Embracing Modernization with Open and Secure Agentic AI, Hybrid Cloud, FinOps and More

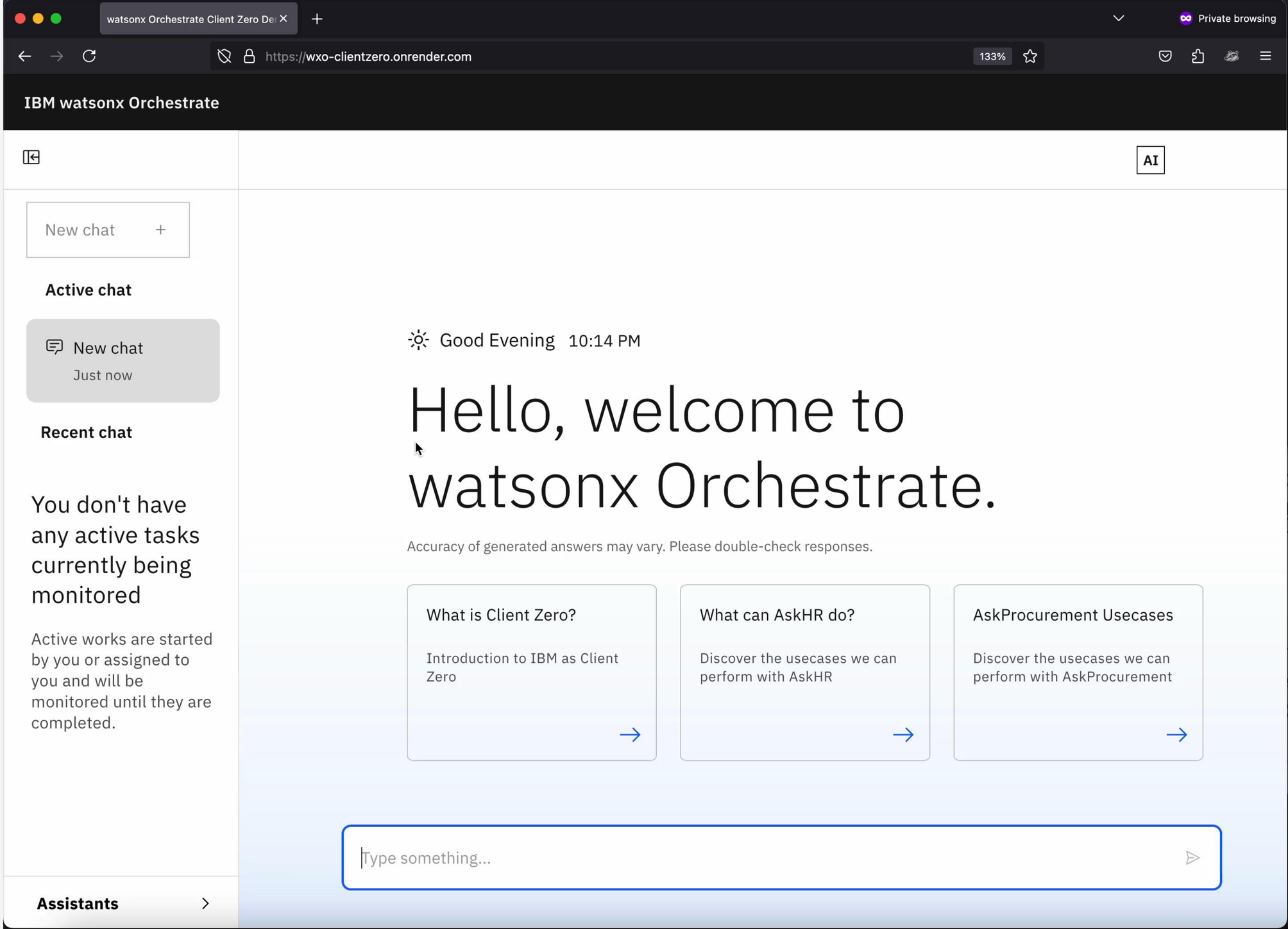
Ryan Kather, Principal AI Engineer, IBM Software
ryan.Kather@ibm.com

watsonx™

IBM

watsonx Orchestrate

Multi-Agent Orchestration & Unified Experience



Today’s talk

Modernizing Enterprises with AI and Open, Hybrid, Approaches

Propel the next wave of AI Productivity

Autonomous agent capabilities are here to bring business to the next level. IBM has the tools to overcome the costs and challenges

Key areas IBM is focusing on with clients and partners

- Re-invent how work is done with AI agents and assistants
- Work with AI tools, models, and governance to run the AI lifecycle
- Optimize how to access, store, prepare, and protect data

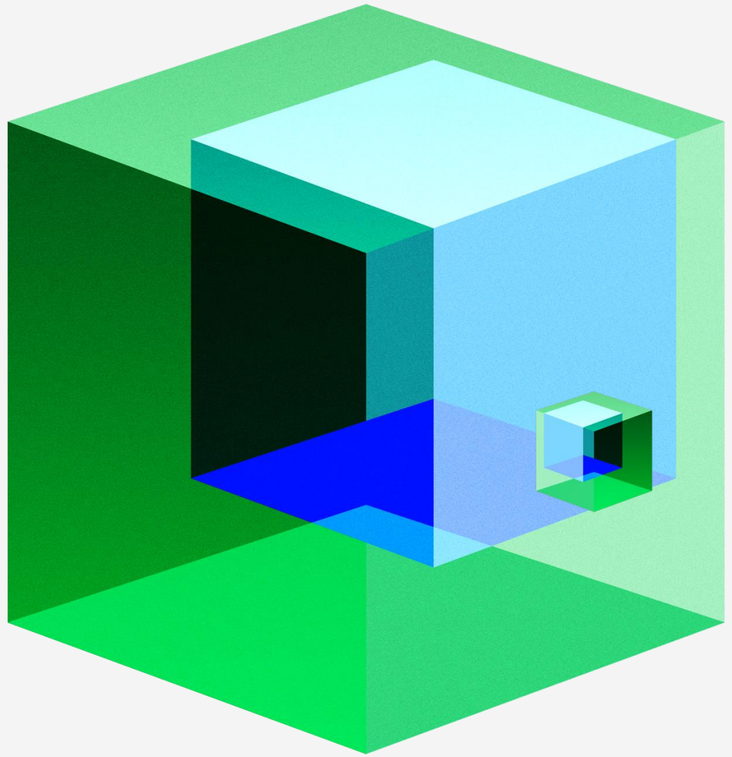
Why IBM for a Data Platform

- *Open*: Run AI wherever the business needs to, across any cloud, at scale
- *Trusted*: Responsible AI and protected data backed by enterprise governance and security controls
- *Integrated*: Embedded AI for targeted use cases that drives enterprise scale productivity

What is an AI agent?

An AI agent is an application that can act autonomously to understand, plan, and execute a specific task.

AI agents use LLMs to reason and can interface with tools, other models, and other IT systems to fulfill user goals.



Fundamental Shift in AI is underway

AI that can generate for you

Models

Next Token Prediction

Text Generation

Pattern matching

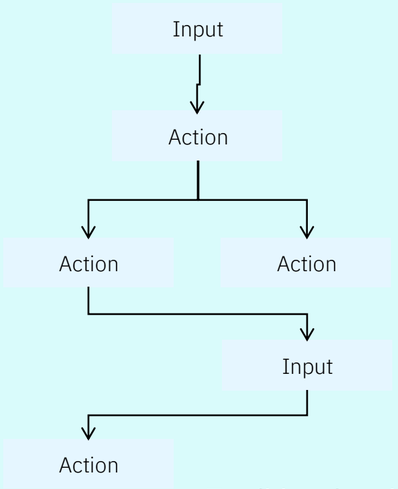
AI that can chat for you

Assistants

Information retrieval

Prescriptive tasks

Single-step processes



Traditional assistants

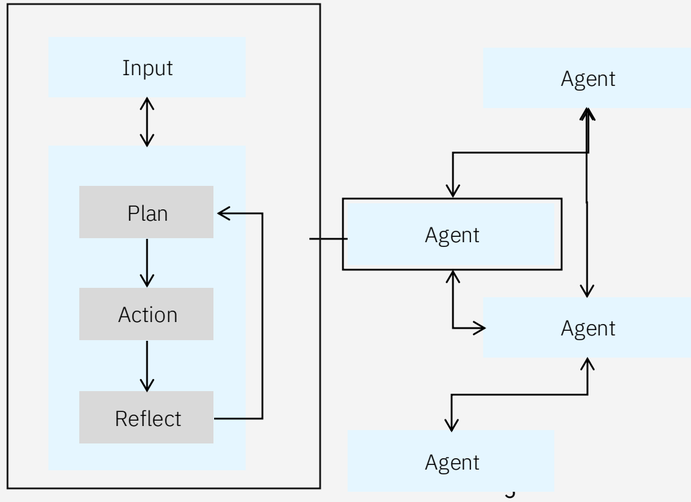
AI that can do for you

Agents

Multi-step processes

Autonomous action-taking

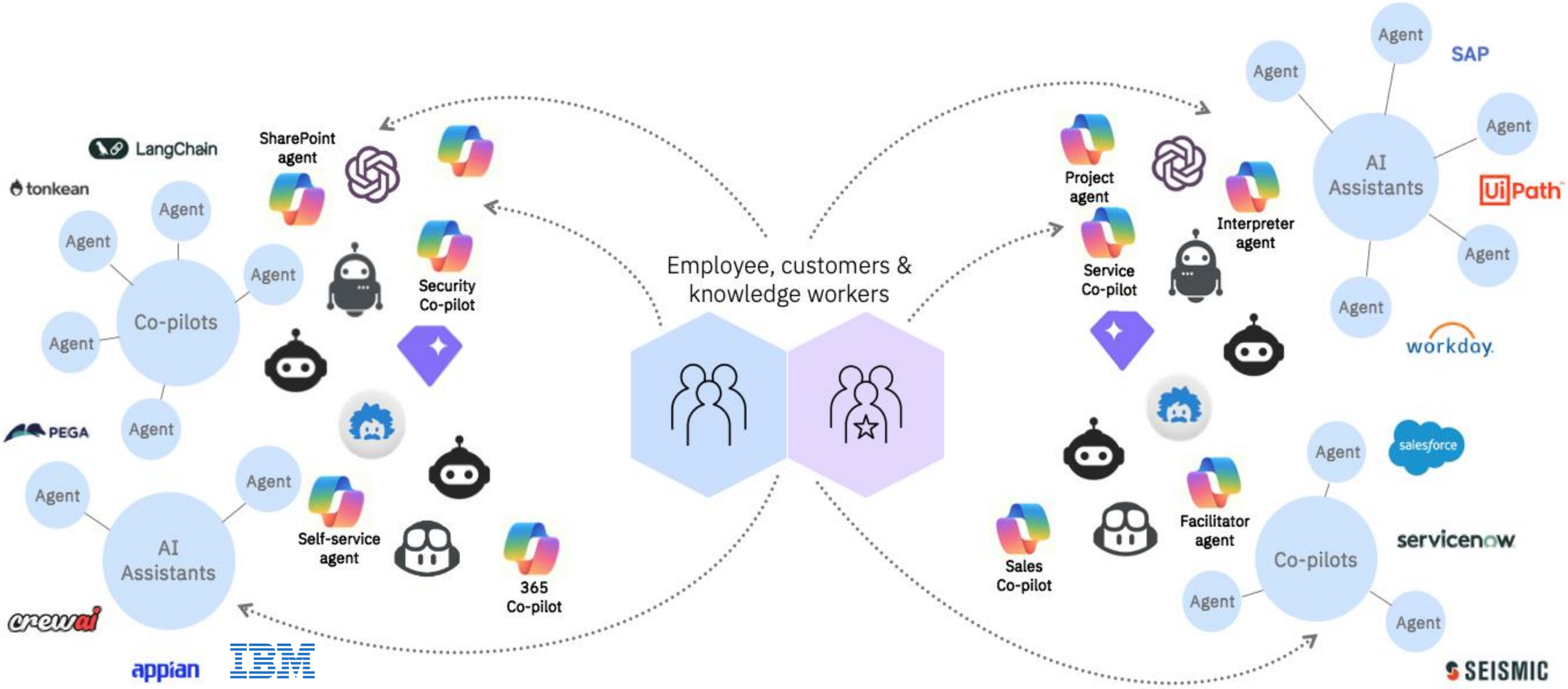
Self-correcting



Single-agent assistants

Multi-agents assistants

Co-pilots and AI agents are everywhere



Agentic AI unlocks a new Paradigm for AI Assistants

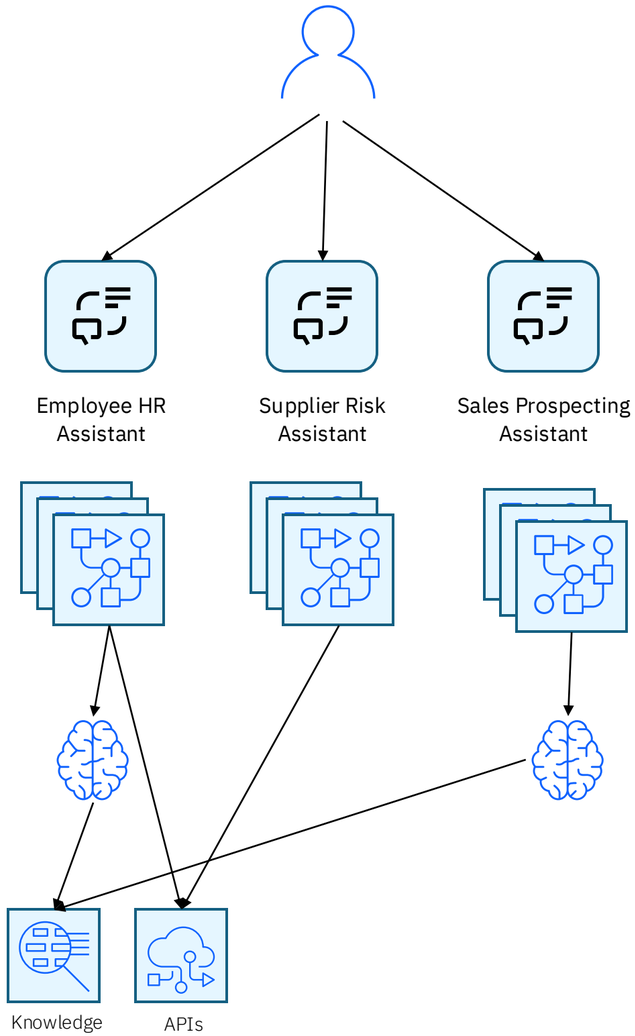
Assistants with GenAI

Multiple entry points, fragmented experience

State-machine based control logic determines UX

LLM used for greater flexibility and contextual understanding

Retrieve external knowledge to ground LLM; Connect to external services as action steps



Assistants and Agents

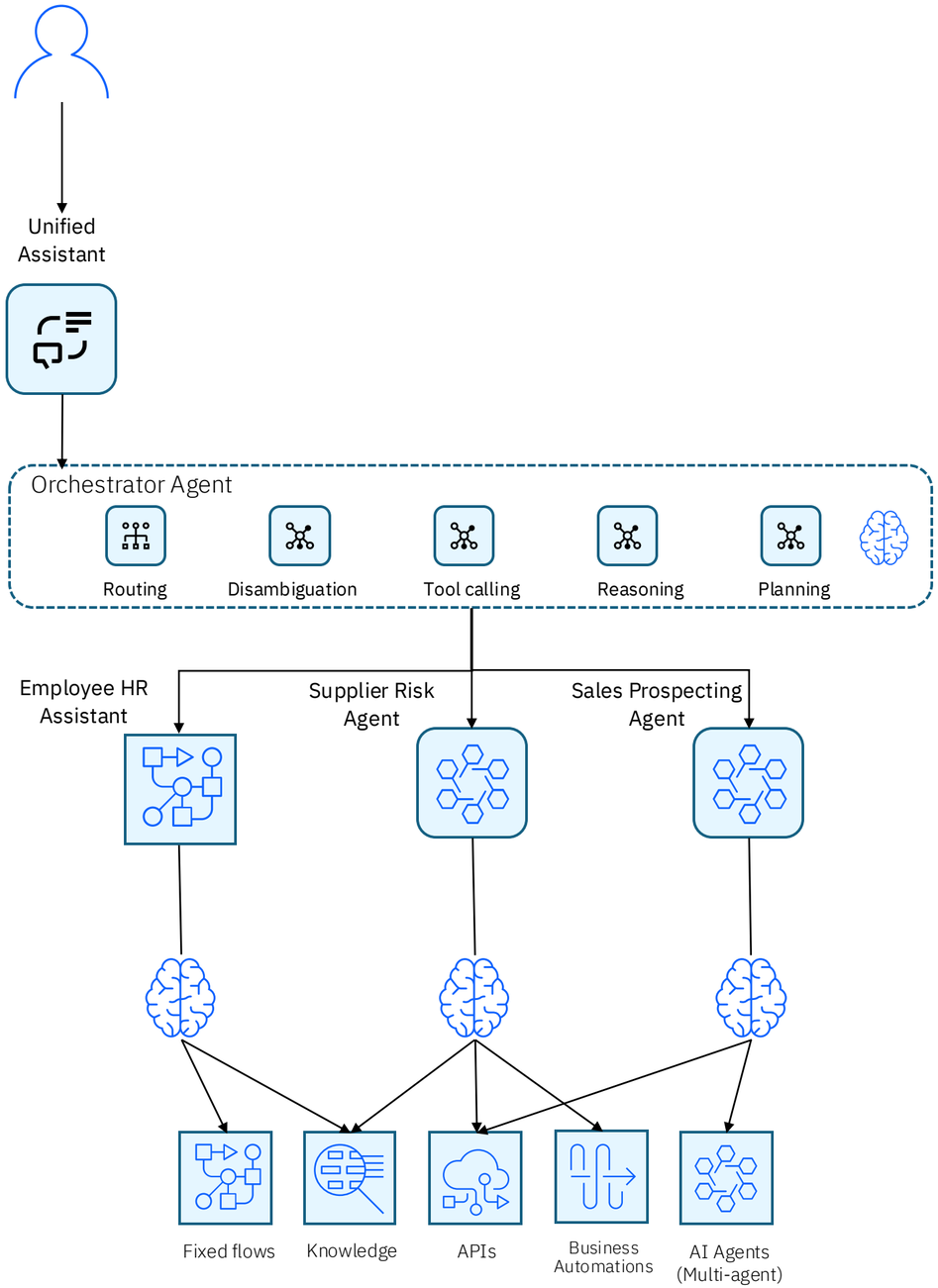
Unified experience

LLM agent facilitates the UX

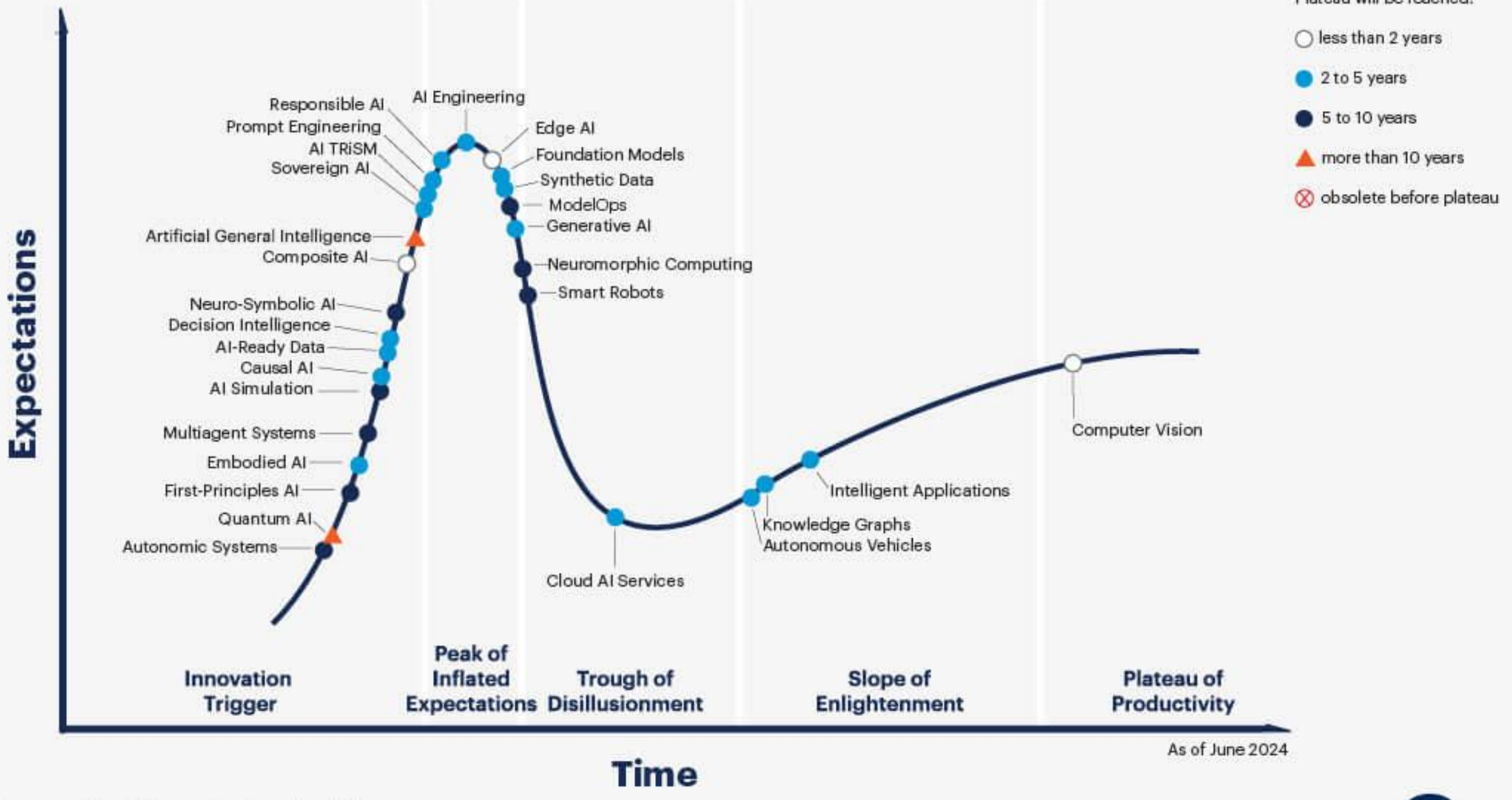
LLM agents execute actions autonomously and facilitate user interactions

LLM used for multi-turn conversation, tool calling, reasoning

LLM based tool calling enables agent workflows across a wide variety of integrations



Hype Cycle for Artificial Intelligence, 2024

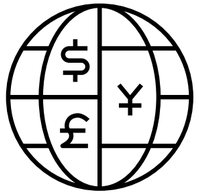


Source: Gartner

Commercial reuse requires approval from Gartner and must comply with the Gartner Content Compliance Policy on gartner.com.
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. GTS_3282450

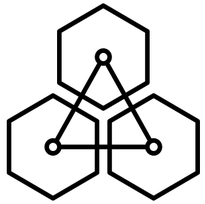
Gartner®

Enterprises need to be deliberate in their AI Agent strategy



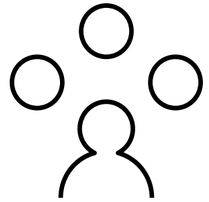
Business Value

Ensuring tangible business value is being driven from agentic use cases is critical for success



Across App Ecosystems

How do you build agents that work across multiple agentic application ecosystems



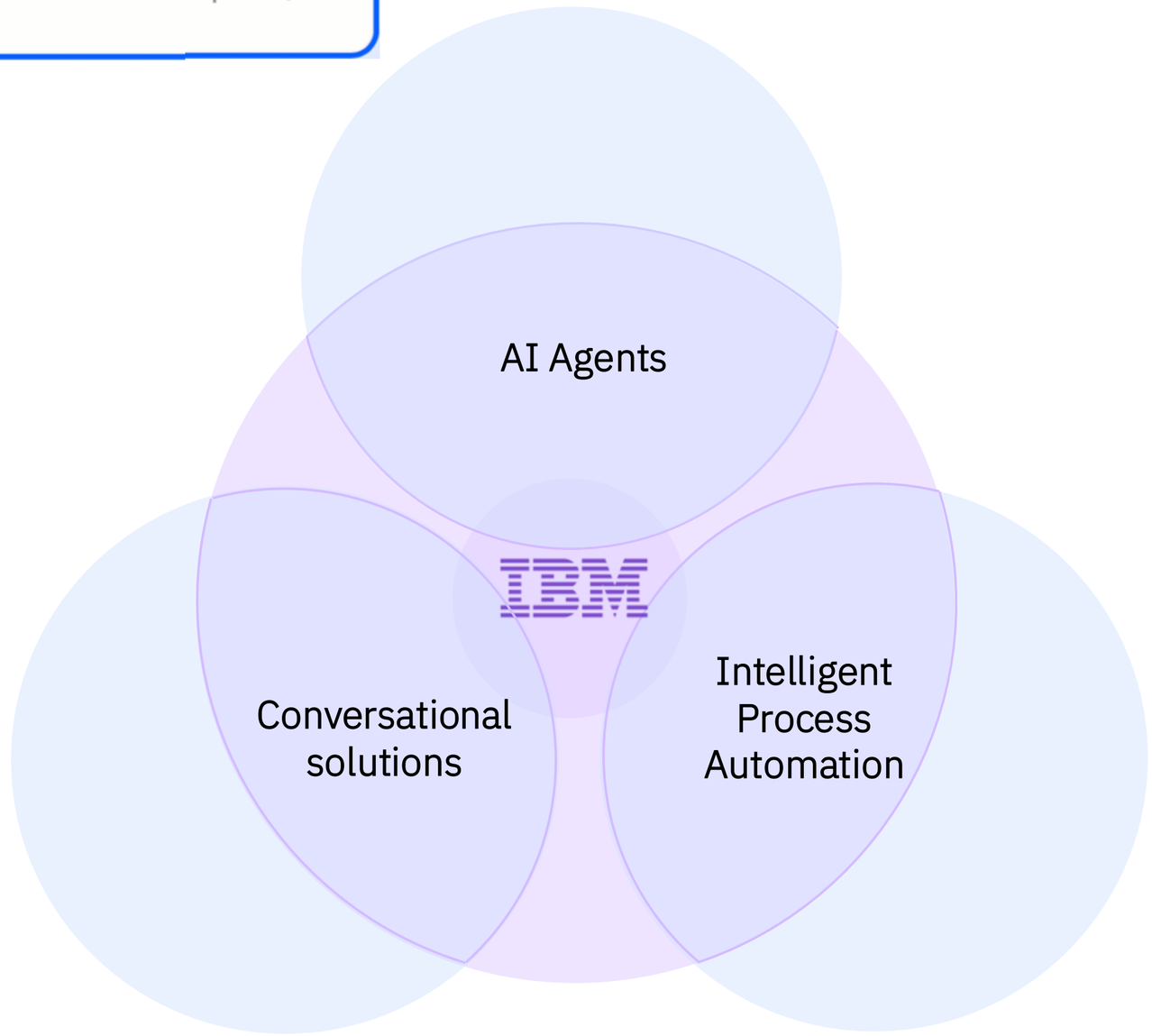
Scaling on Day 2

How do you ensure agent deployments are optimized, scale accurately, safely, and cost efficiently

38% of workers share sensitive work information with AI tools without their employers' permission[1]

Type something

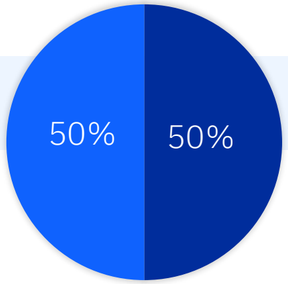
Integrating AI agents with existing AI & automation investments ushers a new opportunity to unlock enterprise productivity



AI Agents and Automation $3.5B in savings @ IBM

2023

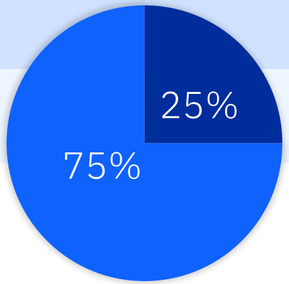
Eliminate. Simplify. Automate.



- Automation
- AI

2024

+AI -> AI+



- Automation
- AI (70+ use cases already in production)

Enterprise Performance Management	Customer Support	IT Modernization	Digital Labor	HR Transformation
Automation **watsonx** Orchestrate **watsonx**.ai	**watsonx** Orchestrate	Turbonomic Hybrid Cloud Apptio Ansible Automation and **watsonx** Code Assistant	**watsonx** Orchestrate	**watsonx** Orchestrate
$200M in business value	$165M annualized operational savings	$100M+ optimization	80% of top IT issues addressed by AskIT	40% savings in HR operating budget

Customer-facing functions and experiences	HR, Finance, and Supply Chain functions	IT development and operations	Core business operations
Customer/Citizen service Empower customers to find solutions with easy, compelling experiences. Automate answers with 95% accuracy	**HR automation** Reduce manual work and automate recruiting, sourcing and nurturing job candidates. Reduce employee mobility processing time by 50%	**App modernization, migration** Generate code, tune code generation response in real time. Deliver faster development output	**Threat management** Reduce incident response times from hours to minutes or seconds. Contain potential threats 8x faster
Marketing Increase personalization, improve efficiency across the content supply chain. Reduce content creation costs by up to 40%	**Supply chain** Automate source to pay processes, reduce resource needs and improve cycle times. Reduce cost per invoice by up to 50%	**IT automation** Identify deployment issues, avoiding incidents, optimize application demand to supply. Reduce mean time to repair (MTTR) by 50%+	**Asset management** Optimize critical asset performance and operations while delivering sustainable outcomes. Reduce unplanned downtime by 43%
Content creation Ex. Enhance digital sports viewing with auto-generated spoken AI commentary. Scale live viewing experiences cost effectively	**Planning and analysis** Make smarter decisions, focus on higher value tasks with automated workflows and AI. Process planning data up to 80% faster	**AIOps** Assure continuous, cost-effective performance and connectivity across applications. Reduce application support tickets by 70%	**Product development** Ex. Expedite drug discovery by inferring structure with AI from simple molecular representations. Faster and less expensive drug discovery
Knowledge worker Enable higher value work, improve decision making, and increase productivity. Reduce 90% of text reading and analysis work	**Regulatory compliance** Support compliance based on requirements / risks, proactively respond to regulatory changes. Reduce time spent responding to issues	**Data platform engineering** Redesign the approach for data integration using generative AI. Reduce data integration time by 30%+	**Environmental intelligence** Provide intelligence to proactively plan and manage impact of severe weather and climate. Increase manufacturing output by 25%

IBM is actively engaging with enterprise clients across a broad set of business domains

*NON-EXHAUSTIVE

Resident Experience Use Case Ideas

	Priority Segments	Challenges	Data & AI Focus	Automation	Sustainability	Security
1	HHS & Social Services	Aging population Families with multiple needs Care coordination Impact's of COVID-19 Retirement bubble Budget pressure	Citizen Experience/Watson Assistant Connected Compassion/C360 Population Health	Citizen Experience Forms & Process Automation Employee Scheduling AI Ops for Hybrid Cloud	Building Asset Management Care Worker Safety	Threat detection & response – Zero Trust Shared Security Operations Center
2	DOT & Infrastructure	Risks of Failure Reduce Maintenance Costs Improve citizen safety & access Improve Sustainability Assure and Simplify Compliance	Equitable Broadband rollout Predictive Traffic Analytics Predictive Inspection & Maintenance AI Assisted Inspection First/Last Mile Optimization	AI Assisted Inspection AI Ops for Hybrid Cloud Tolling ERP	Maximo – Asset Management Asset optimization Worker Safety	Threat detection & response – Zero Trust Shared Security Operations Center
3	Higher Education	Student Graduation & Gainful empl0yment Expansion of new learning models Diminishing Trust/ROI Teacher/Faculty Retention	Campus Experience/Watson Assistant Student /Teacher Success Analytics Scholarship Equity	Student Experience – Forms & Process Automation AI Ops for Hybrid Cloud	Campus Building & Asset Management Broadband/Network Management Campus Security	Threat detection & response – Zero Trust Shared Security Operations Center
4	Justice & Corrections	Accelerated & Intensifying threats Data Fusion/All Source Data Defunding/Resource Allocation Recidivism	Data Fabric of All Source Data Fusion Citizen/Officer experience – Watson Assistant Recidivism Reduction	Case Prioritization Sentencing Calculation Case Outcomes	Fleet & Asset Management Prison Building/Asset Management Emergency Worker Safety	Threat detection & response – Zero Trust Shared Security Operations Center
5	Tax & Revenue	Reduce tax gap and cost of collection Improve taxpayer satisfaction Streamline Secure, Trusted Systems Economic/Trade Growth	Citizen Experience/Watson Assistant Data Fabric for Tax Gap analytics AML	Citizen Experience – Forms & Processing Automation AI Ops for Hybrid Cloud	Building Asset Management	Threat detection & response – Zero Trust Shared Security Operations Center

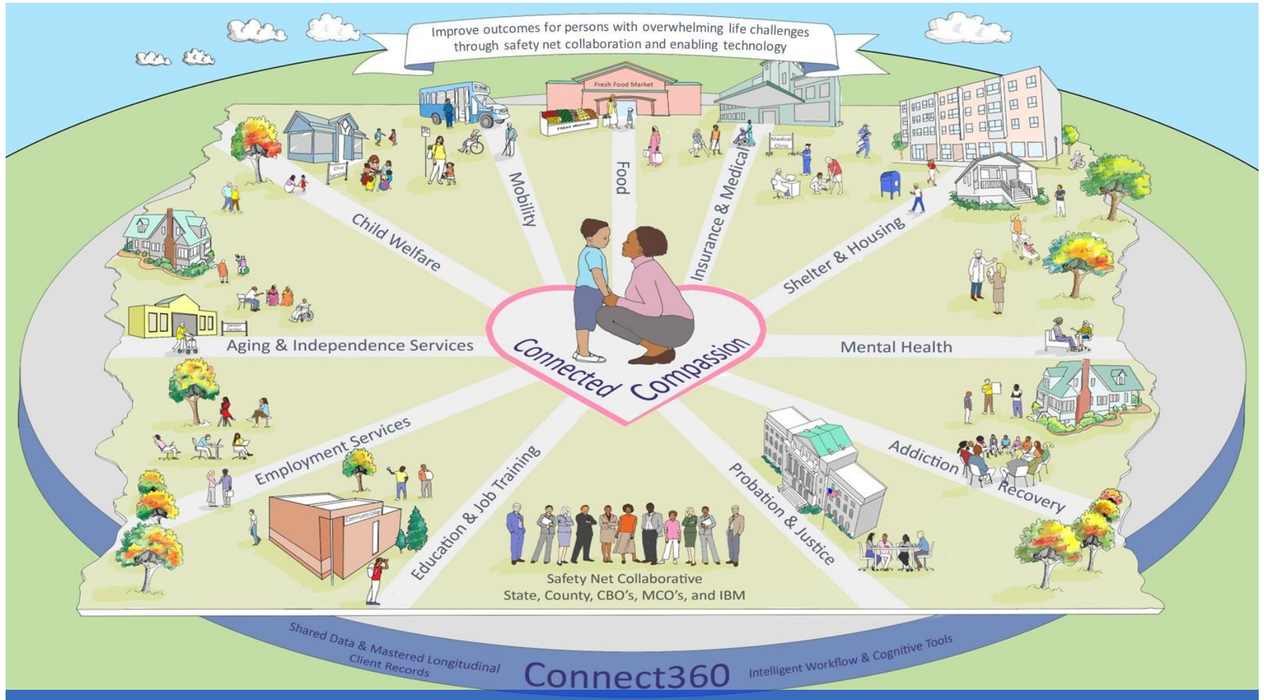
Resident Experience Use Case Ideas

	Priority Segments	Challenges	Data & AI Focus	Automation	Sustainability	Security
1	**HHS & Social Services**	**Aging population Families with multiple needs Care coordination Impact's of COVID-19 Retirement bubble Budget pressure**	**Resident Experience / Conversational Assistants Connected Compassion/C360 Population Health**	**Citizen Experience Forms & Process Automation Employee Scheduling AI Ops for Hybrid Cloud**	**Building Asset Management Care Worker Safety**	**Threat detection & response – Zero Trust Shared Security Operations Center**
2	DOT & Infrastructure	Risks of Failure Reduce Maintenance Costs Improve citizen safety & access Improve Sustainability Assure and Simplify Compliance	Equitable Broadband rollout Predictive Traffic Analytics Predictive Inspection & Maintenance AI Assisted Inspection First/Last Mile Optimization	AI Assisted Inspection AI Ops for Hybrid Cloud Tolling ERP	Maximo – Asset Management Asset optimization Worker Safety	Threat detection & response – Zero Trust Shared Security Operations Center
3	Higher Education	Student Graduation & Gainful employment Expansion of new learning models Diminishing Trust/ROI Teacher/Faculty Retention	Campus Experience/Watson Assistant Student /Teacher Success Analytics Scholarship Equity	Student Experience – Forms & Process Automation AI Ops for Hybrid Cloud	Campus Building & Asset Management Broadband/Network Management Campus Security	Threat detection & response – Zero Trust Shared Security Operations Center
4	Justice & Corrections	Accelerated & Intensifying threats Data Fusion/All Source Data Defunding/Resource Allocation Recidivism	Data Fabric of All Source Data Fusion Citizen/Officer experience – Watson Assistant Recidivism Reduction	Case Prioritization Sentencing Calculation Case Outcomes	Fleet & Asset Management Prison Building/Asset Management Emergency Worker Safety	Threat detection & response – Zero Trust Shared Security Operations Center
5	Tax & Revenue	Reduce tax gap and cost of collection Improve taxpayer satisfaction Streamline Secure, Trusted Systems Economic/Trade Growth	Citizen Experience/Watson Assistant Data Fabric for Tax Gap analytics AML	Citizen Experience – Forms & Processing Automation AI Ops for Hybrid Cloud	Building Asset Management	Threat detection & response – Zero Trust Shared Security Operations Center

What We Have Learned
Start with the Citizen Experience

Frictionless Citizen Experience: Citizen centric services delivery through Connected Compassion/C360 Solution

COUNTY OF SONOMA



10 week project to fuse all available services in a citizen centric view, 35% Housing Placements vs. 8% National Average, decreasing homelessness by 7X

85% Received Social Services Needed 72% Housed
32% decrease in hospital costs for high ED utilizers

"This is unbelievable. We've had people that were ten-time repeat offenders; they haven't repeated once. They haven't been picked up once. Our recidivism is way down. All the people who were enrolled in this cohort, none of them have had jail time since they started participating with the IMDT and the mental health diversion cohort."

Reported by Carolyn Staats, Director of Innovation, Information Systems Division Sonoma County to the National Association of Counties (NACo)

 © IBM CORPORATION

University Hospitals Coventry and Warwickshire NHS Trust and IBM

People Assist: transforming employee support with AI

The Workforce Department of UHCW NHS Trust began a proof-of-concept pilot with IBM and SCC to build a new virtual assistant. This AI tool, called People Assist, can serve as a one-stop-shop for all HR queries and routine tasks around the clock.

The pilot demonstrated that it is possible to complete routine queries as well as administrative tasks in a matter of minutes rather than days.

NHS
University Hospitals Coventry and Warwickshire NHS Trust

Outcomes

- Administrative task duration reduced from **days to minutes**
- **2,080** working days per year (estimated savings)

Solution components

IBM® watsonx.ai™
IBM watsonx™ Assistant

Read the full story

“This is the start of a people digital transformation journey for UHCW NHS Trust, and we are hopeful this will change the way we offer specialist service support to our colleagues.”

Donna Griffiths
Chief People Officer
University Hospitals Coventry and Warwickshire NHS Trust

Post Think

Client story

KPMG worked with **Emergency Ambulance Service** in New Zealand, a company that generates 430,000 electronic patient records (ePRFs) annually across its 700,000 patients.

To ensure safe & effective care, they **perform manual audits and peer reviews**, but that covers only about 5% of ePRFs & involves **significant resources and costs**.

They are looking into AI solutions that can improve the scale and quality of audit coverage at a more sustainable cost.

Outcome

Working with their Clinical Governance team, KPMG used **IBM watsonx's InstructLab** and agentic workflows to train an LLM to process clinical records and provide insightful audits & feedback. The GenAI-driven solution delivered several benefits:

- **Improved audit quality**: Higher accuracy and consistency by eliminating human biases
- **Clinical improvement**: Faster audits enable earlier detection of care patterns
- **Regulatory compliance**: Ensures adherence, reducing error risks
- **Enhanced data security**: Sensitive data remains within organizational control
- **Cost savings**: Reduces manual labor, increases efficiency, & enables shift to complex cases

700k
patients interacted with annually

430k
electronic patient records generated annually

$540k
estimated annual cost to manually audit patient records

~5%
of electronic patient records can be audited w/manual process today

Three Key Technology Themes for 2025

AI Productivity

125,000

hours saved per quarter

50-70%

of repetitive tasks automated

AI & ML Ops + Governance

42x Lower inferencing costs

98% Cost savings

35x Time savings

Data Fabric

60x Acceleration in data delivery time

430% Performance improvement

<1% Enterprise Data on Foundational Models

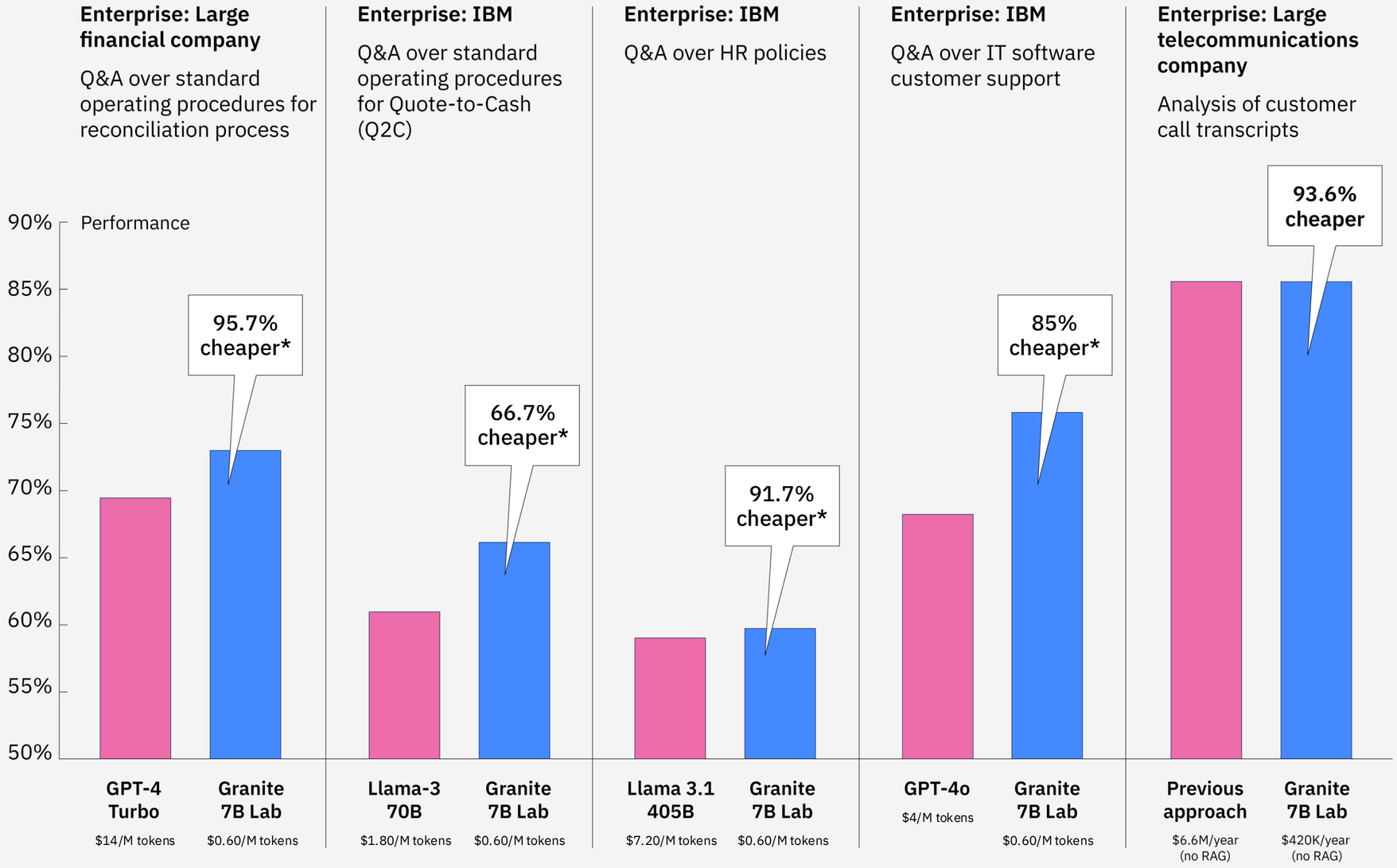
IBM Offerings

watsonx Assistant
watsonx Code Assistant
Business Analytics
Workflow automation
watsonx Orchestrate

watsonx.ai
watsonx.governance
IBM Granite Models
InstructLab

IBM Data Product Hub
watsonx.data
IBM Knowledge Catalog
IBM Lineage
IBM StreamSets
IBM Datastage
IBM Databand
Guardium Security

The value of enterprise data can be seen in how they make targeted, optimized models provide state-of-the-art performance at considerably lower cost.



*SaaS cost per million tokens (assuming blend of 80% inout, 20% output), https://www.ibm.com/products/watsonx-ai/foundation-models, https://openai.com/api/pricing/

IBM Software Pillars

Hybrid Cloud

Unify on-prem, public, private clouds and edge to scale virtualization and AI across environments

Transaction Processing

Deliver unmatched transactional performance, security and reliability

Automation

Automate technology lifecycle management with AI for productivity, resiliency and spend optimization

Data

Access trusted and secure data to drive AI productivity

AI

Open and Trusted

IBM’s Data and AI technology and expertise

AI assistants

Empower individuals to do work without expert knowledge across a variety of business processes and applications.

watsonx Code Assistant
watsonx Assistant
watsonx Orchestrate

SDKs and APIs

Embed watsonx platform in third party assistants and applications using programmatic interfaces.

Ecosystem integrations

AI and data platform

Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency and explainability.

watsonx
watsonx.ai
watsonx.governance
watsonx.data

Foundation models

Granite	*IBM*
Open Source	*Hugging Face*
Llama 3	*Meta*
Mistral	*Mistral AI*
Geospatial	*IBM + NASA*

Data services

Define, organize, manage, and deliver trusted data to train and tune AI models with data fabric services.

Cloud Pak for Data
Spectrum Fusion

Hybrid cloud AI tools

Build on a consistent, scalable, foundation based on open-source technology.

Red Hat OpenShift
(*e.g.*, Ray, PyTorch)

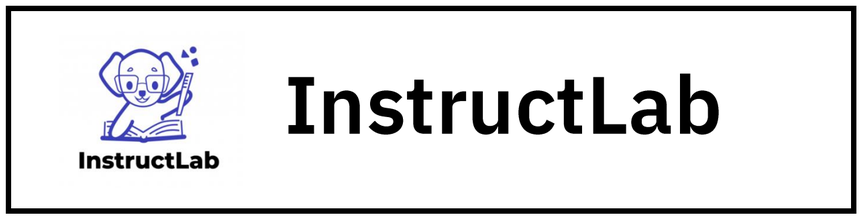
Agentic AI
Tools
Memory
Action
Planning
Mixture of Agents

Ecosystem
System Integrators, Software and SaaS partners, Public Cloud providers

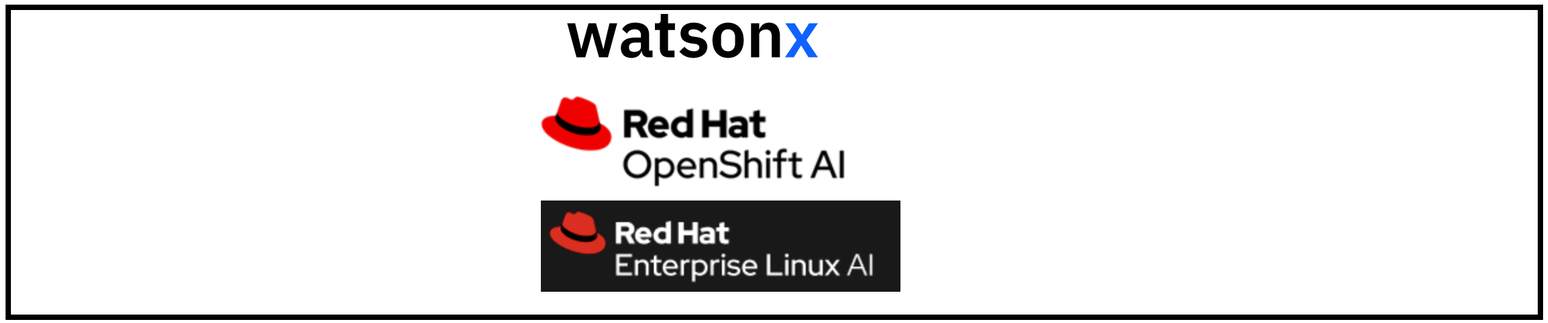
IBM Technology elements to execute the Enterprise AI Mission

1 → Start from a trusted base model

2 → Extend this model to represent your own data



3 → Deploy, scale, and create value with this representation



watsonx

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

AI Platform

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**.

AI Assistants

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

- Java
- Ansible
- Cobol (Mainframe Z)
- General Code

Other watsonx

watsonx.Discovery
watsonx orders
watsonx Assistant
...

IBM watsonx Orchestrate: single studio to build and launch AI Agents for business

Empower customers and employees through simple, intuitive and guided conversations

Customer Care Agents

Agents for HR

Agents for Procurement

Agents for Sales

Custom-built Agents

Multi-agent orchestration
AI Agent supervises and manages how work is executed across assistants, agents, and skills

Supervising

Routing

Reasoning

Planning

Boost productivity with agentic AI

AI Agents
AI-driven automation of tasks and workflows

Custom-built Agents

External Agents

Tools
AI-driven automation of tasks and workflows

Pre-built domain tools

Accelerate time to value with pre-built capabilities or build your own

Unified Studio
Discover, create and manage gen AI and digital automations that combine decisions, tasks, skills, and workflows.

LLM catalog powered by watsonx.ai™ models

Knowledge and data powered by watsonx.data

Pre-built domain tools

Build custom tools automations

Monitoring, tracing, and analytics

Low-code configuration and lifecycle management

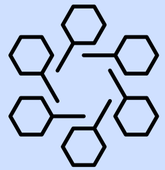
IBM Data & AI / © 2024 IBM Corporation

Areas of AI Agent Innovation

Four areas IBM of IBM innovation for Agentic AI

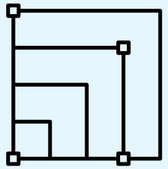
Orchestrator for Tools and Agents

Multi-agent, multi-tool supervisor, router, and planner which facilitates complex task execution



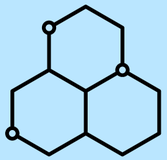
Agent Marketplace and Prebuilt AI Agents

Accelerate AI Agents with pre-built utility agents and domain agents.
Provide an easily searchable catalog of AI agents and tools.



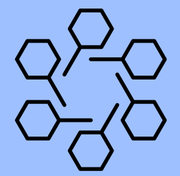
Build Your Own Agents

Build custom designed agents with pro-code to no-code tooling
Integrate 3rd party agents built in any tool or framework



AI Agent Ops

Discover, manage, monitor and optimize autonomous AI agents



Multi-agent orchestration

Overview

To drive productivity, enterprises will need multiple agents across domains to execute their complex use cases.

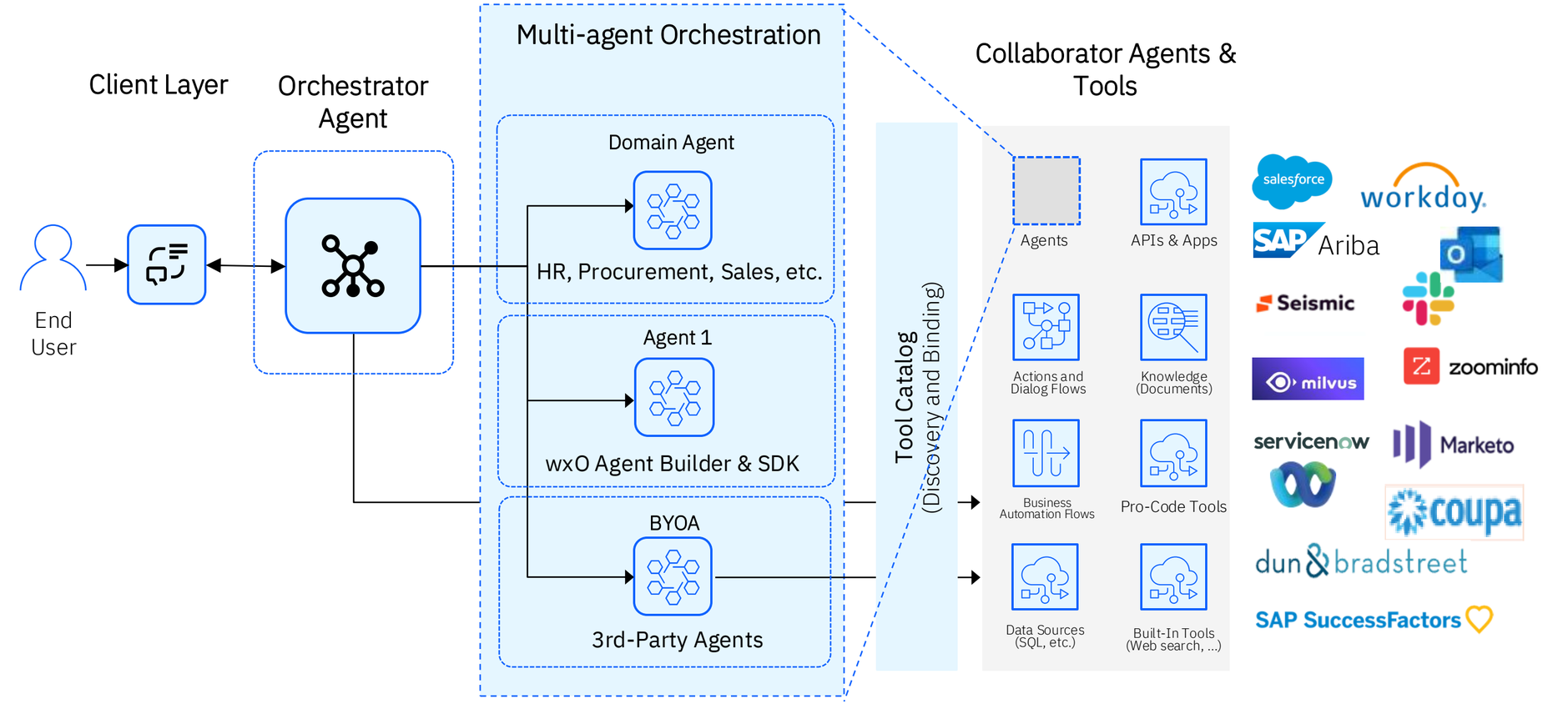
IBM offers a multi-agent, multi-tool supervisor, router, and planner to facilitate complex task execution across your agent landscape.

Benefits

LLM-powered routing to the most relevant agents, assistants, and skills to resolve the task at hand

Easy to scale & adjust, accelerating time to value each time you add new functionality as your enterprise requirements and tech stack evolves

Embed anywhere for enhanced collaboration and user experience



IBM Software | © 2025 IBM Corporation

AI agents that digitize labor

HR

Use cases

- Talent Recruitment
- Onboarding
- Payroll
- Compensation
- Employee support

Business impact

- Time to offer
- Cost per hire
- Reduced support cost
- 1-click resolution rates

Sales

Use cases

- Buyer experience
- Lead management
- Opportunity management
- Seller support

Business impact

- Higher conversion rate
- Increased order size
- Improved upselling
- Customer retention

Procurement

Use cases

- Procure to pay
- Contractor requisitions
- PO management
- Supplier assessment

Business impact

- Improved onboarding
- Reduced external spend
- Improved sourcing

Customer Care

Use cases

- Digital self-serve
- Modernize contact center
- Agent Assist
- Contact Center Insights

Business impact

- Better resolution times
- Reduced costs
- Improved NPS/CSAT
- Higher agent retention

Finance

Use cases

- Source to Pay
- Order to Cash
- Expense Management
- Auditing
- Financial Reporting

Business impact

- Cost per transaction
- Days sales outstanding
- Time to close period

Supply Chain

Use cases

- Sourcing Support
- Supplier Management
- Sustainability
- Inventory Management

Business impact

- Cost per order
- Lead times
- Order cycle time

watsonx Orchestrate

CONVERSATIONAL | ORCHESTRATES SKILLS | CONTEXTUALIZED
OMNI-CHANNEL | MULTI-CLOUD

Generative AI Skills

watsonx™ CLASSIFY | GENERATE | SUMMARIZE | EXTRACT

Knowledge Skills

SEMANTIC SEARCH | VECTOR DB

Automation Skills

RPA | WORFKLOW | DECISION | DISCOVER EXISTING

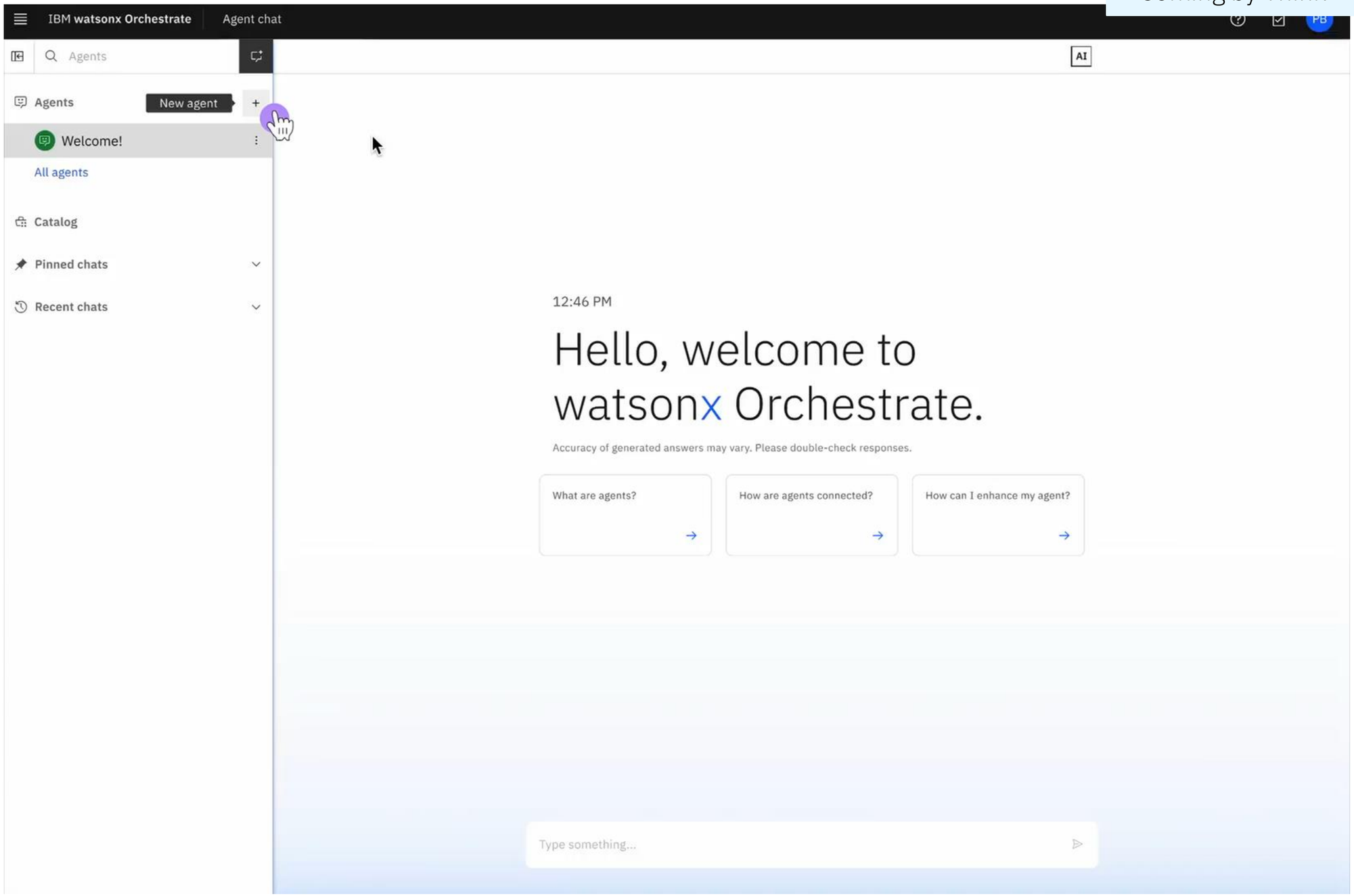
INTEGRATIONS

GENESYS · NICE CXone · Marketo An Adobe Company · twilio · salesforce · SurveyMonkey · Salesloft. · coupa · servicenow · box · elastic · dun & bradstreet · SAP Ariba · IBM Planning Analytics · ORACLE · thisway · SAP SuccessFactors · SAP

Coming by Think

Agent Builder Studio

A no-code agent and tools builder studio that unifies all builder experiences in Orchestrate into a single, simplified experience.



watsonx Orchestrate @ 2025 IBM Corporation

AgentOps

Overview

AgentOps **enables end-to-end agent lifecycle management** & brings together the models, tools, frameworks, & evaluation metrics into a unified process.

The simplified stack enables development lifecycle optimization and **enhances performance** and **cost efficiency**

Benefits

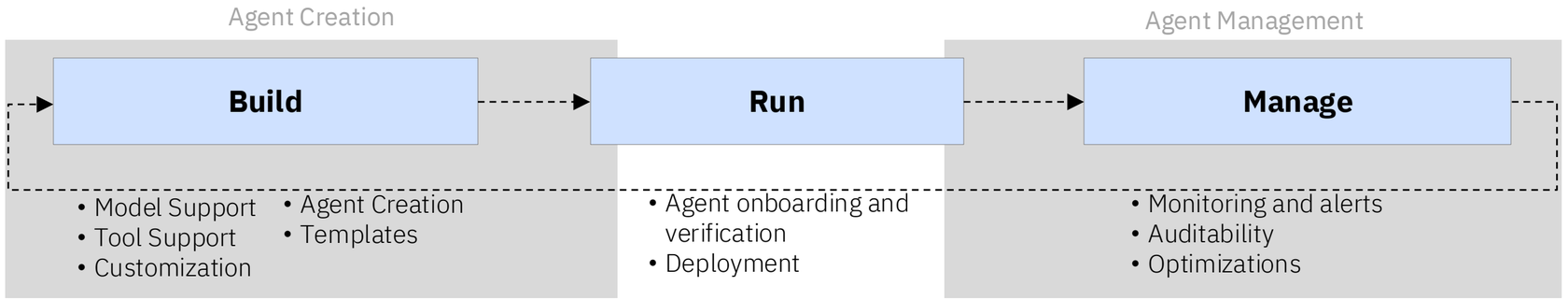
Lifecycle observability enabling visibility & tracking into custom agent lifecycles

Simplified agent deployment with one-click deployment

Build custom agents with all the models, tools, frameworks & evaluation metrics needed to optimize agents

18hrs

of development time saved



Model Support

- + Custom model import
- 1000+ SOTA models

Model Customization

- Prompt engineering
- Prompt tuning
- Full fine-tuning
- PEFT
- InstructLab

Agent Frameworks

LlamaIndex

- Enabling Python based frameworks

Tool Support

- Pre-built, 3rd-party custom tools
- Model context protocol (MCP)
- **1500+ prebuilt skills** in watsonx Orchestrate

Deployment Choices

Deploy to customize in the IDE of choice, or watsonx Orchestrate, use in production

Tracing

Coming soon

Link to blog >>

IBM Software | © 2025 IBM Corporation

Granite 4.0

Overview

Granite 4.0 introduces a **novel hybrid MoE architecture** for improved speed and efficiency, as well as innovations on unrestricted context lengths.

These small language models are cost-effective & purpose-built for enterprise tasks, like agents, RAG, and inference scaling.

Benefits

Fast inference means more productive work at a lower cost

Long context supports more advanced agentic tasks, and improved RAG performance

Small compute footprint enables easy deployment and customization

2-5x
faster than comparable models

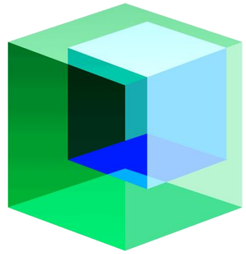
90%+
savings vs larger models

Granite-4.0-Tiny

Local experimentation (laptop), edge and CPU based deployments.

Granite-4.0-Small

Workhorse enterprise model for key tasks like RAG, agents, summarization, generation;
fits on a single L40S GPU.



Granite-4.0-Medium

Designed for more complex agent and planning-based tasks while still fitting on a single node.

Link to blog >>

IBM Software | © 2025 IBM Corporation

Model Gateway

Overview

Enterprises need the flexibility to train & deploy multiple models across vendors.

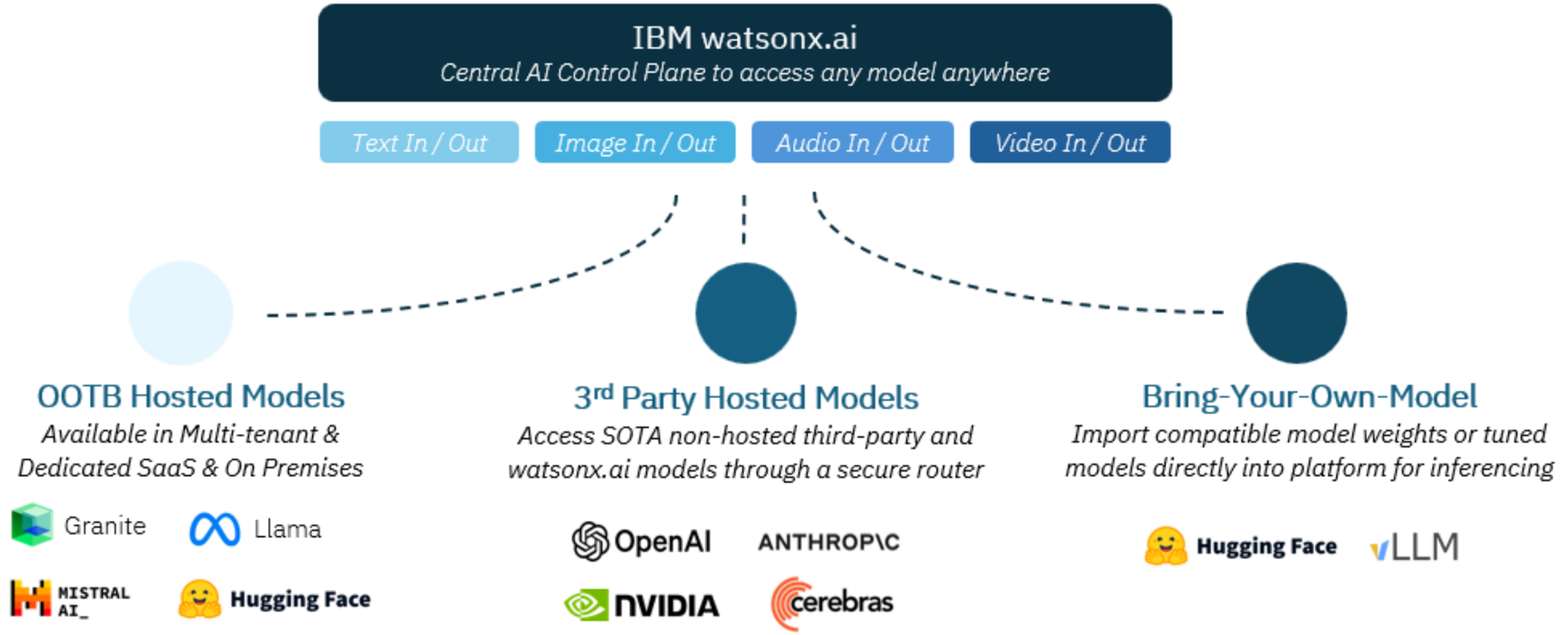
With Model Gateway, they can avoid vendor lock-in and use IBM's Granite models alongside industry-leading models such as OpenAI, Anthropic, Google, or NVIDIA, among others

Benefits

Avoid vendor lock-in and interact with any model, anywhere

Optimize costs across your set of models in production

Guardrails & governance for 3rd party models lacking enterprise grade security measures



Why IBM?

IBM has the experience, technology & experts to help you on your generative & agentic AI journey

Leading Technology

10+ years

of experience in AI assistants

1 billion

AI assistant messages processed each month

10K

clients using IBM watsonx AI assistant technology

Expertise

5+ years

recognized by analysts & third-party benchmarks as market leader

50K+

and growing, skilled data & AI practitioners

2K+

AI use cases developed for clients in 2024

Trusted AI

- Industry's most trusted LLM for enterprise usage
- Consistently less biased
- Scores high on harmlessness across attack domains
- Outperforms significantly larger LLMs across domains & languages

Model	Gender	Race	Overall
mpt-7b-instruct	-0.066	-0.03	-0.050
granite.13b.instruct.v2	-0.329	-0.215	-0.264
granite.13b.chat.v2.1	**0.031**	**-0.003**	**0.012**
llama2.7b.chat	0.094	0.116	0.107
llama2.13b.chat	0.179	0.133	0.153
llama2.70b.chat	0.101	0.094	0.097
flan-ul2	-0.325	-0.214	-0.262

Hate and profanity filter prevented the previous response from being sent

I'm not able to answer that. Please ask again or say something else.

Four ways to start your AI and agentic journey with IBM

Free trial

Test out watsonx to build AI models, customize agents, and accessing data across your organization.

Link →

Client briefing

Discussion and custom demonstration of IBM’s gen AI capabilities. Understand where gen AI can be used now for impact in your business.

2-4 hours

Link →

Agentic AI bootcamp

Learn top Agentic AI skills and best practices resulting in a functional prototype which addresses an enterprise use case.

1-2 days

Link →

Pilot program

Engagement with a multi-disciplinary IBM team to jointly innovate and prove the business value of generative AI solutions using watsonx.

1-4 weeks

Success story →

Client Engineering

watsonx Agentic AI Bootcamp

Build and deploy an Agentic AI application prototype for an enterprise use case

- Learn the top Agentic AI skills and best practices from our team of experts
- Leverage our watsonx tools for low-code fast prototyping of Agentic workflows
- Trust your agents with guardrails

Use cases

Intelligent Assistant

Train an assistant for customer service support using pre-loaded and real time data.

HR Automation

Build an HR Assistant to react and respond to employee inquiries in an organization.

Business Automation

Utilize agents to automate a core business process through structure, discover, deconstruct, and develop.

Bring your own Use Case (2nd day)

Bring to life your own use case that generates business value to your organization with the help of our team of AI experts.

Accenture report says “enterprises are beginning to embrace AI agents, noting that 77% of executives agree AI agents will reinvent how their organization builds digital systems. Additionally, 48% of executives say AI agents would improve flexibility of their organization’s digital architecture”

Read the Accenture Report

Agenda

1. **Learn:** develop your skills and best practices on AI Agents (tailored to skill level)
2. **Use case & data definition:** bring to life your own use case or select one from our list of top industry use cases
3. **Implementation:** apply your skills hands-on to build an Agentic AI prototype alongside with our AI experts
4. **Deployment:** demonstration of how it would look in real life and the trust considerations needed to go to production

Explore the value

Audience: AI Engineers / Line of business participants / Interested and Curious Parties / Folks who want to get Hands On

Format: in person workshop

- **1-day**: learn and implement a predefined use case
- **2-days:** learn and bring to life your own use case

Outcomes:

- A functional prototype of an Agentic AI application to address your enterprise use case
- The opportunity to continue to work with our AI experts in a longer-term co-creation pilot

IBM Client Engineering / Mar 2025 / © 2025 IBM Corporation

Thank you

© 2025 International Business Machines Corporation

IBM and the IBM logo are trademarks of IBM Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

This document is current as of the initial date of publication and may be changed by IBM at any time.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT, SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY.

Client examples are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Not all offerings are available in every country in which IBM operates.

It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Appendix

AI is placing huge demands on IT

100x

more model parameters

7x

greater computational throughput

7x

faster security threat lifecycles

10x

growth in newly-generated AI data

System Requirements of LLMs Can Be Massive

Llama3-405b:

- Training Requirements:
  - 16,384 Nvidia H100 GPUs
  - 30.84 million GPU hours
  - 54 days continuous runtime
  - ~15 Trillion Tokens
  - 8930 Equivalent Tons of CO2

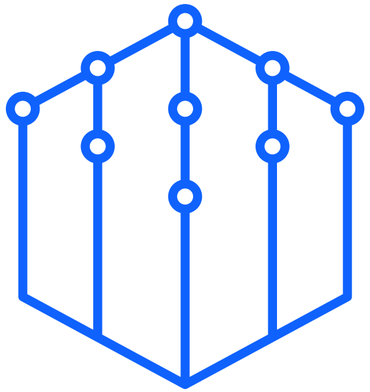
Llama3-405b:

- Inferencing Requirements:
  - 1944 GB VRAM (32 bit mode)
  - 800 GB VRAM (16 bit mode/quantized)
  - 640 GB VRAM (8 bit mode/quantized)
  - Minimum 8 Nvidia A100 GPUs
  - 1.5 TB Total System Memory
  - 4 TB NVME High Speed Storage

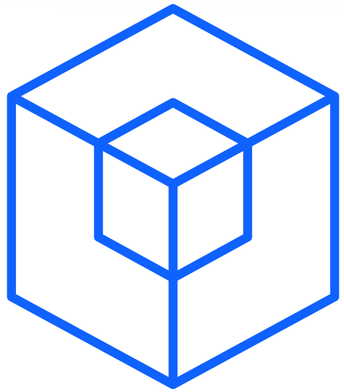
https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

The IBM approach: fit-for-purpose models

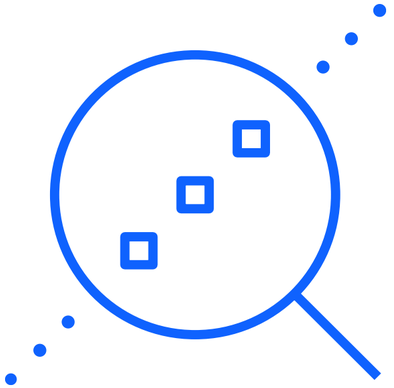
Up to 42x lower inferencing costs



The right data



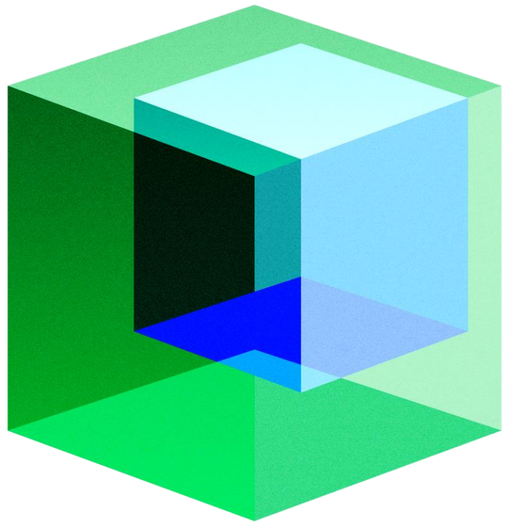
The right model



Targeted use case fine tuning

IBM Granite

A family of open, performant and trusted AI models to accelerate enterprise AI adoption



Open

- Open sourced under Apache 2.0
- Transparency of data, training methods
- Customize with business data

Performant

- Diverse range of fit-for-purpose models
- Designed for scalability

Trusted

- IP indemnification
- Responsible and safe AI
- Guardrails to secure data and mitigate risks

Granite family of models

Large Language Models (LLMs) for enterprise

Granite-3.0-8B-Instruct
Granite-3.0-2B-Instruct

Inference-efficient Mixture of Experts (MoE)

Granite-3.0-3B-A800M
Granite-3.0-1B-A400M

Guardrail models

Granite-Guardian-3.0-8B
Granite-Guardian-3.0-2B

Speculative decoding models

Granite-3.0-8B-Instruct-Accelerator

Time-series models

Geospatial models

IBM Granite is built and optimized for enterprise use cases

- Leveraging innovation for the open-source community and ability to customize with a businesses own data
- Delivering the performance and cost requirements for scale
- Documented AI guardrails and sources for trust

Agentic workflows

Automate tasks, streamline processes, and enhance operational efficiency with AI agents for business

Example: Autonomous HR agents to support Employee Support, Talent Acquisition, and Onboarding

Language-based tasks

Retrieval augmented generation (RAG), summarization, content generation, insight extraction, and classification based on documents or dynamic content

Example: Building a Q&A resource from a broad knowledge base, providing customer service assistance

Code

Optimize the software development lifecycle with code generative tasks, including code generation, code explanation, and code editing

Example: AI-generated code recommendations , IT application modernization from COBOL to Java

Time series

Time-series forecasting to easily analyze current data to make predictions and help make informed decisions

Example: Predicting future customer demand for a given product and period, using historical sales and other data sources

Geospatial

Uncover patterns and trends in geo data

Example: NASA and IBM teamed up to create an AI Foundation Model for Earth Observations using large-scale satellite and remote sensing data

Safety

Safeguard AI with models ensuring enterprise data security and mitigate risks across a variety of user prompts and LLM response

Example: AI compliance with regulatory requirements in financial services, healthcare, and government.

Customizing an LLM with InstructLab to achieve model performance at a fraction of the cost

Up to 98%
cost savings

Add skills and knowledge to the model, generate synthetic data

Use Lab techniques to merge contributions

Deploy customized model

Key data use cases fueling AI initiatives

Real-time analytics and AI

By combining scalable storage and data processing in real-time, underpinned by governance, organizations can unlock faster insights and drive smarter decision-making.

Focus solutions:
Databases
Data Intelligence
Data Integration

Multi modal AI

Make more data usable for generative AI by automating the unification, transformation, enrichment and governance of structured and unstructured data.

Focus solutions:
Databases
Data Intelligence
Data Integration

Privacy, Risk, Governance, & Compliance

Protect sensitive data, identify shadow AI and databases, govern and secure AI in one unified experience and quickly recover from breaches. By mitigating risk and streamlining compliance, reduce regulatory and financial exposure while maintaining brand integrity and AI trust.

Focus solutions:
Data Security
Data Intelligence
AI governance

Mergers and Acquisitions

Merge large volumes of heterogeneous data, manage metadata, and streamline ETL processes while ensuring transparency, consistency, and trust in data during complex consolidation projects.

Focus solutions:
Data Intelligence
Data Integration
Data Security

Real-time transactions

Deliver transaction data and capture changes across systems in real-time to ensure consistent, up-to-date data availability to support immediate decision-making and operational continuity.

Focus solutions:
Databases
Data Integration

Data-as-a-Product (lifecycle)

Streamline the creation, management, and sharing of data products by providing centralized governance, metadata, and lifecycle management, enabling data teams to deliver value-driven insights.

Focus solutions:
Data Intelligence
Databases

Cloud modernization

Accelerate the hybrid cloud journey by transferring data from on premises to cloud data stores with performant engines, smart workload optimization and reusable pipelines.

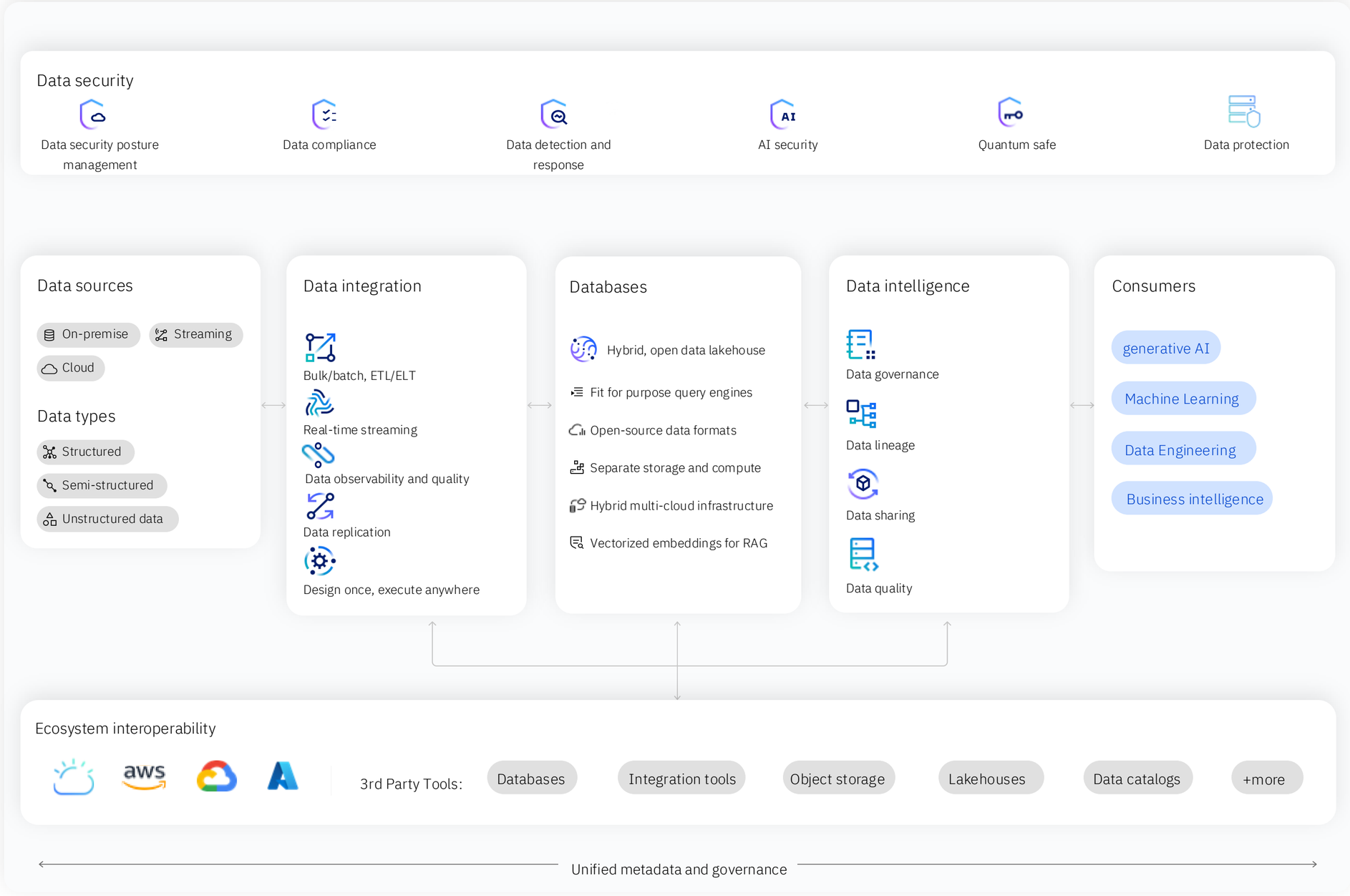
Focus solutions:
Databases
Data Integration

Bridging the skills gap

Simplify data access, storage, and data quality with automated processes: self-service access, no/low code experiences, and pre-built functions. Empower any data user to collaborate and leverage data without technical expertise.

Focus solutions:
Databases
Data Integration
Data Intelligence

Integrate, access, govern, and secure all data types with an open and hybrid data architecture



Pain points for scaling AI

Changing regulations

Multiple stakeholders

Inaccurate documentation

Increased risk

Disparate tools and data

Vulnerable data

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows

Centralized
AI lifecycle governance

Manage, monitor and govern any AI: model, app or agent; across IBM and 3rd party like OpenAI, AWS, and Meta

Proactive
AI risk and security management

Proactively detect and mitigate AI risks, evaluate AI assets, and secure AI deployments with Guardium AI security

Trustworthy and dynamic
compliance

Manage AI for safety and transparency with our regulatory library, automation and industry standards

Platform agnostic: Govern any AI, deployed anywhere

Data

Access trusted and secure data to drive AI productivity

AI Productivity

Reinvent how work is done with AI agents/assistants

AI Assistants

- watsonx Code Assistant™
- watsonx Orchestrate™
- Planning Analytics

AI/ML Ops

Work with AI models, tools and governance that's built for business—engineered to ensure trust and scalability in applications

AI Models

- Granite™
- Meta Llama
- Mistral

AI Tools

- watsonx.ai™

AI Governance

- watsonx. governance™

Data Fabric

Bring all the business data together and optimize how it moves through systems to scale analytics and AI in applications while protecting it

Databases

- watsonx.data™

Data Intelligence

- Data Product Hub
- Knowledge Catalog
- Manta Data Lineage

Data Integration

- DataStage®
- Databand®
- Streamsets

Data Security

- Guardium® Data Security Center

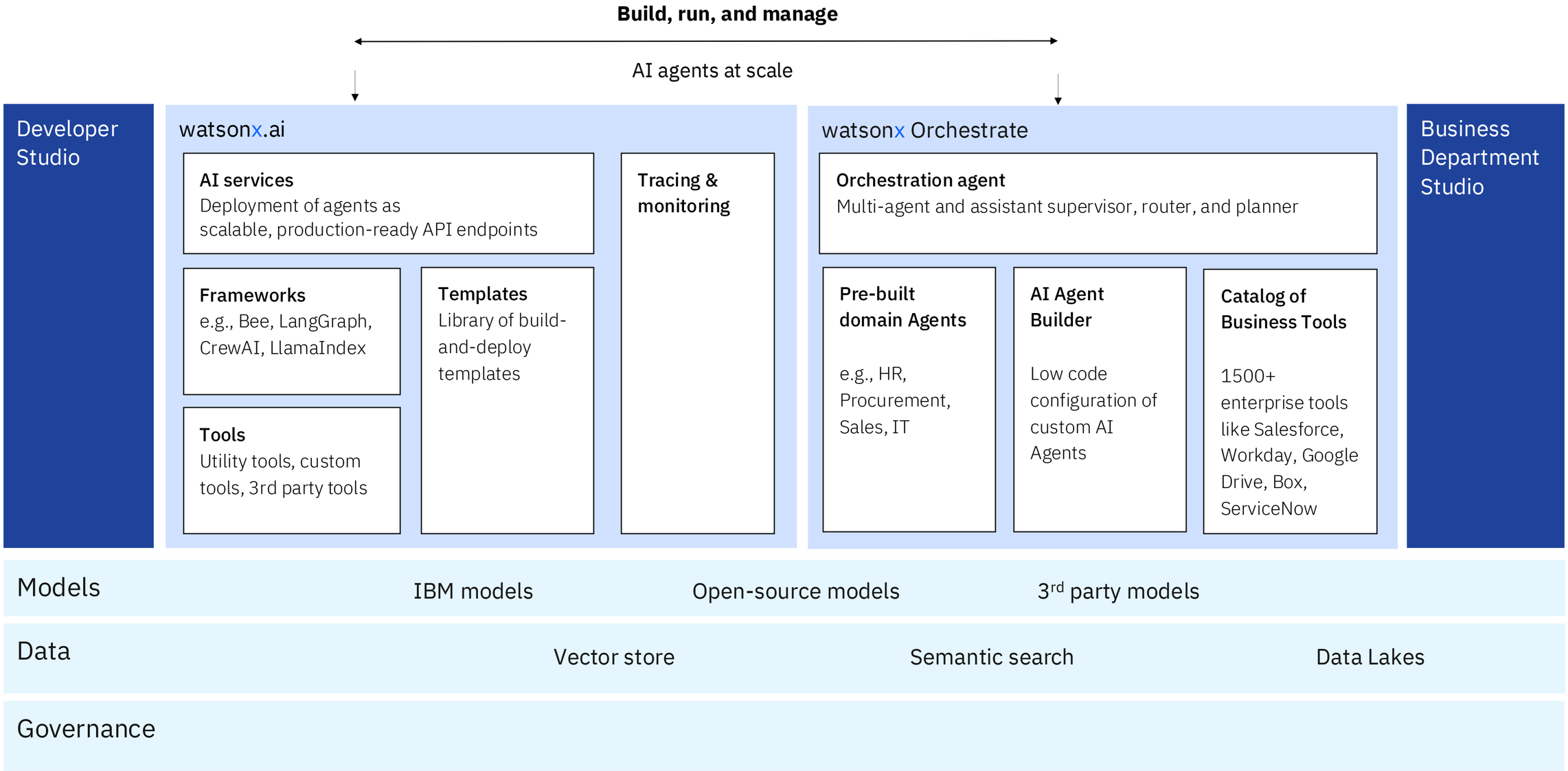
Data Storage

Store data across edge, core, and clouds

Software-defined Storage

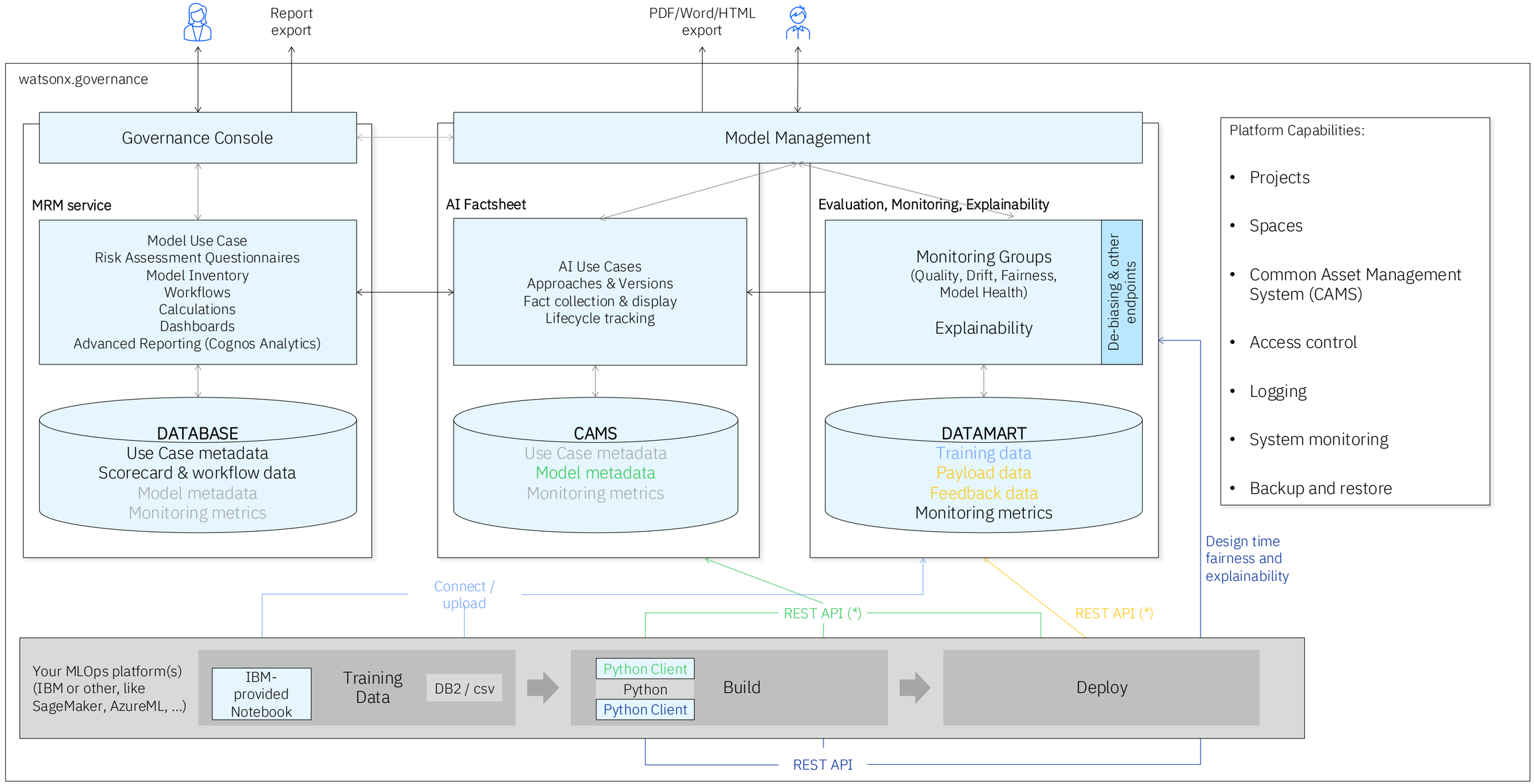
- Storage Ceph®

IBM’s approach: One size does not fit all for building agents



watsonx.governance – functional architecture (software)

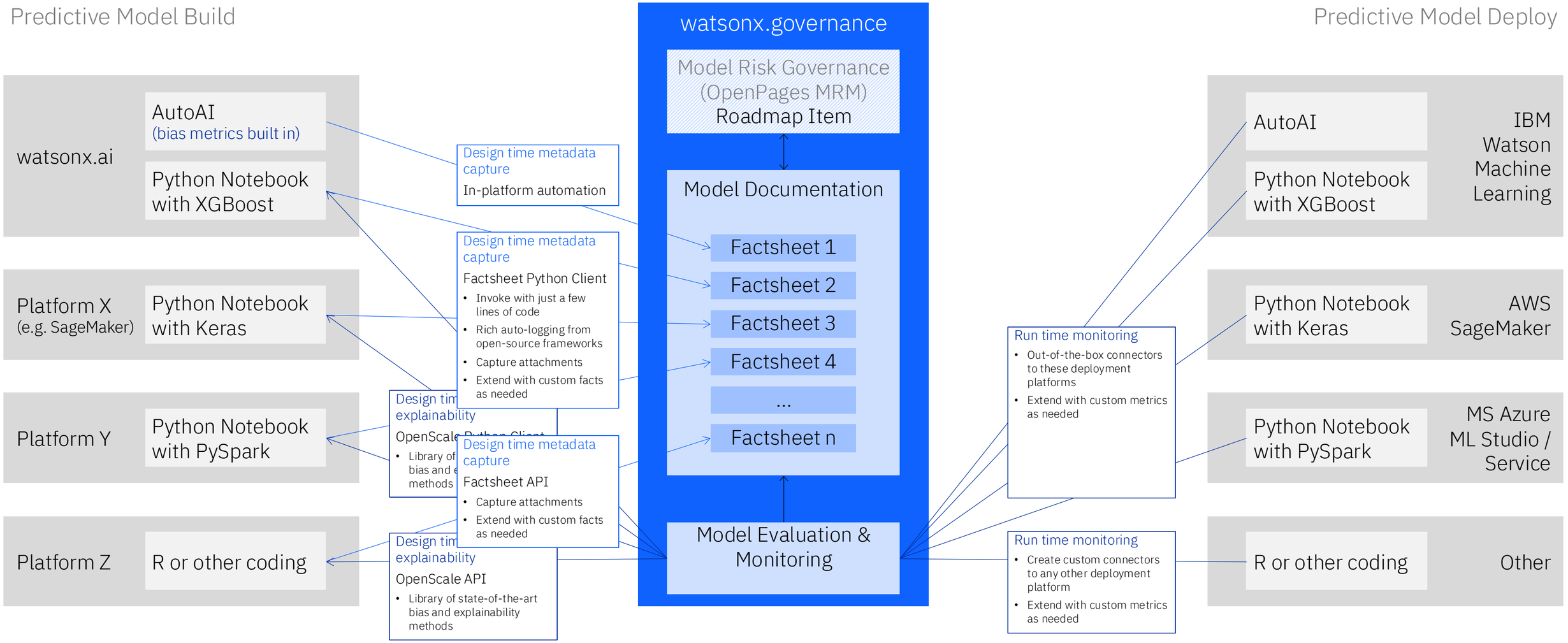
(legend on separate slide)



(c) 2024, IBM Corporation

watsonx.governance

*Integration with different **Predictive** Model Build and Model Deploy platforms (non-generative AI support)*

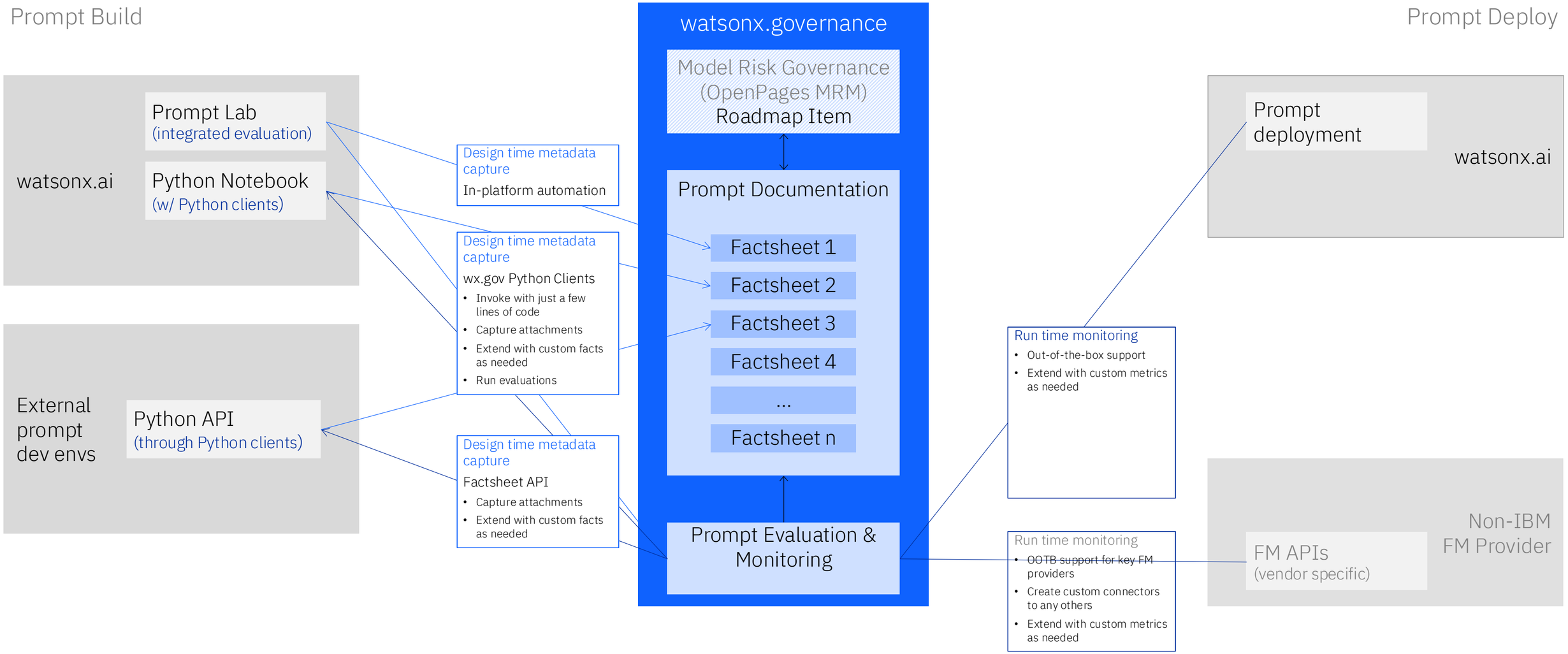


Sample notebooks are available to get you going quickly

(c) 2022-2023, IBM Corporation

watsonx.governance

*Integration with different **Prompt** Build and Deploy platforms (generative AI support)*



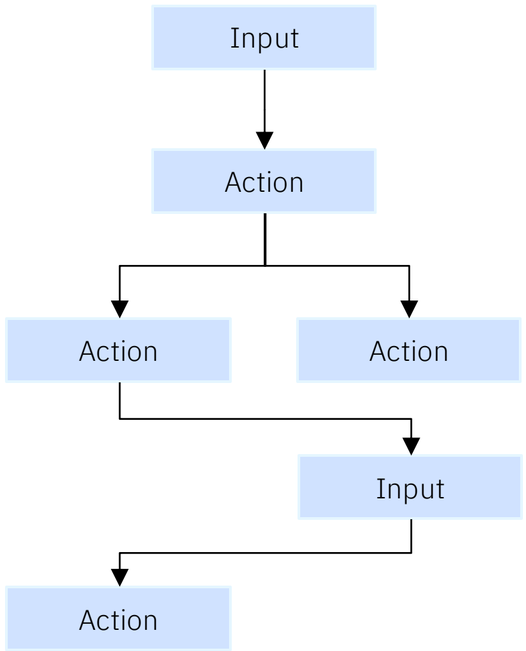
Sample notebooks are available to get you going quickly

(c) 2022-2023, IBM Corporation

Evolution of AI assistants

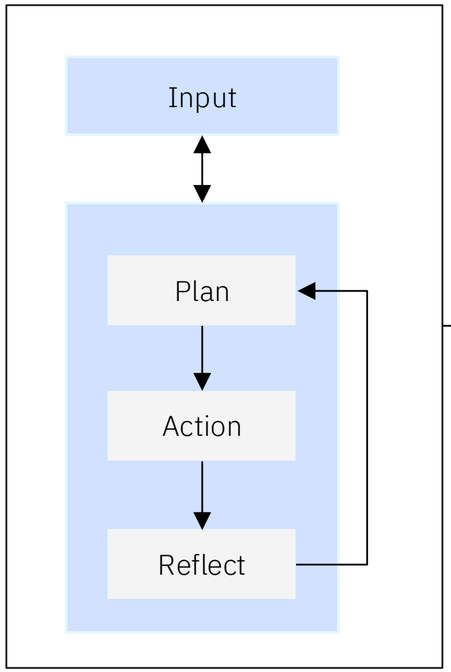
Traditional assistants

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



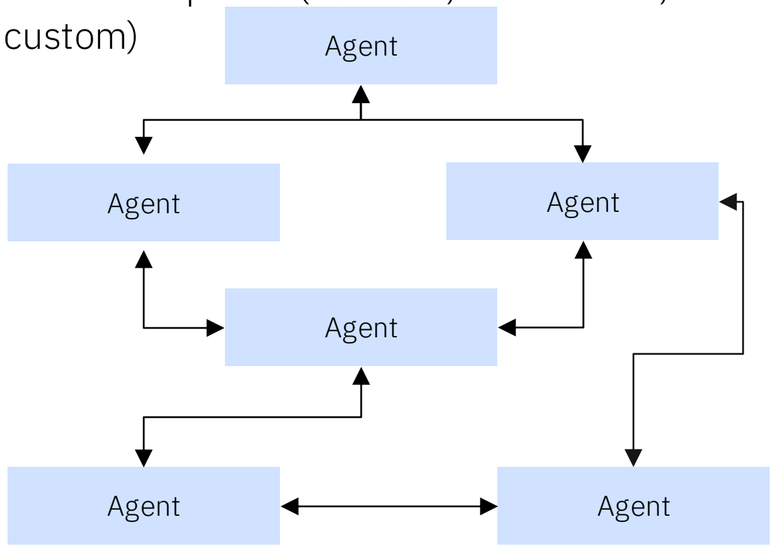
Single-agent assistants

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



Multi-agent assistants

- 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)



IBM watsonx Code Assistant Portfolio

IBM watsonx Code Assistant

Multi-language generative AI for code

IBM watsonx Code Assistant for Red Hat Ansible Lightspeed

Automation for infrastructure via Ansible Playbooks

IBM watsonx Code Assistant for Z

Mainframe lifecycle management and modernization

IBM watsonx Code Assistant delivers enterprise-ready AI for code solutions to address skills gaps and increase developer productivity for targeted, business use cases.

Empower developers to accelerate software development lifecycles, enhance productivity, and improve code for over 116 languages — Java, Python, YAML, COBOL, and more.

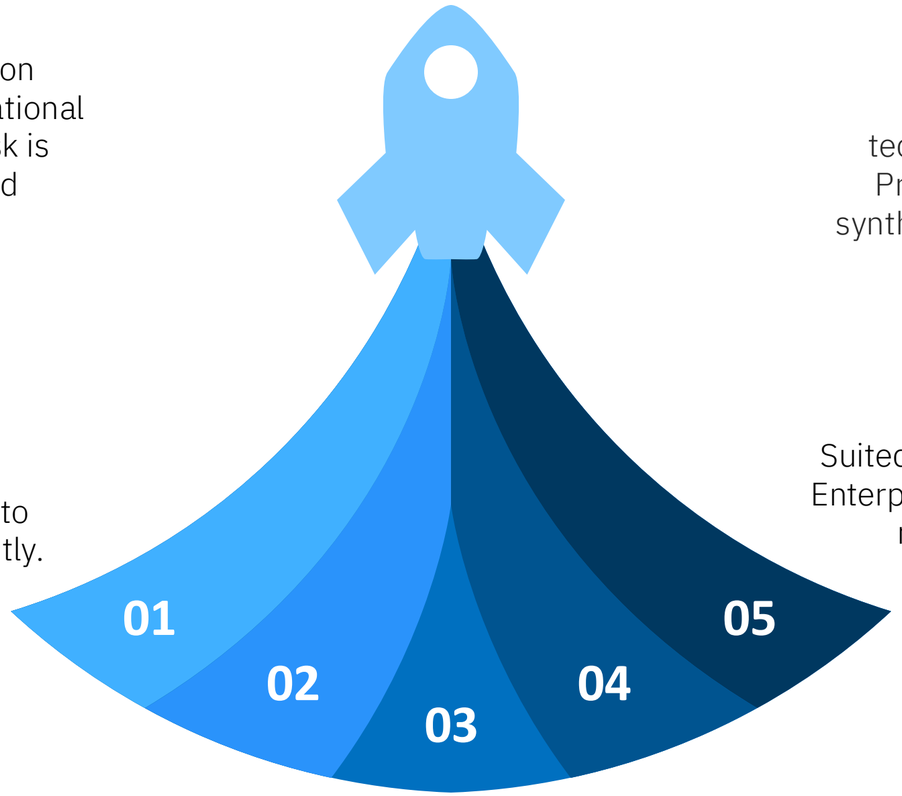
IBM Granite foundation models are built for business.

Granite is IBM's flagship brand of open and proprietary large language model (LLM) foundation models, spanning multiple modalities.

Augmenting Models with Enterprise Data

01
Prompt Tuning
Most suitable for quick adaptation tasks, especially when computational resources are limited, or the task is closely related to the pre-trained model

02
PEFT (LoRA, QLoRA)
Ideal for resource-constrained environments or when needing to fine-tune multiple tasks efficiently.



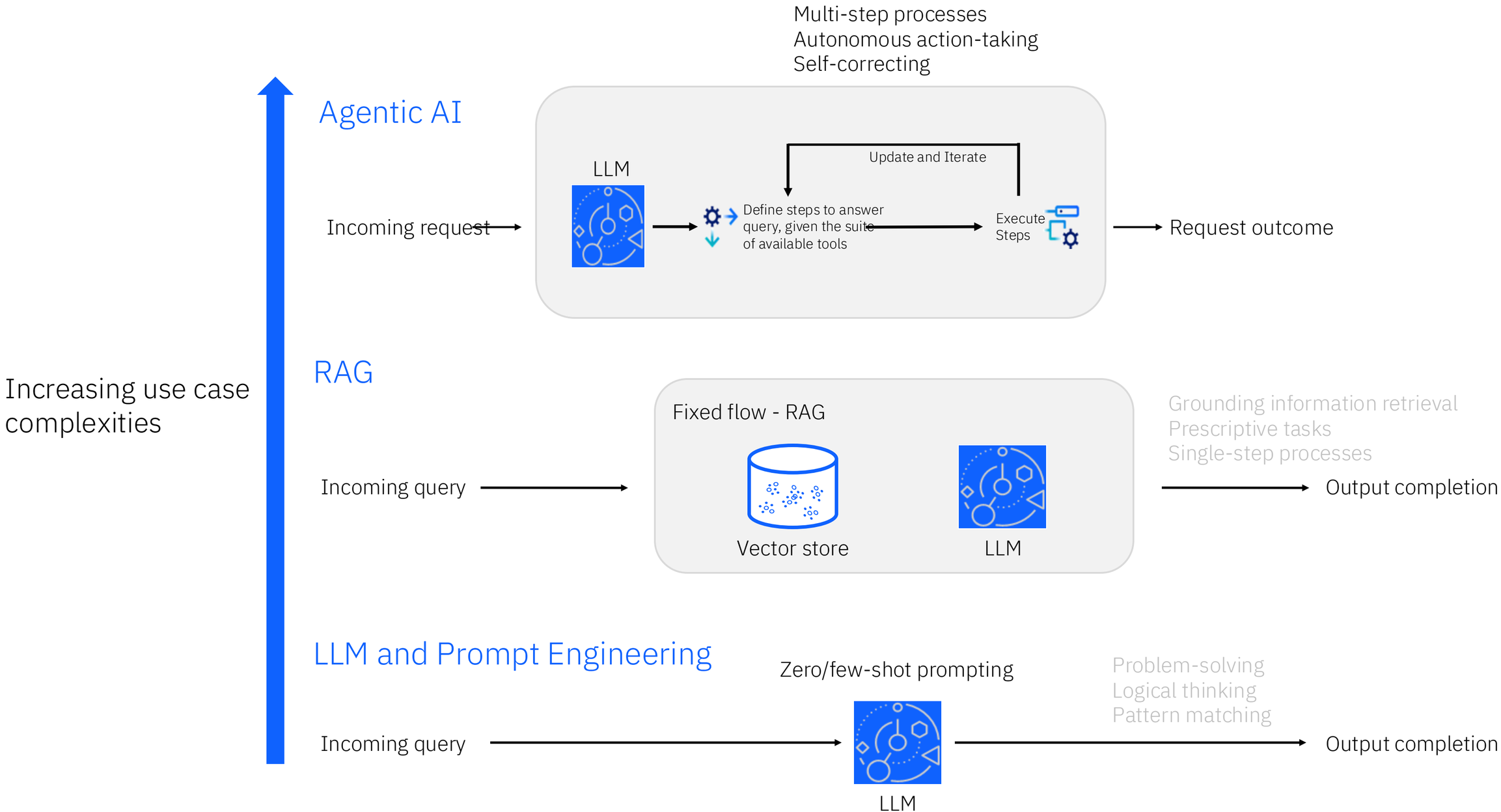
03
Full fine-tuning
Best for scenarios where maximum accuracy and task specific adaptation are critical and resources are abundant

05
★ InstructLab
Much more than just a tuning technique. Doesn’t need deep technical expertise and data prep. Provides a collaborative platform, synthetic data generation and better accuracy of models

04
RAG
Suited for dynamically changing data. Enterprise data not represented in the model. Does not improve model

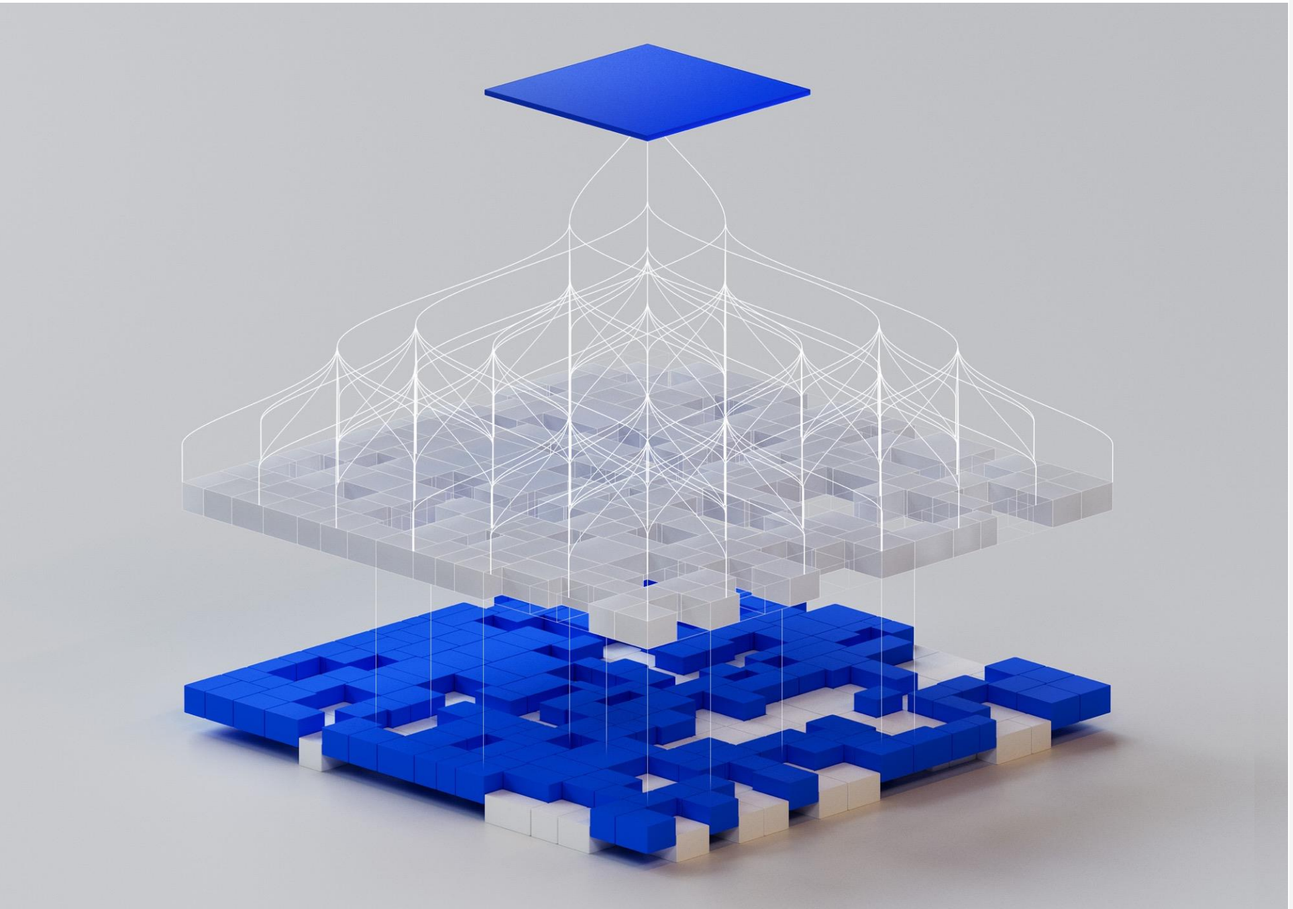
IBM Data & AI | © 2025 IBM Corporation

From Q&A to Agentic AI



IBM watsonx Code Assistant

Enterprise-ready AI for code solutions to address skills gaps and increase productivity for targeted, business use cases.



AI-powered coding tasks

Code generation, explanation, unit test creation, AI-derived code documentation, and more.

Enterprise Java modernization

Generative AI and automation-assisted modernization for Java enterprise applications.

Chat for Code integrated UI

Free-form AI conversational chat for planning applications and writing code.

Enhanced prompt engineering

Pre-built chat commands with optimized prompts for specific tasks.

Integrated directly into IDEs

Available directly within VS Code and Eclipse, integrating seamlessly with your workflow.

Trust, transparency, and privacy

Code similarity checks, IP protections, deploy either via SaaS on cloud or on-premises.