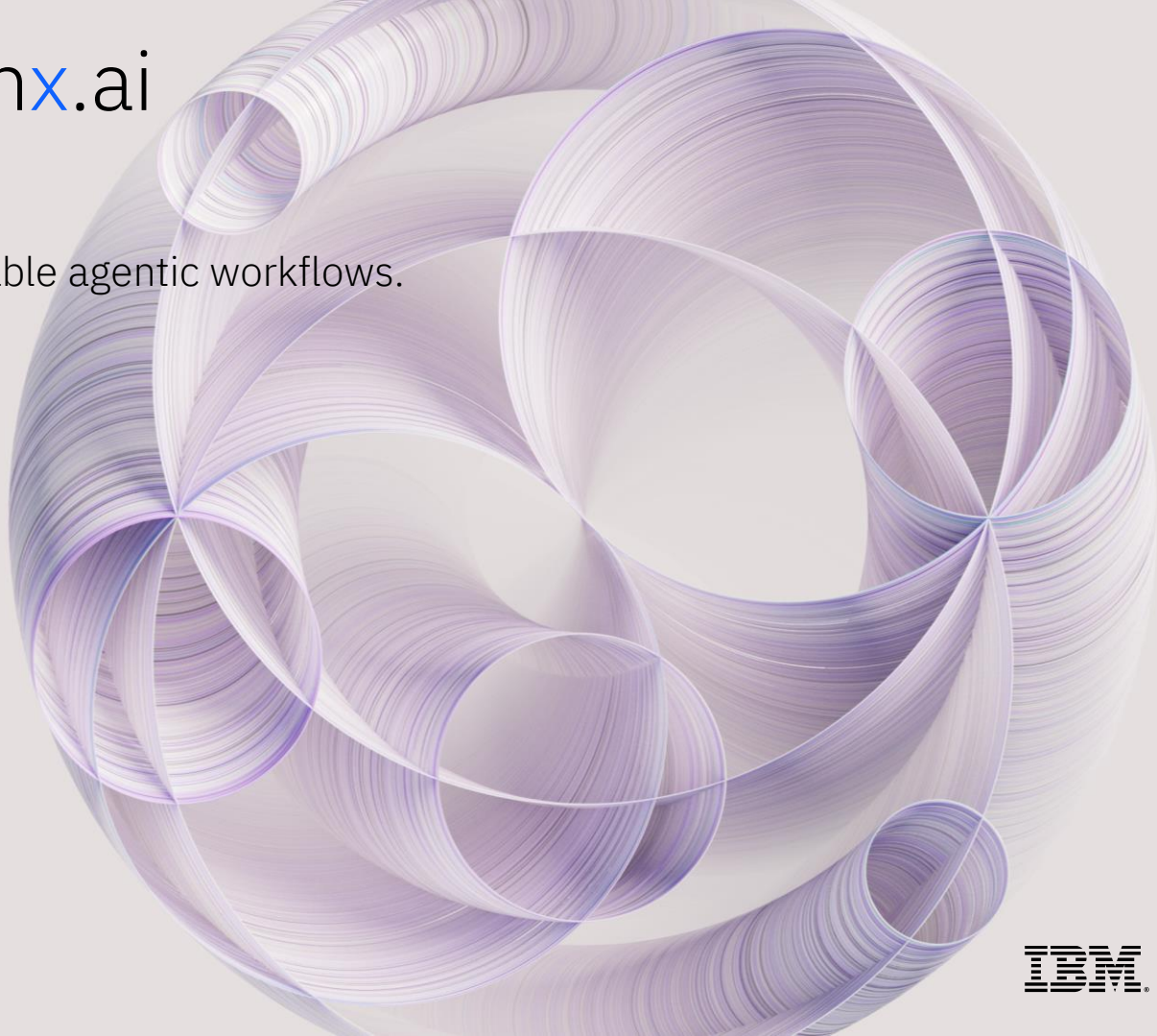


Agents on watsonx.ai

Building useful, reliable, and scalable agentic workflows.



Seller guidance and legal disclaimer

IBM and Business Partner
Internal Use Only

Slides in this presentation marked as **"IBM and Business Partner Internal Use Only"** are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

© IBM Corporation 2024.

All Rights Reserved.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

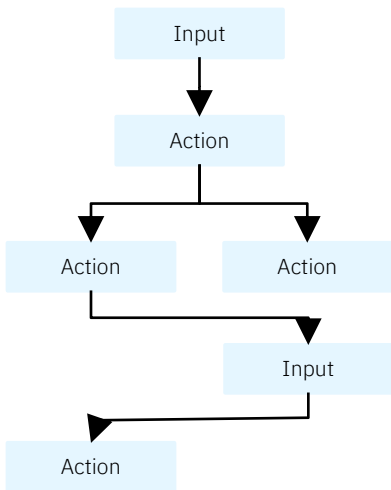
References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results, they may have achieved. Actual environmental costs and performance characteristics may vary by client.

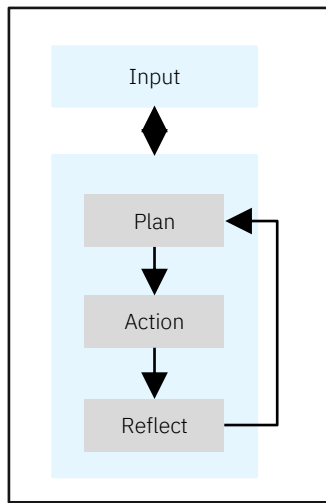
Evolution of assistants

Traditional assistants → Single-agent assistants → Multi-agent assistants

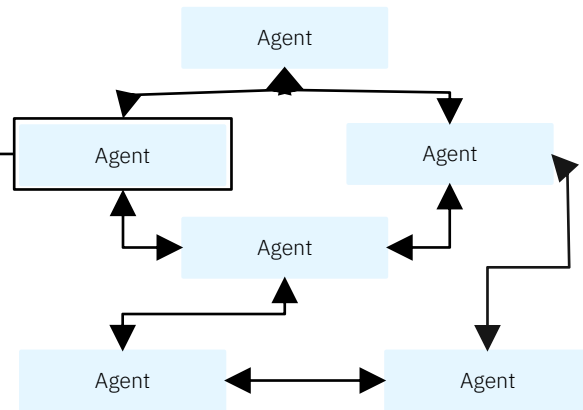
- Rule based (if x, do y)
- Predefined action paths



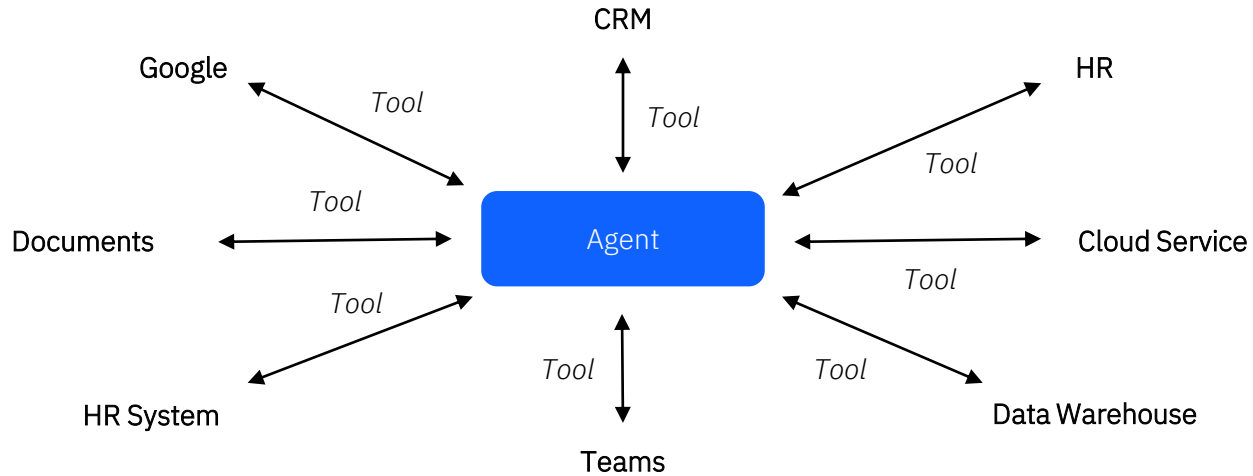
- Task based (e.g., flight booker)
- Performance constraints as scope of task increases
- Limited control



- Domain-based (e.g. travel agent)
- Specialized agents (planner, flight booker, hotel booker, etc.) work together, improving system performance
- Control over how agents communicate
- Multiple architecture options (network, hierarchical, supervisor, custom)



Agents interact with external systems to retrieve data and complete actions



Recruiter example

Diana is a recruiter who works on large-scale hiring initiatives in IBM Software

- 100s of hires across the organization
- Diana is assigned a set of reqs to work on

Great target user for productivity assistance:

- Highly repetitive work
- Involves many tools, systems of record
- Requires a lot of copy and paste, checking
- Tracking is ad hoc

Diana knows *how* to perform her work. She would mostly benefit from *speeding it up*. She is unlikely to have the time or skills to fully automate her work. However, she has the support of a non-technical process excellence person who would be a good target user for developing deeper automations (custom assistant for recruiters).

Workflow sample

Go to Box folder for req approval numbers by hiring manager

Create reqs, copy and paste from spreadsheet

Go to BrassRing

Check for prior position to copy from, otherwise use a blank form

Pull approval number, data, job responsibilities, preferred etc. band level, how many headcount approved

Input these while double-checking

Check with hiring manager if it doesn't make sense - e.g., job description

Go to w3 People for division code, etc.

Go to comp tool for comp range based on job code, band, region

Check job code is correct - Workday job catalog

(Etc.)

Tools used in opening a req

Outlook

Box (approval numbers for reqs)

BrassRing (reqs)

Workday job catalog (check job codes)

w3 People (division code, etc.)

Comp spreadsheet (salary range)

Word (write/format job description)

Notepad (notes)

Webex

Org funding tool

Wombat (reqs from hiring managers)

Select for IBM (interviews)

LinkedIn

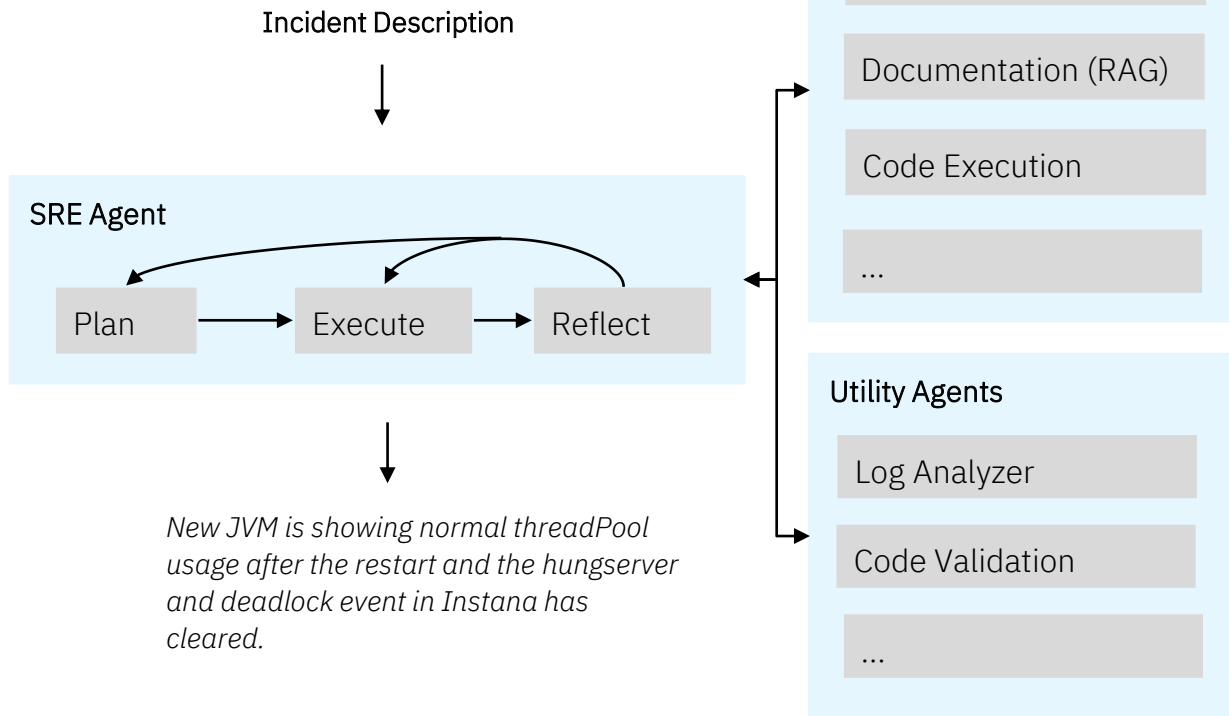
SmashFly

Agents will unlock the next
wave of productivity gains for
the enterprise...

Site Reliability Engineer (SRE) Agent powered by watsonx.ai

Agent performs intelligent root cause analysis and remediation.

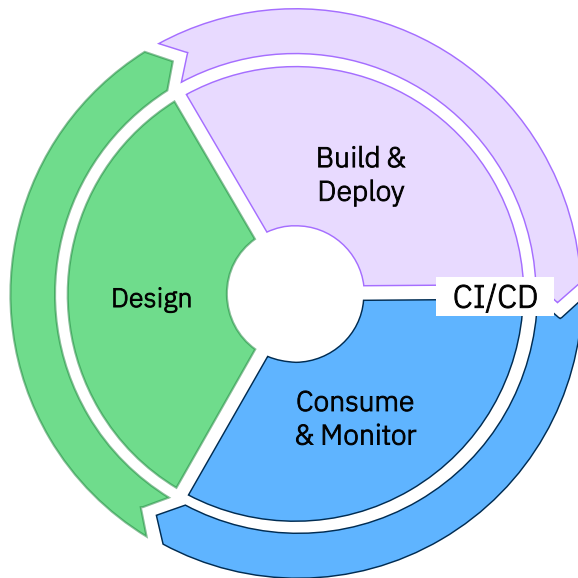
- Dynamic orchestration of tools including Instana API and documentation RAG
- Adaptive workflow and continuous improvement
- Enhanced expertise through specialized agent roles and collaboration



...but they also have their own
unique set of operational
challenges

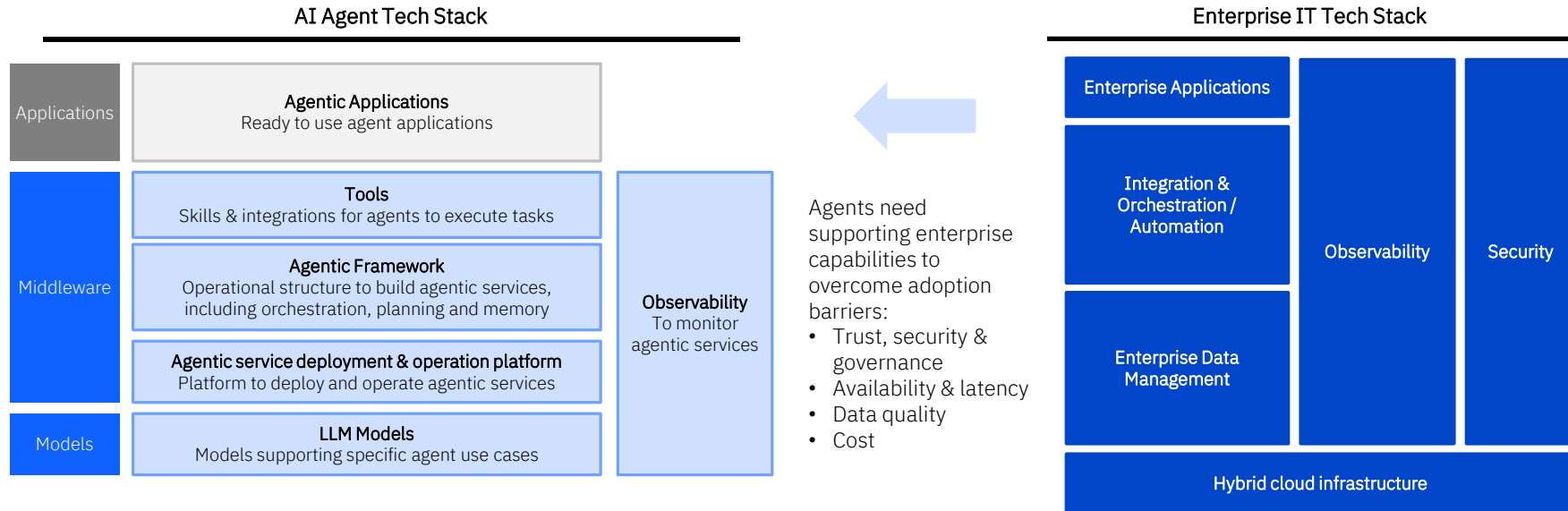
There are many “jobs-to-be-done” along the development lifecycle

1. Define agent use case, detailed workflow and KPIs to align with business goal
2. Identify data sources (tools) available to validate feasibility of project
3. Select/fine-tune appropriate model to suit the agentic workflow
4. Define appropriate architecture & patterns (framework & libraries) to enable reasoning, planning, self-improvement, tool usage
5. Design underlying infrastructure to optimise cost effectiveness



1. Integrate agentic workflow with LLM inference provider
 2. Integrate service with data sources (tools) across environments
 3. Simulate and debug service behaviour
 4. Guardrail actions and outputs
-
1. Deploy agentic workflow as API endpoint
 2. Ensure access control and security
 3. Integrate agentic workflow with application services (UI, etc.)
 4. Monitor agentic workflow KPIs & logs to ensure optimised results, provide transparency & explainability

AI agents need supporting enterprise capabilities to overcome adoption barriers and be deployed at scale



Our goal at IBM is to bring
**useful, reliable, secure, and
scalable** agentic workflows to
the Enterprise.

Three areas of agentic innovation

Accelerate AI agent
deployment

Build custom
designed agents

Manage all agents
in one place

Pre-built
agents



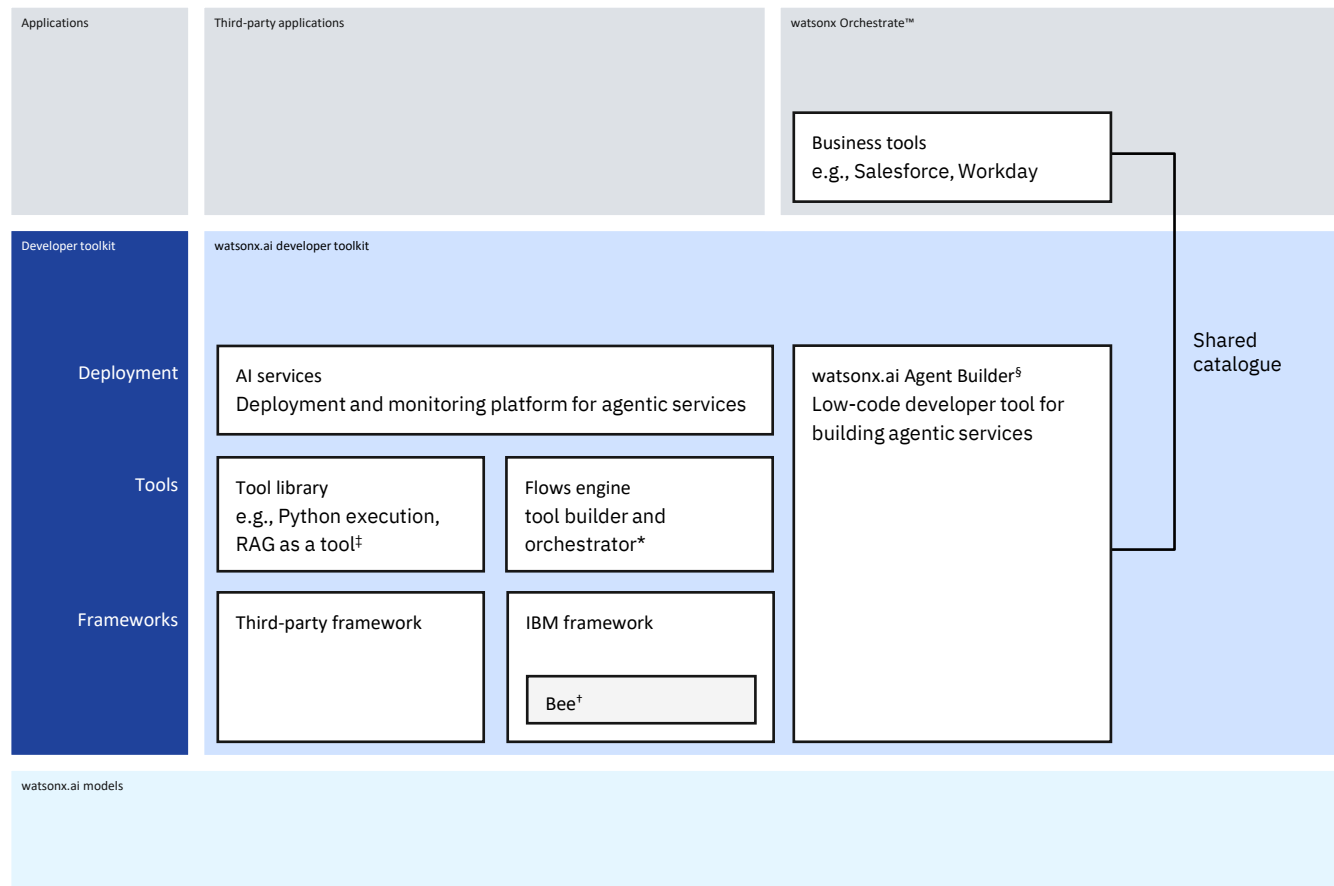
Custom-built
agents



Multi-agent
orchestration



One size does not fit all for building agents




















* Stand-alone in tech preview; to be integrated into watsonx.ai
† Experimental, open-source project

‡ In discovery
§ Coming soon

watsonx.ai Models

Power agentic services using our library of third-party and Granite models suitable for agentic workflows.

 granite-13b-chat-v2 <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: InstructLab	 granite-13b-instruct-v2 <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: Provided model	 granite-20b-code-instruct <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: Provided model	 granite-20b-multilingual <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: InstructLab	 granite-34b-code-instruct <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: Provided model	 granite-3b-code-instruct <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: Provided model	 granite-7b-lab <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: InstructLab	 granite-8b-code-instruct <small>The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...</small> Provider: IBM Type: Provided model
 llama-2-13b-chat <small>Llama-2-13b-chat is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 llama-2-70b-chat <small>Llama-2-70b-chat is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 llama-3-1-70b-instruct <small>Llama-3-1-70b-instruct is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 llama-3-1-8b-instruct <small>Llama-3-1-8b-instruct is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 llama-3-405b-instruct <small>Llama-3-405b-instruct is Meta's largest open-sourced foundation model to date, with 405 billion parameters, optimized for dialogue us...</small> Provider: Meta Type: Provided model	 llama-3-70b-instruct <small>Llama-3-70b-instruct is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 llama-3-8b-instruct <small>Llama-3-8b-instruct is an auto-regressive language model that uses an optimized transformer architecture.</small> Provider: Meta Type: Provided model	 mistral-large <small>Mistral Large, the most advanced Large Language Model (LLM) developed by Mistral AI, is an exceptionally powerful model. Than...</small> Provider: Mistral AI Type: Provided model
 mixtral-8x7b-instruct-v01 <small>The Mixtral-8x7B Large Language Model (LLM) is a pretrained generative Sparse Mixture of Experts.</small> Provider: Mistral AI Type: Provided model							

Leverage trusted, performant and cost-effective models optimised for agentic workflows.

Trusted

IBM Granite was trained on enterprise relevant content that meet rigorous data governance, regulatory and risk criteria defined and enforced by IBM AI Ethics code and Chief Privacy Office.

Performant

Improved accuracy for targeted enterprise business domains like Finance and agentic use-cases.

Cost-effective

Competitively priced model with less infrastructure requirement, IP indemnification, and easy-to-use toolkit for model customization and application integration.

Granite Dense Models

granite-8b

- Base & Instruct
- 4K context ([Oct 2024](#))
- 128K context ([Coming soon](#))



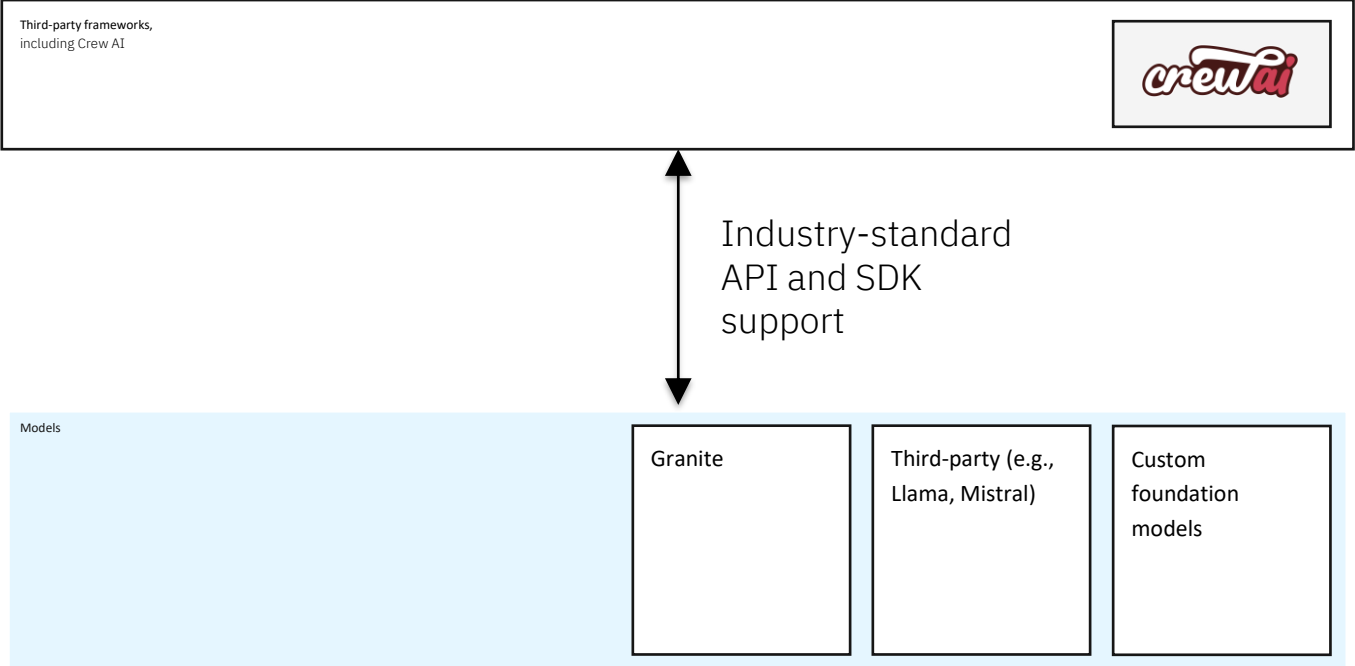
Granite MoE Models

Large mixture-of-experts (MoE) models with strong performance on tool calling.



Build, deploy and monitor
agentic services using **3rd party**
frameworks

Build agents with third-party frameworks through industry-standard API support



Integrate agents with our
extensive library of
enterprise-ready tools.

Web Search

Empower your AI agents with the ability to search the web in real-time. Our Web Search API provides fast, relevant results to keep your applications informed with the latest information from across the internet.

Document Search (RAG as a tool)

Efficient indexing and retrieval of your organization's documents.

Code Execution

Empower your AI agents with the ability to execute Python code in real-time.

Data Connectors

Seamless access to your organization's databases and data warehouses.

Custom

Build your own tool to integrate watsonx.ai with external services.

Coming soon (Oct 2024)

Deploy agents

Framework-agnostic solution for deploying agents.

- Available
- Scalable
- Secure

Coming soon

Monitor agents

Monitor KPIs & logs to ensure optimized results, provide transparency & explainability

The screenshot displays the IBM watsonx AI Services web interface. At the top, there's a navigation bar with the 'IBM watsonx' logo, a search bar, and user account information (IBM account, Dallas, and a KB icon). Below the navigation bar, the breadcrumb trail shows 'Projects / watsonx.ai Demo Project / ab_test_tune1 (1) /'. The main content area is titled 'test' with a green 'Deployed' status and an 'Online' indicator. A blue button labeled 'Open in the Prompt Lab' is in the top right corner. The interface is divided into two tabs: 'API reference' (selected) and 'Test'. Under the 'API reference' tab, there's a section 'Endpoints for inferencing' with a sub-section 'Private endpoint'. It shows 'Text endpoint URL' and 'Stream endpoint URL' fields with example URLs. A 'Bearer <token>' field is also present with a dropdown menu showing 'IAM'. Below this, there's a 'Public endpoint' section with similar URL fields. A link 'Learn more about the 2021-05-01 version query parameter' is provided. The 'Code snippets' section is expanded, showing a 'cURL' snippet with a note about setting the \$API_KEY and a curl command for authentication. Below the curl command, there's a note about manually defining and passing values to be scored, followed by another curl command for a POST request to the generation stream endpoint.

Build, deploy and monitor
agents using **Agent
Builder**

Coming soon (Q4 2024)

watsonx.ai Agent Builder

Accelerate your development lifecycle by building and deploying agentic services using our intuitive visual interface.

- Real-time testing and debugging capabilities
- “One-click” deployment as an API endpoint (watsonx.ai AI Service)

The screenshot displays the IBM watsonx AI Agent Builder interface. At the top, the header includes the IBM watsonx logo, an 'Upgrade' button, a help icon, a notification bell, and user account information for 'Dejan Glozic's Account' with a location dropdown set to 'Dallas'. Below the header, the breadcrumb navigation shows 'Projects / Document based grounding / Prompt Lab'. The main interface is divided into three sections: a left sidebar with icons for chat, structured prompts, and freeform prompts; a central chat area; and a right sidebar with settings and tool management.

In the central chat area, a user message at 01:33 PM asks to 'Analyze the main competitors in the golf ball market and create a detailed report in tabular format.' The watsonx agent responds at 01:33 PM with a detailed report in tabular format analyzing the main competitors in the golf ball market:

Manufacturer	Market Share	Product Features	Target Market
Callaway Golf	22.1%	High-quality golf balls with advanced technology, such as Chrome Soft and Apex	Professional and amateur golfers
Titleist	20.5%	Premium golf balls with precise distance and control, such as Pro V1 and AVX	Professional and serious amateur golfers
TaylorMade Golf	17.3%	Innovative golf balls with speed and distance, such as TP5 and Project (a)	Professional and amateur golfers
Bridgestone Golf	12.1%	High-quality golf balls with advanced technology, such as Tour B X and e6	Professional and amateur golfers
Srixon	8.5%	High-performance golf balls with advanced technology, such as Z-Star and Q-Star	Professional and serious amateur golfers
Others	19.5%	Various golf ball manufacturers with different product features and target markets	Amateur golfers and beginners

Below the table, a note states: 'Note: The market share percentages are approximate and based on various reports and sources. The product features and target markets are general descriptions and may vary depending on the specific product and manufacturer.'

The right sidebar contains the 'Agents' section with 'IBM Bee' selected, a note about tool augmentation, and a 'Selected tools' list including Context, Wikipedia, DuckDuckGo, LLM, SDXL Turbo, and Weather. The 'Execution plan' section shows a 'Finished' phase with three steps: 'Searching for golf ball market information', 'Searching for golf ball market competitors', and 'Generating final answer'.

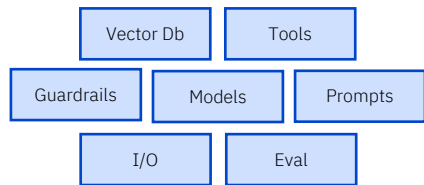
Build, deploy and monitor agents using watsonx.ai **Flows Engine** *

** Standalone in Tech Preview, to be integrated into watsonx.ai*

A declarative approach to unlock enterprise-grade functionality.

*In Tech Preview **

Easily develop agentic services with an intuitive flow language.



Useful
abstractions

```
ragApp = query
| retrievedContext
| promptTemplate
| modelAnswer
| hallucinationScore
```

That can be
composed in a
declarative way

In Discovery

Tool builder and **lightweight agentic framework** which enables flexible *and* trustworthy AI-mediated interaction with enterprise IT systems.

- Build custom tools to quickly integrate your enterprise systems
- Lightweight agentic framework to enable reasoning
- Chat UI widget that can be integrated into any 3rd party application

** Standalone in Tech Preview, to be integrated into watsonx.ai*

Build, deploy and monitor agents using **Bee** *

** Experimental, open-source project*



Important internal note

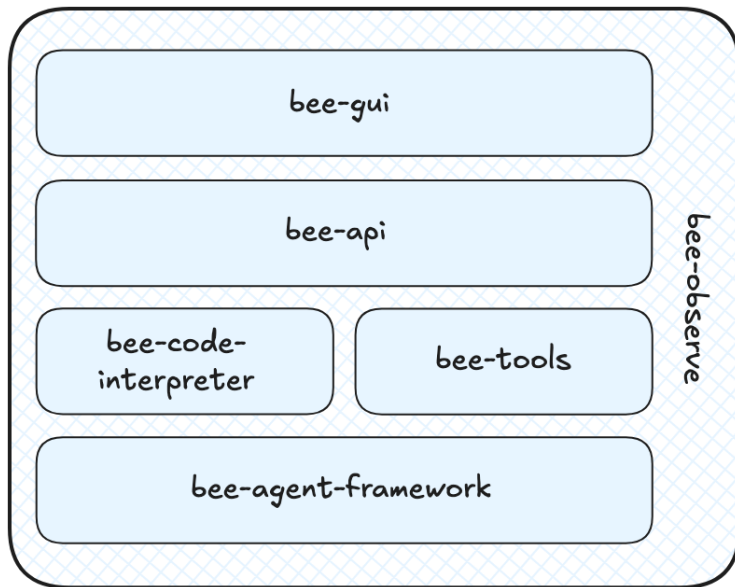
The **bee stack** is in its early stages of development and remains an experimental project. Value prop and roadmap priorities will likely pivot based on early feedback.

We welcome the feedback and appreciate external contributions and integrations to Bee with an understanding that Bee as it stands today does not have any longer-term support or maintenance guarantees.

Additionally, the current version of the bee-agent-framework is unstable, and we may introduce breaking changes in subsequent releases.



The Bee open-source framework was released in early September 2024 to gain feedback from the developer community.



github.com/i-am-bee

Available now

- 🔗 **/bee-agent-framework:** a model-agnostic framework for building scalable agentic workflows.
- 🔗 **/bee-code-interpreter:** a code execution sandbox for safely run LLM-generated and third-party python code.

Coming soon

- 🔗 **/bee-tools:** a library of community contributed tools.
- 🔗 **/bee-api:** a production-grade API to deploy agentic workflows.
- 🔗 **/bee-gui:** a self-hosted chat UI to serve agents to users with built-in transparency, explainability, and user controls.
- 🔗 **/bee-observe:** logging and debugging tooling for agent traces.

The **watsonx.ai** developer toolkit offers solutions for every use-case

Solution	Integration	Benefits
Build, deploy and monitor agents using 3rd party frameworks	<ol style="list-style-type: none">1. Power with watsonx.ai models via SDKs2. Integrate with watsonx.ai Tool Library3. Deploy & monitor as watsonx.ai AI Service	<ul style="list-style-type: none">• Open-source• Deploy existing agents built with 3rd party frameworks
Build, deploy and monitor agents using Agent Builder	<ol style="list-style-type: none">1. Power with watsonx.ai models2. Integrate with watsonx.ai Tool Library3. Deploy & monitor as watsonx.ai AI Service	<ul style="list-style-type: none">• Low-code• Accelerate time-to-market
Build, deploy and monitor agents using Flows Engine*	<ol style="list-style-type: none">1. Power with watsonx.ai models using SDKs2. Deploy & monitor on watsonx.ai	<ul style="list-style-type: none">• Intuitive flow language• Deep integration with IBM products
Build, deploy and monitor agents using Bee**	<ol style="list-style-type: none">1. Power with watsonx.ai models using SDKs2. Integrate with watsonx.ai Tool Library	<ul style="list-style-type: none">• Open-source• Model agnostic

** Standalone in Tech Preview, to be integrated into watsonx.ai*

*** Experimental, open-source project*

Getting started

Developer Hub

Our developer hub provides developers with a library of template and guides to help them get started.

[Coming soon](#)

API Documentation

Access our developer documentation to understand how to build with our SDKs

<https://cloud.ibm.com/apidocs/watsonx-ai>

Free trial

Sign up to our free trial and access 50,000 free tokens.

<https://dataplatform.cloud.ibm.com/registration/stepone>

Granite

[Learn more about Granite](#)

Bee

[Learn more about Bee](#)

Put AI to work with **watsonx**.

