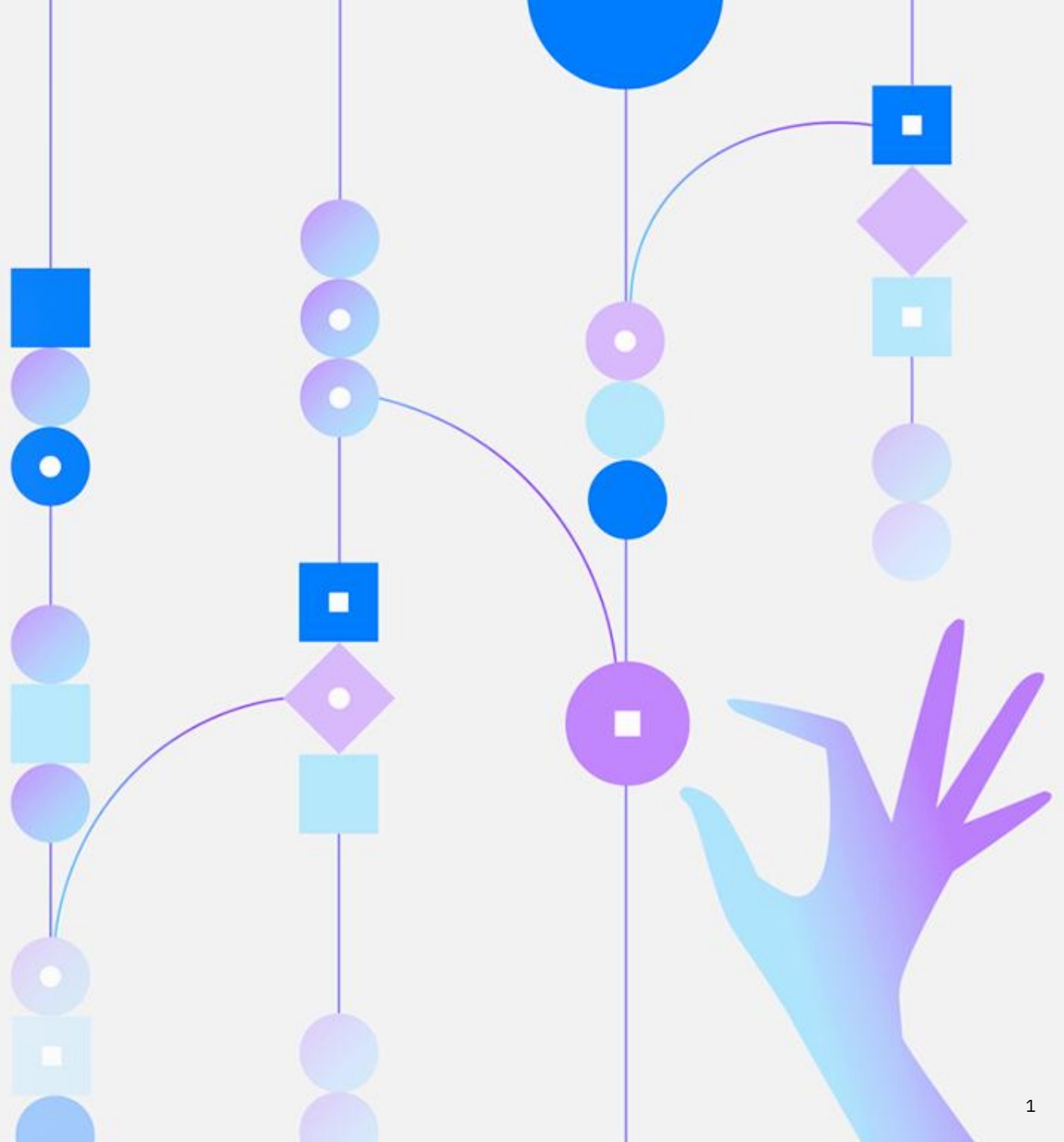
Self-Service-**AI**

Next gen AI chat for self service

Customer Service
HR
Procurement
Knowledge Search
...

IBM Consulting, July 2025



The case for change in Conversational AI

The Challenges | Hurdles in Self Service chat bots of yesterday

Rule Based Bot

Rule-based bots operate within fixed flow and frameworks (intent identification and dialog flow), leading to unmet expectations.

Difficulty in managing intricate scenarios

Conventional chatbots often falter with complex inquiries and necessitate human support. Not able to adapt during chat flow.

Challenges in scaling & maintenance:

The rigid structure of rule-based systems makes it challenging to adapt to new use cases. Also difficult to maintain as complexity grows.

The Impact | Lost Revenue, Poor Employee and Customer Experience

Deflection Rates are very low

Most conversational AI solutions do not deliver the expected containment

Subpar Customer Experience

A higher rate of incorrect responses, frequent follow-up inquiries, and a deficiency in human empathy contribute to user frustration.

Higher Cost of Operation

Higher dependency on technical team and hence increases the cost of operation

The Need and Opportunity |

With the use of Agentic AI and sophisticated engineering, IBM Consulting is able to provide a solution which solves the challenges, and delivery the efficiency and effectiveness to make a self service chat solution truly next generation.

1. **Length**
   Understands and adapts to wide range of queries, answers with higher confidence and accuracy
   Value: High Deflection and reduce operational cost

2. **Breadth**
   Built to scale instantly, across languages, processes and regions.
   Value: Lower cost of build/manage, faster time to value

3. **Depth**
   Remembers. Understands. Solves. Just like your best agent
   Value: Superior Customer Experience, better NPS

Self-Service-AI

The Agentic Conversational solution for complex Enterprise Workflows

The solution enables scalable, cost-efficient, and hyper-personalized conversation across every business function, channel, and language

App Overview

- Powers intelligent conversations across business functions like HR, Finance, IT, and Customer Service
- Understands complex queries by accessing structured and unstructured data across systems
- Executes tasks autonomously, integrates with enterprise tools and APIs to take action
- Reduces cost through hybrid model orchestration and efficient RAG pipelines
- Scales across channels, personas, and languages with configurable workflows
- Supports video, image, online search. Easily configure data sources and modes.

Use Cases

Customer Service Cross Industry

Knowledge Worker Agent

Capabilities

Language Agnostic Multi Sources Retrieval
Extracts relevant answers from enterprise data, documents, and APIs

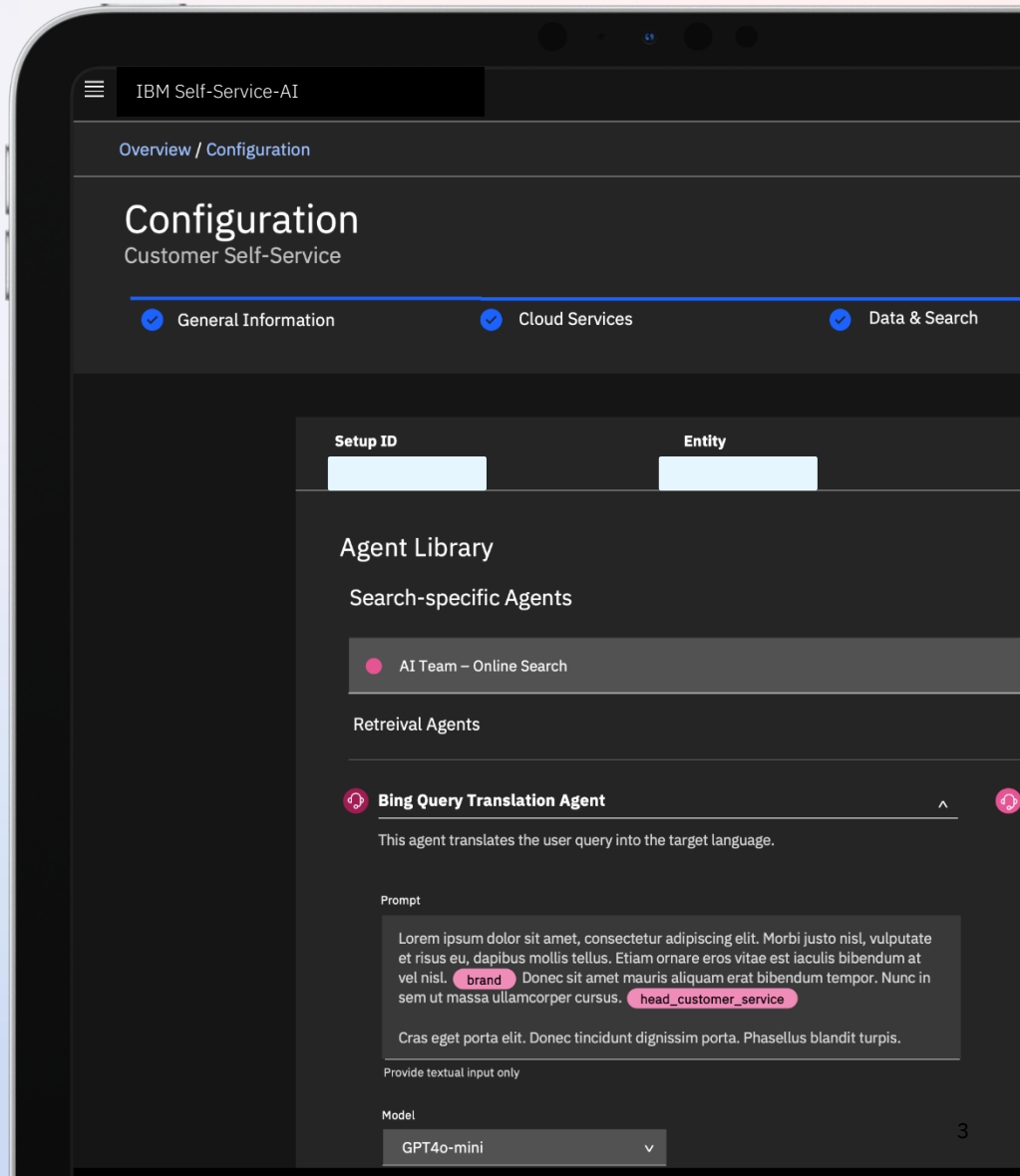
Muti-Modal Agentic RAG
Combines hybrid search, real-time queries, and cached knowledge

Flexible Model Usage
Hybrid and configurable use of models

Reliable and Highly Accurate
Agentic AI with recursive reasoning and fact-checking for reliability

Configurability
Fully Configurable use cases, topics and retrieval methods.

Guardrails and Enterprise Compliance
Built-in monitoring, privacy filters, logging, and governance guardrails

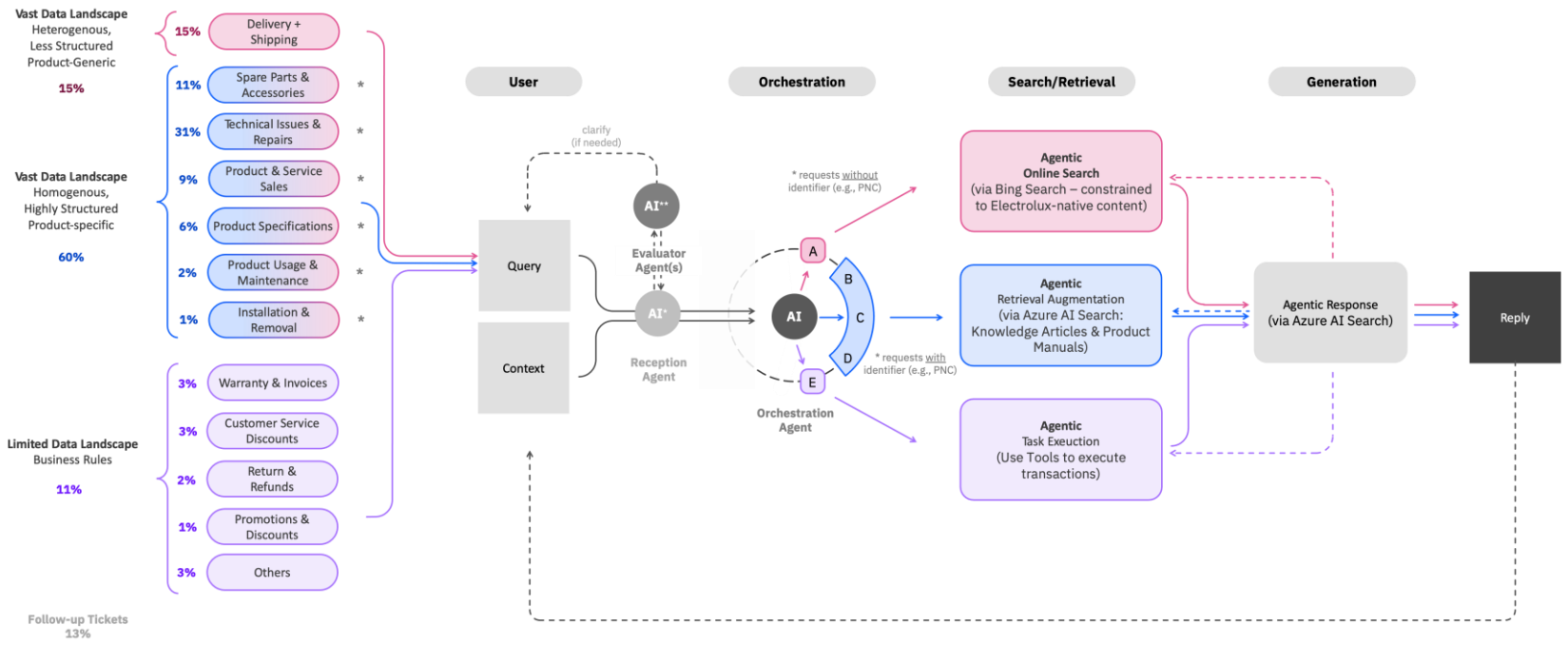


Self-Service-**AI**

Inside Self-Service-AI Architecture: No Chatbots, just Agents at Work

This end-to-end agentic solution replaces traditional chatbots with a fully autonomous agent framework. Every user interaction is managed by an orchestration layer that dynamically invokes three specialized agent teams.

Customer Service Example



Benefits & Value

2.5X	Increase in Containment
75%	Reduction in time to Market for new markets
>40%	Operational cost savings
20%	Reduction in Service Warranty Costs

Accelerate Scaling across organization
Apply to various business functions (internal and external) with same proven asset. (Customer Service, HR, Finance, IT, Safety, Procurement…)

Inclusion of safety guardrails to provide confidence in AI Governance

Features and Innovations

Highly configurable – Use case agnostic

- Seamlessly configure query types, retrieval strategies, and response logic
- High configurability ensures broad coverage across products, services, and use cases

Low code, prompt-based Configuration

- Empower business users and citizen developers to handle change requests using natural-language interfaces
- Respond rapidly to evolving business needs—no deep technical expertise required

Diverse Retrieval Mechanisms for broader Coverage

- Advanced Agentic RAG Pipeline combining multi-agent orchestration, online, and cached search and handle multi-modal (text, images, videos)
- Hybrid Search: Integrates structured and unstructured enterprise knowledge
- Cached FAQ: Enables ultra-fast responses via prompt-based pre-caching

Advanced Conversational AI

- Adaptive, fluent, and context-aware interactions
- Agentic state flow for intelligent query routing
- Recursive retrieval and LLM-based validation
- Multilingual support for global applicability

Architected for Performance

- Intelligent orchestration engine (LLM as runtime OS)
- Dynamic state routing across agents
- Parallelized execution for low-latency performance at scale

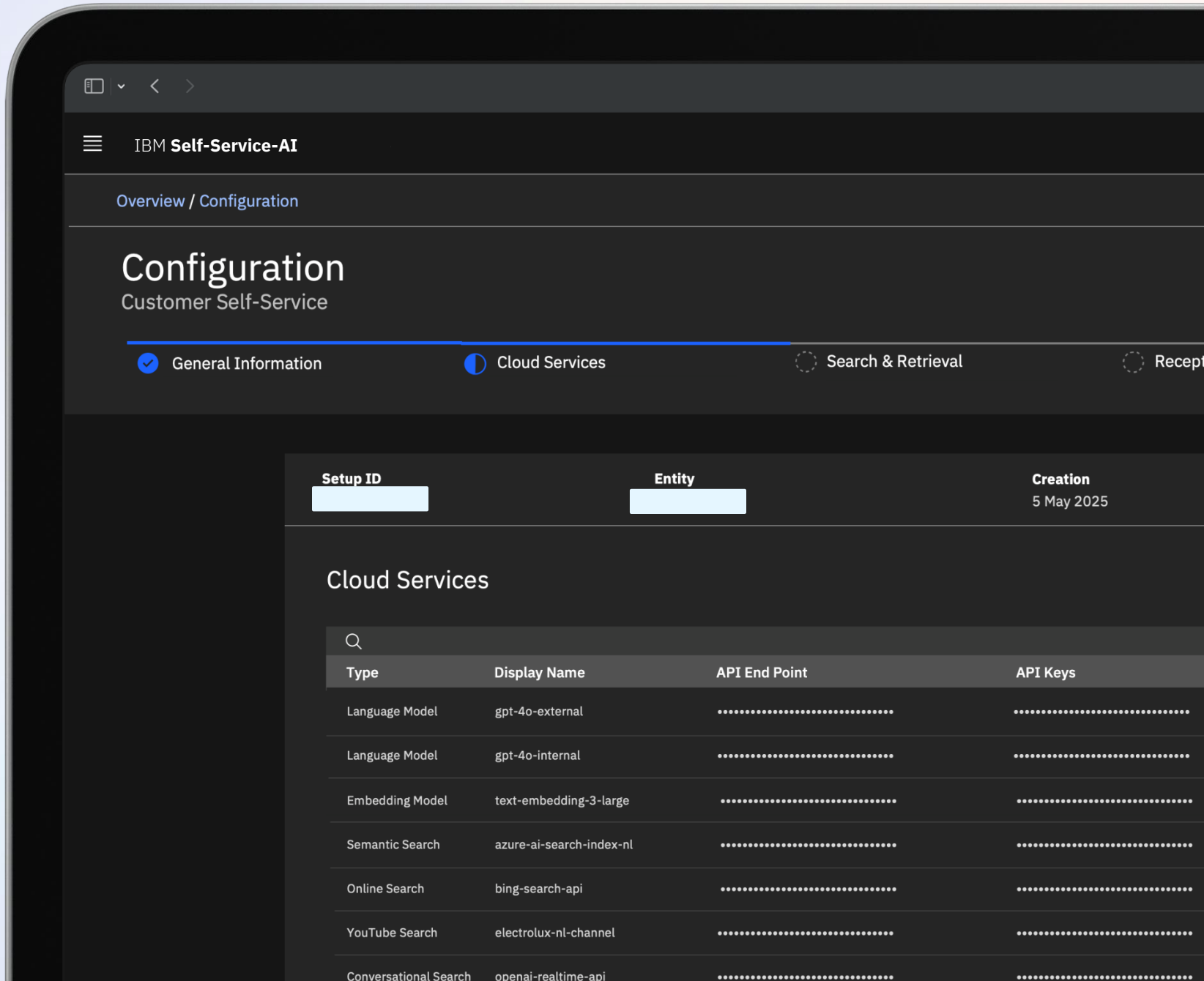
AI Governance by Design

- Enterprise-grade content filtering and risk mitigation
- Configurable guardrails and policy enforcement
- Privacy, security, and regulatory compliance built in
- Full observability with logging and traceability

Self-Service-AI

LLM, Data Source, mode settings

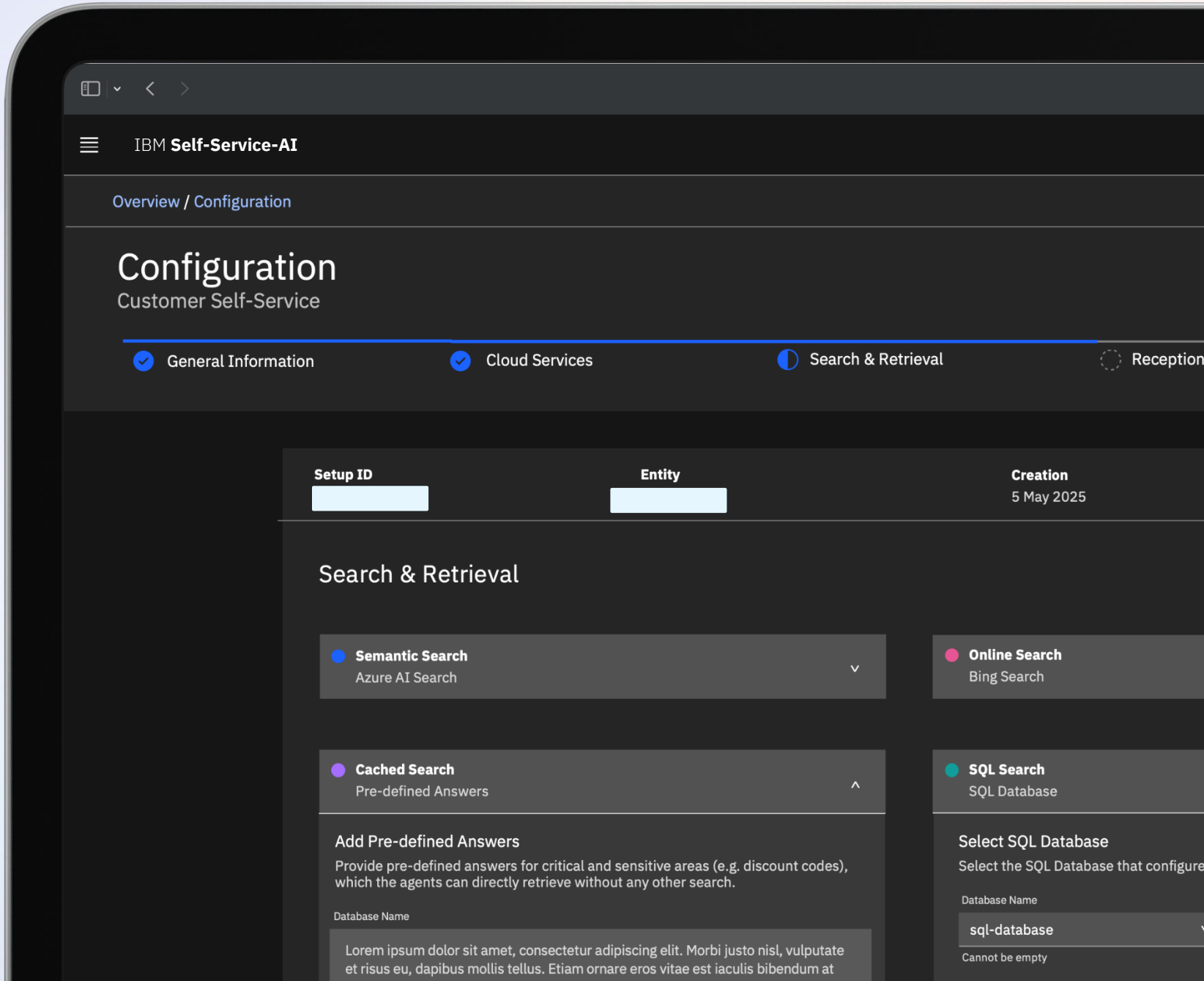
- Select and configure LLMs based on task, accuracy, latency, or cost requirements
- Define retrieval strategies including hybrid, online, semantic, and cached search
- Map data sources structured or unstructured into a unified knowledge pipeline
- Connect external APIs and cloud services to enrich agent capabilities
- Enable online, video and image search options.



Self-Service-AI

Agent Configuration

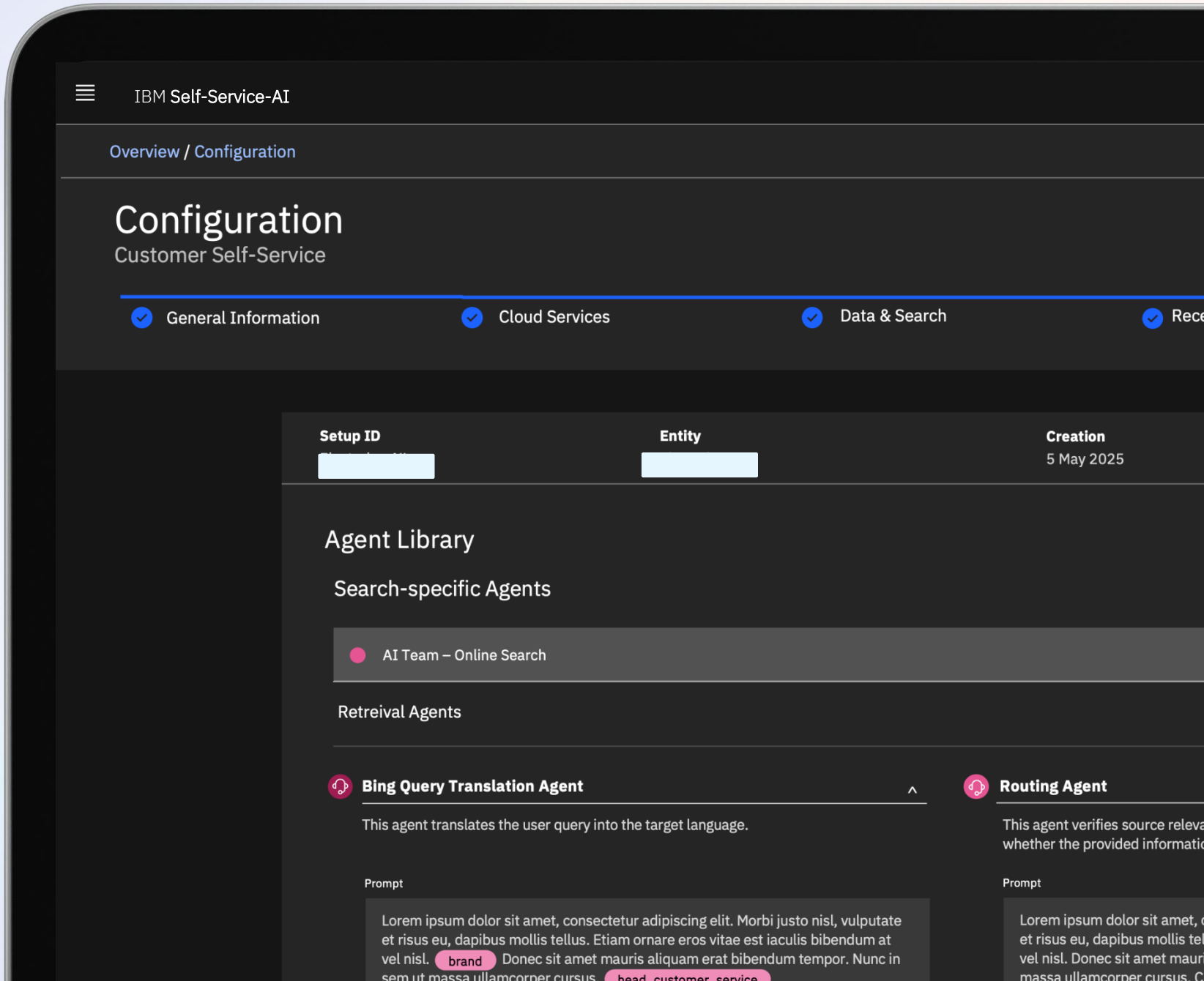
- Select appropriate retrieval agents (semantic, SQL, cached, or online) based on use case
- Map each agent to the relevant knowledge source or database
- Customize fallback strategies to control agent behavior when answers are ambiguous or unavailable
- Define and load pre-defined answers for critical topics requiring deterministic responses



Self-Service-AI

Advanced Agent setup

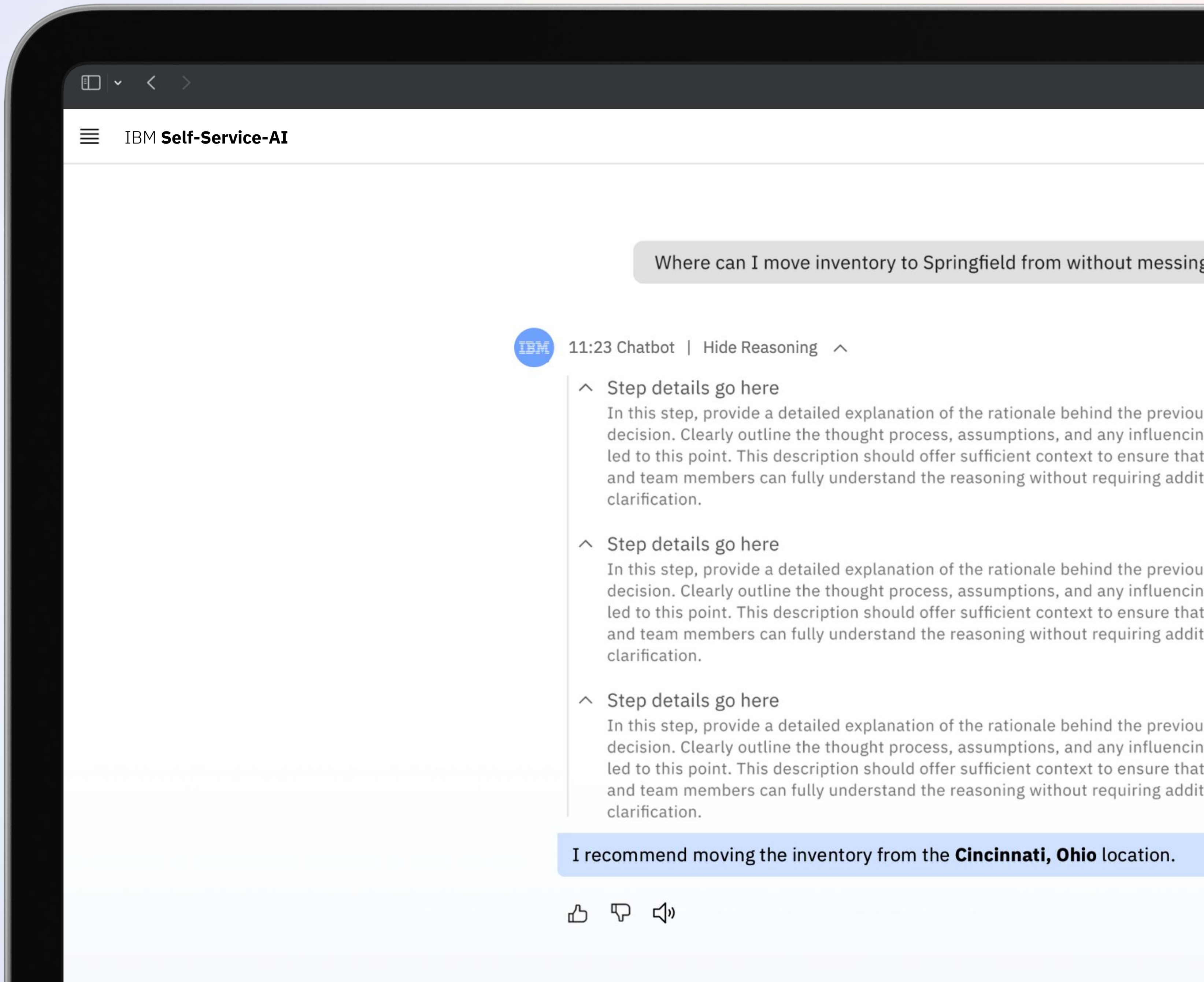
- Design intelligent retrieval pipelines by chaining specialized agents (e.g. translation, routing, filtering)
- Configure agent behavior with tailored prompts, thresholds, and response logic for each query type
- Enable dynamic language handling with translation agents that preprocess multilingual user input
- Route queries smartly across retrieval strategies based on topic, confidence, or source availability
- Add custom logic filters to prioritize factuality, safety, or regulatory compliance
- Test multi-agent flows to validate relevance, latency, and answer quality in real-world scenarios



Self-Service-AI

Test and Tune

- Validate agent output across multiple query types, domains, and languages
- Analyze reasoning steps to understand how decisions are made and identify potential logic gaps
- Evaluate factual accuracy, actionability, and tone of responses using human-in-the-loop feedback
- Solution has ability to refine agent prompts using advanced ML based on user feedback.



Agentic AI for Customer Service for a large CPG client

Business Opportunity

Large CPG organization aspired to grow direct to consumer digital channel revenue from 2% to 40%, while also driving significant productivity gain across customer services operations. Additional goal was also to cut operational cost of contact center.

IBM Solution

IBM Consulting proposed to build an Enterprise Platform for Agentic AI to improve

1. Virtual agent for customer service
2. AI agent assistant for human agents
3. Automation of post call and back office task automation.

The scope covered **consumer service journeys across 14 different markets and 17 different languages to drive digital channel engagement and adoption.**

Benefits achieved

40+%

Operational cost reduction

10%

Increase in conversation rate

20%

Reduction in Service Warranty Costs

Agentic AI Impact

3x

Increase in Containment within first month

75%

Reduction in time to deploy AI solution for new markets

Indicative Delivery Timeline

1 to 3 weeks		8 weeks				
Assessment	**Blueprint**	**Build**				
▪ Confirm Sponsoring Organization, Key Users Mapping ▪ Baseline business processes for transformation ▪ Identify Strategic & Ops KPIs against specific goals ▪ Identification of associated AI Risks and their impact on the use case	▪ Define key customization aspects, upstream and downstream integrations, any specific functional agents, any UI customization or UI/UX integrations. ▪ Define deployment plan and number of environments (dev, test, production) ▪ Structure and plan a 8-week MVP for the prioritized features. ▪ Define test plan for 2-3 document types	**Setup Sprint** Deploy the Agentic App. Get access to data sources Gather initial data set Set and test with sample data	**Dev Sprint** **Dev Sprint** **Dev Sprint** Build, Test & Playback MVP: implement the business process and validate through weekly playbacks with business users Create and integrate required Data Products Test out core capabilities Test with MVP data set	**Test Sprint** Train users on agentic app and required governance knowledge, refine AI Assistant-human collaboration, and start KPI tracking Prepare production deployment	**Prepare for Scale** Structure engagement plan for scale	

GenAI Assistants, Tools, Assets

- Agentic AI Readiness Score Framework
- ICA Assistant to classify process activities
- Mapping process mining metrics to Agentic AI Readiness Scores.

- ICA Assistant to generate first mapping to agentic architecture
- Agentic app Reference architecture

- Base App deployment code and containers
- GIT Repos
- Documentation and Deployment Guide

Delivery Squad

- Delivery Executive
- Agentic AI technical Lead
- Technical Architect
- UX Designer
- Data Scientist
- AI Engineer
- Full Stack Developer
- DevOps
- Test Specialist

Assumptions and Dependencies

1 **Setup and onboarding:**
Environment access prior to project kickoff

- Asset will be deployed in one development environment and availability zone
- Azure resources provisioned
- Required knowledge store finalized

IBM development team onboarded to environment within first week of project

2 **SME time commitment:**
SME input needed:

- Data set knowledge
- As is process and expected results
- UI validation

SME feedback within three (3) business days

3 **Base deployment:**

- Will cover core base features of the app within 8-10 week MVP
- Following query types will need additional development effort
  - Tool integration like raising ticket in workflow tool like service or any other type of integration with external API
  - Hyper personalization by connecting to user data
  - Queries based structured data sources like data base
  - Complex queries which require custom calculation before responding to user question

Get in touch and find out more

Self-Service-AI

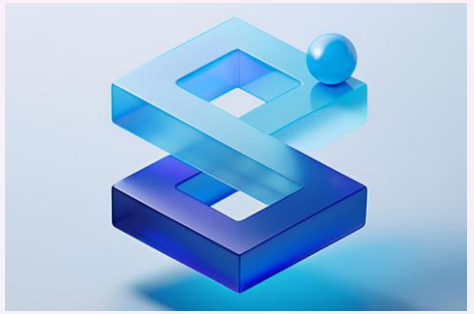
Find out more and request a demo in ICA4AA Agentic App Library

Click here

GitHub

Read more about the technical documentation in our GitHub

Coming Soon



Visit AI Integration Services space on Content+

Click here