.

TIME TO IMPACT: 2 Days

> *50% of companies are trying to bring GenAI capabilities in, but only 10% have actually implemented them. To just think that we did something in two days on top of our data, that's a pretty big deal."*

– Jas Pannu, SVP Data & Analytics

carelon

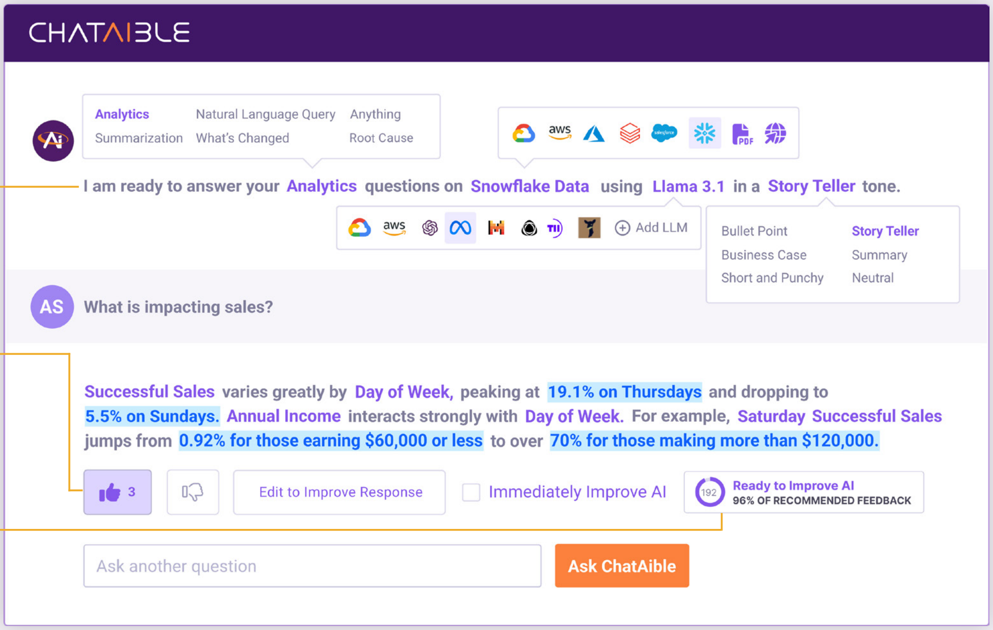
1 Business-User-Led Rapid Prototyping to Derisk Upfront

- Business users can create their own GenAI prototypes based on any kind of enterprise data in minutes
- End-users can immediately start providing feedback at scale that can be used to improve the GenAI models

Users can ask any kind of question on any kind of data, all in their cloud

Users can provide easy feedback or even edit the response of the SLM

One-click few-shot-learning and end-to-end automated fine-tuning



I am ready to answer your Analytics questions on Snowflake Data using Llama 3.1 in a Story Teller tone.

What is impacting sales?

Successful Sales varies greatly by Day of Week, peaking at 19.1% on Thursdays and dropping to 5.5% on Sundays. Annual Income interacts strongly with Day of Week. For example, Saturday Successful Sales jumps from 0.92% for those earning $60,000 or less to over 70% for those making more than $120,000.

2 Built-in Enterprise Guardrails to Secure By Design

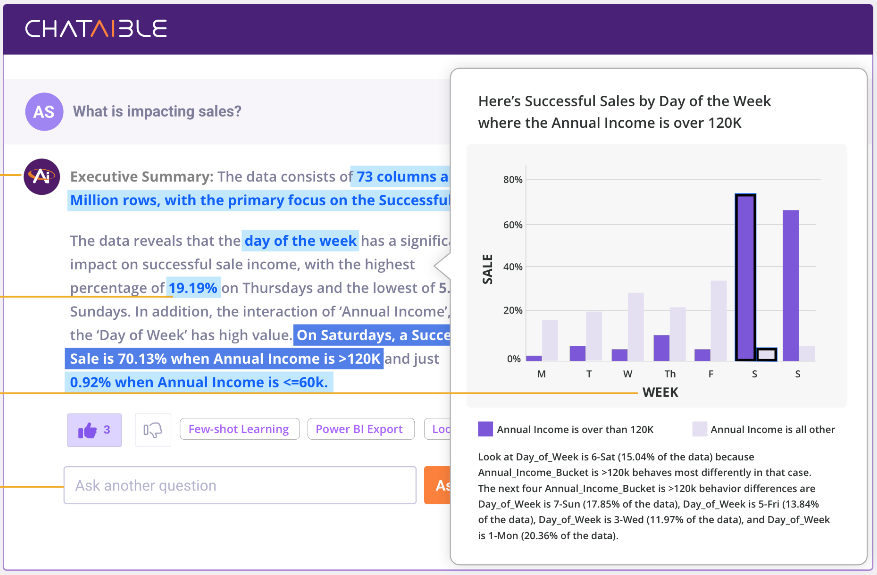
- Implemented fully in the customer's cloud (including models) with no data egress
- Deterministic double-checking for hallucinations that traces back to source documents
- Can segregate VectorDBs and fine-tuned models by role, use-case, user

Because GenAI is only as good as the questions users ask, ChatAible asks millions of questions automatically at very low cost and auto generates an executive summary as a starting point (no user question needed)

ChatAible double checks the LLM response: If it's blue, it's true (completely deterministic, not generative)

Users can triple-check the summary via tool tips showing relevant charts

Detects and reports Anomalous Prompts



CHATAIBLE

AS What is impacting sales?

Executive Summary: The data consists of **73 columns a**
Million rows, with the primary focus on the Successful

The data reveals that the **day of the week** has a signific
impact on successful sale income, with the highest
percentage of **19.19%** on Thursdays and the lowest of 5.
Sundays. In addition, the interaction of 'Annual Income',
the 'Day of Week' has high value. **On Saturdays, a Succe**
Sale is 70.13% when Annual Income is >120K and just
0.92% when Annual Income is <=60k.

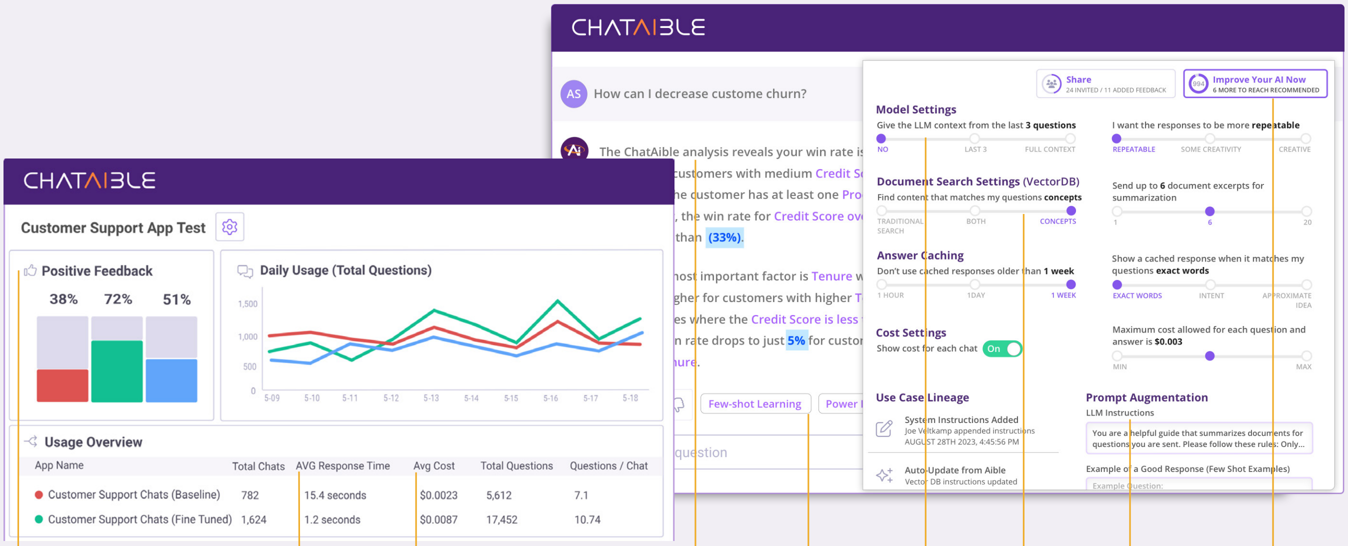
Here's Successful Sales by Day of the Week where the Annual Income is over 120K

■ Annual Income is over than 120K ▢ Annual Income is all other

Look at Day_of_Week is 6-Sat (15.04% of the data) because Annual_Income_Bucket is >120k behaves most differently in that case. The next four Annual_Income_Bucket is >120k behavior differences are Day_of_Week is 7-Sun (17.85% of the data), Day_of_Week is 5-Fri (13.84% of the data), Day_of_Week is 3-Wed (11.97% of the data), and Day_of_Week is 1-Mon (20.36% of the data).

3 Iterative Automated Improvement

- One-click few-shot learning and end-to-end automated fine-tuning (from user feedback collection to model deployment)
- Serverless A-Z testing for rapid experimentation in parallel. Understand cost & time savings
- Adjusts at scale without end-user disruption as new technology emerges



CHATAIBLE

Customer Support App Test

Usage Overview

App Name	Total Chats	AVG Response Time	Avg Cost	Total Questions	Questions / Chat
● Customer Support Chats (Baseline)	782	15.4 seconds	$0.0023	5,612	7.1
● Customer Support Chats (Fine Tuned)	1,624	1.2 seconds	$0.0087	17,452	10.74

CHATAIBLE

AS How can I decrease custome churn?

The ChatAible analysis reveals your win rate is
customers with medium Credit S
he customer has at least one Pro
, the win rate for Credit Score ov
than (33%).

most important factor is Tenure w
gher for customers with higher T
es where the Credit Score is less
n rate drops to just 5% for custor
ure.

Model Settings

Give the LLM context from the last **3 questions**: ● NO ○ LAST 3 ○ FULL CONTEXT

I want the responses to be more **repeatable**: ● REPEATABLE ○ SOME CREATIVITY ○ CREATIVE

Document Search Settings (VectorDB)

Find content that matches my questions **concepts**: ○ TRADITIONAL SEARCH ○ BOTH ● CONCEPTS

Send up to **6** document excerpts for summarization: ○ 1 ● 6 ○ 20

Answer Caching

Don't use cached responses older than **1 week**: ○ 1 HOUR ○ 1DAY ● 1 WEEK

Show a cached response when it matches my questions **exact words**: ● EXACT WORDS ○ INTENT ○ APPROXIMATE IDEA

Cost Settings

Show cost for each chat: On

Maximum cost allowed for each question and answer is **$0.003**: MIN ● MAX

Use Case Lineage

System Instructions Added
Joe Veltkamp appended instructions
AUGUST 28TH 2023, 4:45:56 PM

Auto-Update from Aible
Vector DB instructions updated

Prompt Augmentation

LLM Instructions

You are a helpful guide that summarizes documents for questions you are sent. Please follow these rules: Only...

Example of a Good Response (Few Shot Examples)

Example Question:

Feedback

Speed

Cost

Augmented Analytics

Few-shot Learning

Model Settings

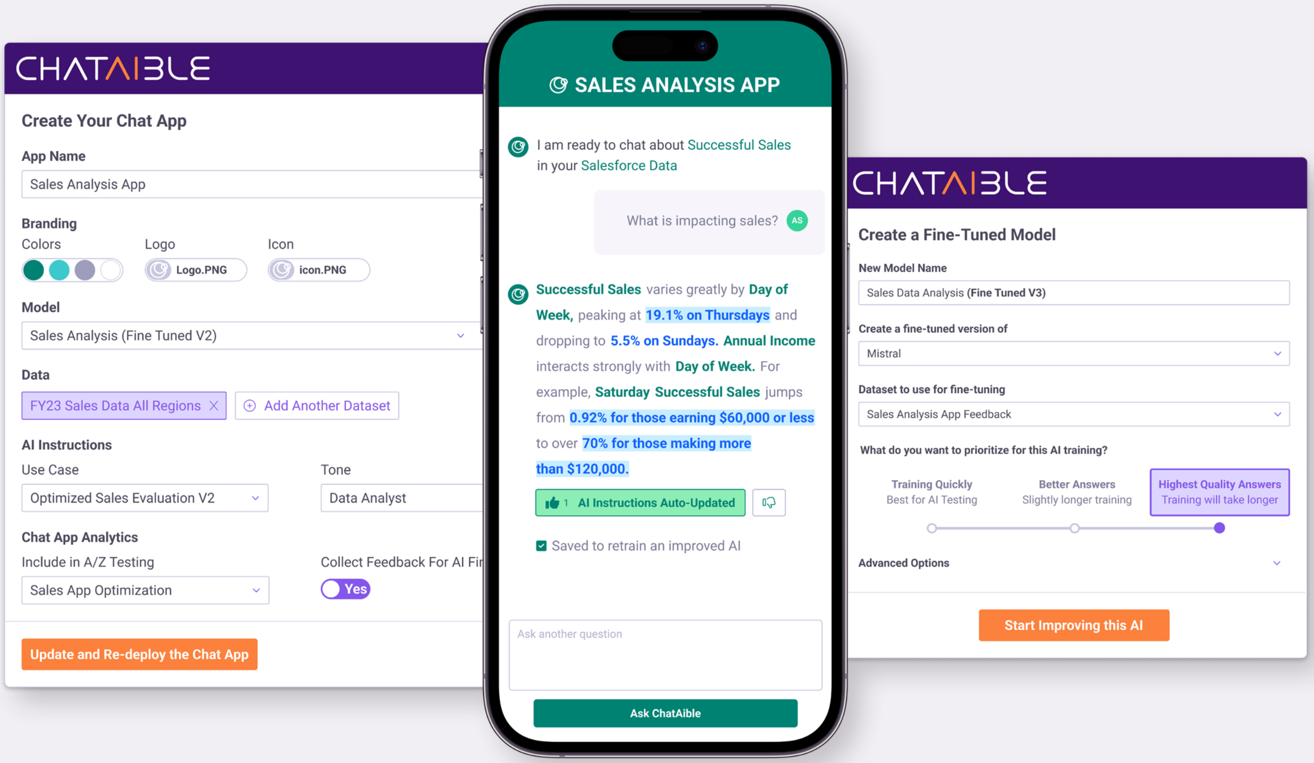
Vector DB

Prompt Engineering

Fine Tuning

4 Automated Deployment & Integration

- Leveraging serverless techniques, automatically deploys the VectorDB and models in the customer's own cloud in a scalable way without having to size the servers
- Complete automation that delivers all DevOps steps including creating and deploying necessary data processing/transformation code, setting up secure gateways, etc. in less than 10 minutes
- Includes out-of-the-box integrations for the model with platforms like Salesforce, Looker, PowerBI, Tableau, Boomi, etc. as well as any place where Google Maps can be integrated



Create Your Chat App

App Name
Sales Analysis App

Branding
Colors
Logo: Logo.PNG
Icon: icon.PNG

Model
Sales Analysis (Fine Tuned V2)

Data
FY23 Sales Data All Regions

AI Instructions
Use Case: Optimized Sales Evaluation V2
Tone: Data Analyst

Chat App Analytics
Include in A/Z Testing: Sales App Optimization
Collect Feedback For AI Fi: Yes

I am ready to chat about Successful Sales in your Salesforce Data

What is impacting sales?

Successful Sales varies greatly by Day of Week, peaking at 19.1% on Thursdays and dropping to 5.5% on Sundays. Annual Income interacts strongly with Day of Week. For example, Saturday Successful Sales jumps from 0.92% for those earning $60,000 or less to over 70% for those making more than $120,000.

☑ Saved to retrain an improved AI

Create a Fine-Tuned Model

New Model Name
Sales Data Analysis (Fine Tuned V3)

Create a fine-tuned version of
Mistral

Dataset to use for fine-tuning
Sales Analysis App Feedback

What do you want to prioritize for this AI training?

Training Quickly	Better Answers	Highest Quality Answers
Best for AI Testing	Slightly longer training	Training will take longer
○	○	●

Advanced Options

Final Consideration in Times of Rapid Change

GenAI technologies are changing rapidly. It makes no sense to obsess over which is the 'best' model, VectorDB, etc. because the best model will change constantly. You need to design your solution to be flexible. Aible is designed so that end-user experience can be improved without disruption as new technology changes. For example, when a new model arrives, the Aible Chat Templates need to be adjusted so that the prompt augmentation and default settings change based on the needs of the model. The user does not need to learn how to write new prompts for the new model. Aible translates their questions to the appropriate prompts for each model and each use case. The same flexibility that enables rapid iteration, enables rapid adjustment to change.

Learn How Aible Delivers Enterprise GenAI Projects in 30 Days