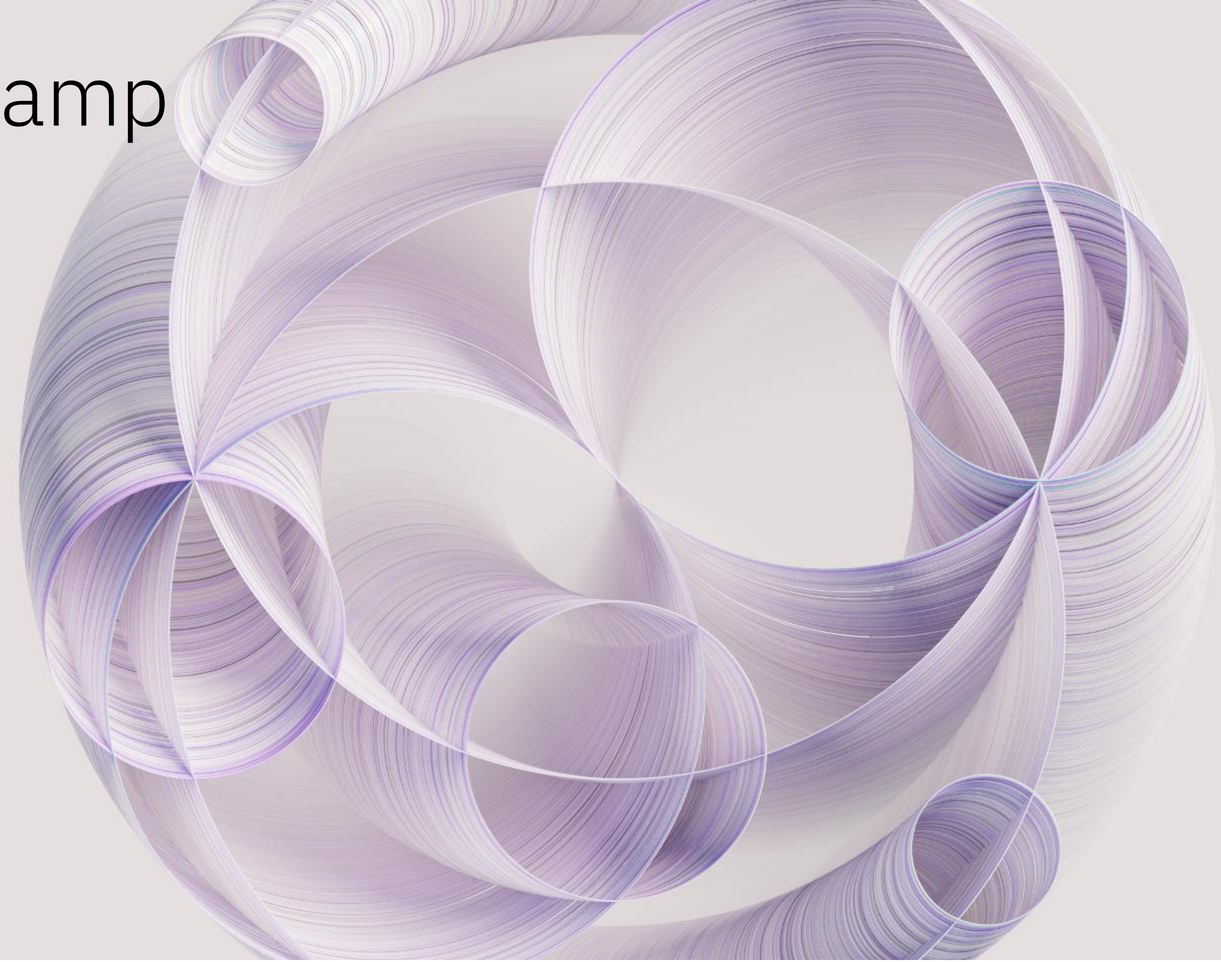
AI Agentic Bootcamp



The future of AI is about agency and productivity

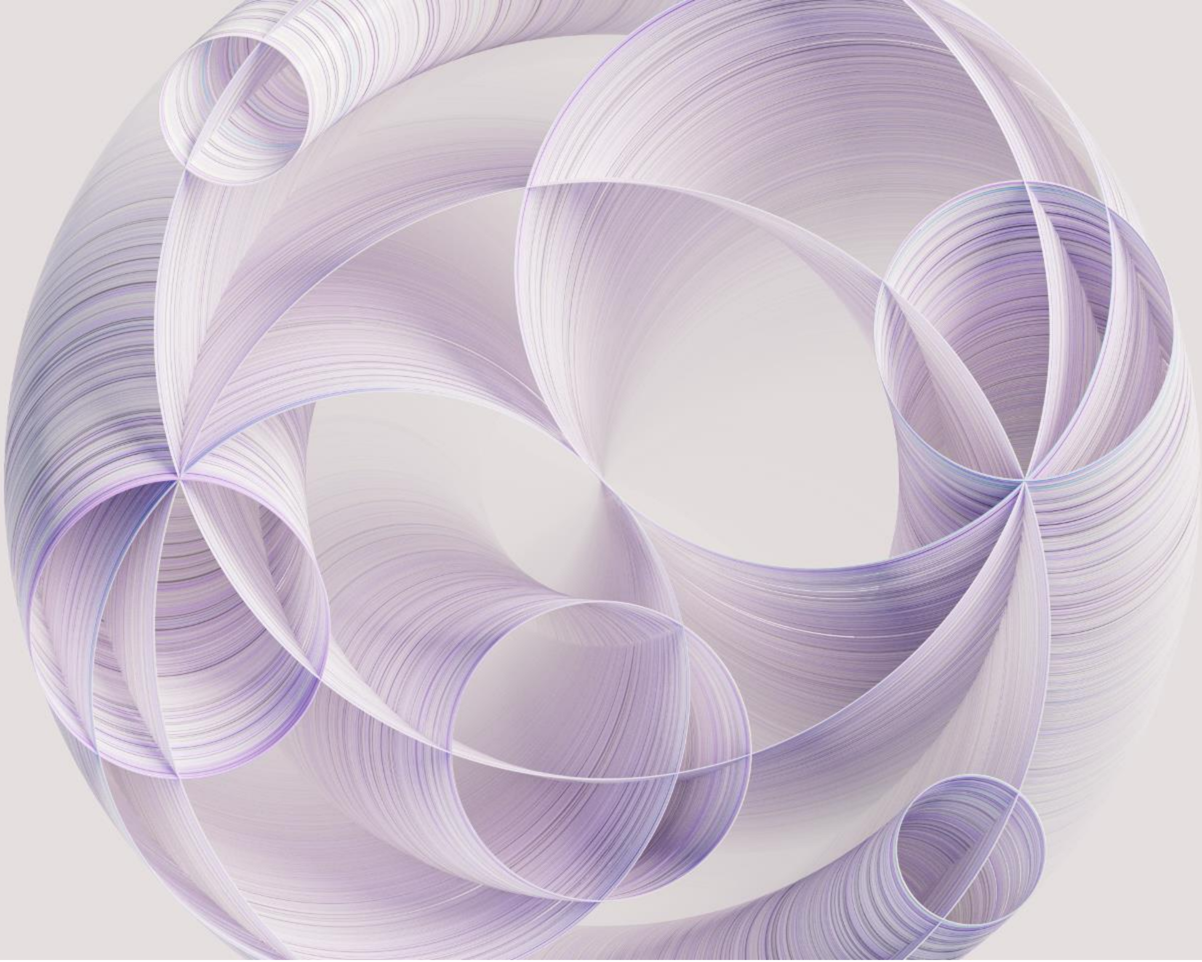
Sebastián Fripp & Facundo Iraola

Agenda

- AI Agents
- Wastonx.ai
- Watsonx Orchestrate
- Labs

AI Agents

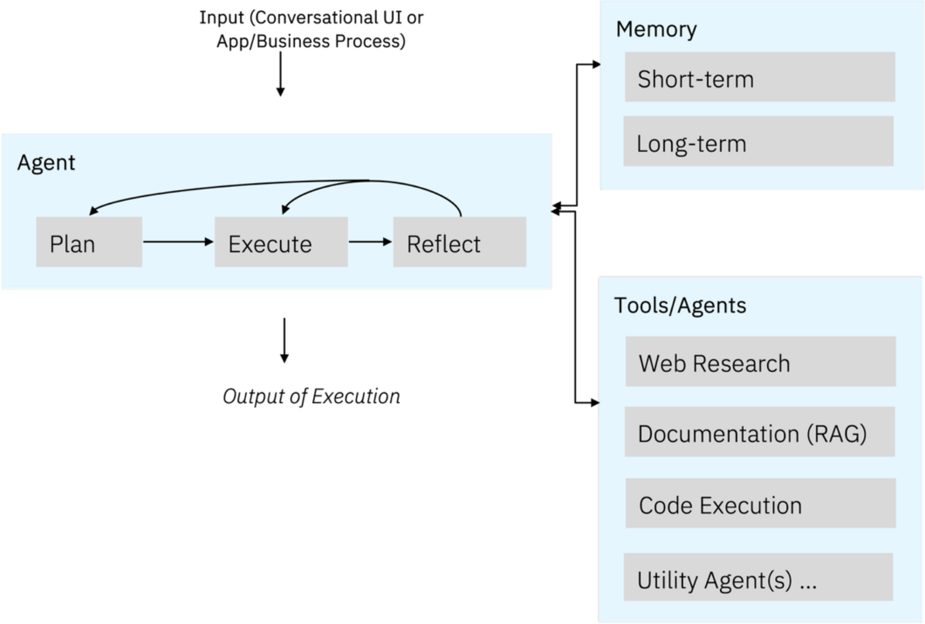
The future of AI is about agency and productivity



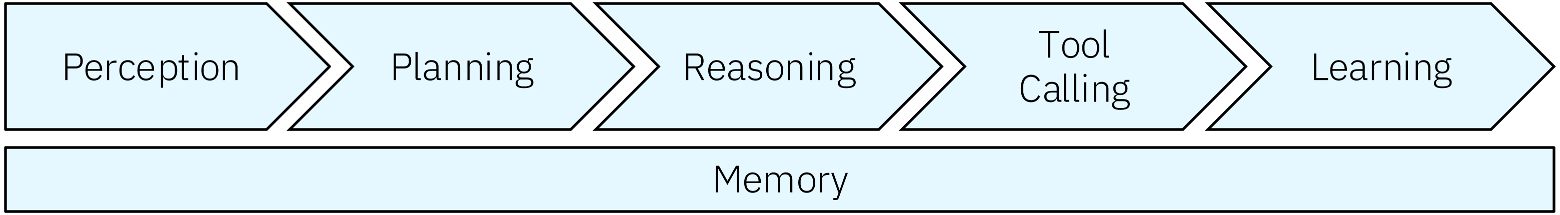
Sebastián Fripp

AI Agents

An AI agent is an autonomous system that can use tools and collaborate with other agents to plan and act on tasks. After it acts, the agent reflects on the results of its actions, learning iteratively and refining its approach to better align with its defined objectives.



AI Agents Main Components



Perception

Planning

Reasoning

Tool Calling

Learning

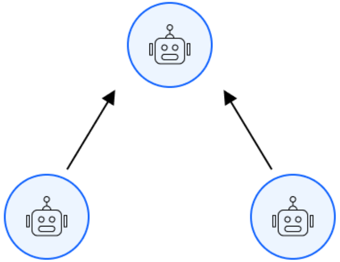
Memory

Agentic Architectures

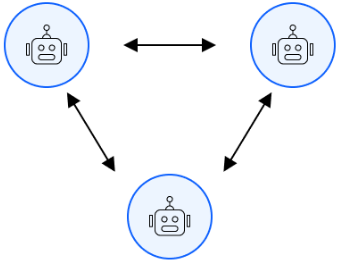
Single agent architecture

Multi agent architecture

Vertical architecture



Horizontal architecture



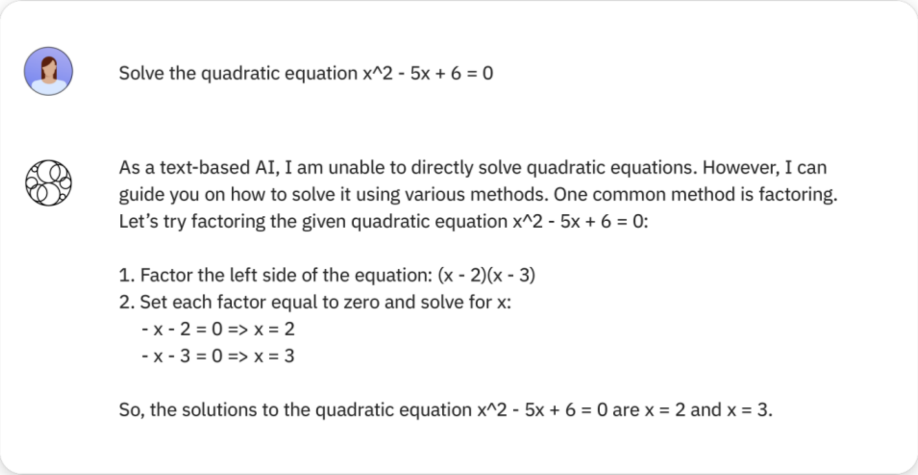
Single vs Multi Architecture

	Single Agent	Multi-agent	
		Vertical	Horizontal
Structure	Single AI agent operates independently to achieve a goal	Leader agent oversees subtasks and decisions, with agents reporting back for centralized control	Agents work as equals in a decentralized system, collaborating freely to solve tasks
Key features	Autonomy	Hierarchy, Centralized communication	Distributed collaboration, Decentralized decisions
Strengths	Simplicity, Predictability, Speed, Cost	Task efficiency, Clear accountability	Dynamic problem solving, Parallel processing
Weaknesses	Limited scalability, Rigidity, Narrow	Bottlenecks, Single point of failure	Coordination challenges, Slower decisions
Best use case	Simple chatbots, Recommendation systems	Workflow automation, Document generation	Complex problem solving

Chain of Thought

Chain of thought (CoT) is a prompt engineering technique that enhances the output of large language models (LLMs), particularly for complex tasks involving multistep reasoning. It facilitates problem-solving by guiding the model through a step-by-step reasoning process by using a coherent series of logical steps.

Chain of Thought

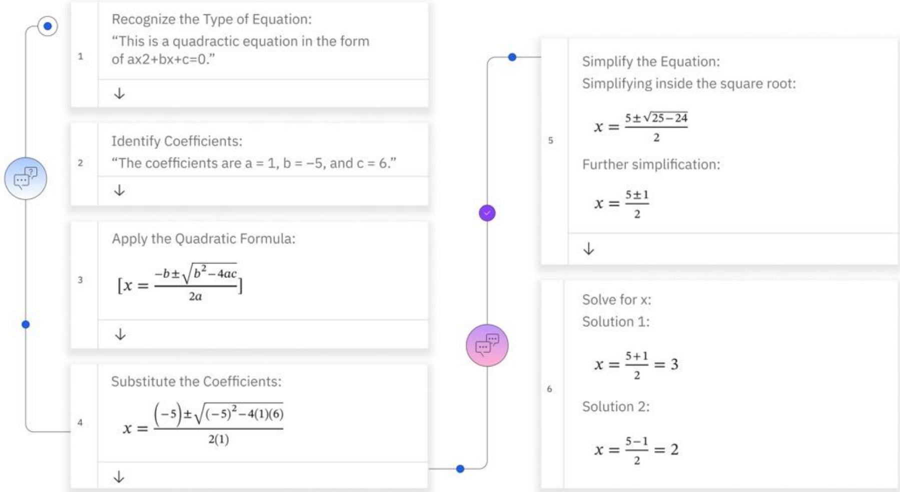


Solve the quadratic equation x^2 - 5x + 6 = 0

As a text-based AI, I am unable to directly solve quadratic equations. However, I can guide you on how to solve it using various methods. One common method is factoring. Let’s try factoring the given quadratic equation x^2 - 5x + 6 = 0:

1. Factor the left side of the equation: (x - 2)(x - 3)
2. Set each factor equal to zero and solve for x:
   - x - 2 = 0 => x = 2
   - x - 3 = 0 => x = 3

So, the solutions to the quadratic equation x^2 - 5x + 6 = 0 are x = 2 and x = 3.



1 Recognize the Type of Equation:
“This is a quadractic equation in the form of ax2+bx+c=0.”

2 Identify Coefficients:
“The coefficients are a = 1, b = −5, and c = 6.”

3 Apply the Quadratic Formula:

$$[x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}]$$

4 Substitute the Coefficients:

$$x = \frac{(-5) \pm \sqrt{(-5)^2-4(1)(6)}}{2(1)}$$

5 Simplify the Equation:
Simplifying inside the square root:

$$x = \frac{5 \pm \sqrt{25-24}}{2}$$

Further simplification:

$$x = \frac{5 \pm 1}{2}$$

6 Solve for x:
Solution 1:

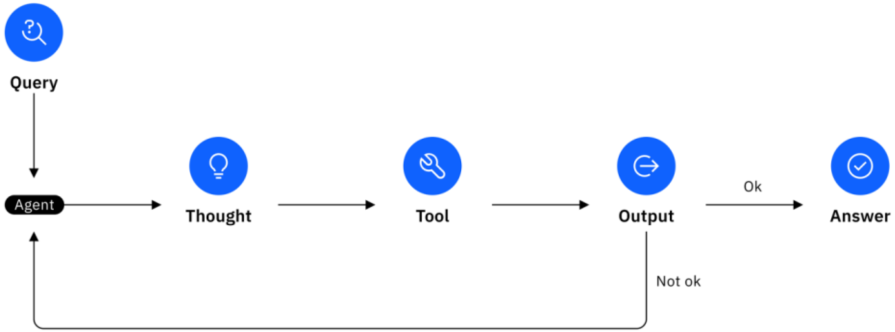
$$x = \frac{5+1}{2} = 3$$

Solution 2:

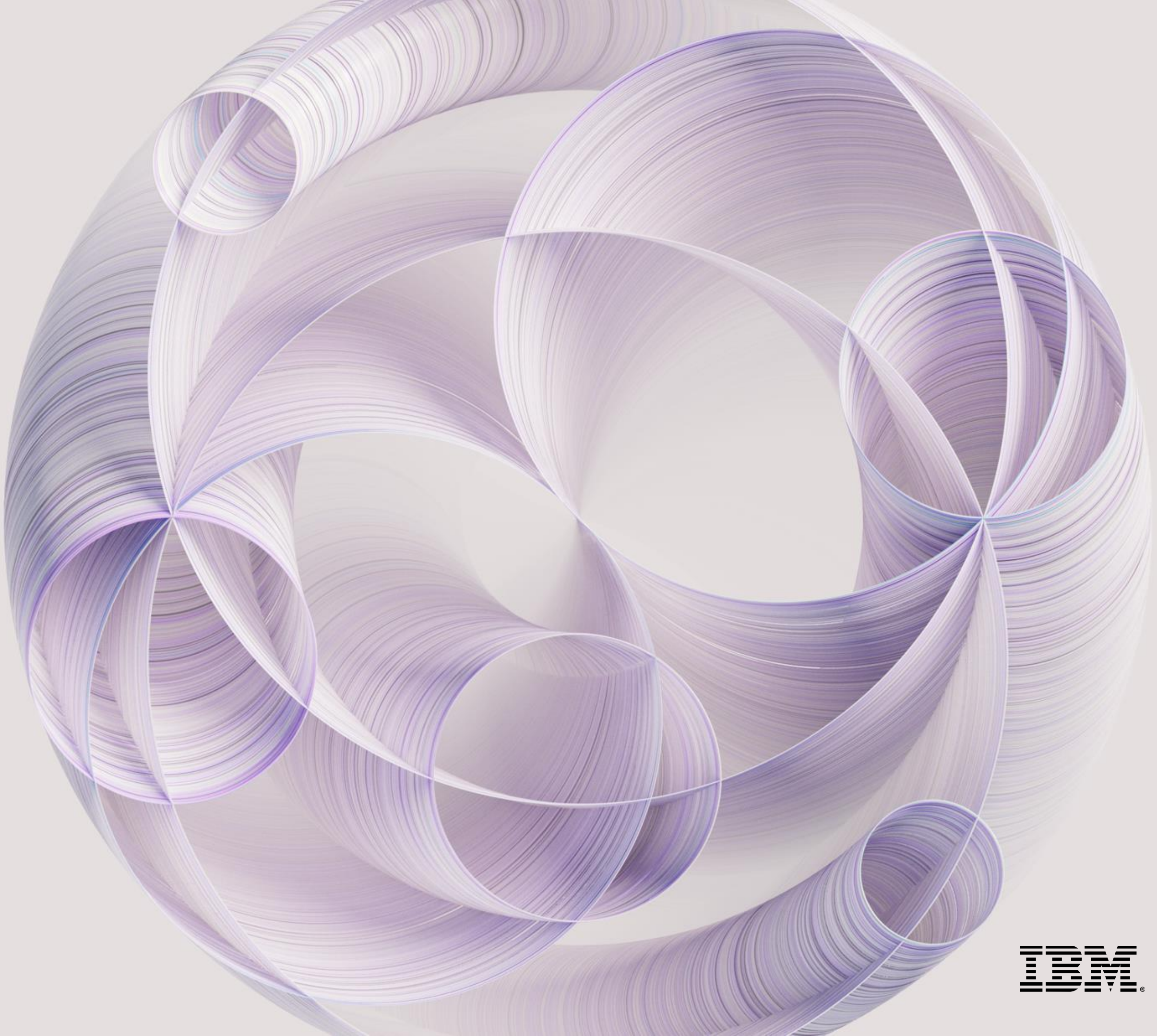
$$x = \frac{5-1}{2} = 2$$

ReAct Agents

A ReAct agent is an AI agent that uses the “reasoning and acting” (ReAct) framework to combine chain of thought (CoT) reasoning with external tool use.



Watsonx.ai



watsonx

A portfolio of AI products that accelerates the impact of generative AI in core workflows to drive productivity.

watsonx.ai

Enterprise-grade AI studio that helps AI builders innovate with all the APIs, tools, models, and runtimes to build AI solutions

Featuring **IBM Granite,** and popular third-party models including **Mixtral, Llama** series

watsonx.data

The **hybrid, open data lakehouse** to power AI and analytics with all your data, anywhere

watsonx.governance

End-to-end toolkit for AI governance to manage **risk and compliance across the entire AI lifecycle**.

watsonx Orchestrate

An enterprise-ready solution that helps create, deploy, and manage AI assistants and agents to automate processes and workflows.

watsonx Code Assistant

Accelerate development, **application modernization, and assist with IT Operations**

Why IBM watsonx for scaling enterprise AI to drive productivity

Open

→ Offers choice to train the right foundation models, including open-source models, and the choice of data, tools, and frameworks to achieve desired business outcomes.

→ Run AI wherever the business needs to, across any cloud, at scale.

Trusted

→ Built with open and transparent technology to give enterprises confidence in their AI and meet regulatory compliance demands.

→ Responsible AI and protected data backed by enterprise governance and security controls.

Integrated

→ Integrates technology seamlessly into existing infrastructures, systems, and processes with choice of cloud to transform the enterprise and drive productivity from within.

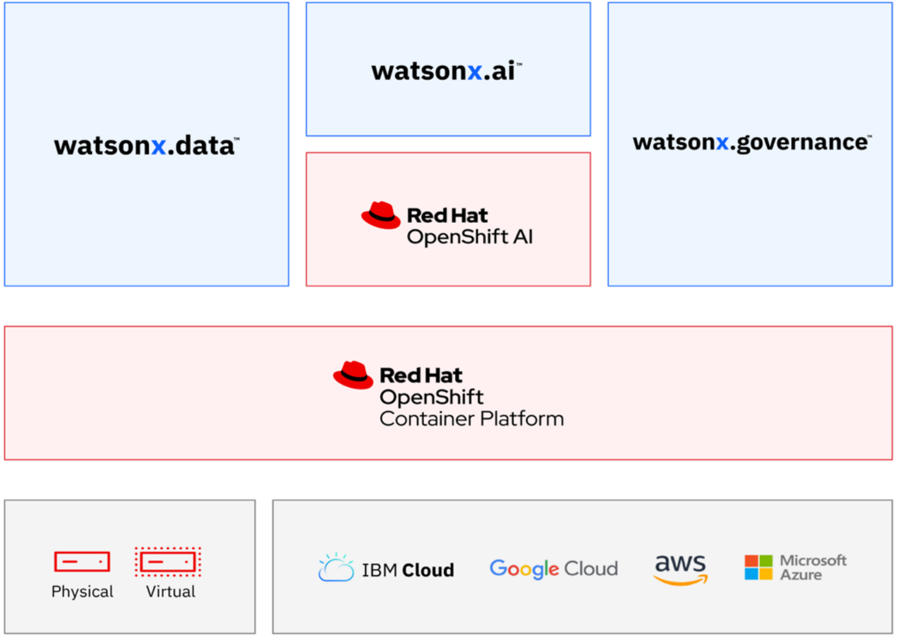
→ Embedded AI for targeted use cases that drives enterprise scale productivity.

watsonx.ai and Red Hat

Watsonx.ai integrated with **Red Hat OpenShift AI** & **Red Hat OpenShift Container Platform** delivers an accelerated time to value through an optimized generative AI and ML workflow.

Key benefits:

- A complete platform for managing AI workloads, training models and scaling AI deployments
- Access to high quality, governed, and trusted AI through curated IBM-developed, third-party, and open-source foundation models
- Deploy to on-premises environments, private or public clouds
- Integrations to the wider watsonx and Cloud Pak for Data platforms.



Creating Generative AI Solutions

Model Selection

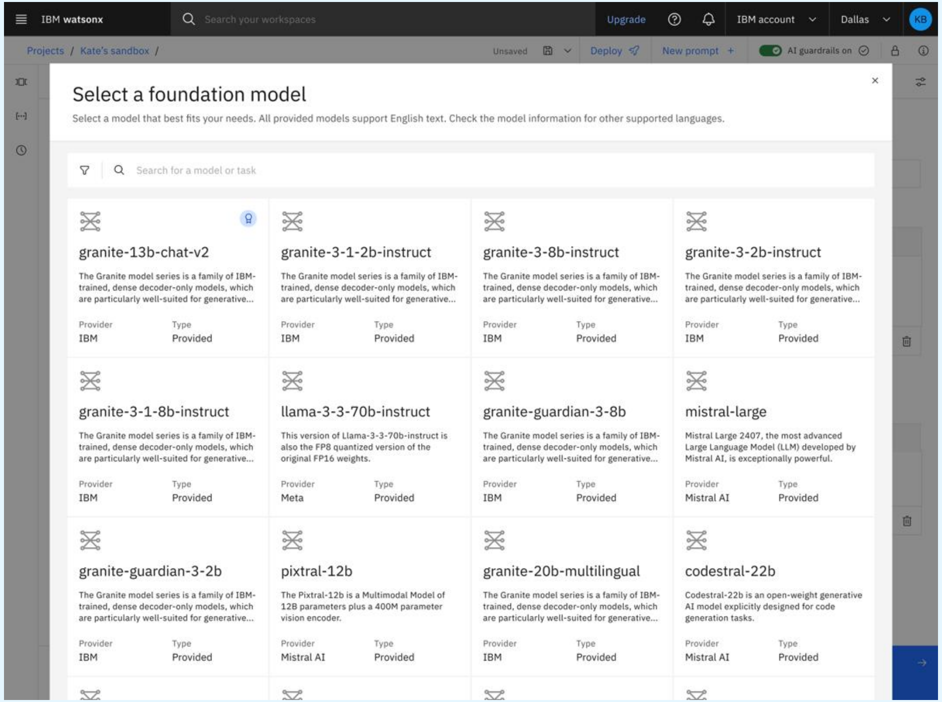
watsonx.ai Provided Models

A highly curated library with IBM-built, open-source and third-party foundation models that are pre-deployed and ready to use in watsonx.ai for a wide range of business requirements and budgetary considerations. These models are deployed on multi-tenant GPU's.

Key benefits:

- Great for experimentation and prototyping to determine which model is best for your use case
- Pay only for the tokens that are consumed, no time commitment
- Ideal for clients that need periodic or infrequent use of foundation models
- Indemnification offered for IBM and select models

Learn more in [documentation]()



Select a foundation model

Select a model that best fits your needs. All provided models support English text. Check the model information for other supported languages.

granite-13b-chat-v2
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

granite-3-1-2b-instruct
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

granite-3-8b-instruct
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

granite-3-2b-instruct
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

granite-3-1-8b-instruct
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

llama-3-3-70b-instruct
This version of Llama-3-3-70b-instruct is also the FP8 quantized version of the original FP16 weights.
Provider Meta
Type Provided

granite-guardian-3-8b
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

mistral-large
Mistral Large 2407, the most advanced Large Language Model (LLM) developed by Mistral AI, is exceptionally powerful.
Provider Mistral AI
Type Provided

granite-guardian-3-2b
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

pixtral-12b
The Pixtral-12b is a Multimodal Model of 12B parameters plus a 400M parameter vision encoder.
Provider Mistral AI
Type Provided

granite-20b-multilingual
The Granite model series is a family of IBM-trained, dense decoder-only models, which are particularly well-suited for generative...
Provider IBM
Type Provided

codestral-22b
Codestral-22b is an open-weight generative AI model explicitly designed for code generation tasks.
Provider Mistral AI
Type Provided

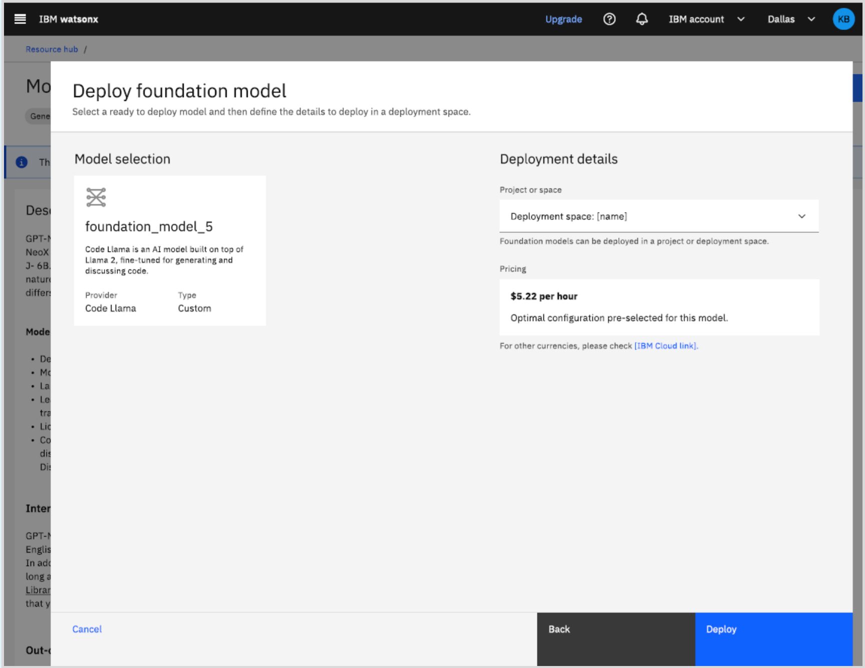
Deploy on Demand Models

A curated collection of high quality and popular foundation models that can be easily deployed to a dedicated GPU with just a few clicks.

Key benefits:

- Faster interactions with a dedicated, single-tenant deployment, hosted 24/7 until de-provisioned by client
- Predictable hourly hosting price compared to variable token-based pricing
- Supports full context length of the model
- No rate limits on the inference requests per second (vs. 8 req/s limit on pre-deployed models)

Learn more in [documentation]()



Deploy foundation model

Select a ready to deploy model and then define the details to deploy in a deployment space.

Model selection

foundation_model_5

Code Llama is an AI model built on top of Llama 2, fine-tuned for generating and discussing code.

Provider	Type
Code Llama	Custom

Deployment details

Project or space: Deployment space: [name]

Foundation models can be deployed in a project or deployment space.

Pricing

$5.22 per hour

Optimal configuration pre-selected for this model.

For other currencies, please check [IBM Cloud link].

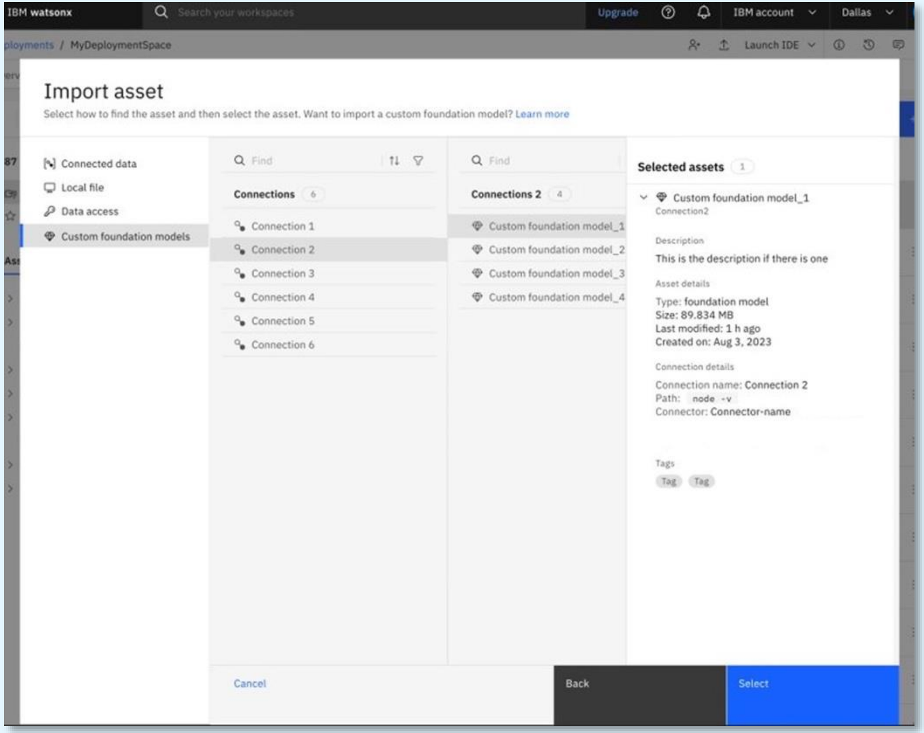
Custom Foundation Models

Import custom foundation models for maximum flexibility in how generative AI solutions are developed.

Key benefits:

- Leverage externally fine-tuned LLMs that support a specific language or are customized for an industry or business domain
- Import a model that is not already provided by IBM.
- Compatible with 1000’s of models from repositories like Hugging Face, which provides access to a huge selection of open-source foundation models

Learn more in the blog, tutorial or documentation



Import asset

Select how to find the asset and then select the asset. Want to import a custom foundation model? Learn more

Custom foundation model_1
Connection2

Description
This is the description if there is one

Asset details
Type: foundation model
Size: 89.834 MB
Last modified: 1 h ago
Created on: Aug 3, 2023

Connection details
Connection name: Connection 2
Path: `node -v`
Connector: Connector-name

Prompt Lab

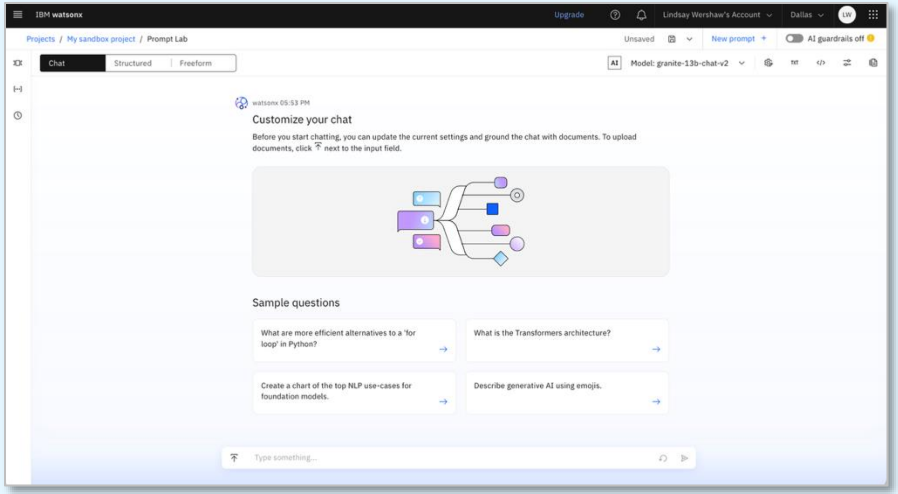
watsonx.ai Prompt Lab

Experiment with foundation models through an interactive user interface or an API.

- Easy to use chat, structured and free form prompt building interfaces
- Experiment with zero-shot, one-shot, or few-shot prompting to get the best results
- Adjustable AI guardrails
- Includes prompt examples for various use cases and tasks
- Save and share prompts for team collaboration

- Wide selection of foundation models to meet any task requirement
- Adjust model parameters to optimize results such as sampling, min/max tokens, stop sequences, repetition penalties, and more
- Export prompts and settings to a notebook to jump start development

Learn more in documentation



Customize your chat

Before you start chatting, you can update the current settings and ground the chat with documents. To upload documents, click ⤒ next to the input field.

Sample questions

What are more efficient alternatives to a 'for loop' in Python? →

What is the Transformers architecture? →

Create a chart of the top NLP use-cases for foundation models. →

Describe generative AI using emojis. →

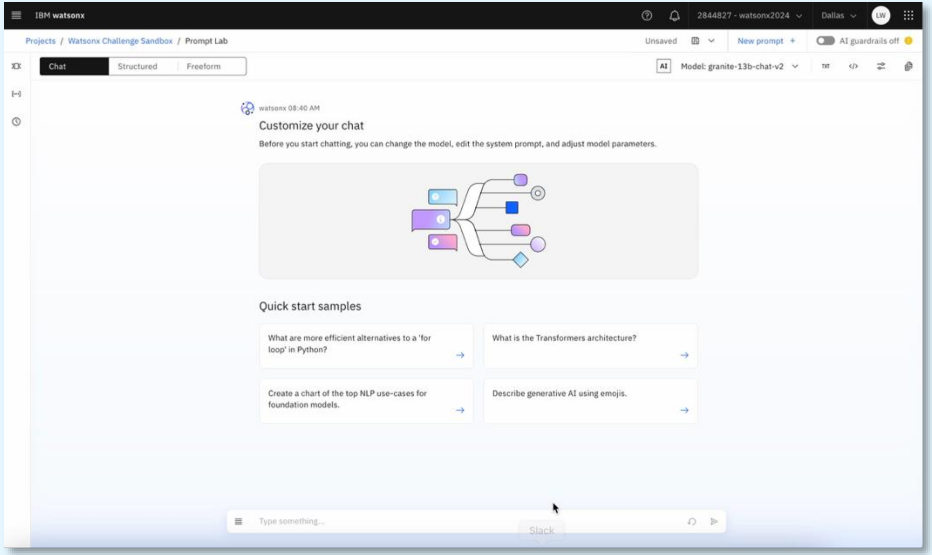
Chat with Documents

The Chat with Documents feature, within the Prompt Lab, provides an easy way to support RAG use cases.

Key benefits:

- Dramatically improve model output by grounding with relevant content
- Fast prototyping of your RAG use cases (e.g. test different LLMs, iterate on document index/collections, change embedding models, set chunk size and overlap, and more)
- Accelerate development by exporting the prompt as a python function deployed as a REST API
- Save prompts and settings as a template for team collaboration.

Learn more in [documentation]()



watsonx 08:40 AM

Customize your chat

Before you start chatting, you can change the model, edit the system prompt, and adjust model parameters.

Quick start samples

What are more efficient alternatives to a 'for loop' in Python? →

What is the Transformers architecture? →

Create a chart of the top NLP use-cases for foundation models. →

Describe generative AI using emojis. →

Chat with Images

The Chat with Images feature, within the Prompt Lab, provides an easy way to convert visual information into text.

Example use cases:

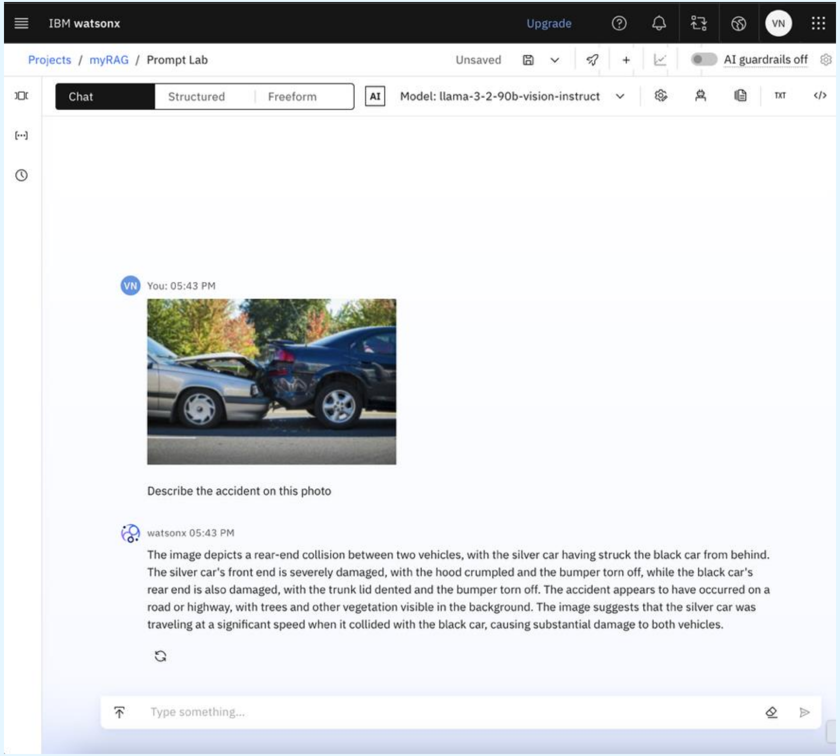
- Automate the generation of alternative text for images to meet accessibility requirements

- Summarize photos to support use cases such as insurance claims (and many others)

- Convert images from a document into text before the document is used as grounding information for a RAG use case.

Chat with Images is supported by **Llama 3.2** (11b and 90b vision models).

Learn more in the documentation and release blog.



Model: llama-3-2-90b-vision-instruct

You: 05:43 PM

Describe the accident on this photo

watsonx 05:43 PM

The image depicts a rear-end collision between two vehicles, with the silver car having struck the black car from behind. The silver car's front end is severely damaged, with the hood crumpled and the bumper torn off, while the black car's rear end is also damaged, with the trunk lid dented and the bumper torn off. The accident appears to have occurred on a road or highway, with trees and other vegetation visible in the background. The image suggests that the silver car was traveling at a significant speed when it collided with the black car, causing substantial damage to both vehicles.

Type something...

RAG Solutions

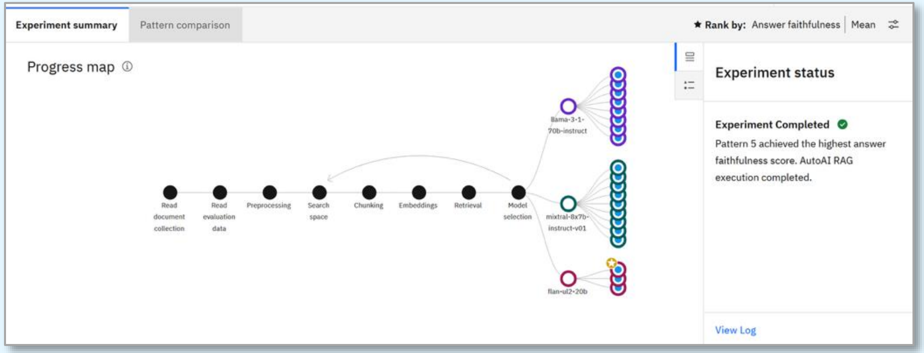
watsonx.ai AutoAI RAG

RAG prototypes are easy to build but hard to productionalize, which can require a team of experts and months of effort. AutoAI RAG helps to resolve this.

Key benefits:

- Accelerates design and deployment of optimized RAG systems based on client data and use case
- Quickly run various experiments to evaluate a constrained set of configuration options (models, chunk size, and more)
- Re-evaluate and modify recommended configurations when something changes (e.g. a new model version is released, quality of model responses change).

Learn more in documentation.



Experiment Completed ✓

Pattern 5 achieved the highest answer faithfulness score. AutoAI RAG execution completed.

AutoAI RAG tests a range of parameters (models, chunk size, etc.) in a series of experiments to automatically find the most optimal combination. It has 3 layers:

- Efficient RAG **hyper-parameter optimization** algorithm with end-to-end automation
- Best-of-breed RAG **evaluation metrics** and **benchmarking tools**
- **Parameterized RAG pipelines** for creating embeddings and for retrieval-based inference

watsonx.ai Text Extraction API

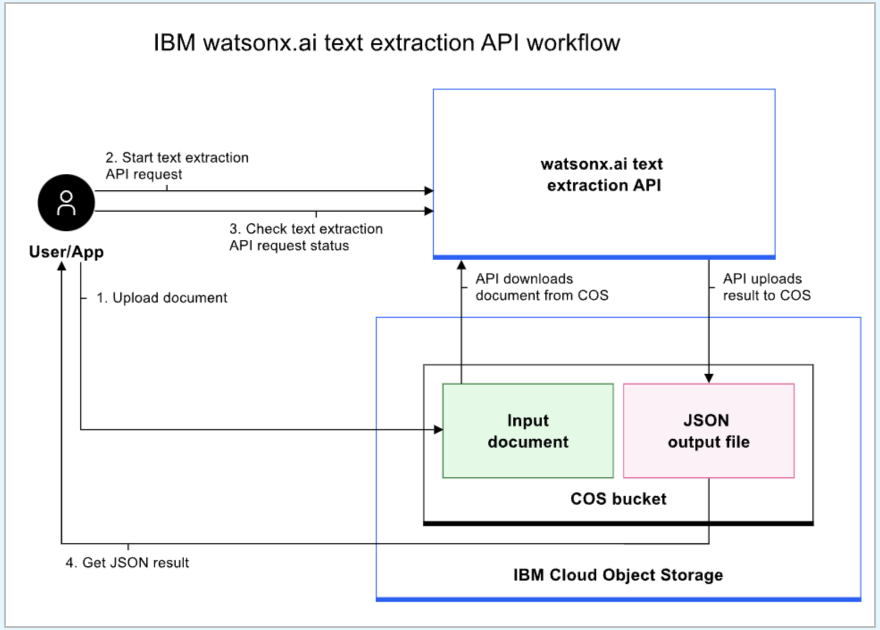
Extract content from documents which is essential for RAG use cases.

- Convert files with tables, diagrams and images into an AI-model friendly JSON file format
- Process the following file types: GIF, JPG, PDF, PNG, TIFF. Including scanned, hand-written documents
- Supports multiple languages

Technology:

- IBM's Natural Language Understanding (NLU) Service. See Watson Document Understanding.
- Optical Character Recognition (OCR) to extract text from images.

Learn more in the documentation.


IBM watsonx.ai text extraction API workflow
2. Start text extraction API request
3. Check text extraction API request status
User/App
watsonx.ai text extraction API
API downloads document from COS
API uploads result to COS
1. Upload document
Input document
JSON output file
COS bucket
4. Get JSON result
IBM Cloud Object Storage

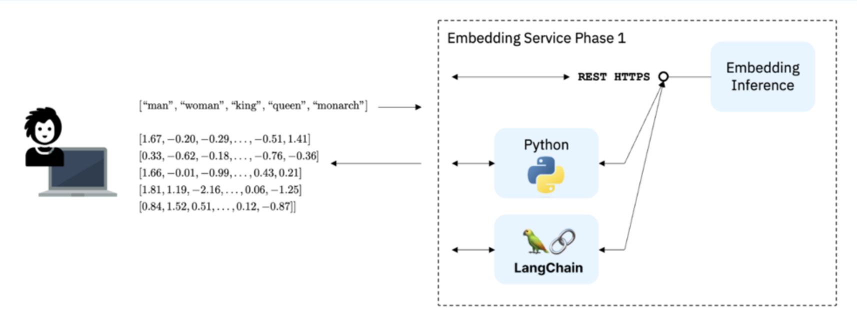
watsonx.ai Embeddings API

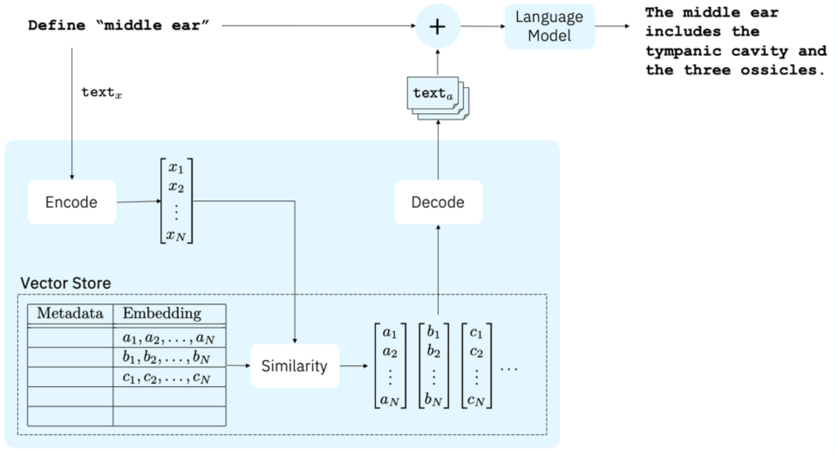
Convert text into dense vector embedding representations. Embeddings capture nuanced semantic and syntactic relationships between words and passages which are then stored in a vector database for retrieval.

Key benefits:

- Support RAG patterns with contextual grounding when utilizing embedding models for query and passage vectorization
- Improve RAG performance with semantically faithful representation of content, especially when compared to basic keyword-based alternatives in classic NLP modeling
- Efficient storage and compute profiles of embeddings make them easily infusible into generative AI applications

Learn more in documentation. See list of supported embedding models.


Embedding Service Phase 1
REST HTTPS
Embedding Inference
Python
LangChain
[“man”, “woman”, “king”, “queen”, “monarch”]
$[1.67, -0.20, -0.29, \ldots, -0.51, 1.41]$
$[0.33, -0.62, -0.18, \ldots, -0.76, -0.36]$
$[1.66, -0.01, -0.99, \ldots, 0.43, 0.21]$
$[1.81, 1.19, -2.16, \ldots, 0.06, -1.25]$
$[0.84, 1.52, 0.51, \ldots, 0.12, -0.87]]$



watsonx.ai Text Rerank API

Add precision to RAG use cases by improving the order of retrieved information with advanced re-ranking algorithms.

The Text Rerank API helps ensure relevant embeddings are used as context within a RAG pipeline, significantly enhancing performance.

- Reorder document passages (from most-to-least likely to answer) based on their similarity to a specified query
- Powered by the ms-marco-minilm-l-12-v2 reranker model (see details)
- Fully integrated with the watsonx.ai REST API

Learn more in documentation.

watsonx.ai Chat API

Create dynamic and adaptive conversational interfaces including chat-based application that require grounding documents, images, and tool calling for agent-driven applications.

- Build multi-user conversational workflows that use foundation models to generate answers.
- Easily identify different message types, such as a system prompt, user inputs, and foundation model outputs, including user-specific follow-up questions and answers.
- Supports granite, llama and mistral models (see supported list)
- Fully integrated with the watsonx.ai API & SDK

Learn more about adding chat functions and building agents in documentation. See the developer hub for examples.

Agents

Propel the next wave of AI productivity through the developer

Innovation

1/3

of gen AI interactions will use action models and autonomous agents by 2028[1]

Scale

1B

net new applications will be fueled by gen AI by 2028[2]

Opportunity

72%

of top performing CEOs say their competitive advantage rests in advanced generative AI[3]

[1] Gartner Press Release, 11 March 2024.
[2] 2024 Developer Survey, Stack Overflow, May 2024.
[3] 6 Hard Truths CEOs Must Face, IBM Institute for Business Value, May 2024

Three areas of agentic innovation

Pre-built agents

watsonx Orchestrate
Accelerate AI agent deployment.

Get started quickly with prebuilt AI agents powered with business logic and seamless integration to the tools that power your business.

Custom-built agents

watsonx.ai
Build custom-designed agents.

Design, deploy and manage AI agents with ease using pro- and low-code options.

Multi-agent orchestration

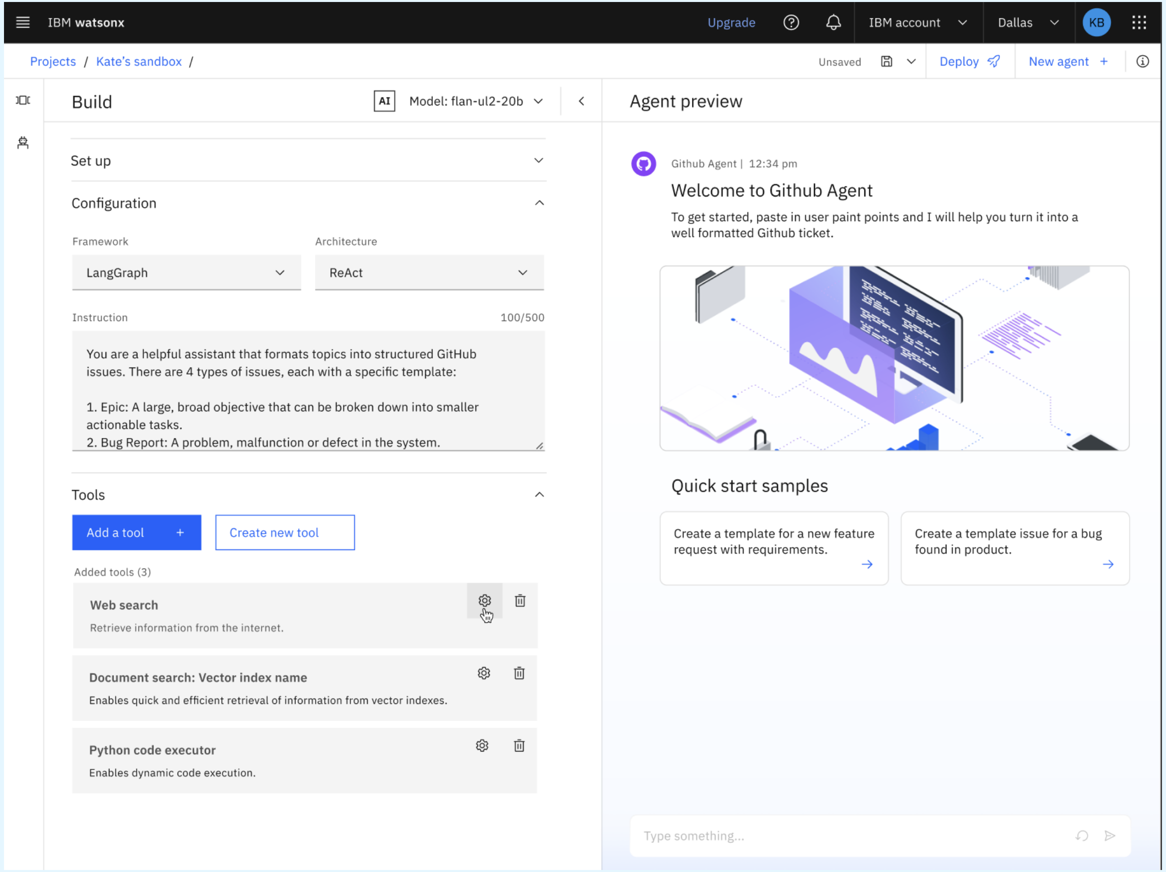
watsonx Orchestrate
Manage all agents in one place.

Easily deploy and manage any agent for any task within a simple and unified user experience optimized to scale.

watsonx.ai Agents

What's available today?

- Support for Chat API, Tool Calling, JSON returns together with providing the foundation for agentic support
- AI Services that provide the foundation for deployment of python based agentic solutions
- End-to-end template for the development and deployment of a custom Langraph agents
- Introduction of Agent Lab (Beta), enabling a low-code experience to build agentic solution with 1-click deployment as a production-ready API endpoint



Model: flan-ul2-20b

Framework: LangGraph

Architecture: ReAct

Instruction 100/500

You are a helpful assistant that formats topics into structured GitHub issues. There are 4 types of issues, each with a specific template:

1. Epic: A large, broad objective that can be broken down into smaller actionable tasks.
2. Bug Report: A problem, malfunction or defect in the system.

Added tools (3)

Web search
Retrieve information from the internet.

Document search: Vector index name
Enables quick and efficient retrieval of information from vector indexes.

Python code executor
Enables dynamic code execution.

Github Agent | 12:34 pm

Welcome to Github Agent

To get started, paste in user paint points and I will help you turn it into a well formatted Github ticket.

Create a template for a new feature request with requirements.

Create a template issue for a bug found in product.

watsonx.ai SDKs

Build agentic services using popular open-source agentic frameworks through industry-standard API and SDK support.

LangGraph

crewai

LangChain

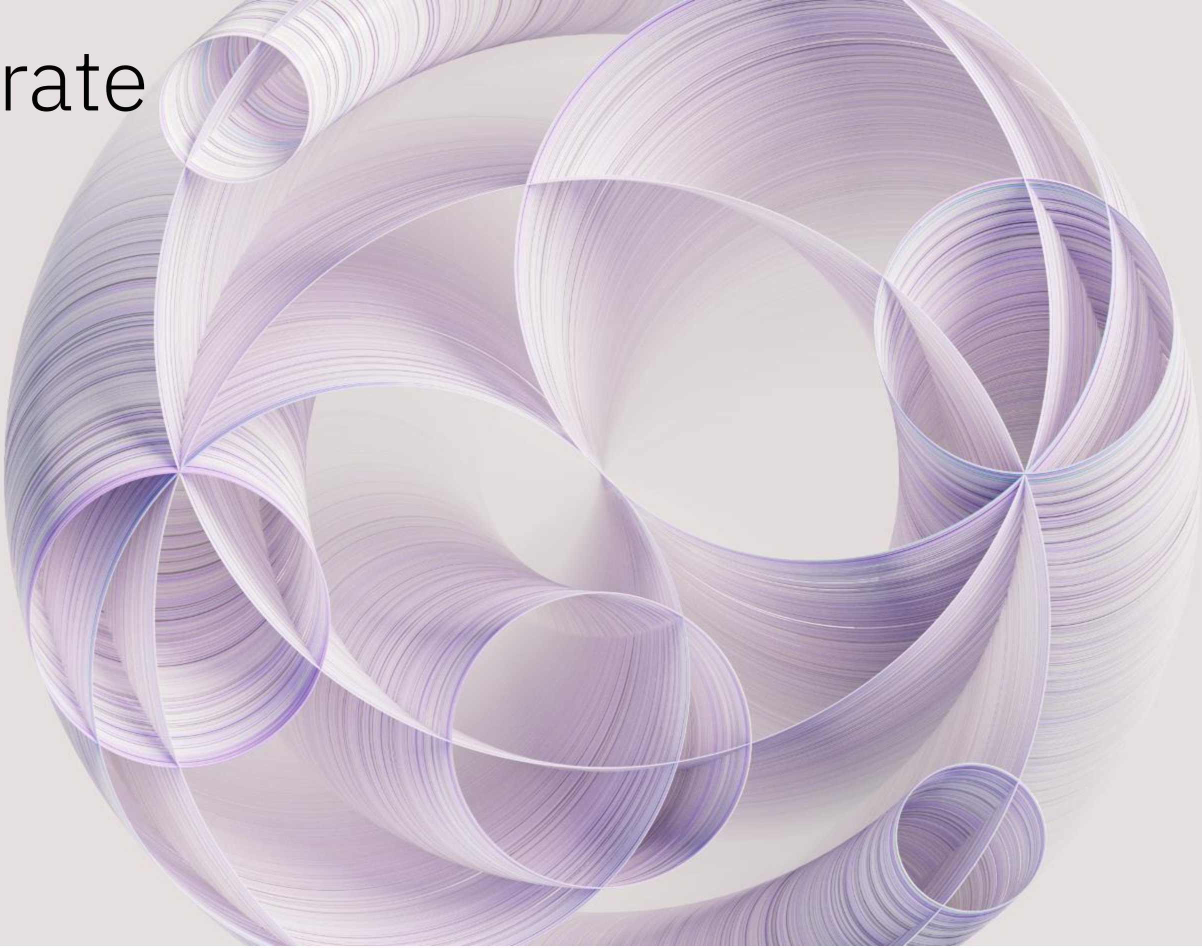
LlamaIndex

Bee

Lab 1

Watsonx Orchestrate

The future of work with
AI Assistants and Agents

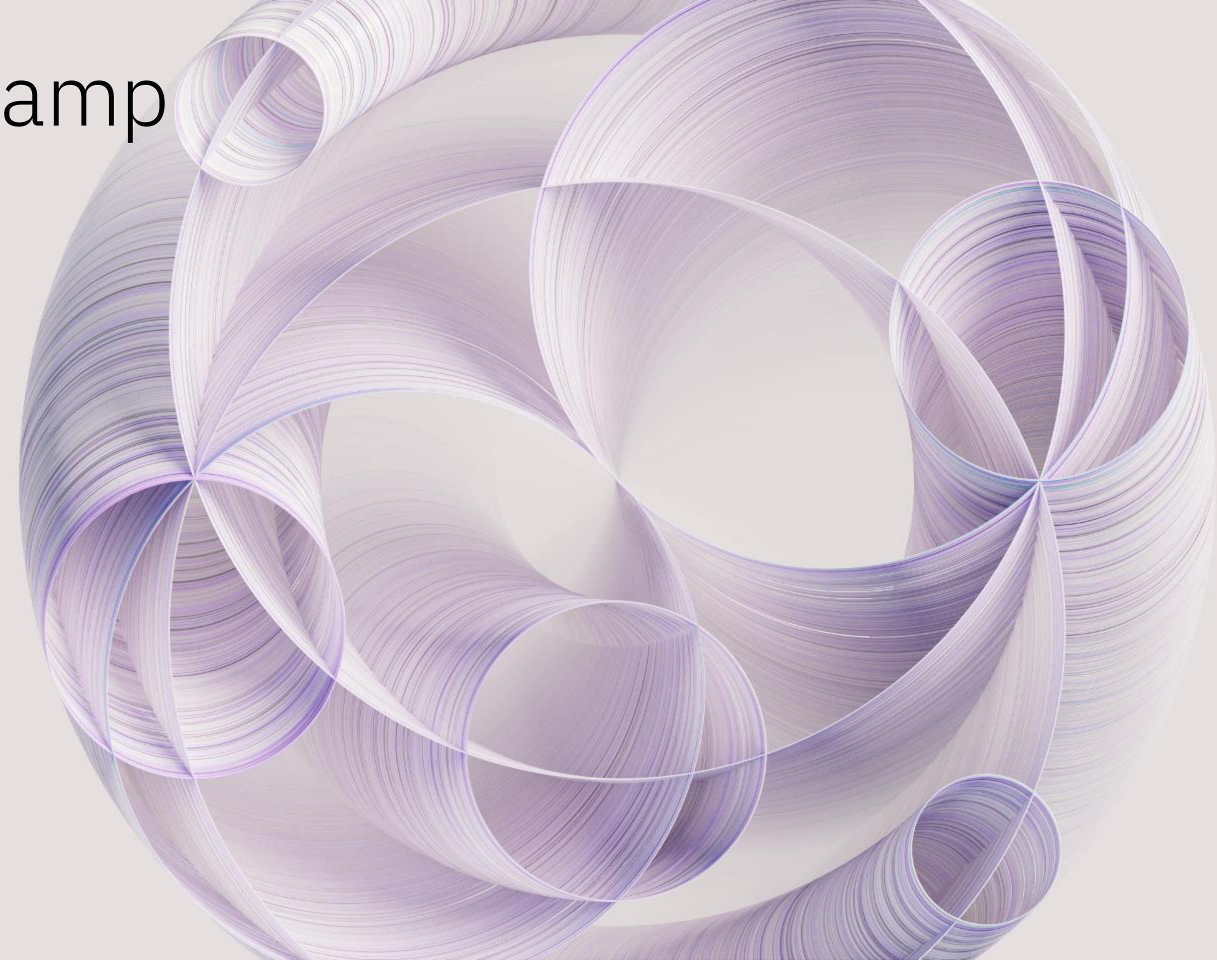


Sebastián Fripp

IBM

Lab 2

AI Agentic Bootcamp



DAY 2

Sebastián Fripp & Facundo Iraola

IBM Client Engineering
Ecosystem

Contáctanos

Gabriela Retamosa

BTL | Client Engineering – Ecosystem
SSA + MX
gabyretamosa@uy.ibm.com

Sebastián Fripp

AI Engineer | CE-Ecosystem
SSA + MX
sfripp@ibm.com

Facundo Iraola

AI Engineer | CE-Ecosystem
SSA + MX
fid@ibm.com