IBM Canada Newsroom

IBM Introduces Granite 3.0: High Performing AI Models Built for Business

• New Granite 3.0 8B & 2B models, released under the permissive Apache 2.0 license, show strong performance across many academic and enterprise benchmarks, able to outperform or match similar-sized models

• New Granite Guardian 3.0 models deliver IBM's most comprehensive guardrail capabilities to advance safe and trustworthy AI

• New Granite 3.0 Mixture-of-Experts models enable extremely efficient inference and low latency, suitable for CPU-based deployments and edge computing

• New Granite Time Series model achieved state-of-the-art performance in zero/few-shot forecasting, outperforming models 10 times larger

• IBM unveils next generation of Granite-powered watsonx Code Assistant for general purpose coding; Debuts new tools in watsonx.ai for building and deploying AI applications and agents

• Announces Granite will become the default model of Consulting Advantage, an AI-powered delivery platform used by IBM's 160,000 consultants to bring new solutions to clients faster

ARMONK, NY – October 21, 2024 –Today**,** at IBM's (NYSE: IBM) annual TechXchange event the company announced the release of its most advanced family of AI models to date, Granite 3.0. IBM's third-generation Granite flagship language models can outperform or match similarly sized models from leading model providers on many academic and industry benchmarks, showcasing strong performance, transparency and safety.

Consistent with the company's commitment to open-source AI, the Granite models are released under the permissive Apache 2.0 license, making them unique in the combination of performance, flexibility and autonomy they provide to enterprise clients and the community at large.

IBM's Granite 3.0 family includes:

- General Purpose/Language: **Granite 3.0 8B Instruct**, **Granite 3.0 2B Instruct, Granite 3.0 8B Base, Granite 3.0 2B Base**
- Guardrails & Safety: **Granite Guardian 3.0 8B**, **Granite Guardian 3.0 2B**
- Mixture-of-Experts: **Granite 3.0 3B-A800M Instruct**, **Granite 3.0 1B-A400M Instruct, Granite 3.0 3B-A800M Base**, **Granite 3.0 1B-A400M Base**

The new Granite 3.0 8B and 2B language models are designed as 'workhorse' models for enterprise AI, delivering strong performance for tasks such as Retrieval Augmented Generation (RAG), classification, summarization, entity extraction, and tool use. These compact, versatile models are designed to be fine-tuned with enterprise data and seamlessly integrated across diverse business environments or workflows.

While many large language models (LLMs) are trained on publicly available data, a vast majority of enterprise data remains untapped. By combining a small Granite model with enterprise data, especially using the revolutionary alignment technique InstructLab – introduced by IBM and RedHat in May – IBM believes businesses can achieve task-specific performance that rivals larger models at a fraction of the cost (based on an observed range of 3x-23x less cost than large frontier models in several early proofs-of-concept[1]).

The Granite 3.0 release reaffirms IBM's commitment to building transparency, safety, and trust in AI products. The Granite 3.0 technical report and responsible use guide provide a description of the datasets used to train these models, details of the filtering, cleansing, and curation steps applied, along with comprehensive results of model performance across major academic and enterprise benchmarks.

Critically, IBM provides an IP indemnity for all Granite models on watsonx.ai so enterprise clients can be more confident in merging their data with the models.

Raising the bar: Granite 3.0 benchmarks

The Granite 3.0 language models also demonstrate promising results on raw performance.

On standard academic benchmarks defined by Hugging Face's OpenLLM Leaderboard, the Granite 3.0 8B Instruct model's overall performance leads on average against state-of-the-art-performance of similar-sized open-source models from Meta and Mistral. On IBM's state-of-the-art AttaQ safety benchmark, the Granite 3.0 8B Instruct model leads across all measured safety dimensions compared to models from Meta and Mistral.[2]

On the core enterprise tasks of RAG, tool use, and tasks in the Cybersecurity domain, the Granite 3.0 8B Instruct model shows leading overall performance on average compared to similar-sized open-source models from Mistral and Meta.[3]

The Granite 3.0 models were trained on over 12 trillion tokens on data taken from 12 different natural languages and 116 different programming languages, using a novel two-stage training method, leveraging results from several thousand experiments designed to optimize data quality, data selection, and training parameters. By the end of the year, the 3.0 8B and 2B language models are expected to include support for an extended 128K context window and multi-modal document understanding capabilities.

Demonstrating an excellent balance of performance and inference cost, IBM offers its Granite Mixture of Experts (MoE)

Architecture models, Granite 3.0 1B-A400M and Granite 3.0 3B-A800M, as smaller, lightweight models that could be deployed for low latency applications as well as CPU-based deployments.

IBM is also announcing an updated release of its pre-trained Granite Time Series models, the first versions of which were released earlier this year. These new models are trained on 3 times more data and deliver strong performance on major time series benchmarks, with an average performance better than models 10 times larger from Google and Alibaba. The updated models also provide greater modeling flexibility with support for external variables and rolling forecasts. [4]

Introducing Granite Guardian 3.0: ushering the next era of responsible AI

As part of this release, IBM is also introducing a new family of Granite Guardian models that permit application developers to implement safety guardrails by checking user prompts and LLM responses for a variety of risks. The Granite Guardian 3.0 8B and 2B models provide the most comprehensive set of risk and harm detection capabilities available in the market today.

In addition to harm dimensions such as social bias, hate, toxicity, profanity, violence, jailbreaking and more, these models also provide a range of unique RAG-specific checks such as groundedness, context relevance, and answer relevance. In extensive testing across over 19 safety and RAG benchmarks, the Granite Guardian 3.0 8B model has higher overall accuracy on harm detection on average than all three generations of Llama Guard models from Meta. It also showed on par overall performance in hallucination detection on average with specialized hallucination detection models WeCheck and MiniCheck.[5]

While the Granite Guardian models are derived from the corresponding Granite language models, they can be used to implement guardrails alongside any open or proprietary AI models.

Availability of Granite 3.0 models

The entire suite of Granite 3.0 models and the updated time eries models are available for download on HuggingFace under the permissive Apache 2.0 license. The instruct variants of the new Granite 3.0 8B and 2B language models and the Granite Guardian 3.0 8B and 3B models are available today for commercial use on IBM's watsonx platform. A selection of the Granite 3.0 models will also be available as NVIDIA NIM microservices and through Google Cloud's Vertex AI Model Garden integrations with HuggingFace.

To help provide developer choice and ease of use and support local, edge deployments, a curated set of the Granite 3.0 models are also available on Ollama and Replicate.

The latest generation of Granite models expand IBM's robust open-source catalog of powerful LLMs. IBM has collaborated with ecosystem partners like AWS, Docker, Domo, Qualcomm Technologies, Inc. via itsQualcomm® AI Hub, Salesforce, SAP, and others to integrate a variety of Granite models into these partners' offerings or make Granite models available on their platforms, offering greater choice to enterprises across the world.

Assistants to Agents: realizing the future for enterprise AI

IBM is advancing enterprise AI through a spectrum of technologies – from models and assistants, to the tools needed to tune and deploy AI specifically for companies' unique data and use-cases. IBM is also paving the way for future AI agents that can self-direct, reflect, and perform complex tasks in dynamic business environments.

IBM continues to evolve its portfolio of AI assistant technologies – from watsonx Orchestrate to help companies build their own assistants via low-code tooling and automation, to a wide set of pre-built assistants for specific tasks and domains such as customer service, human resources, sales, and marketing. Organizations around the world have used watsonx Assistant to help them build AI assistants for tasks like answering routine questions from customers or employees, modernizing their mainframes and legacy IT applications, helping students explore potential career paths, or providing digital mortgage support for home buyers.

Today IBM also unveiled the upcoming release of the next generation of watsonx Code Assistant, powered by Granite code models, to offer general-purpose coding assistance across languages like C, C++, Go, Java, and Python, with advanced application modernization capabilities for Enterprise Java Applications.[6] Granite's code capabilities are also now accessible through a Visual Studio Code extension, IBM Granite.Code.

IBM also plans to release new tools to help developers build, customize and deploy AI more efficiently via watsonx.ai – including agentic frameworks, integrations with existing environments and low-code automations for common use-cases like RAG and agents.[7]

IBM is focused on developing AI agent technologies which are capable of greater autonomy, sophisticated reasoning and multi-step problem solving. The initial release of the Granite 3.0 8B model features support for key agentic capabilities, such as advanced reasoning and a highly-structured chat template and prompting style for implementing tool use workflows. IBM also plans to introduce a new AI agent chat feature to IBM watsonx Orchestrate, which uses agentic capabilities to orchestrate AI Assistants, skills, and automations that help users increase productivity across their teams.[8] IBM plans to continue building agent capabilities across its portfolio in 2025, including pre-built agents for specific domains and use-cases.

Expanded AI-powered delivery platform to supercharge IBM consultants with AI

IBM is also announcing a major expansion of its AI-powered delivery platform, IBM Consulting Advantage. The multi-model platform contains AI agents, applications, and methods like repeatable frameworks that can empower 160,000 IBM consultants to deliver better and faster client value at a lower cost.

As part of the expansion, Granite 3.0 language models will become the default model in Consulting Advantage. Leveraging Granite's performance and efficiency, IBM Consulting will be able to help maximize the return-on-investment for the generative AI projects of IBM clients.

Another key part of the expansion is the introduction of IBM Consulting Advantage for Cloud Transformation and Management and IBM Consulting Advantage for Business Operations. Each includes domain-specific AI agents, applications, and methods infused with IBM's best practices so IBM consultants can help accelerate client cloud and AI transformations in tasks, like code modernization and quality engineering, or transform and execute operations across domains, like finance, HR and procurement.

To learn more about Granite and IBM's AI for Business strategy, visit https://www.ibm.com/granite.

---

[1] Cost calculations are based on API cost per million tokens pricing of IBM watsonx for open models and openAI for GPT4 models (assuming blend of 80% inout, 20% output) for customer proofs-of-concept.

[2] IBM Research technical paper: Granite 3.0 Language Models

[3] IBM Research technical paper: Granite 3.0 Language Models

[4] The Tiny Time Mixer: Fast Pre-Trained Models for Enhanced Zero/Few Shot Forecasting on Multivariate Time Series

[5] Evaluation results published in Granite Guardian GitHub Repo

[6] Planned availability for Q4 2024

[7] Planned availability for Q4 2024

[8] Planned availability for Q1 2025
