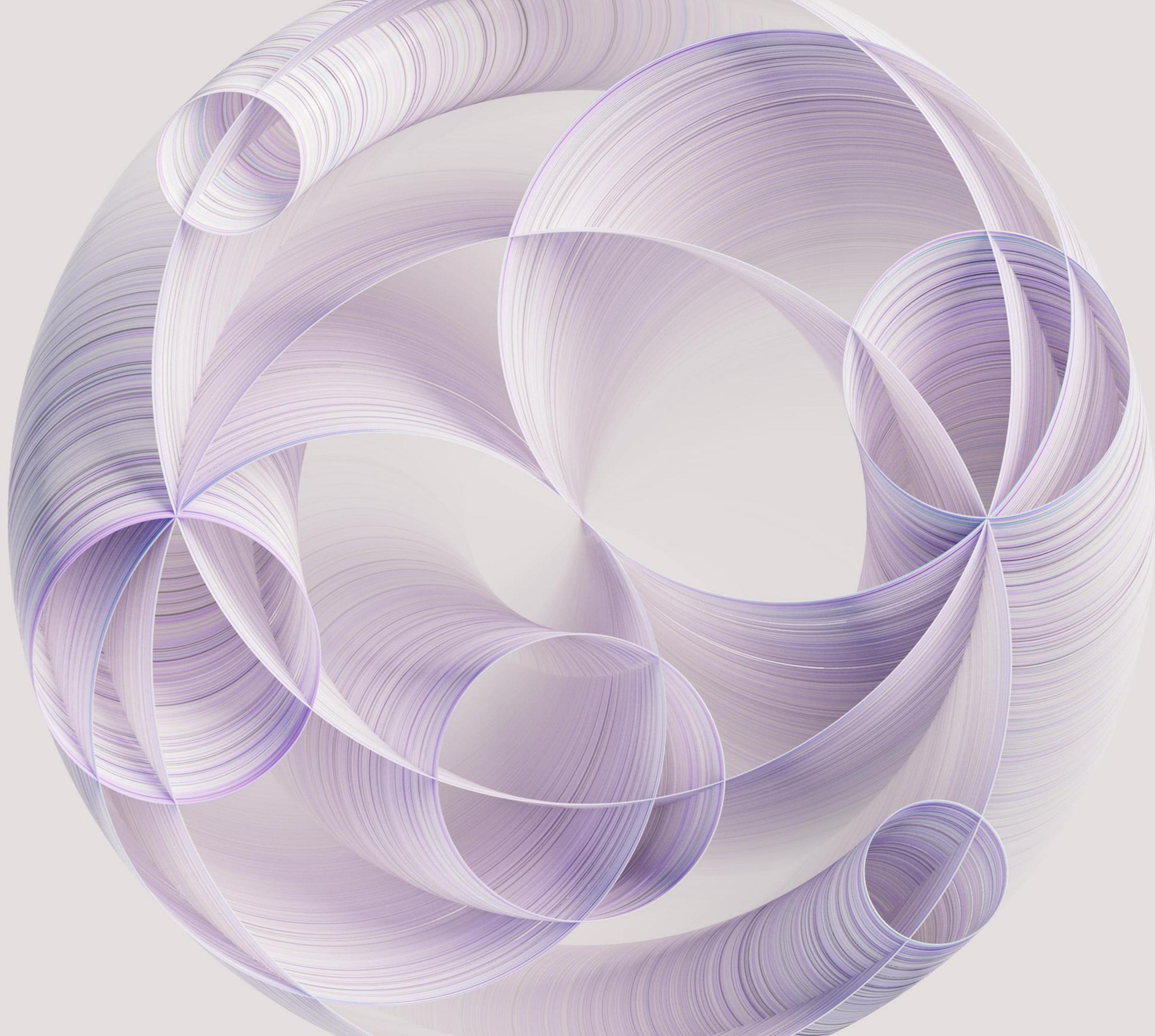


watsonx.ai

Q4 Product Updates

October 17, 2024



Disclaimer

© 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.

Agenda

TechXchange Preview 8 min, Lindsay

watsonx.ai Futures 20 min, Alex/AJ

- Agents & RAG
- InstructLab

Chat with Images & Model Updates 10 min, Nisarg

- Third models
- Granite models
- Chat with image

Product Updates 8 min, AJ

- Text extraction
- Import custom foundation models (SaaS)
- Fine tuning

Sales Motion Updates (Michael) 15 min, Luv/Michael

- Mistral Large (Add-on)
- BYOL for non-ELA (live on 10/10)
- HCI Fusion on Passport Advantage and sales configurator support

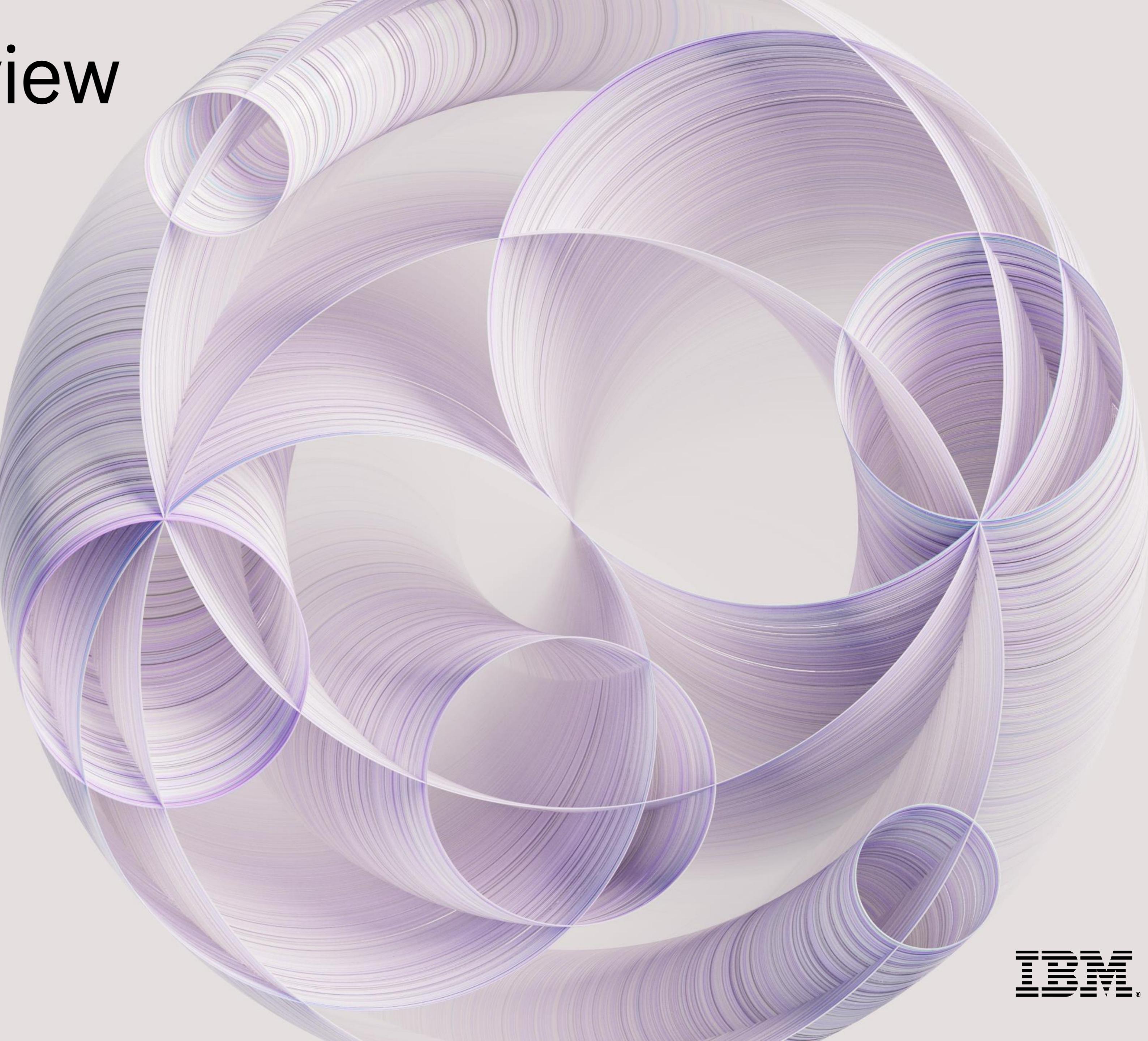
Call to Action 5 min, Luv

Q&A

TechXchange Preview

Lindsay Wershaw
Senior Product Marketing Manager
IBM watsonx.ai

watsonx.ai



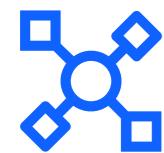
IBM®

Key Messages:

- 1. Enabling greater developer productivity** with new tooling to build autonomous agents, customize LLMs, and optimize RAG-based workflows with low-code based solutions
- 2. IBM's approach to models: fit-for-purpose.** Lead the market and drive AI relevancy with new Granite models and top benchmarks with best-in-class performance, cost-efficiency, transparency and safety
- 3. Trust in AI:** Continue to drive our market leadership and differentiation in AI Governance with new model benchmarks, evaluations, and AI guardrails, along with new, integrated AI Security

AI @ TechXchange

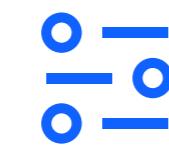
Major NFI and NPIs coming to life onsite and online



watsonx.ai Developer Toolkit

Enterprise-grade AI studio that helps AI builders innovate with all the APIs, tools, models, and runtimes to simplify and scale the development and deployment of AI applications

[Featured:](#) AI Spotlight/ Focused sessions



Granite 3.0

IBM Granite offers businesses a family of open, performant, and trusted AI models to accelerate enterprise AI adoption. New models deliver high performance, outperforming or on par with leading industry models across benchmarks

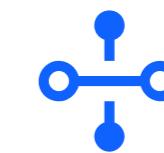
[Featured:](#) AI Spotlight / Focused sessions



watsonx Code Assistant

Enterprise grade coding assistant that leverages genAI and automation to boost productivity and accelerate the end-to-end software development lifecycle for enterprise teams.

[Featured:](#) AI Spotlight / Focused sessions



watsonx.governance - Model Evaluation and Guardrails

Summarizes and pulls out key pieces of data from documents so that you can quickly and easily [ask questions about your content and get answers](#).

[Featured:](#) BA Spotlight / Focused Sessions

AI Track Spotlight Flow

Accelerate innovation with IBM watsonx: From code to business impact



Ritika Gunnar,
GM Product Management,
Data and AI
IBM Software



Steven Astorino,
General Manager, Product Development,
Data, AI & Sustainability

Steve
2:00 mins

Scaling AI for business
Product leader opens with an introduction.

Market trends – Gen AI will add \$4.4T to the global economy

That is, if we find the right answers to serious challenges – such as the rising costs of training AI models

And that same rapid development is coming with AI agents.

1/3 of gen AI interactions will use action models and autonomous agents by 2028

AI is changing at a rapid pace, transforming how we interact with technology – entering a new AI agent era

Steve
5:00 mins

AI Building Blocks for the future – starting with the right model foundation

The building blocks to AI require the right strategy... As you move from experimentation to deployment, those building blocks include – *models, data, assistants, agents, and governance tools*

Model building block – sets the stage for IBM's approach to models, fit-for-purpose, cost-effective, efficient and performant

- **NFI** – Meet Granite 3.0, highlighting 2B/8B models available on watsonx and Apache 2.0 license.

These models offer a smaller IT footprint, lower inference costs, with differentiated performance compared to larger size models on top enterprise tasks.

Steve
4:00 mins

Introducing new AI tools to customize models with your enterprise data

Data building block – Your enterprise data is the fuel for an effective AI strategy

- **NFI** – watsonx.ai developer toolkit – IBM is introducing a new AI tooling stack for AI developers, meeting developers where they are with all the APIs, models, tools, and runtimes to build AI services, i.e., IBM's approach to model customization, with

InstructLab in watsonx.ai (*coming soon*) demo included and IBM's approach to RAG with **demo of AutoAI-RAG** (*in Tech Preview*) w/ watsonx.data Milvus

Ritika & guests (live demos)
18:00 mins

Developers are leaning on new AI assistants & agents to unlock productivity gains
AI Assistants building block – helping developers unlock immense productivity gains and automation for the enterprise

- **NPI** – watsonx Code Assistant. Introducing WCA, built from the latest Granite 8B model, designed to accelerate software development, generating high quality code based on natural language prompts. **Ruchir demo live**
- **NFI** – watsonx.governance – Demonstrate new **evaluation studio** and **guardrails** features to help developers select the right model for the task and evaluate the models in production (2 demos included)

Agent building block – sets the stage for IBM's POV on agents

- **Building agentic workflows** in watsonx.ai (with 2 demos embedded highlighting *low-code Agent Builder service and pro-code API-driven approach with Crew AI integration*)
- **MARC demo** – future-driven, open-source approach (**Ratish demo live**)

Steve
4:00 mins

Good governance is good business

Governance building block – Now is time to put your models and applications to work

- **NFI** – watsonx.governance – Demonstrate new **evaluation studio** and **guardrails** features to help developers select the right model for the task and evaluate the models in production (2 demos included)

Prove AI
10:00 mins

Customer POV & demonstration

Prove AI (formerly known as Casper Labs) on stage to discuss and demonstrate LIVE their watsonx.ai, watsonx.governance and integrated AI model story (demo Granite)

Mrinal Manohar, CEO

Henry Guo, VP Product Management

Steve
2:00 mins

What's next

Plans for Q4 2024 and beyond – promote Granite playground and new watsonx Developer Hub for developers to access these new tools and features.

Stay up to date with GA of these new capabilities

Close with call-outs for TechXchange spotlight sessions (IA and BA – further details on IBM Agent POV), plus WCA session and governance session – to see relevant topics at hand.

The enterprise-grade AI developer studio

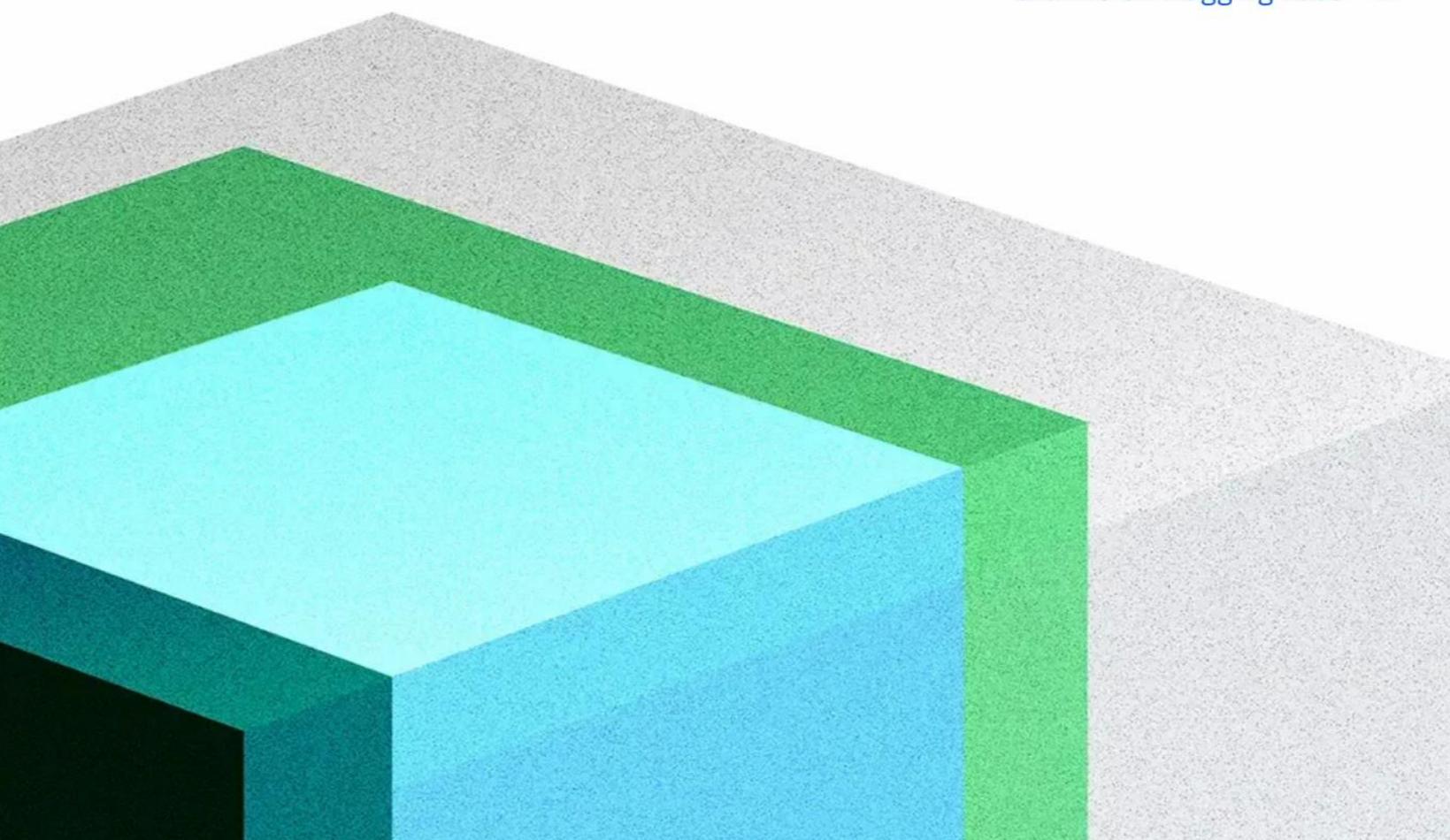
Assistants		Applications		
Templates	Templates			
Developer toolkit	Agents	Tools	Making tuning easy	Making RAG easy
		API/SDK enhancements – Node.js SDK – One unified watsonx.ai™ API – SDK documentation enhancement – Quick start for IDE developer	– InstructLab MVP – Enhanced fine-tuning capabilities	– Chat with documents – Custom APIs, e.g., text extraction, re-rank, embedding – AutoAI: RAG
Platform services	Frameworks	Integrations		Quality of service Model evaluation and benchmarks
	– Open-source frameworks (LangChain, LlamaIndex) – Flows engine	Third-party data sources and tooling		
AI runtimes	Collaboration, access control, user interface		Machine learning	
	– Projects and assets – Spaces and deployments – Jobs		Predictive and prescriptive analytics	
AI infrastructure		Model choices		
		– IBM Granite™ – Third party (Mistral AI, Meta) – Open source (Hugging Face) – Custom foundation models		

Get started with Granite

Explore the use cases.
Run and download the
Granite models to your
laptop.

ibm.com/granite/docs

IBM Granite



Granite

Granite is a family of AI models purpose-built for business, engineered from the ground up to ensure trust and scalability in AI-driven applications.

[Granite on Hugging Face →](#)

Models

- Granite Code
- Granite Time Series

Run Granite

- Mac

How-to guides

- Fine-tuning
- Quantization

Use cases

- Granite Code on Replicate
- VS Code Assistant
- Forecast Energy Demand
- Text-to-shell
- All Cookbooks

Granite Playground

Granite Hugging Face

Granite GitHub

[ibm.com/granite/docs](#)

Granite

The Granite Code models family are decoder-only models designed for code generative tasks, trained with code written in 116

Cookie Preferences

watsonx developer hub

Access a library of
templates and guides
to get started building
gen AI applications.

ibm.com/watsonx/developer

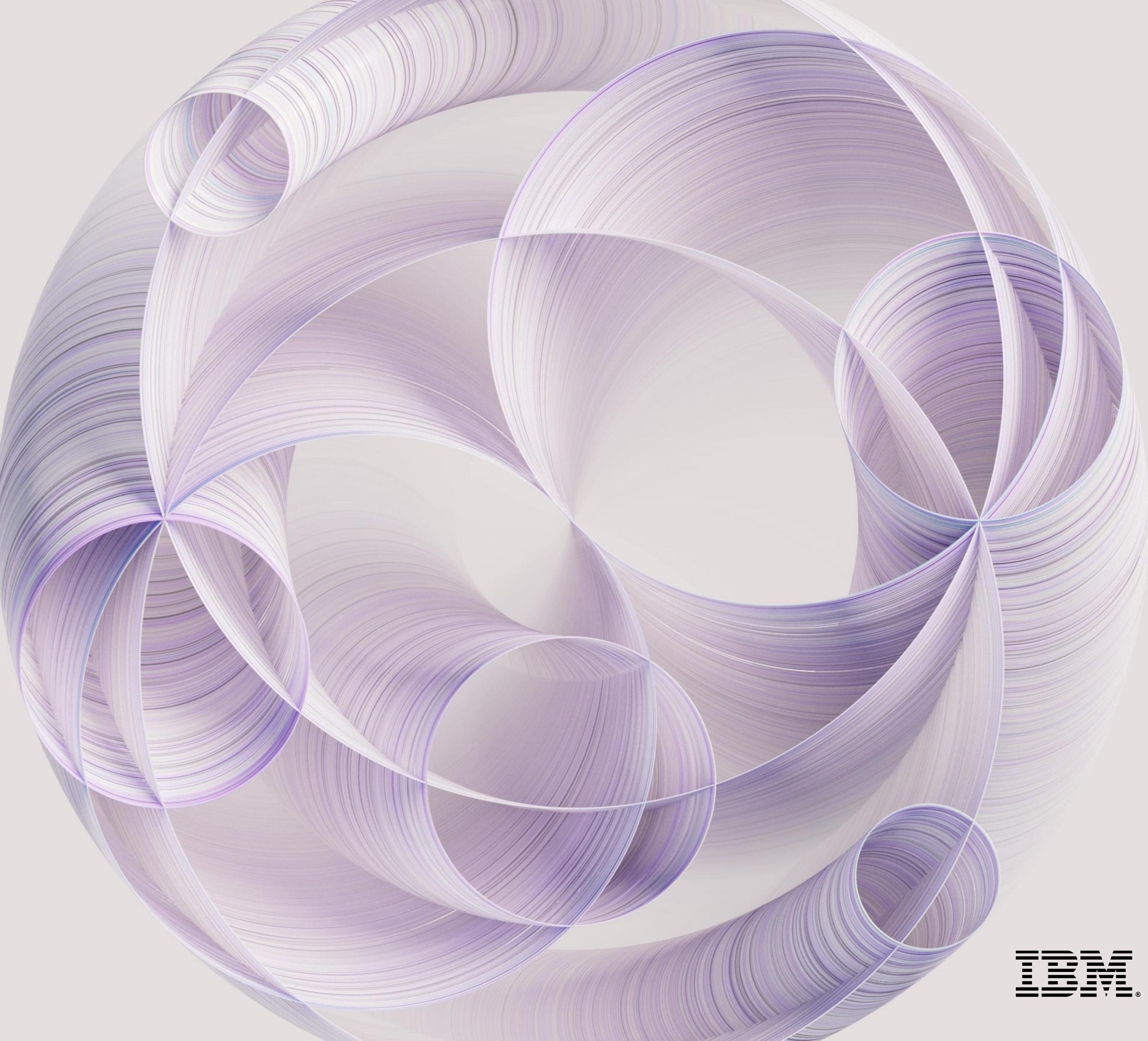
The screenshot shows the IBM WatsonX Developer hub homepage. The left sidebar contains a navigation menu with sections like Get started (Overview, Quick start, Models, Libraries, Capabilities), Endpoints (Chat API, Generation API, Embeddings API, Re-rank API, Extraction API, AutoRAG API, AI services API), Guides (Prompt engineering), Support (IBM community), and external links (watsonx.data, watsonx.governance, Discord). The main content area features a large banner with the WatsonX logo and developer hub text. Below the banner are several cards: 'Quick start' (with curl, Node, Python tabs and a code snippet), 'Begin developing' (with a grid icon and text about supporting projects), 'Models' (with a gear icon and text about choosing a model), 'Embeddings API' (with a chart icon and text about capturing sentence meaning), 'Text generation' (with a document icon and text about prompting models), 'Libraries' (with a code icon and text about integrations), and 'Retrieval augmented generation' (with a document icon and text about improving LLM output).

What's Coming

Agents and RAG

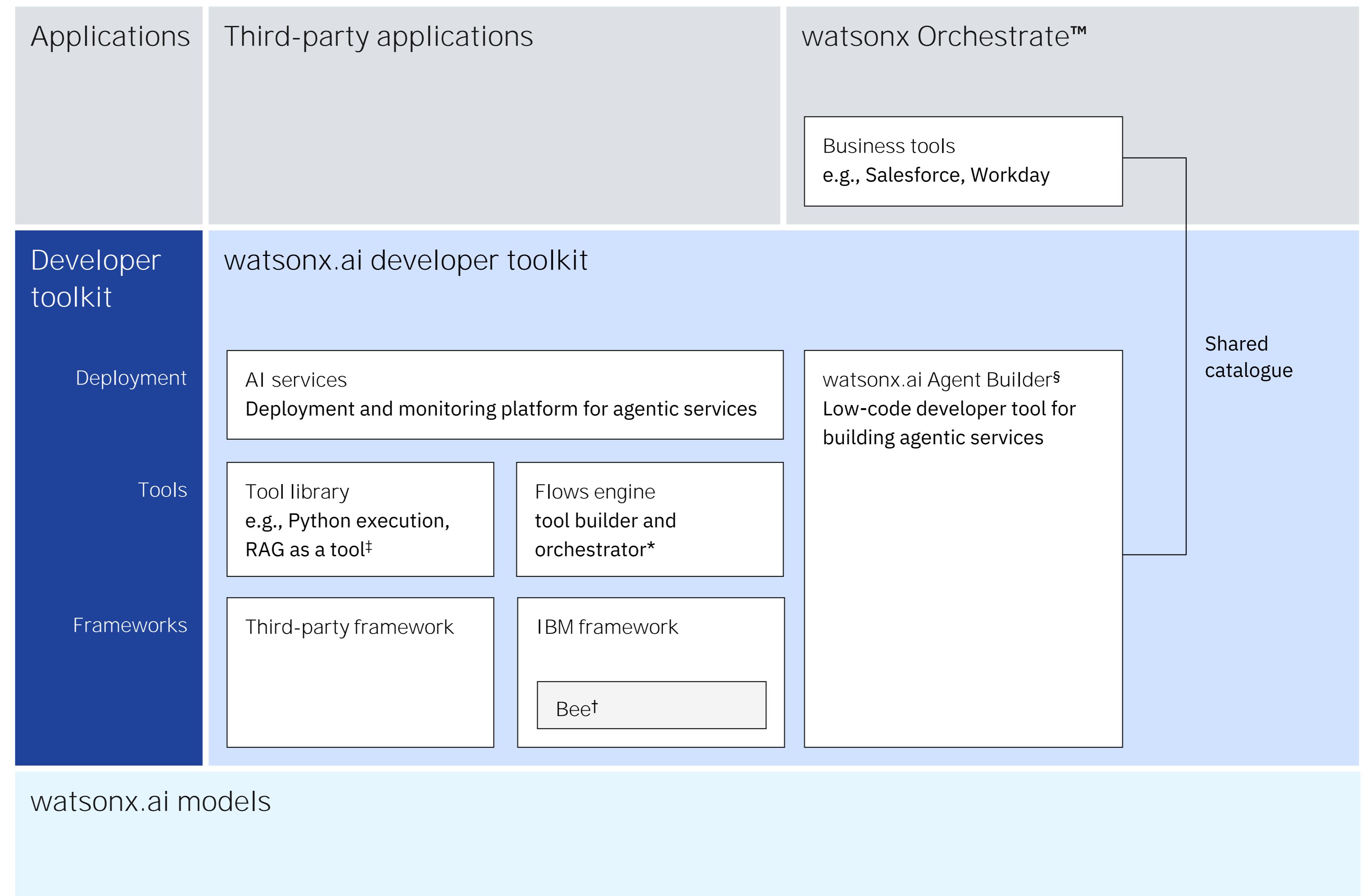
Alex Seymour
Product Manager, Developer Toolkit
IBM watsonx.ai

watson**x**.ai



IBM

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

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*)



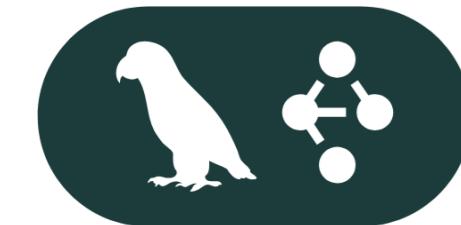
Granite MoE Models

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



watsonx.ai SDKs

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



LangGraph



LangChain



Llamaindex



Bee

watsonx.ai AI Services

Deploy agents

Framework-agnostic solution for deploying agents.

- Available
- Scalable
- Secure

Monitor agents

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

The screenshot shows the IBM WatsonX API reference for a deployment named "test". The deployment status is "Deployed" and "Online". The "API reference" tab is selected. Under "Endpoints for inferencing", there are two sections: "Private endpoint" and "Public endpoint". Each section has a "Text endpoint URL" and a "Stream endpoint URL". The "Text endpoint URL" for the private endpoint is https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/test_12345/text. The "Stream endpoint URL" for the private endpoint is https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/test_12345/stream. The "Text endpoint URL" for the public endpoint is https://us-south.ml.cloud.ibm.com/ml/v1/deployments/test_12345/text. The "Stream endpoint URL" for the public endpoint is https://us-south.ml.cloud.ibm.com/ml/v1/deployments/test_12345/stream. Below the endpoints, there is a "Code snippets" section with a "cURL" tab. The cURL code is as follows:

```
# NOTE: you must set $API_KEY below using information retrieved from your IBM Cloud account (https://dataplatform.cloud.ibm.com/docs/content/ws\_j/analyze-data/ml-authentication.html)
curl --insecure -X POST --header "Content-Type: application/x-www-form-urlencoded" --header "Accept: application/json" --data-urlencode "grant_type=urn:ibm:params:oauth:grant-type:apikey" \
--data-urlencode "apikey=$API_KEY" "https://iam.cloud.ibm.com/identity/token"
# the above CURL request will return an auth token that you will use as $IAM_TOKEN in the scoring request below
# TODO: manually define and pass values to be scored below
curl -X POST --header "Content-Type: application/json" --header "Accept: application/json" --header "Authorization: Bearer $IAM_TOKEN" -d '{ "input": "" }' "https://private.us-south.ml.cloud.ibm.com/ml/v1/deployments/test\_12345/text/generation\_stream?version=2021-05-01"
```

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 shows the IBM Watsonx Agent Builder interface. At the top, there's a navigation bar with 'IBM watsonx' on the left, followed by 'Upgrade', a user profile, 'Dejan Glozic's Account', 'Dallas', and a settings icon. Below the navigation is a toolbar with icons for 'Projects', 'Document based grounding', 'Prompt Lab', 'Unsaved', 'New prompt +', 'AI guardrails off', and a gear icon.

The main area is a 'Chat' window. On the left, there are three tabs: 'Chat' (selected), 'Structured', and 'Freeform'. On the right, there are several sections: 'Agents' (with 'IBM Bee' selected), 'Selected tools' (listing 'Context', 'Wikipedia', 'DuckDuckGo', 'LLM', 'SDXLTurbo', and 'Weather'), and 'Execution plan' (showing a three-step process: 1. Searching for golf ball market information, 2. Searching for golf ball market competitors, 3. Generating final answer). A note at the bottom 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.'

In the center, a message from 'You' at 01:33 PM reads: 'Analyze the main competitors in the golf ball market and create a detailed report in tabular format.' A response from 'watsonx' at 01:33 PM follows: 'Here is 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

At the bottom, there's a search bar with the placeholder 'Type something...' and a blue search icon.

One size does not fit all for building RAG solutions

Making RAG easy

watsonx.ai developer toolkit

Tailored, high-performance RAG systems targeting complex use cases via custom APIs

Text extraction

Embeddings

watsonx.data™
(Milvus)

Re-rank

Text generation

Evaluations

Streamline the RAG pipeline development and application process

Chat with documents

AutoAI: RAG

watsonx Assistant™

watsonx Discovery

watsonx Orchestrate

Prebuilt RAG
solutions

watsonx.ai models

watsonx.ai AutoAI- RAG (API beta)

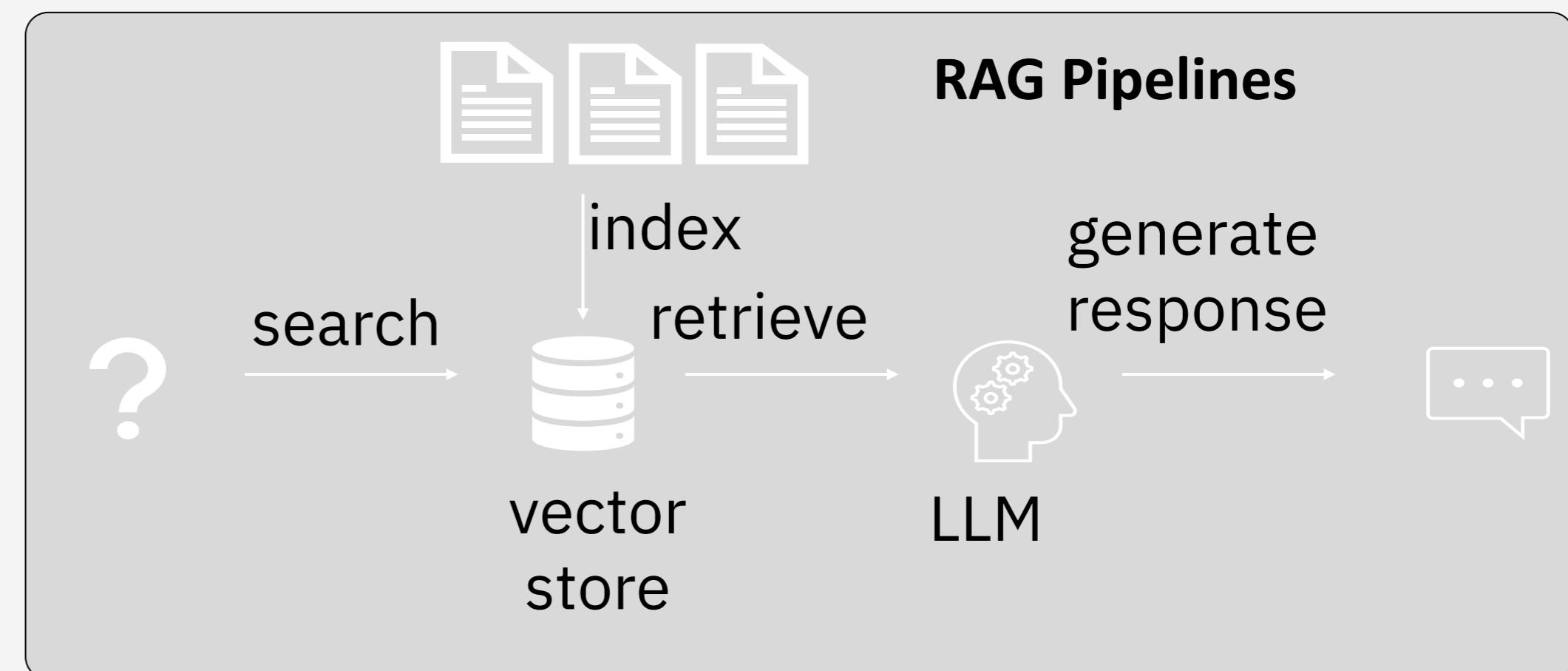
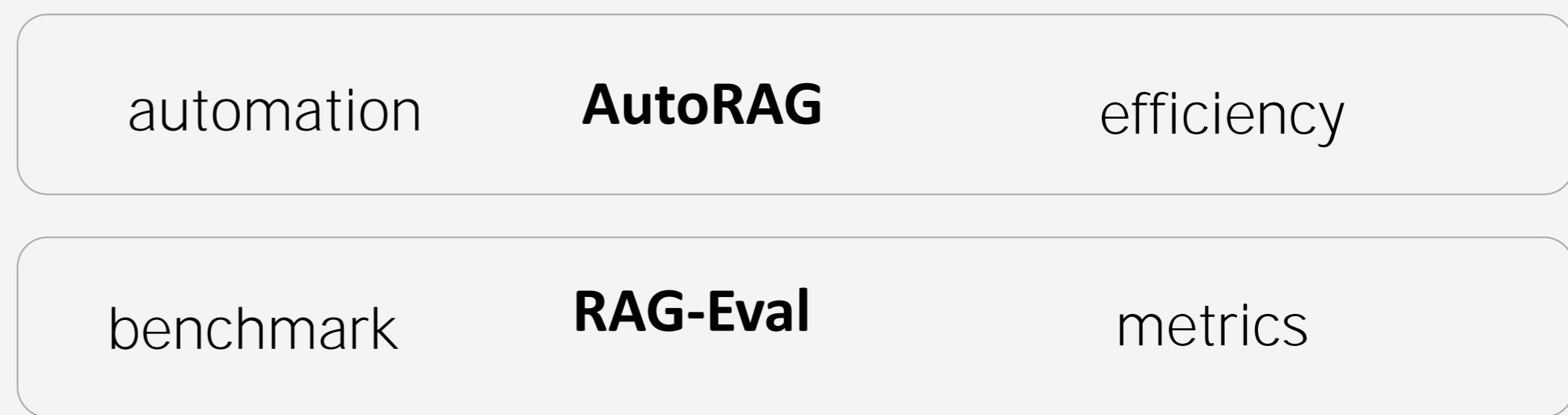
RAG prototypes are easy to build.
However, it is hard to deliver a production
ready solution which can require a team of
experts and months of effort.

AutoAI-RAG: Automate design and deployment
of optimized, production-quality, RAG systems
based on client data and use case.

Allow full exploration or evaluate a constrained
set of configuration options

Re-evaluate and modify recommended
configuration when something changes (e.g., new
model is released, quality of responses change).

Learn more in the [documentation](#).



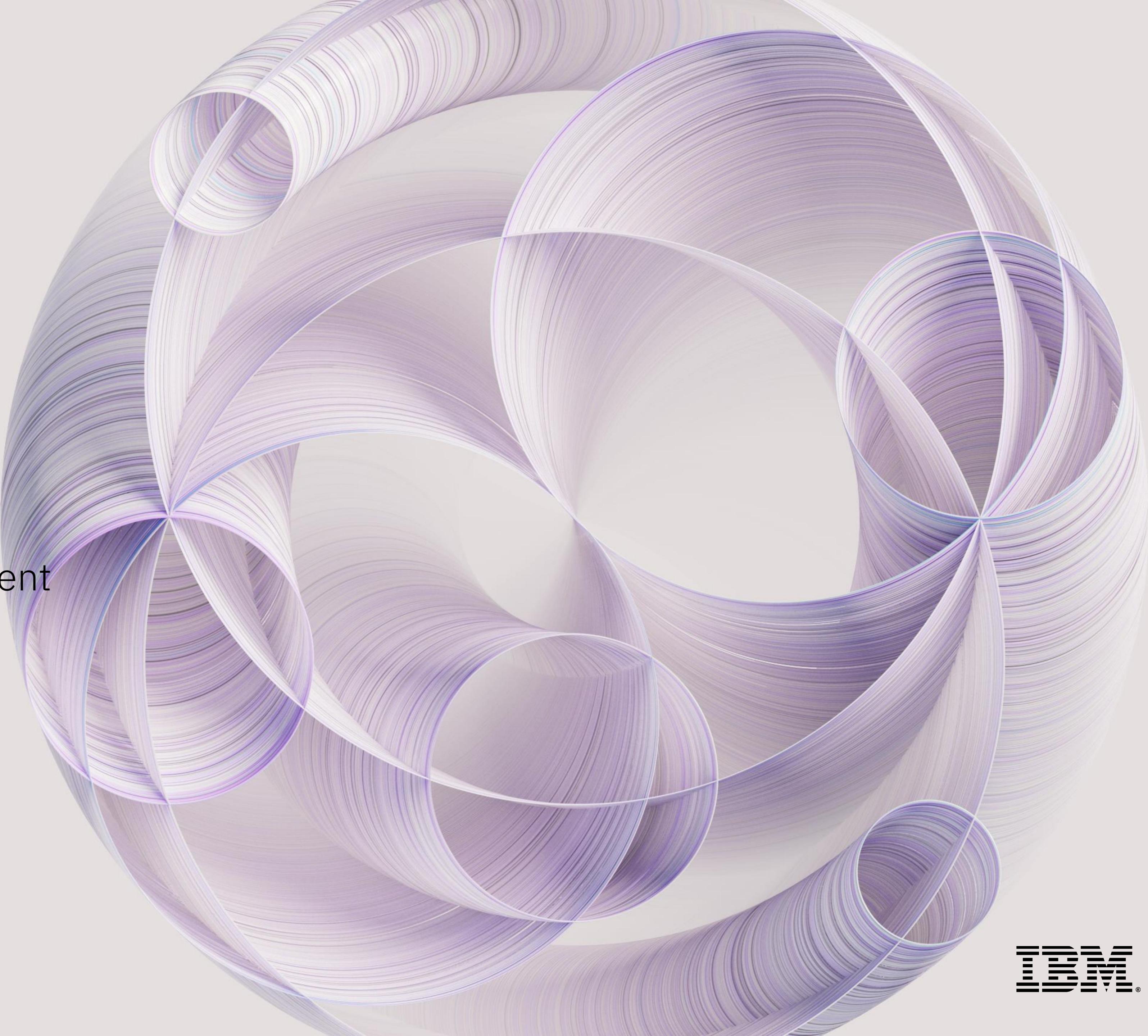
AutoAI-RAG has 3 layers:

- 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

What's Coming InstructLab

Angela Jamerson
Program Director, Product Management
IBM watsonx.ai

watson**x**.ai



IBM
®

One size does not fit all for customizing LLMs

Making model customizations easy

watsonx.ai developer toolkit

Prompt Lab
Context customization

Prompt engineering

Tuning Studio
Tune foundation models with labeled data

Prompt tuning

Fine-tuning

InstructLab

External customization options

Parameter-efficient fine-tuning (PEFT)*

Custom pretraining

watsonx.ai models

InstructLab in watsonx.ai

Provide a platform for collaborative language model enhancements to build the future of generative AI.

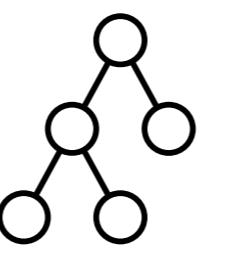
Share the client presentation on [Seismic](#)

IBM watsonx.ai adds enterprise workflows to InstructLab model alignment with a streamlined developer experience

watsonx.ai value adds



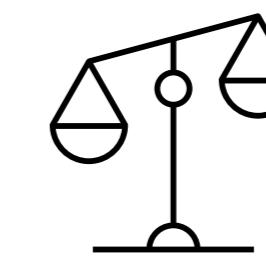
Advanced data ingestion



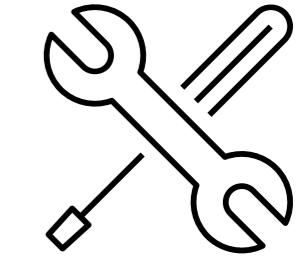
Data curation and lineage



Data and model evaluation

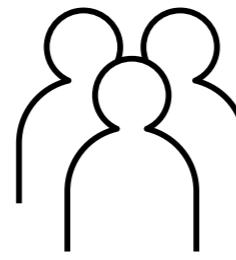


Audit readiness

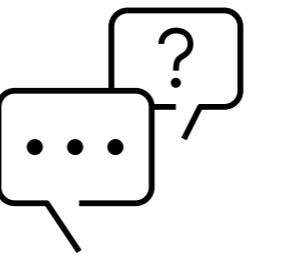


Developer AI toolkit

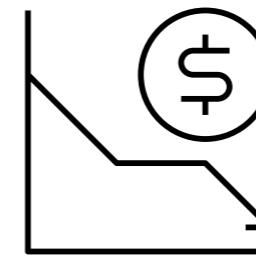
Customer value



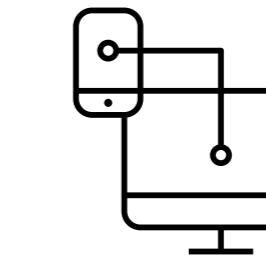
Collaborative contribution



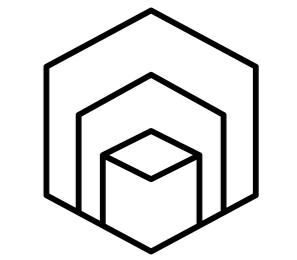
Rapid iteration



Cost savings



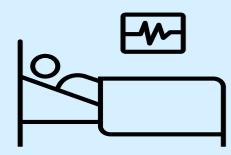
Unified experience



Out-of-the-box integrations

Sl No	Requirements	Details	RHEL AI	OSAI	watsonx.ai
1	Deployment	1. Deployment type directly affects time for SDG and tuning time	Single node SDG, tuning Self-Managed Only	Distributed tuning and SDG	Distributed tuning and SDG
2	Platforms supported	1. Cloud environments 2. On-Prem environments	Bare Metal PaaS on Hyperscalers	Managed and self-managed OpenShift environments (OS Data Foundation, HCI, Portworx, NFS, NetApp Trident)	IBM Cloud, AWS, Dell PowerEdge, Fusion HCI , Portworx ,NFS, NetApp Trident
3	Pdf to md conversion	1. Basic open-source based data ingestion	✓	✓	✓
4	Out of the box connectors	1. Connection to enterprise storage connectors like watsonx.data, Amazon S3, Box, Azure and more	✗	✗	✓
5	Data ingestion pipeline	1. Enterprise grade chunking of grounding documents 2. HAP and PII filtering	✗	✗	✓
6	Versioning	1. Versioning of synthetically generated data and fine-tuned model	✗	✗	✓
7	Data lineage	1. Tracking of taxonomy to synthetic data to fine-tuned model	✗	✗	✓
8	Audit readiness	1. Data and model traceability 2. Access control and monitoring 3. Compliance and regulatory requirements	✗	✗	✓
9	Model governance	1. Model performance 2. Explainability, interpretability 3. Drift, bias, fairness	✗	✗	✓
10	Open-source evaluation benchmarks	Loss function, MMLU, MT-Bench results	✓	✓	✓
11	Customer defined and proprietary benchmarks	1. IBM Bluebench benchmarks 2. IBM proprietary benchmarks around text classification, sentiment analysis, entity extraction, summarization 3. Customer defined benchmarks	✗	✗	✓
12	Model deployment and inference	1. Ability to deploy and infer the model	✗	✗	✓
13	Enterprise RAG	Build an enterprise RAG solution with in-built scalability and governance	✗	✗	✓
14	Agentic framework	LLM Agents crafted to execute intricate tasks demanding reasoning, planning and memory capabilities	✗	✗	✓

GenAI Use Cases Improved by InstructLab



EMERGENCY MEDICAL SERVICES

Use case: A large hospital needs to automate processing of emergency medical records and be capable of:

- Case classification (green, orange or red flag)
- Peer medical review recommendation (e.g. does patient require further review)
- Clinical compliance and guideline deviation (e.g. identify report discrepancies compared to guidelines)

Knowledge: hospital compliance data

Skill: case classification and compliance deviation

watsonx ai



RAG

Model inference



CALL TRANSCRIPT PROCESSING

Use case: Large North American Telco needs to summarize incoming support call transcripts to get information based on a set of 80 questions, e.g.:

- Did the customer want to upgrade their plan?
- Did the customer complain about bandwidth issues?

Knowledge: call transcripts
Skill: interpreting human writing styles

watsonx ai



RAG

Model inference



PERSONALIZED RETAIL RECOMMENDATIONS

Use case: A retailer needs a personalized recommendation engine to suggest products in-stock based on user's dietary preferences.

Knowledge: product nutritional information

Skill: recommend products based on ingredient classification

Agents: understand inventory, client preferences and purchase history

watsonx ai

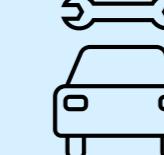
watsonx governance



RAG

Model inference

Agents



INTELLIGENT AUTO CLAIMS PROCESSING

Use case: An insurance provider needs to process images of an auto accident and suggest insurance coverage based on the active policy.

Knowledge: policy details

Skills: accident severity classification

Agents: understand driver history, determine prevailing policies

watsonx ai



RAG

Model inference

Agents

Multi modal

watsonx.ai InstructLab Roadmap

2Q 2024

Pre MVP

- Taxonomy explorer
 - Skills tree
 - Knowledge tree
 - Seeds examples
- InstructLab models
 - Granite-7b-lab
 - Granite-20b-multilingual
 - Merlinite-7b
 - Granite-13b-chat-v2

4Q 2024

InstructLab MVP on SaaS

- Enterprise taxonomy workflow
 - Support innersource across an organization
- Skills and Knowledge submission UI (via Tuning Studio)
- Scalable data ingestion for large sets of data or documents
- Model evaluation benchmarks (Loss function, MMLU, MT-Bench)
- Seamless deployment and inference of InstructLab-tuned models (GUI and API accessible)
- InstructLab CLI

1Q 2025

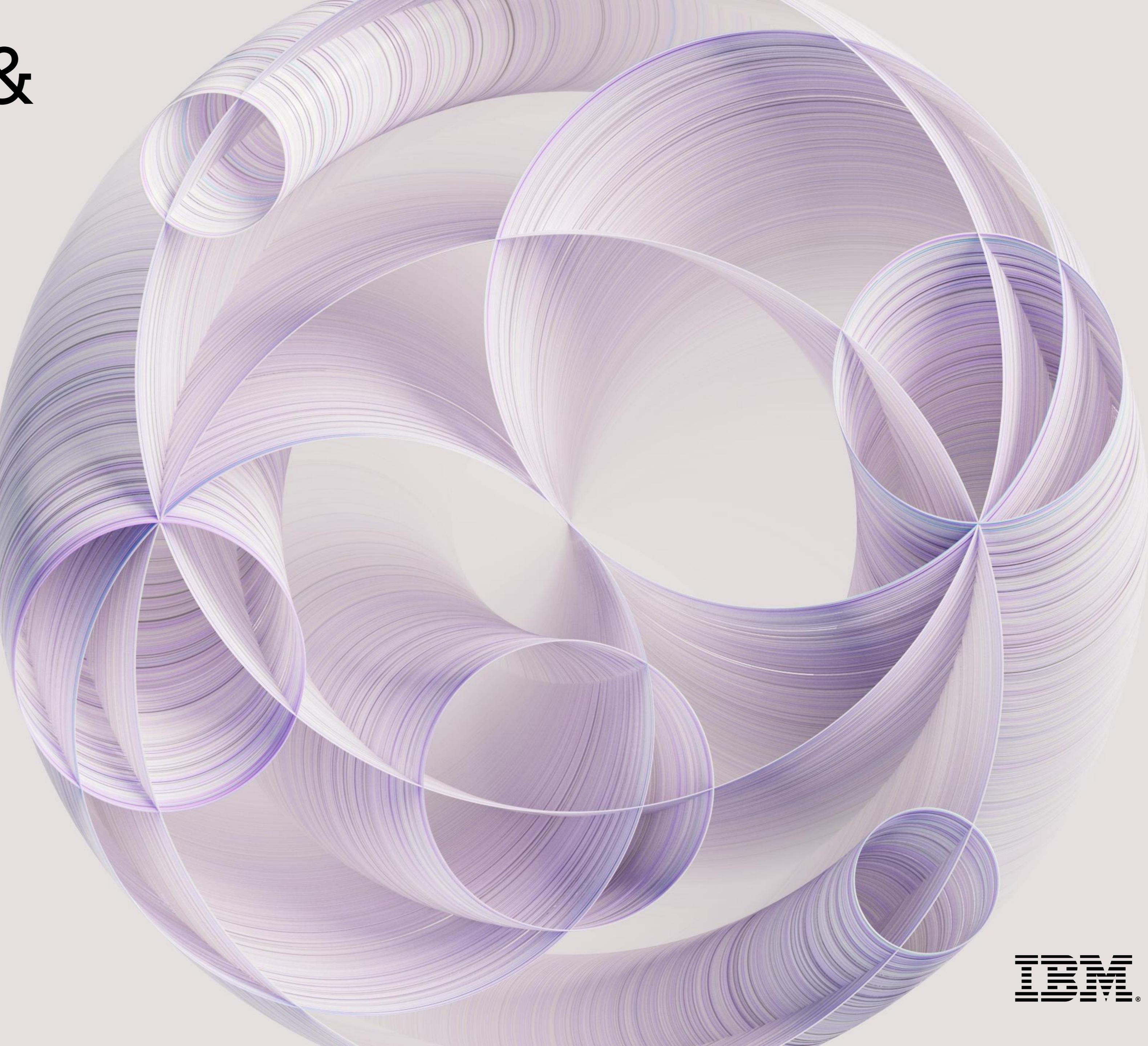
InstructLab Enhancements

- Taxonomy management
 - Search
 - Edit data lineage
 - Delete skills and knowledge
 - Taxonomy versioning
- Advanced data ingestion
 - Pipelines, HAP/PII Filtering, Masking, Chunking, OOTB data source connectors
- Additional benchmarks
 - Proprietary model evaluation benchmarks
 - Customer defined model eval benchmarks
 - Synthetic data eval benchmarks

Chat with Images & Model Updates

Nisarg Patel
Product Manager, Foundation Models
IBM watsonx.ai

watson**x**.ai



IBM
®

Chat with Image

Convert visual information into text to help with the following types of tasks:

- Automate the generation of alternative text for images to meet accessibility requirements
- Summarize photos of property damage that accompany insurance claims
- 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). Mistral AI's **Pixtral-12b** upcoming on 10/21.

Learn more in the [documentation](#) and [release blog](#).

The screenshot shows a chat interface with the following elements:

- Header:** Chat, Structured, Freeform, Model: llama3-llava-next-8b-hf, and file icons.
- User Message:** MM You: 03:33 PM (with a photo of a small, shaggy dog).
- AI Response:** watsonx 03:33 PM (with a detailed description of the dog's appearance and setting).
- Input Field:** Type something... with a send icon.

The AI response reads:

The image shows a small dog with a shaggy, curly coat that appears soft and matted in places. The dog has a round, woolly head with pronounced eyebrows and a slightly tilted, alert expression. Its eyes are hidden by the curls of fur, giving it a somewhat comical or endearing look. The dog's ears are small and lie close to its head, and it has a short muzzle and a black nose. It is wearing a collar with a tag, suggesting it is a pet. The dog is standing on a wooden floor, and there is a window with a curtain in the background, which is out of focus. The overall style of the photo is candid and appears to be taken indoors.

Meta-Llama 3.2

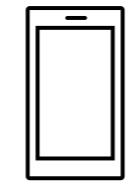
LLMs are the foundation of watsonx.ai, and as such we aim to provide clients with as much **choice** as we can



Vision

The new multimodal Llama 3.2 vision LLMs (11B and 90B) can reason on high resolution images up to **1120x1120 pixels**, enabling their use for **computer vision** tasks:

- *Image classification*
- *Object detection*
- *Image-to-text transcription (including handwriting) through optical character recognition (OCR)*
- *Contextual Q&A*
- *Data extraction and processing*,
- *Image comparison*
- *personal visual assistance.*



Edge

The lightweight 1B and 3B models are meant for local, **edge use cases**.

These models empower developers to:

- build personalized, on-device agentic applications with **strong privacy**
- Develop applications with **quick response times** due to on-device processing
- Take advantage of traditional capabilities of larger models like large context lengths, with multilingual text generation and tool calling abilities.



Guardrails

Llama Guard 3 Vision is a Llama-3.2-11B pretrained model, fine-tuned for **content safety** classification.

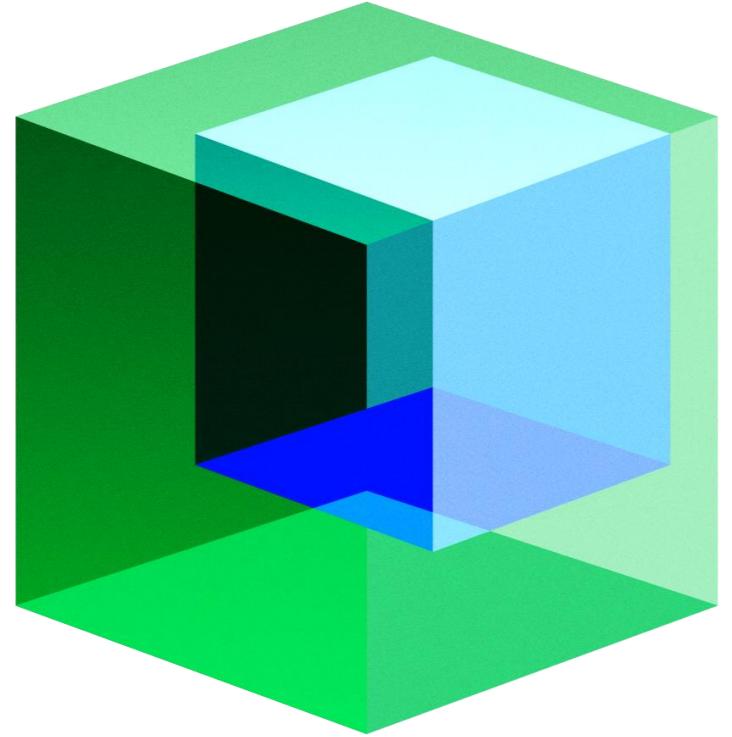
It supports image reasoning use cases and was optimized to **detect harmful multimodal (text and image) prompts** and text responses to these prompts.

Hazard categories

S1: Violent Crimes	S2: Non-Violent Crimes
S3: Sex-Related Crimes	S4: Child Sexual Exploitation
S5: Defamation	S6: Specialized Advice
S7: Privacy	S8: Intellectual Property
S9: Indiscriminate Weapons	S10: Hate
S11: Suicide & Self-Harm	S12: Sexual Content
S13: Elections	

Granite 3.0

IBM's third generation of Granite LLMs, built for enterprise



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

- New Granite-Dense architecture
- Over 12T tokens training data across 12 languages and 116 programming languages
- Available under Apache 2.0

Coming October 21st:

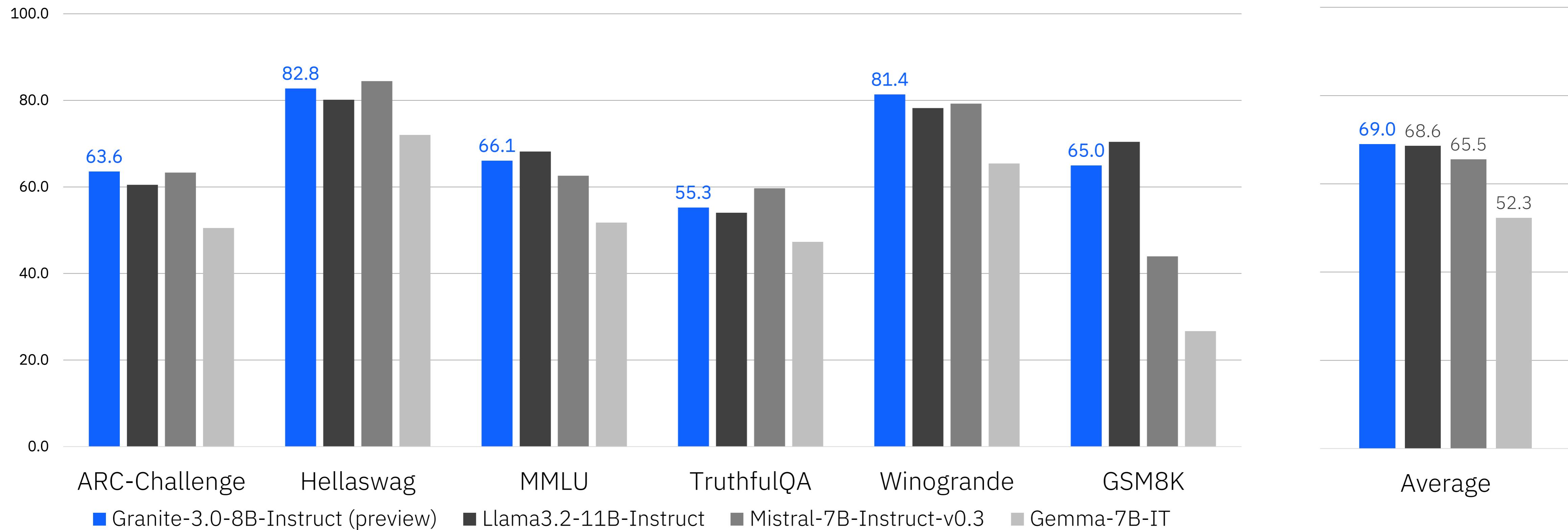
- Language: RAG, summarization, entity extraction, classification
- Code: code generation, code explanation, code editing
- Agents: tool use, json prompting
- Safety: available with Granite Guardian companion model

By end of Year:

- Extension to 128K Context Length
- Multi-modal adapters for document-understanding tasks (chart understanding, scanned pdfs, GUI screens, etc.)

Granite-3.0-8B-Instruct meets or exceeds SOTA performance across top academic benchmarks

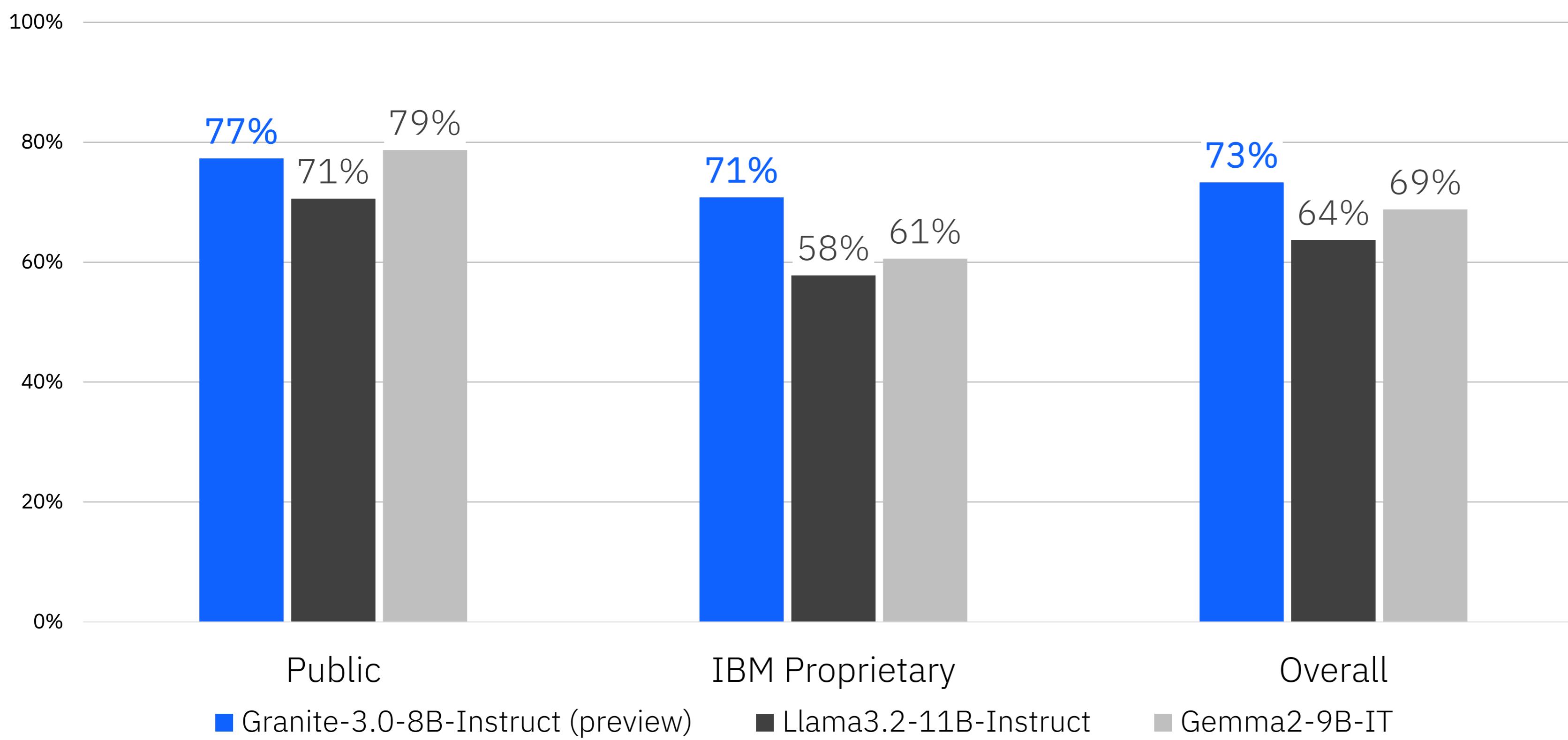
Comparison on Hugging Face's OpenLLM V1 Leaderboard Benchmarks



Based on preliminary evaluations, numbers subject to change

Granite-3.0-8B-Instruct demonstrates differentiated performance on top enterprise tasks, like Cybersecurity

Comparison on Public and IBM Proprietary Cybersecurity Benchmarks



Public Benchmarks Include

- SecMMLU
- CyberMetric-80
- CyberMetric-500
- Linkedin Sec Test Assessment
- CTI Bench RCM
- CTI Bench MCQ
- SecQA-v1
- SecQA-v2
- SecEval
- CISSP Test

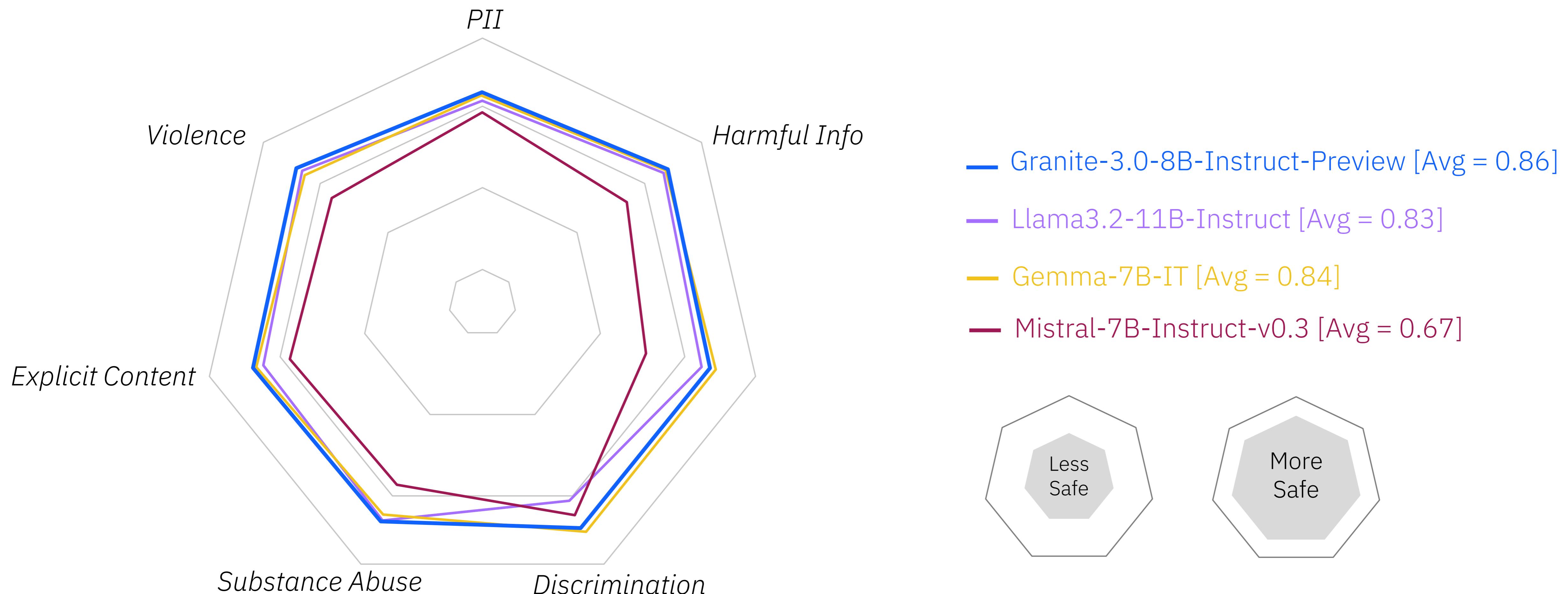
IBM Proprietary Benchmarks Include

- MITRE MC
- SIEM Rule TTP Mapping
- CWE Impact Classification
- BRON MC/A-B Classification
- BRON Yes/No Tasks
- CWE Summarization

Based on preliminary evaluations, numbers subject to change

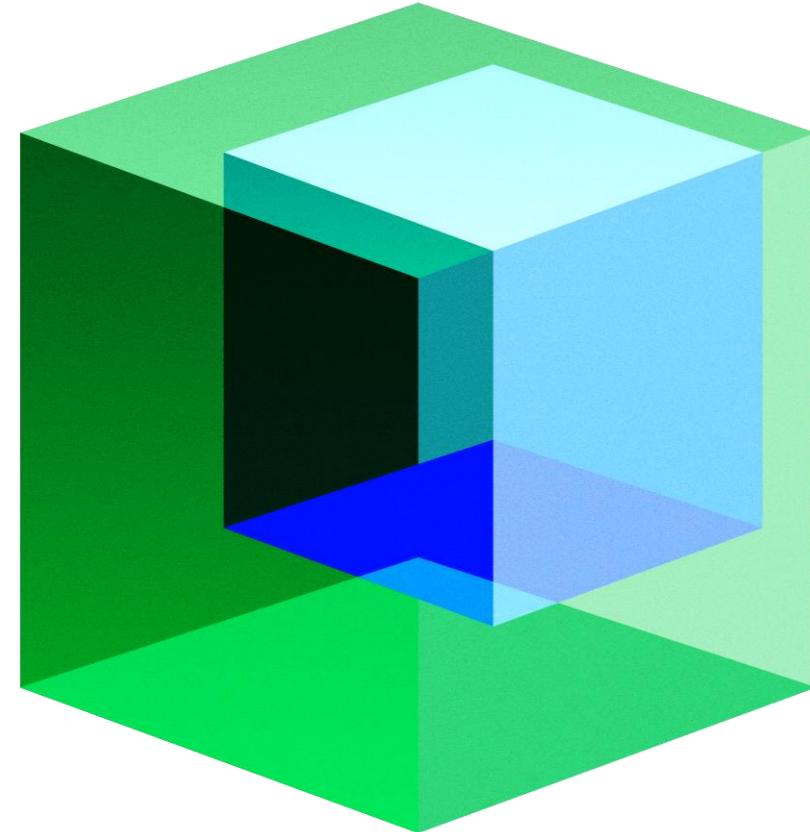
Granite-8B-Instruct is aligned to maximize safety and support trusted deployments

Comparison on IBM's Red-teaming Benchmark, AttaQ



Granite Guardian

The Next Step in Responsible AI



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

- LLM based input-output guardrail
- Available under Apache 2.0

Coming October 21st:

- Includes robust RAG metrics including RAG faithfulness, RAG relevance, and RAG groundedness not included in Llama Guard evaluations.
- Trained to detect jailbreaking, bias, violence, profanity, sexual content, and unethical behavior

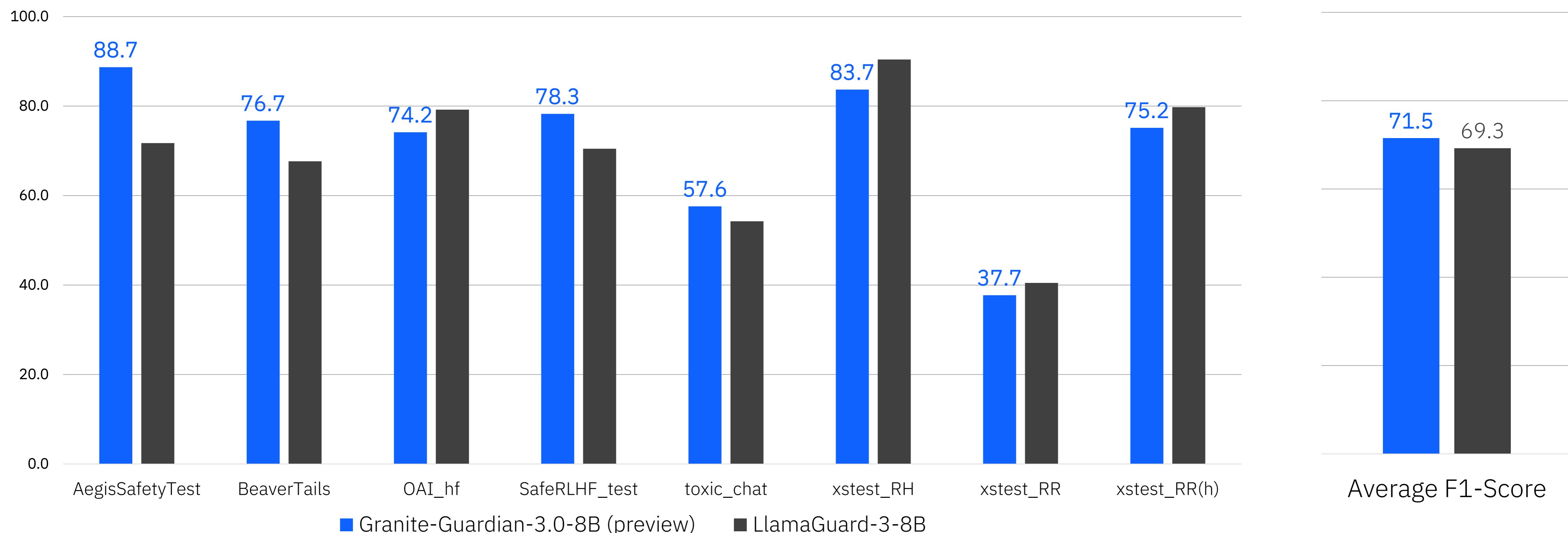
By end of year:

- Small Granite-Guardian(s) for efficiency (e.g. 1B and 400M)
- Additional pre-built detections on areas like PII and function calling hallucination detection
- Improved performance on custom risk detection

Granite Guardian

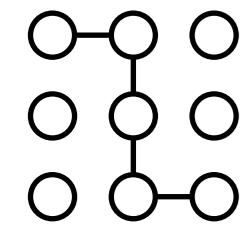
The Next Step in Responsible AI

Comparison of average F1-score on risk detection public benchmarks



Based on preliminary evaluations, numbers subject to change

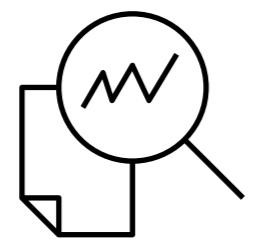
IBM Granite Everywhere



Open-Source communities (Hugging Face & GitHub)

For generative AI model testing & experimentation directly on your laptop or workstation to collaborate on models, datasets, or applications

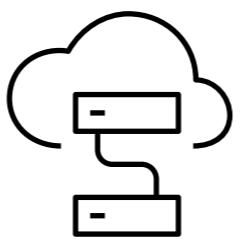
Access new modalities before anyone else, all Granite models released under Apache 2.0 license i.e., code, time series, language, and geospatial



InstructLab

Open-source, model-agnostic toolkit, enables individuals to contribute knowledge and skills to foundation models, facilitating a new “cheaper” approach to alignment tuning that leverages both human and model evaluations for quality assurance

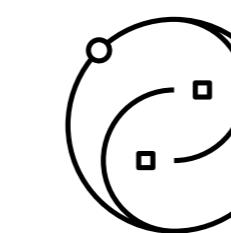
Easily generate synthetic data using a teacher model to train future version of the base model



Red Hat Enterprise Linux AI (RHEL AI)

Brings together the Granite family of open sourced-licensed LLMs, InstructLab model alignment toolkit, a bootable image of Red Hat Enterprise Linux, and enterprise-grade technical support

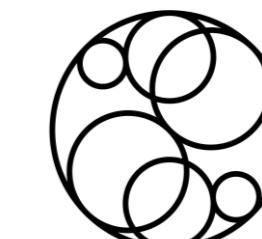
Model IP indemnification for use of Granite open sourced-licensed models provided by Red Hat accessible in RHEL AI



watsonx.ai

Enterprise-grade AI toolkit underpinned by a middleware foundation to deploy AI applications at enterprise scale with lifecycle governance, customizable model choices, and tooling to build AI assistants and agents

IP indemnification for use of Granite models on watsonx.ai, including proprietary Granite language models to watsonx.ai



watsonx Assistants

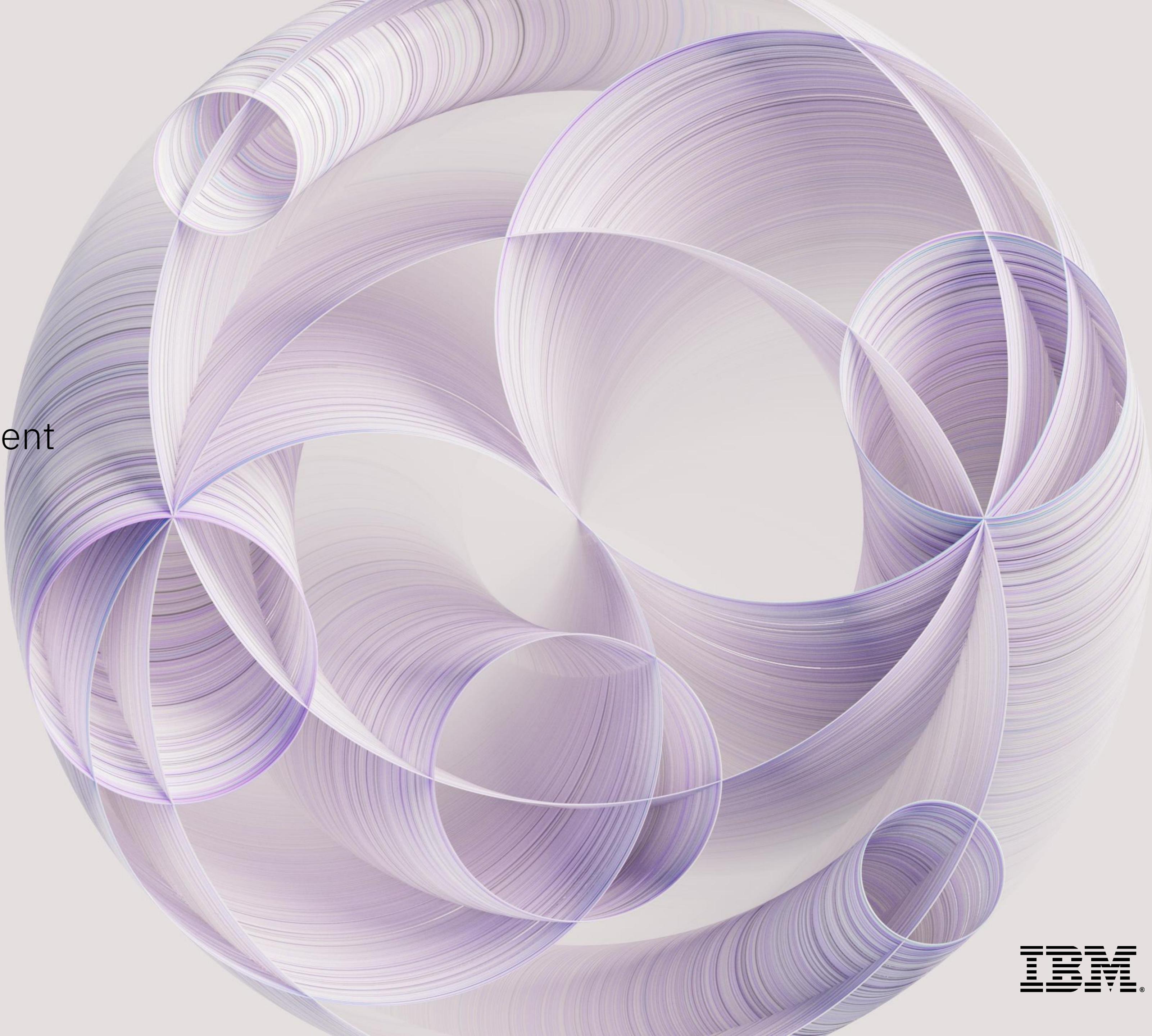
Access Granite specialized models that are fine-tuned and embedded directly into AI applications to solve specific use cases i.e., code generation from COBOL to Java, code completion, etc. for a range of supported languages

Granite models cannot be downloaded or manipulated outside the application

Platform Updates

Angela Jamerson
Program Director, Product Management
IBM watsonx.ai

watson**x**.ai



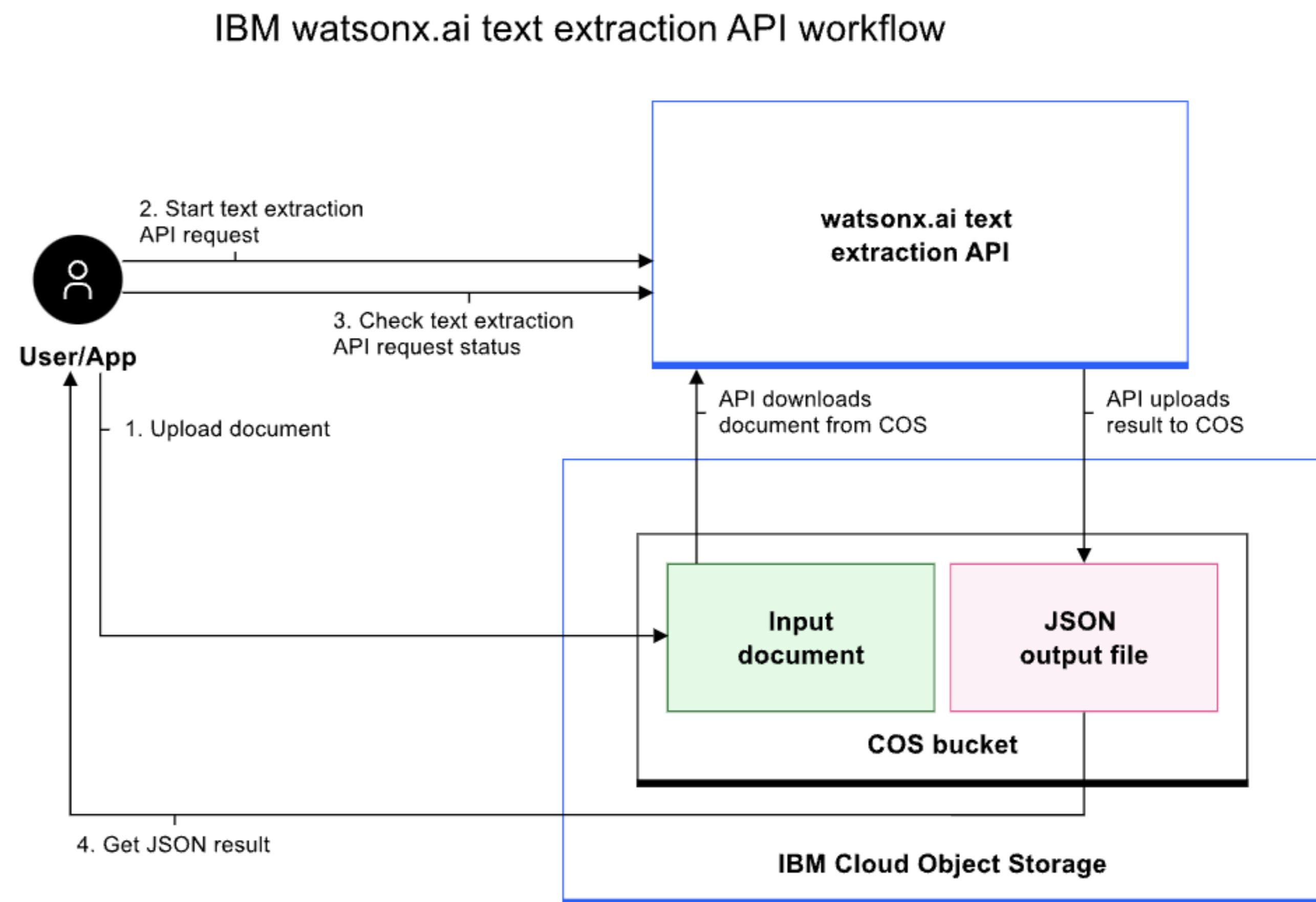
IBM®

Text Extraction REST API - Beta

- Essential for RAG use cases where complex documents are required
- 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](#).

Text Extraction (post GA)

Pricing and Packaging

- Available with **Essentials** and **Standard** Plans (excluded from trial/lite plan)
- Billing is based on the number of **Pages** processed (no charge while in beta).
- Each document page, image file or .tiff frame is considered 1 Page.
 - Else, 1800 characters will be considered 1 Page
- Each Page is charged at a flat rate

Text Extraction Price per Page	Text Extraction Part Name
\$0.038 USD/Page	IBM watsonx.ai Document Understanding Category 1 Essentials
\$0.03 USD/Page	IBM watsonx.ai Document Understanding Category 1 Standard

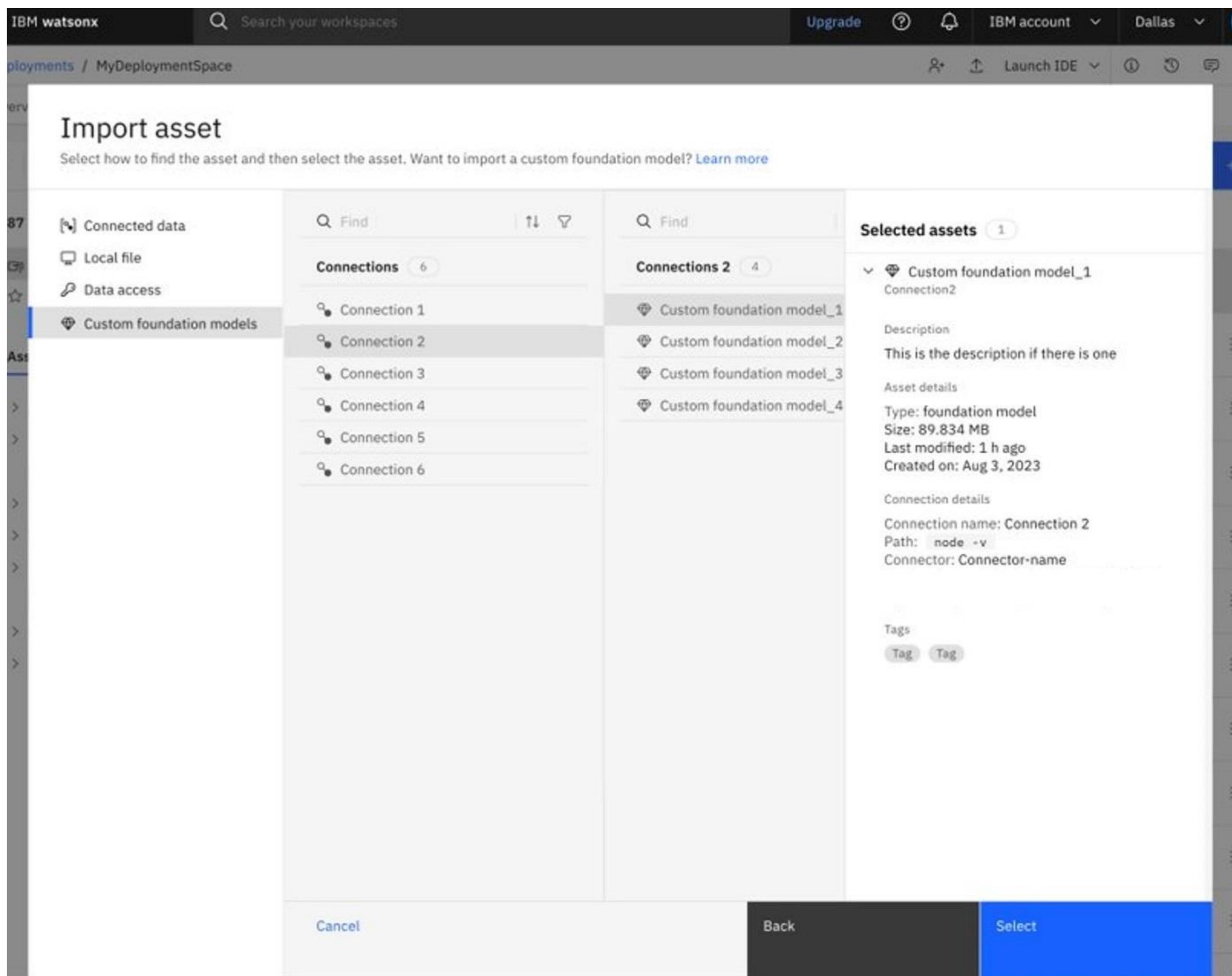
Custom Foundation Models

Why import a custom foundation model?

- Provides greater flexibility in how generative AI use cases are developed
- Leverage externally fine-tuned LLMs, e.g.
 - Trained for specific languages
 - Customized for an industry or business domain
- Leverage models from repositories like Hugging Face, which provides access to a huge selection of open-source foundation models

Available in Software and on IBM Cloud in Dallas for **Standard Plan** clients

Learn more in the [blog](#), [tutorial](#) or [documentation](#)



Custom Foundation Models

Pricing and Packaging

- Available for **Standard Plan** only
- **Hourly** hosting rate; includes both hosting and inference
- Client can deploy **4 small** or **2 medium** or **1 large** sized foundation model

Three configurations are available to support custom foundation model deployment:

Hosting Price per Hour	Configuration (client facing)	Infrastructure Environment (not displayed to client in the UI)	Part Name (not displayed to client in the UI)
\$5.22	Small	1 GPU (<i>currently A100 only</i>)	IBM watsonx.ai Model Hosting Category 1
\$10.40	Medium	2 GPU (<i>currently A100 only</i>)	IBM watsonx.ai Model Hosting Category 2
\$20.85	Large	4 GPU (<i>currently A100 only</i>)	IBM watsonx.ai Model Hosting Category 3

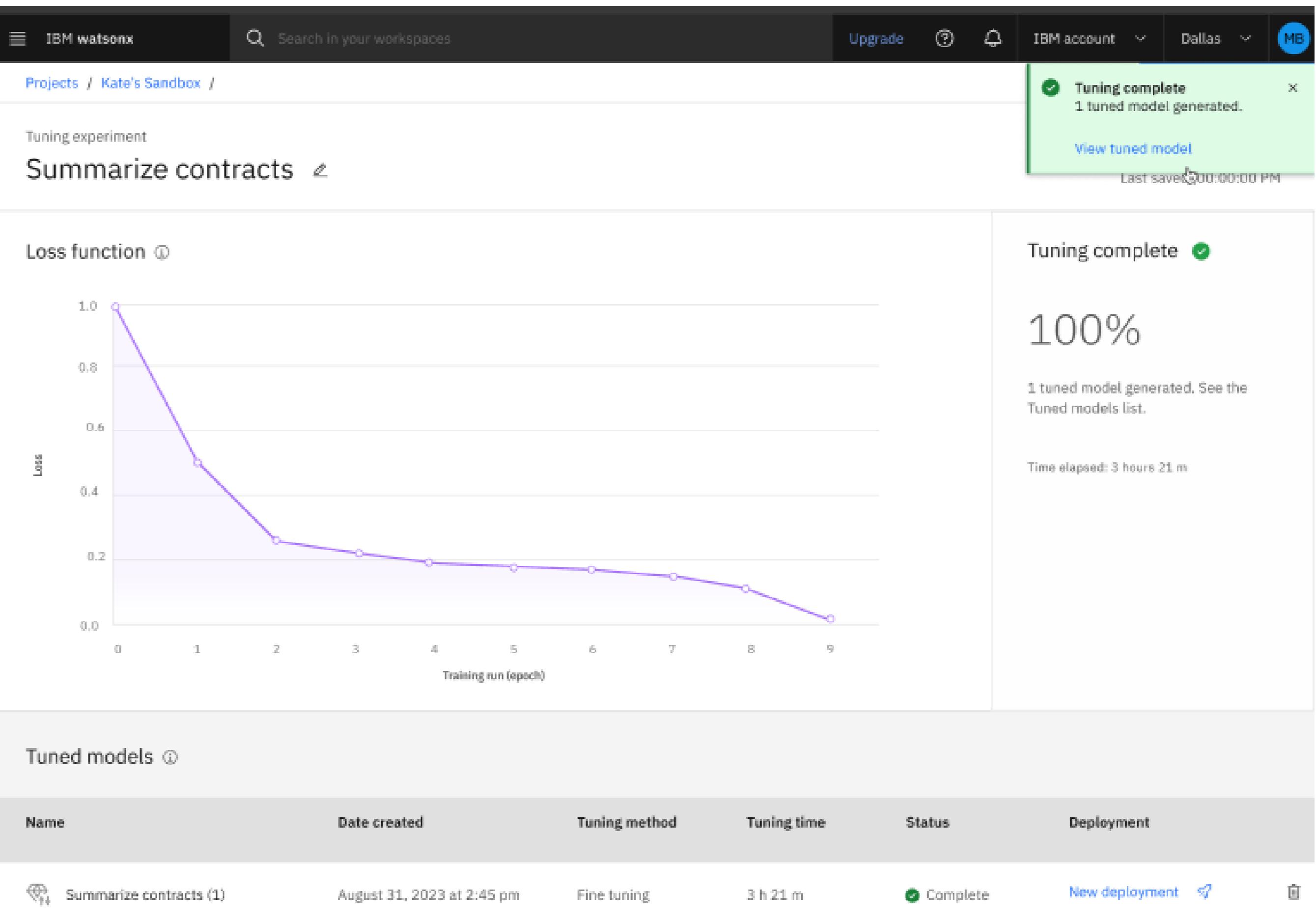
Tuning Studio: Fine Tuning

Unlock the full potential of foundation models by customizing it for a specific task, use case or business domain with fine tuning

- Fine-tune with data sets that are task or domain specific to achieve:
 - Improved performance and accuracy
 - Reduced Bias from original training data
- Easy to use interface
- Supports structured training data in files or connected data sources
- Adjust parameters and view performance metrics

Available in Software v2.0.3, Coming soon to IBM Cloud

Learn more in the [FAQ](#), [Software documentation](#) or technical deepdive [playback](#).



watsonx.ai v2.0.3

GA: Friday, September 27

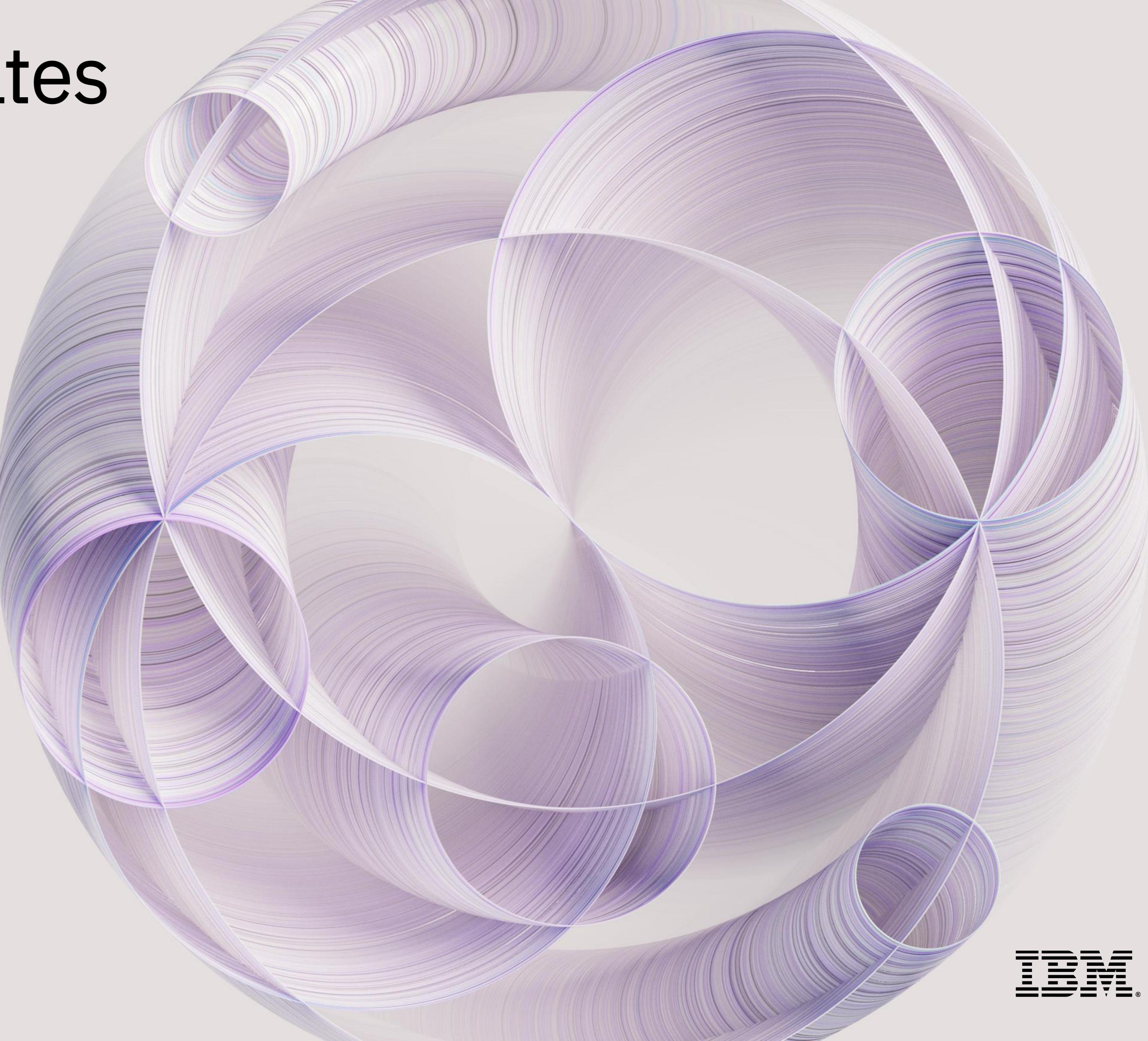
Visit the watsonx.ai [“what’s new”](#) page
in IBM software documentation to learn more

Sales Motion Updates

Luv Aggarwal
Worldwide Sales Leader
IBM watsonx.ai

Michael Connor
GTM Product Management
IBM watsonx

watson**x**.ai



IBM®

Mistral Large 2 on watsonx.ai

Large Enough

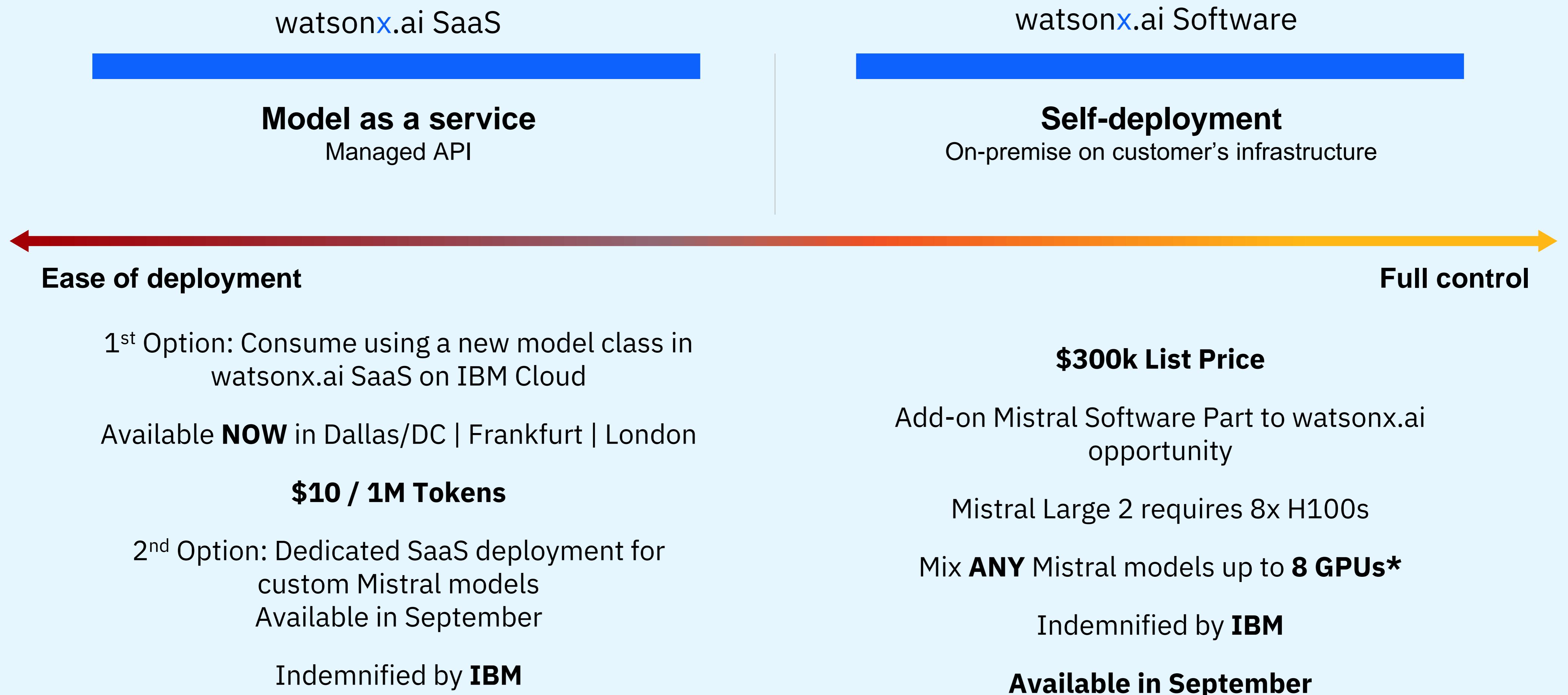
Today, we are announcing Mistral Large 2, the new generation of our flagship model. Compared to its predecessor, Mistral Large 2 is significantly more capable in code generation, mathematics, and reasoning. It also provides a much stronger multilingual support, and advanced function calling capabilities.

Mistral Large 2 has a **128k context window** and supports dozens of languages including **French, German, Spanish, Italian, Portuguese, Arabic, Hindi, Russian, Chinese, Japanese, and Korean**, along with 80+ coding languages including Python, Java, C, C++, JavaScript, and Bash.

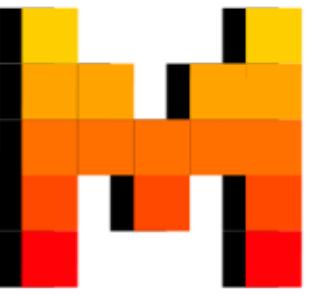
	Average	C++	Bash	Java	TypeScript	PHP	C#
Mistral Large 2 (2407)	74.4%	84.5%	51.9%	84.2%	86.8%	77.6%	61.4%
Mistral Large 1 (2402)	58.8%	67.1%	36.1%	70.3%	71.7%	61.5%	46.2%
Llama 3.1 405B (measured)	73.4%	82.0%	58.2%	82.9%	83.6%	73.9%	59.5%
Llama 3.1 405B (paper)	73.7%	82.0%	57.6%	80.4%	81.1%	76.4%	64.4%
Llama 3.1 70B	66.8%	70.2%	51.3%	74.7%	76.7%	73.3%	54.4%
GPT-4o	75.3%	85.7%	54.4%	82.9%	89.3%	79.5%	60.1%

Mistral Large 2 is designed for **single-node** inference with long-context applications in mind – its size of 123 billion parameters allows it to run at large throughput on a single node.

Mistral's full deployment flexibility fits perfectly into IBM hybrid-cloud deployments offering



*Limited to a total of 640GB of VRAM



Why Mistral + IBM

We are well positioned to support market need of scaling POCs to production globally

Scale your use cases while keeping **full control** over your **LLMs & Data**



For most companies, Leveraging private/blackbox LLM for POC was a fast and easy way to show ROI ...

... but there's an increase in concerns over letting companies and customers' data out in the field for scaled use cases

... Mistral enables customers to deploy On-Prem or leverage WatsonX

Some industries are even more data sensitive such as:
Financial Services, Defense, Government, Pharma, ...

Scale your use cases globally with the state of the art **multilingual model**



Our models over-perform the market in non-english languages while being smaller

Example for german	Mistral Large 2 123B	Llama 405B
arc_challenge	62%	➤ 59%
hellaswag	78%	➤ 77%
mmlu	82%	= 82%

Scale your use cases under a **legal framework**



Contracting with Mistral AI offers legal protection, liability to secure your use cases and indemnification offered by IBM

If a customer is sued, IBM will defend in court at its own expense. there is no cap.

Only if client loses or settles, then IBM will also pay the judgement for them, up to the license amounts paid.

Scale your use cases without compromising **cost & performance**



Mistral Models are well positioned in terms of Cost to Performance ratio. Offering state of the art accuracy on each model size

watsonx.ai Mistral ELA Attach

PID: 5900-B14 (as of 9/25/24)

UT30: Core Platform Ecosystem (30A1K)

Software parts available

- Subscription license
- Monthly license
- Parallel Z parts
- Production & non-production
- No perpetual parts

Dependency/pre-req: Yes, watsonx.ai entitlement

Charge metric: RU = 640 GB RAM

Enables flexibility to deploy 1 Mistral Large or several Mistral Small models (once available)

About Mistral Large 2:

- 128k context window
- Dozens of languages, incl: French, German, Spanish, Italian, Portuguese, Arabic, Hindi, Russian, Chinese, Japanese, and Korean,
- 80+ coding languages including Python, Java, C, C++, JavaScript, and Bash.

IBM is the exclusive provider of Mistral Large 2 on premises!

\$300k subscription per node per year

Receive 100% attainment for Mistral add-on part

Enable your customer to leverage a *GPT4o level performing model* while ensuring privacy and transparency, maintaining control, providing indemnification, and containing costs with Mistral Large on watsonx.ai

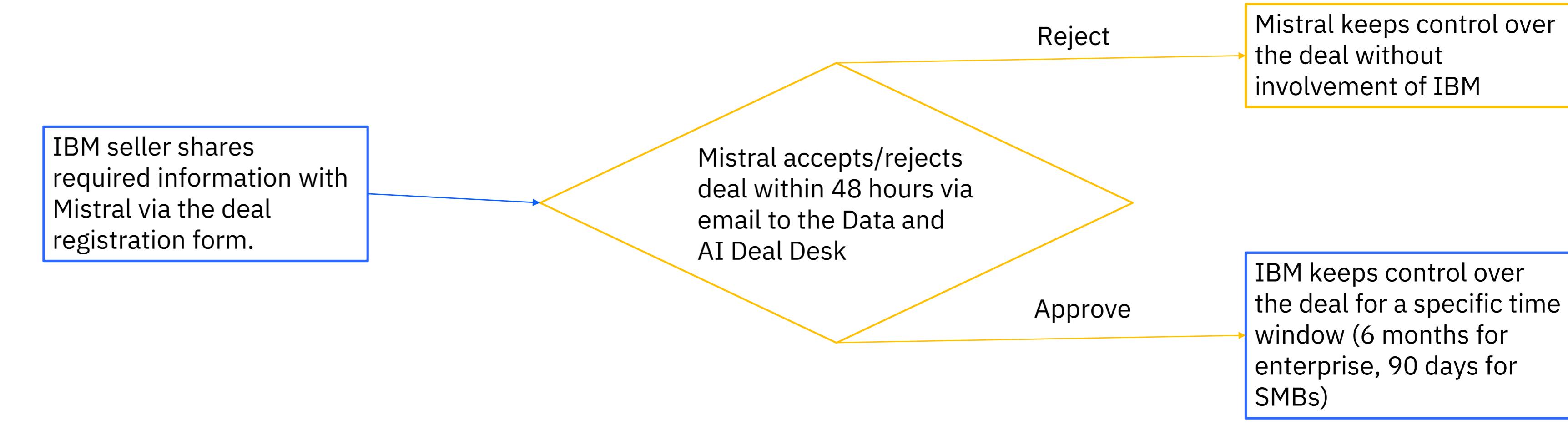
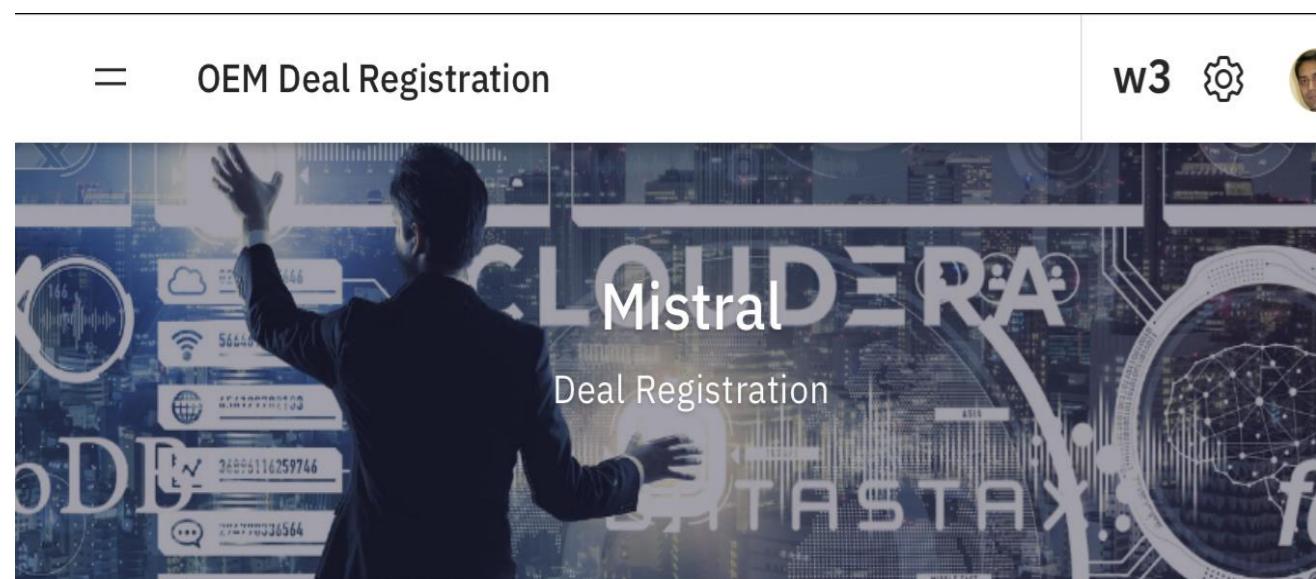
Ideal Client Profile:

- Competing against OpenAI/Azure OpenAI
- Requires a GTP4o performing model
- Public sector and **not authorized to deploy llama models**
- Use case focused on: RAG/Q&A, European language translation, code translation
- Requires an **indemnified model**

Part Number	Product Description	SAP Price Data Entry
D0QN4ZX	Mistral AI with IBM Resource Unit Monthly License	\$ 31,250.00
D0QN5ZX	Mistral AI with IBM Resource Unit Subscription License	\$ 25,000.00
D0QN6ZX	Mistral AI with IBM for IBM Z Resource Unit Monthly License	\$ 31,250.00
D0QN7ZX	Mistral AI with IBM for IBM Z Resource Unit Subscription License	\$ 25,000.00
D0QN8ZX	Mistral AI with IBM Non-Production Resource Unit Monthly License	\$ 15,625.00
D0QN9ZX	Mistral AI with IBM Non-Production Resource Unit Subscription License	\$ 12,500.00
D0QNAZX	Mistral AI with IBM Non-Production for IBM Z Resource Unit Monthly License	\$ 15,625.00
D0QNBZX	Mistral AI with IBM Non-Production for IBM Z Resource Unit Subscription License	\$ 12,500.00

Mistral Deal Registration Process

<https://w3.ibm.com/w3publisher/oem-deal-registration-portal/mistral>



Required Information for the Deal Registration to be sent to Mistral

Attributes	Description
IBM Opportunity Number	ISC Opportunity #
Opportunity Description	Description of the opportunity, business requirements and project details.
Customer Company Name	Name of organisation and specific entity/ sub-entity that intends to deploy the models
Business Unit(s)	Name of the business units and specific entities that intends to deploy the models
Customer Company Address	City, State, Country
Enterprise/SMB customer	IBM controls deals for 6 months on enterprise companies and 90 days for SMB
Target Solution (Use Case)	Description of use cases and value for customers
Anticipated Close Date	Date (Month,Year) at which the deal is expected to close
IBM Estimated Annual Contract Value (ACV)	Potential contract value, e.g. estimated budget that the customer is able to spend on Mistral model(s)
Target infrastructure: (GPU types) and # nodes	e.g H100, 8 production nodes, 2 non-production nodes
Term	Indicate length of time in month or years
IBM Seller Details	Name, E-mail, Phone number, Direct/IBM Business Partner
Expansion to an Existing Customer:	Yes / No

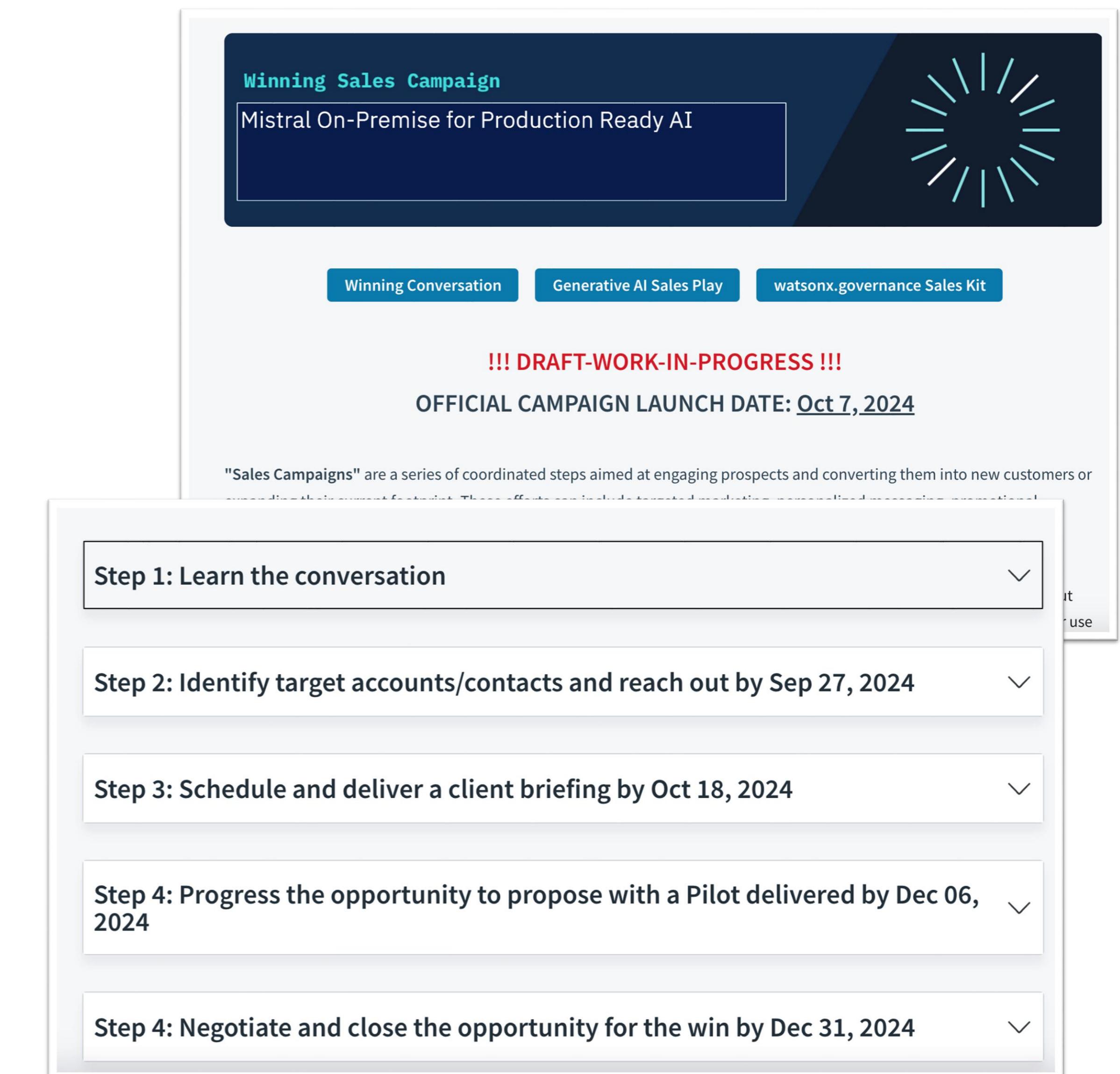
Mistral on watsonx.ai Sales Campaign

Financial Services / Healthcare / Telcom
OpenAI customers

- Customers in regulated industries
- OpenAI GPT4o performance across various use cases
- Maintain control of your data and workload
- Contain ballooning GenAI cloud costs by deploying workloads on software

watsonx.ai & OpenAI customers

- Upsell existing watsonx.ai customers
- Customers experimenting with OpenAI
- Provide customers with industry leading LLM to supplement watsonx platform and productionalize use cases at enterprise scale



The screenshot shows a dark-themed sales campaign landing page. At the top, it says "Winning Sales Campaign" and "Mistral On-Premise for Production Ready AI". Below that is a large button with three sub-links: "Winning Conversation", "Generative AI Sales Play", and "watsonx.governance Sales Kit". To the right is a decorative graphic of radiating teal lines. In the center, there's a red warning message: "!!! DRAFT-WORK-IN-PROGRESS !!!" and "OFFICIAL CAMPAIGN LAUNCH DATE: Oct 7, 2024". A note below explains: "'Sales Campaigns' are a series of coordinated steps aimed at engaging prospects and converting them into new customers or...". The main content area is a vertical stack of five expandable steps, each with a downward arrow icon:

- Step 1: Learn the conversation
- Step 2: Identify target accounts/contacts and reach out by Sep 27, 2024
- Step 3: Schedule and deliver a client briefing by Oct 18, 2024
- Step 4: Progress the opportunity to propose with a Pilot delivered by Dec 06, 2024
- Step 4: Negotiate and close the opportunity for the win by Dec 31, 2024

watsonx 4Q 2024 sales motions

Need	Available Now Motion	Prepare	Propose
Need AI Now but On Premise as sensitive data requirements	Fusion HCI with watsonx	Determine Application AI, Data, and Governance use case	On-Premise appliance - watsonx with Fusion HCI
Leverage ELA funds to move to AI	ELA Reserved to watsonx SaaS	Determine Unused watsonx in ELA/ACEL after year 1	Declare and park unused watsonx licenses. purchase watsonx SaaS
Business pressure to do more with AI while managing costs	Non ELA BYOL to watsonx SaaS	Determine Unused watsonx.ai and watsonx.governance VPC'	License driven discount lowers cost of watsonx SaaS
	CPD Trade-Up or Add-On to watsonx	Determine Unused CPD or ML only CPD VPC's	Trade-up or SL Add-on to watsonx.ai, .gov

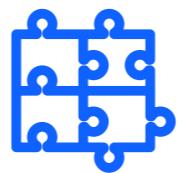
Fusion HCI enabled in Passport Advantage & coming in Sales Configurator?

Data & AI Sellers can easily package watson^x on a Fusion HCI system in Passport Advantage via quoting and ordering

Data & AI Sellers receive compensation each time they sell Fusion HCI in Passport Advantage – includes Fusion software and watsonx software

Sellers can configure watson^x software on an IBM Fusion HCI system in Sales Configurator

Sales Configurator can auto-generate the Bill of Materials into Software Sales, Quote and Order (SQO) system based on the configuration



Streamlined

The processes in Sales Configurator and Passport Advantage are already optimized for sales motion.



ELAs

Selling HCI from Passport Advantage allows Fusion HCI to be included in Enterprise License Agreements.



Compensation

Passport Advantage allows automatic and accurate compensation to the appropriate sellers.



Familiarity

Data & AI sellers already use these tools, so the only new learning at this point is the Fusion HCI.

IBM watsonx and Fusion HCI are better together

Delivering the most robust, reliable, and performance optimized OpenShift platform

AI and Models

- Address ML and GenAI development, RAG summarization, and inferencing
- IBM models responsive, lower costs with up to 96 x H100* or L40S GPUs

Data Lakehouse - for AI

- Lower cost, fully integrated, [better than Databricks at 60% cost](#)
- Address RAG use cases with vector store
- Quickly join data from different sources for rapid analytics
- Data lake, Hadoop augmentation

Governance

- Act as governance hub to assure truth, transparency, and explainability end-to-end
- Manage model lifecycles
- Increase visibility, evaluate risks



IBM Website

[IBM Fusion](#)

See Sales Kit

[IBM Fusion Sales Kit](#)

Level 1 Course

[Data Foundation for IBM Fusion](#)

Fusion HCI Benefits for watsonx

Speed of Deployment

Accelerate **watsonx** deployment on bare metal OpenShift – from months to days

Production Ready

Enterprise-grade, application aware
Backup, HA & DR

Highly available Hyper Converged Infrastructure

Compute, GPU, Storage and Network

Built-in Support Experience

Support for both **watsonx** and Fusion HCI

Simplify Day 2 Operation

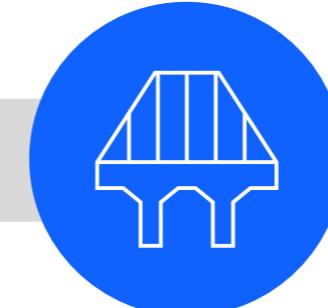
Automated lifecycle management

Reserved use concept



Client Managed Software

- Typically, large up-front capital expenditure investment for software license
- Purchased infrequently over longer periods of time (e.g. >2 years)



Flexibility

- Allows customers to **leverage existing ELA Cap catalog investment**
- Provides flexibility for many on-prem customers seeking SaaS benefits
- Allows customer "safe" exploration of SaaS as a deployment model
- Allows client to park watsonx licenses from an OTC Acel or SL Acel
- S+S used for immediate watsonx SaaS purchase
- Client reports reserved use as consumption against their full Catalog and put licenses on the shelf



SaaS

- Subscription or "pay-as-you-go" models
- Pricing typically based on consumption
- Requires minimal upfront investment by shifting capital expenditure to operational expenditure (CAPEX to OPEX)
- Sales receives SaaS signing credit/commissions

Reserved use detail



watsonx Reserved Use:

- Means watsonx OTC ACELs or SL ACELs that are reported against the catalog cap, that are not going to be in Active Use (these are parked/shelved licenses) and are placed in Reserved Use
- The unbilled watsonx S&S (annually) is for an IMMEDIATE purchase of watsonx SaaS
- Can be used for OTC ACEL Catalogs as well as SL ACEL Catalogs

When to offer:

- Client has undeployed watsonx ACEL or SL programs in a catalog and is finding it difficult to deploy those programs on prem; and
- Client wants to try a watsonx SaaS offering but does not have available funds

See sales kit update:
[Reserved Usage](#)



➤ Process

- Client must have expressed a desire to deploy watsonx on cloud and be unwilling to make an incremental purchase
- Available with approval from GEO ELA Sales Lead (identified in the ELA Guidelines)
- Sales Concession approval is required if Reserved Use is not in the existing contract
- Contract Amendment required
 - annual reserved rate for watsonx needs to be calculated from the original approved DMSS
 - Refer to ELA Guidelines section on **ACEL Annual Opt out for SaaS**

BYOL to SaaS (Non ELA) concept



Client Managed Software

- Typically, large up-front capital expenditure investment for software license
- Purchased infrequently over longer periods of time (e.g. >2 years)



Flexibility

- Allows customers to **leverage non ELA licenses**
- Bridges subscription and SaaS product offering
- Allows customers to leverage software license with equivalent SaaS product
- Provides flexibility for many on-prem customers seeking SaaS benefits
- Allows customer "safe" exploration of SaaS as a deployment model



SaaS

- Subscription or “pay-as-you-go” models
- Pricing typically based on consumption
- Requires minimal upfront investment by shifting capital expenditure to operational expenditure (CAPEX to OPEX)
- Sales opportunity longer term by focusing on extending the SaaS environment

BYOL to SaaS (Non ELA) detail



Traditional BYOL

Licenses:

- watsonx.ai and/or watsonx.governance must be active and remain active through the discount

When to offer:

- Client has undeployed watsonx.ai and/or watsonx.governance licenses and is finding it difficult to deploy those programs on prem; and
- Client wants to try a watsonx SaaS offering and wants to leverage offered discount to better utilize available funds(currently 32%)



BYOL to SaaS

➤ Process

- From watsonx Cloud environment, customer registers for discount
- Offering, discount are confirmed
- SD Addendum documenting terms is accepted
- Customer obtains discount upon next billing

The screenshot shows a web-based application for managing discounts in the IBM Watson environment. A modal window titled 'Add discount' is open, prompting the user to 'Select a cloud service' and 'Select a software program'. The 'Cloud service' dropdown is set to 'watsonx.ai - Watson Studio Professional' and the 'Software program' dropdown is set to 'IBM watsonx.ai'. The modal also indicates that a 32% discount will be applied to CUH. The background shows the main Watson dashboard with various service icons and account information.

[Internal Blog](#)

See sales kit update: [BYOL to SaaS](#)

Tradeup and upgrade options to watsonX

CP4D licenses

From	To (Subscription Add on or Perpetual Trade-up)	Availability
Cloud Pak for Data Enterprise Edition	watsonx.ai watsonx.governance (Model Management)	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling
Cloud Pak for Data Standard Edition	watsonx.ai watsonx.governance (Model Management)	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling

Governance Licenses on CP4D

Openpages Cartridge	watsonx.governance Risk and Compliance Foundation	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling
OpenPages Applications Model Risk Governance, Operational Risk Management, Regulatory Compliance Management	watsonx.governance Model Risk Governance, Operational Risk Management, Regulatory Compliance Management	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling
AI Governance Install	watsonx.governance AI Governance	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling

Advanced Analysis and Decision Optimization Licenses

SPSS Modeler, Modeler Server, Collaboration and Deployment Services, Analytic Server	watsonx.ai (TU only)	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling
ILOG Cplex Optimizer, Optimization Studio	watsonx.ai (TU Only)	Perpetual Tradeups and SL Upgrades are in Modernization Engine in Partner Guided Selling

- Prod, Non-Prod, zParts included where existing
- Various editions included – see excel mapping file for full list on [Modernization Central](#)
- VPC's and Install metric offerings transition 1-1, other metrics transition to VPC's at predefined multipliers
- If Perpetual Transition, customers stay on Transition Edition moving forward
- Pricing calculations performed in the [Modernization Engine in Partner Guided Selling Tool](#)
- **Mandatory TEL service requirement**

Clients with CP4D Perpetual licenses – trade-up to watsonx.ai

Trade-up at Renewal



Example

- Client has 128 VPCs of Cloud Pak for Data Enterprise Edition: Runs on 128 cores of CP4D EE
- Client can **trade-up** to a maximum of 128 VPCs of the watsonx.ai Transition Edition
- If client buys **trades up** 64 VPCs of watsonx.ai Transition Edition:
 - Client may deploy those 64 cores to watsonx.ai Client out of compliance if they deploy an 65th core to watsonx.ai
 - Client keeps 64 VPCs of CP4D EE as a perpetual license

Key Characteristics

- watsonx.ai Transition Edition is a **bundle** of: CP4D Enterprise Edition Transition and watsonx.ai
- Allows the customer to deploy either Cloud Pak for Data or watsonx or both
- Customer stays on Transition Edition for S&S Moving forward

Sales Process

- The CP4D perpetual license is traded up to a x.ai Transition Edition Perpetual License
- **Transition Edition is only available as a trade-up from Cloud Pak for Data**, net-new sales should use regular watsonx.ai parts
- Standard perpetual trade-up motion, includes 12 months of S&S for watsonx.ai so should be sold at renewal time

Pricing Calculation

- Modernization Engine provides pricing calculations in Partner Guided Selling Tool

Cloud Pak for Data
Enterprise Edition
128 VPC's, 128 Cores

Customer keeps the remaining VPCs
of CP4D EE as a perpetual license

Customer can trade-up
1 – 128 Wx.ai Transition VPC's

Cloud Pak for Data
Enterprise Edition
64 VPC's, 64 Cores

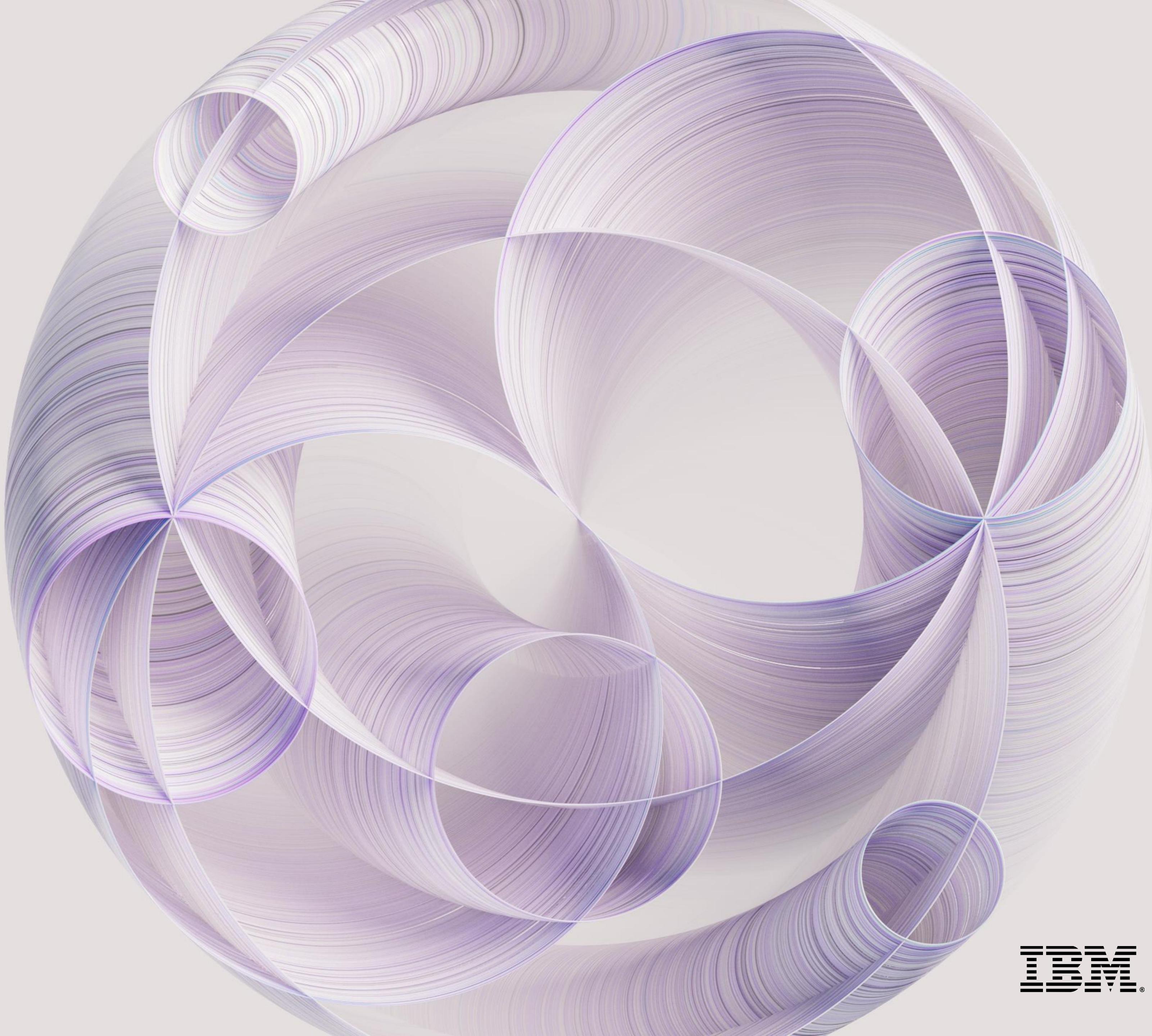
watsonx.ai
Transition Edition
64 VPC's, 64 Cores

See sales kit update:
[CPD to wx Transition deck](#)
[CPD to wx Parts list](#)

Call to Action

Luv Aggarwal
Worldwide Sales Leader
IBM watsonx.ai

watson**x**.ai



IBM®

watsonx Quickstarts v4 – September 2024 – Coming soon!

The screenshot shows the watsonx Quickstarts landing page. At the top, there's a decorative background with purple and white swirls and the text "watsonx Quickstarts". Below this, a section titled "What is a Quickstart?" explains that Quickstarts are prepackaged offerings with multiple products and services to help you prospect, identify, progress and win opportunities faster. It lists categories like Customer Experience, Digital Labor, IT Productivity, Data Transformation, and AI Governance. Under "Customer Experience", there's a card for "Customer Experience Quickstarts" from 2024. Under "Digital Labor", there's a card for "Digital Labor Quickstarts" from 2024.

Assets per Quickstart:

- Executive summary
- SW Subscription License
- SaaS T-shirt pricing
- Bill of materials
- Business Value Assessment
- Demos **New!**

Assets per Use Case:

- Expert Labs offerings: Enablement, workshops, full implementations, pilot to production, alpha development.
- Business Value Tree
- Fusion HCI sample configurations

Deprecated Use cases:

- Watsonx.data with Planning Analytics
- Watsonx.ai with Business Analytics Enterprise

watsonx Quickstarts

Customer Experience:

- Digital Assistant for Revenue Growth
- Modernize IVR
- Agent Assist
- Contact Center Insights
- Speech Transcription

Digital Labour:

- Employee Experience
- Expert Assist
- Intelligent Document Processing
- Enterprise Search

IT Productivity:

- Watsonx Code Assistant for Ansible
- Watsonx Code Assistant Platform **New!**

Data Transformation:

- RAG with watsonx.data and watsonx.ai
- Mainframe Modernization
- Lakehouse for Data Warehouse Optimization

AI Governance:

- Model Management
- Operationalize AI use cases with Risk Atlas
- Optimize AI Governance integrated with Amazon SageMaker **New!**

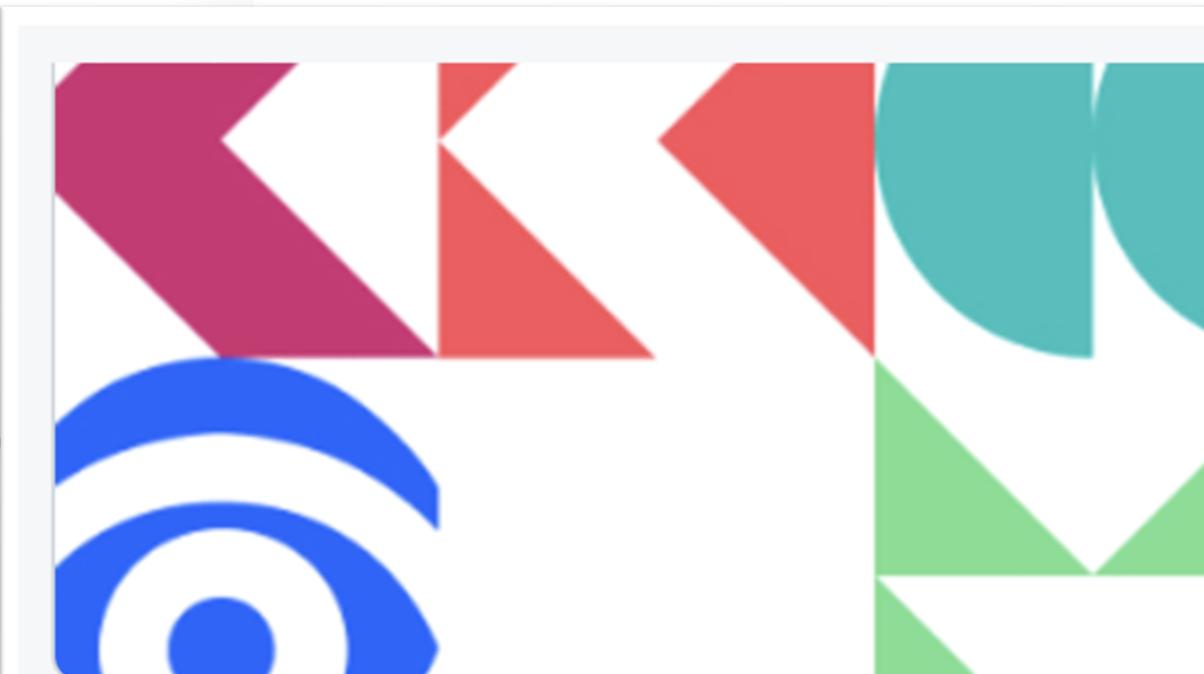
Fuel Your AI at the ultimate IBM learning event

IBM TechXchange Conference

October 21-24, 2024

Mandalay Bay – Las Vegas

Register now →



IBM TechXchange Conference 2024 - Asset Hub

Get instant access to all conference assets, invites, and deals in one place. Whether you're a speaker, sponsor, or attendee, this page is your go-to ...

[See details](#)



Registration
dashboard



Invite clients



Invite partners



Speaker
conference
essentials



Get social



Sandbox
MeetingZone

<https://ibm.biz/Bdar8M>

Q&A

