IBM Newsroom

IBM Unveils Next Chapter of watsonx with Open Source, Product & Ecosystem Innovations to Drive Enterprise AI at Scale

• Releases a family of IBM Granite models into open source, including its most capable and efficient Code LLMs that can out-perform larger code models on many industry benchmarks

• Jointly with Red Hat, launches InstructLab, a first-of-its-kind model alignment technique, to bring open-source community contributions directly into LLMs

• Unveils new vision and momentum for new data and automation capabilities; Accelerates generative AI infusion into IBM assistants, automation, infrastructure, resource management products and consulting services

• Strengthens collaborations with AWS, Adobe, Meta, Microsoft, Mistral AI, Palo Alto Networks, SAP, Salesforce, and SDAIA to expand capabilities and offer model choice, flexibility and governance through watsonx

ARMONK, N.Y., May 21, 2024 – Today, at its annual THINK conference, IBM (NYSE:IBM) announced several new updates to its watsonx platform one year after its introduction, as well as upcoming data and automation capabilities designed to make artificial intelligence (AI) more open, cost effective, and flexible for businesses. During his opening keynote, CEO Arvind Krishna will share the company's plans to invest in, build and contribute to the open-source AI community as a core part of IBM's strategy.

"We firmly believe in bringing open innovation to AI. We want to use the power of open source to do with AI what was successfully done with Linux and OpenShift," said Krishna. "Open means choice. Open means more eyes on the code, more minds on the problems, and more hands on the solutions. For any technology to gain velocity and become ubiquitous, you've got to balance three things: competition, innovation, and safety. Open source is a great way to achieve all three."

IBM released a family of Granite models into open source and launched InstructLab, a first-of-its-kind capability, in collaboration with Red Hat

Furthering its commitment to the open-source AI ecosystem, IBM has now open-sourced a family of its most advanced and

performant language and code Granite models. By open sourcing these models, IBM is inviting clients, developers and global experts to build on these strengths and push the boundaries of what AI can achieve in enterprise environments.

Available today under Apache 2.0 licenses on Hugging Face and GitHub, the open-source Granite models stand out for their development process, quality, transparency and efficiency. The Granite code models range from 3B to 34B parameters and come in both base and instruction-following model variants, which are suitable for tasks such as complex application modernization, code generation, fixing bugs, explaining and documenting code, maintaining repositories and more. The code models, trained on 116 programming languages, consistently reach state-of-the-art performance among open-source code LLMs across various code-related tasks[i]:

- Testing by IBM found that the Granite Code models overall show very strong performance across all model sizes and benchmarks, often outperforming other open-source code models that are twice as large compared to Granite.[ii]
- Testing by IBM on benchmarks including HumanEvalPack, HumanEvalPlus, and reasoning benchmark GSM8K showed Granite code models have strong performances on code synthesis, fixing, explanation, editing, and translation across most major programming languages, including Python, JavaScript, Java, Go, C++, and Rust.[iii]
- The 20B parameter Granite base code model was used to train IBM watsonx Code Assistant (WCA) for specialized domains. It also powers watsonx Code Assistant for Z — a product designed to help enterprises transform monolithic COBOL applications into services optimized for IBM Z.
- The 20B parameter Granite base code model was tuned to generate SQL from natural language questions for the purpose of transforming structured data and extracting insights from that data. IBM demonstrated leadership in natural language to SQL, an important industry use case, as benchmarked by BIRD's independent leaderboard, which ranks models according to Execution Accuracy (EX) and Valid Efficiency Score (VES).[iv]

IBM and Red Hat also recently launched InstructLab — a revolutionary approach to advancing true open-source innovation around LLMs.

The InstructLab methodology allows for continuous development of base models through constant incremental contributions, much like software development has worked in open source for decades. With InstructLab, developers can build models specific to their business domains or industries with their own data, so that they can see the direct value of AI rather than just the model providers seeing the value. IBM plans to harness these open-source contributions to bring additional value to its clients through integrations with watsonx.ai and the new Red Hat Enterprise Linux AI (RHEL AI) solution.

RHEL AI comprises an enterprise-ready version of the InstructLab, IBM's open-source Granite models, and the world's leading enterprise Linux platform to simplify AI deployment across hybrid infrastructure environments. *Read this blog to learn more about Instruct Lab and watsonx.ai.*

IBM Consulting is also launching a practice to help clients leverage InstructLab with their own proprietary data to train purpose-specific AI models that can be scaled to better fit the cost and performance requirements of an enterprise's business needs. *Read this blog to learn more about how IBM Consulting is helping enterprises adopt AI.*

IBM unveils a new class of watsonx assistants

This new wave of AI innovation has the potential to unlock an estimated $4 trillion in annual economic benefits across industries.[v] However, IBM's annual Global AI Adoption Index recently found that while 42% of enterprise-scale companies (> 1,000 employees) surveyed have implemented AI in their business, another 40% of those companies that are exploring or

experimenting with AI have yet to deploy their models. For the companies stuck in the sandbox, 2024 is the year of overcoming barriers to entry such as the skills gap, data complexity and – perhaps most importantly – trust.

To address these challenges, IBM is announcing several upcoming updates and enhancements to its family of watsonx assistants, along with an upcoming capability in watsonx Orchestrate to help clients build their own AI Assistants across domains.

The new AI Assistants include watsonx Code Assistant for Enterprise Java Applications (planned availability in October 2024), watsonx Assistant for Z to transform how users interact with the system to quickly transfer knowledge and expertise (planned availability in June 2024), and an expansion of watsonx Code Assistant for Z Service with code explanation to help clients understand and document applications through natural language (planned availability in June 2024). *Read more on watsonx Code Assistant for Enterprise Java Applications, watsonx Assistant for Z, and Watsonx Code Assistant for Z.*

IBM is expanding its NVIDIA GPU offerings to now offer NVIDIA L40S and NVIDIA L4 Tensor Core GPUs, as well as support for Red Hat Enterprise Linux AI (RHEL AI) and OpenShift AI to help enterprises and developers address the needs of AI and other mission-critical workloads. Additionally, to help clients accelerate time to value for AI, IBM is using deployable architectures for watsonx to enable quick AI deployment while empowering enterprises with security and compliance capabilities to help them protect their data and manage compliance controls. *Read this blog to learn more about IBM Cloud capabilities*

Additionally, IBM has announced several new and upcoming generative AI powered data products and capabilities to augment how organizations observe, govern, and optimize their increasingly robust and complex data for AI workloads. Learn more about the upcoming IBM Data Product Hub (planned availability in June 2024), Data Gate for watsonx (planned availability in June 2024), and a host of the latest and planned updates on watsonx.data. *Read this blog to learn more about these data capabilities.*

IBM previews new vision and capabilities for AI-powered automation

Hybrid cloud and AI are transforming how companies do business. The average enterprise today manages multiple cloud environments — public and private — and around 1,000 apps,[vi] each with multiple dependencies. They also deal with petabytes of data. With generative AI expected to drive up to 1 billion apps by 2028,[vii] automation is no longer an option – it is how businesses will save time, solve problems, and make decisions faster.

IBM is addressing these challenges by delivering a set of AI-powered automation capabilities that will allow CIOs to move from proactive management of their IT environments to AI-powered predictive automation. AI-powered automation will be an essential tool for driving the speed, performance, scalability, security, cost efficiency of an enterprise's infrastructure.

Today, IBM's portfolio of automation, networking, data, application, and infrastructure management products help businesses manage their increasingly complex IT environments. For technology business management, Apptio enables organizations to make informed, data-driven decisions about their investments by driving clarity on technology spend and how it drives business value — enabling organizations to quickly respond to changing market conditions. Clients can also combine Apptio with the power of Instana for automated observability and Turbonomic for performance optimization to help clients efficiently allocate resources and control IT spend through enhanced visibility and real-time insights, allowing them to focus more time on deploying and scaling AI to drive new innovative initiatives.

To complement these products, IBM recently announced its intent to acquire HashiCorp, which helps organizations automate

multi-cloud and hybrid environments with Infrastructure Lifecycle Management and Security Lifecycle Management with products including Terraform, Vault and others. With HashiCorp, clients can easily move to and operate across multi-cloud and hybrid cloud environments.

Now, at THINK, IBM is continuing to advance its state-of-the-art in automation portfolio by previewing a new generative AI-powered tool called IBM Concert, which will be generally available in June 2024. IBM Concert will serve as the 'nerve center' of an enterprise's technology and operations.

Powered by AI from watsonx, IBM Concert will provide generative AI-driven insights across clients' portfolios of applications to identify, predict, and suggest fixes for problems. The new tool integrates into clients' existing systems, using generative AI to connect with data from their cloud infrastructure, source repositories, CI/CD pipelines and other existing application management solutions to build out a detailed view of their connected applications.

By allowing clients to eliminate unnecessary tasks and accelerate others, Concert is designed to make teams more informed so they can be fast and responsive in addressing issues and solving problems before they happen. Concert will initially focus on helping application owners, SREs and IT leaders gain insights about, pre-empt and more quickly address issues around application risk and compliance management. *Read this blog to learn more about IBM Concert.*

IBM expands ecosystem access to watsonx, adds third-party models

IBM continues to foster a strong ecosystem of partners to offer clients choice and flexibility through bringing third-party models onto watsonx, enabling leading software companies to embed watsonx capabilities into their technologies, and offering IBM Consulting expertise for enterprise business transformation. IBM Consulting has rapidly expanded its global generative AI expertise, with more than 50,000 practitioners certified in IBM and strategic partner technologies. Our ecosystem of partners, large and small, are helping clients adopt and scale tailored AI across their businesses.

- **AWS:** IBM and AWS are partnering to bring together Amazon SageMaker and watsonx.governance on AWS. Available now, this product will equip Amazon SageMaker clients with advanced AI governance capabilities for their predictive machine learning and generative AI models. Clients can now govern, monitor, and manage their models across platforms, simplifying risk management and compliance processes for their AI operations.
- **Adobe:** IBM and Adobe are collaborating on hybrid cloud and AI, bringing Red Hat OpenShift and watsonx to Adobe Experience Platform and are exploring making watsonx.ai and Adobe Acrobat AI Assistant available on-prem and private cloud. IBM is also introducing a new consulting service to advance client adoption of Adobe Express. These capabilities are expected to become available in 2H24.
- **Meta:** IBM has announced the availability of Meta Llama 3 — the next generation of Meta's open large language model — on watsonx to help enterprises innovate on their AI journeys. The addition of Llama 3 builds on IBM's collaboration with Meta to advance open innovation for AI. The two companies launched the AI Alliance — a group of leading organizations across industry, startup, academia, research and government — late last year, and it has since grown to more than 100 members and collaborators.
- **Microsoft:** IBM is announcing that the watsonx AI and data platform is supported by IBM to run on Microsoft Azure and available to purchase through IBM and our business partner ecosystem as a customer-managed solution on Azure Red Hat OpenShift (ARO).
- **Mistral:** IBM is announcing its intent to create a new strategic partnership with Mistral AI to bring its latest commercial models to the watsonx platform, including the leading Mistral Large model, which the company plans to make available in 2Q24. IBM looks forward collaborating with Mistral AI on open innovation, building on both companies' work in the open-

source community.

- **Palo Alto Networks**: IBM has expanded its partnership with Palo Alto to jointly deliver AI-powered security offerings and several initiatives to improve security outcomes for clients. For more, read the full press release.
- **Salesforce:** IBM and Salesforce are exploring making the IBM Granite model series available later this year for use across the Salesforce Einstein 1 platform, with the aim to provide clients access to more models to enhance decision making for AI CRM use cases.
- **SAP:** IBM Consulting and SAP are also collaborating to find ways to help more customers accelerate their cloud journeys leveraging RISE with SAP so they can realize the transformative benefits of generative AI for business in the cloud. This work seeks to expand on IBM and SAP's collaboration around embedding IBM Watson AI technology into SAP solutions. As part of this initiative with SAP**,** the IBM Granite Model Series is expected to be accessible for use across SAP's portfolio of cloud solutions and applications – which is underpinned by the generative AI hub in SAP AI Core.
- **SDAIA:** IBM has launched the Saudi Data and Artificial Intelligence Authority (SDAIA) 'ALLaM' Arabic model on watsonx, adding new language capabilities to the platform, including the ability to understand multiple Arabic dialects.

To see all THINK news, visit the IBM Newsroom.

Access a suite of social assets to share across your
business and personal channels with our creative toolkit.

Media Contact:

Amy Angelini
alangeli@us.ibm.com

To read IBM recent client & partner testimonials, see below:

Broadridge Financial: Leveraging Generative AI to Transform Financial Services

"As a trusted global technology partner sitting at the intersection of financial services, application sustainability and advancement is critical to our business strategy to safely and confidently embrace new technologies and deliver advanced services to clients to help them better operate, innovate, and grow," said Tyler Derr, CTO, Broadridge Financial Solutions. "Working with IBM, we are looking for ways to leverage AI to improve developer experience and productivity and to help understand and document our existing environments."

Casper Labs: Responsible AI Scaling with Blockchain-Enabled Data Governance Solution
"Trustworthy data is key to realizing AI's full potential, particularly in the current regulatory climate" said Mrinal Manohar, CEO at Casper Labs. "Without visibility into data flows, enterprises face unacceptable levels of risk, including intellectual property crossover and unnecessary data sharing across organizational lines. As an ecosystem partner, we're working with IBM Consulting to create Prove AI, a blockchain-enabled governance solution that pairs our tamper-resistant, highly serialized ledger with IBM watsonx.governance to provide enterprises with an end-to-end AI and data management platform that can safeguard sensitive data and ensure version control."

Department for Work and Pensions: Transforming Service Delivery with Generative AI

“As the UK’s largest public service department, we’re using the power of generative AI to make a positive impact on the lives of the millions of citizens we serve every day,” said Rich Corbridge, Director General and Chief Digital Information Officer, Department for Work and Pensions. “Our collaboration with IBM is instrumental in helping us identify and test the most promising AI use cases, build a robust partner ecosystem, develop innovative governance and risk management approaches, and cultivate the skills needed to scale generative AI successfully across the Department.” He added, “At the heart of our approach are human-centred values, while maintaining a forward-looking focus. We prioritise creating simplified, reusable components which we are building through our partnership with IBM. Together, DWP and IBM are transforming the support systems the UK's most vulnerable people interact with every single day.”

Dun & Bradstreet: AI Adoption with Trusted Data and Co-Created Solutions

“Enterprises need partners who can help them scale AI adoption, which starts with trusted data,” said Gary Kotovets, Chief Data & Analytics Officer at Dun & Bradstreet. “We’re bringing together Dun & Bradstreet and IBM’s nearly 300 years of combined experience to create generative AI solutions that harness Dun & Bradstreet Data Cloud and IBM watsonx to fuel generative AI adoption, starting with Ask Procurement. By leveraging IBM watsonx, D&B Ask Procurement simplifies, accelerates, and reduces the cost of certain essential procurement decisions. Ask Procurement is just one of many examples of how, together, we can help clients drive savings, optimize business processes, and improve risk management with generative AI and reliable data.”

Elevance Health: Driving an AI-first Digital-first Experience in Healthcare Customer Service

“Elevance Health’s purpose is to improve the health of humanity. We have the privilege of serving over 117M people, and are dedicated to making healthcare more proactive, predictive and personalized,’ said Kumar Gudavalli, Tech Strategy Lead & Chief Architect, Elevance Health. “Our longstanding and trusted relationship with IBM Consulting and the capabilities of IBM watsonx Assistant have helped us achieve both great member and provider experience and strong business outcomes. Driving an AI-first Digital-first experience, we were able to move up a third of interactions to messaging, and AI handles complex healthcare interactions with 60% effectiveness, while delivering much higher customer satisfaction and first contact resolution.”

Medtronic: IBM, Microsoft and Medtronic Drive AI-Powered Automation

“We are so fortunate to live in a time of limitless potential, a time where we can experience first-hand the power of collaboration. Together with IBM and Microsoft, we are transforming challenges into groundbreaking analytics solutions, setting new industry standards through innovation and excellence,” said Luciano Miranda, Vice President, Advanced Analytics Global Operations & Supply Chain at Medtronic.

Sevilla FC: Scouting Reimagined with AI-powered Insights

“At Sevilla FC, we’re proud to be at the forefront of innovation in football, and our engagement with IBM is a testament to that commitment,” said José María del Nido Carrasco, Sevilla FC's President. “Our team of 20+ expert scouts, former players and coaches, have curated a treasure trove of 400,000 player-game reports and 200,000 scouting notes, and now using IBM watsonx, we can unlock the full potential of this data, combining human insight with AI-driven analysis to make more informed, data-driven recruitment decisions. This game-changing work with IBM is a bold step forward in the future of football scouting.”

Virgin Money wins ‘Best application of AI in Financial Services’

“A year ago, we launched Redi, our Microsoft AI-powered virtual assistant, and we’re thrilled to see over a million of our credit card customers engaging with this innovative technology,” said Adam Paice, Head of Digital, Proposition Virgin Money. “Our customers rave about their interactions with Redi in our Credit Card app, and we believe that's because we are creating digital experiences for them that are as useful as they are delightful. Our collaboration with IBM Consulting has been instrumental in helping us unlock the full potential of generative AI, striking the perfect balance between innovation and control.”

Westfield Insurance Boosts Developer Productivity by 80% during pilot

“As a leading provider of property casualty insurance, we rely on advanced technology to help balance ever-changing regulations and economic uncertainties. Our goal is to meet customer service expectations while navigating these complexities,” said Don Hurrle, IT Infrastructure, AI, and Agile Transformation Leader, Westfield Insurance. “During an eight-week pilot with IBM, we used watsonx deployed on IBM Cloud to simplify the explanation and documentation of code faster, resulting in an 80%-time reduction for a developer to understand an application. This significant time savings enables our developers to focus on higher-value tasks, improving overall productivity. IBM watsonx Code Assistant for Z also accelerated the understanding of the application by performing analysis, generating reports, and identifying the impact of code changes.”

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

About IBM

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain the competitive edge in their industries. More than 4,000 government and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity and service. Visit www.ibm.com for more information.

---

[i] IBM. “Granite Code Models: A Family of Open Foundation Models for Code Intelligence”

[ii] IBM. “Granite Code Models: A Family of Open Foundation Models for Code Intelligence”

[iii] IBM. “Granite Code Models: A Family of Open Foundation Models for Code Intelligence”

[iv] For data-centric enterprise workloads, generating SQL from natural language questions is one of the most significant and challenging code generation tasks. On May 14, 2024, IBM Granite-based models achieved the top accuracy and efficiency scores on the BIRD ((BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation) leaderboard for text-to-SQL generation, demonstrating IBM's leadership in this critical enterprise use case.

[v] McKinsey. “The Economic Potential of Generative AI”

[vi] Salesforce. “Third Edition State of IT Report.”

[vii] IDC: Gary Chen, Jim Mercer. "1 Billion New Logical Applications: More Background.”

---

https://stage.mediaroom.com/ibmnewsroom/2024-05-21-IBM-Unveils-Next-Chapter-of-watsonx-with-Open-Source,-Product-Ecosystem-Innovations-to-Drive-Enterprise-AI-at-Scale