IBM Newsroom

IBM Expands Granite Model Family with New Multi-Modal and Reasoning AI Built for the Enterprise

- Granite 3.2 – small AI models offering reasoning, vision, and guardrail capabilities with a developer friendly license

- Updated Granite time series models that offer long-range forecasting with less than 10M parameters

ARMONK, N.Y., Feb. 26, 2025 /PRNewswire/ -- IBM (NYSE: IBM) todaydebuted the next generation of its Granite large language model (LLM) family, Granite 3.2, in a continued effort to deliver small, efficient, practical enterprise AI for real-world impact.

All Granite 3.2 models are available under the permissive Apache 2.0 license on Hugging Face. Select models are available today on IBM watsonx.ai, Ollama, Replicate, and LM Studio, and expected soon in RHEL AI 1.5 – bringing advanced capabilities to businesses and the open-source community. Highlights include:

- A new **vision language model (VLM)** for document understanding tasks which demonstrates performance that matches or exceeds that of significantly larger models – Llama 3.2 11B and Pixtral 12B – on the essential enterprise benchmarks DocVQA, ChartQA, AI2D and OCRBench[1]. In addition to robust training data, IBM used its own open-source Docling toolkit to process 85 million PDFs and generated 26 million synthetic question-answer pairs to enhance the VLM's ability to handle complex document-heavy workflows.
- **Chain of thought** capabilities for enhanced reasoning in the 3.2**2B and 8B models**, with the ability to switch reasoning on or off to help optimize efficiency. With this capability, the **8B model** achieves double-digit improvements from its predecessor in instruction-following benchmarks like ArenaHard and Alpaca Eval without degradation of safety or performance elsewhere[2]. Furthermore, with the use of novel inference scaling methods, the Granite 3.28B model can be calibrated to rival the performance of much larger models like Claude 3.5 Sonnet or GPT-4o on math reasoning benchmarks such as AIME2024 and MATH500.[3]
- Slimmed-down size options for **Granite Guardian** safety models that maintain performance of previous Granite 3.1 Guardian models at 30% reduction in size. The 3.2 models also introduce a new feature called verbalized confidence, which offers more nuanced risk assessment that acknowledges ambiguity in safety monitoring.

IBM's strategy to deliver smaller, specialized AI models for enterprises continues to demonstrate efficacy in testing, with the Granite 3.1 8B model recently yielding high marks on accuracy in theSalesforce LLM Benchmark for CRM.

The Granite model family is supported by a robust ecosystem of partners, including leading software companies embedding the LLMs into their technologies.

"At CrushBank, we've seen first-hand how IBM's open, efficient AI models deliver real value for enterprise AI – offering the right balance of performance, cost-effectiveness, and scalability," said David Tan, CTO, CrushBank. "Granite 3.2 takes it further with new reasoning capabilities, and we're excited to explore them in building new agentic solutions."

Granite 3.2 is an important step in the evolution of IBM's portfolio and strategy to deliver small, practical AI for enterprises. While chain of thought approaches for reasoning are powerful, they require substantial compute power that is not necessary for every task. That is why IBM has introduced the ability to turn chain of thought on or off programmatically. For simpler tasks, the model can operate without reasoning to reduce unnecessary compute overhead. Additionally, other reasoning techniques like inference scaling have shown that the Granite 3.2 8B model can match or exceed the performance of much larger models on standard math reasoning benchmarks. Evolving methods like inference scaling remains a key area of focus for IBM's research teams.[4]

Alongside Granite 3.2 instruct, vision, and guardrail models, IBM is releasing the next generation of its TinyTimeMixers (TTM) models (sub 10M parameters), with capabilities for longer-term forecasting up to two years into the future. These make for powerful tools in long-term trend analysis, including finance and economics trends, supply chain demand forecasting and seasonal inventory planning in retail.

"The next era of AI is about efficiency, integration, and real-world impact – where enterprises can achieve powerful outcomes without excessive spend on compute," said Sriram Raghavan, VP, IBM AI Research. "IBM's latest Granite developments focus on open solutions demonstrate another step forward in making AI more accessible, cost-effective, and valuable for modern enterprises."

To learn more about Granite 3.2, read thistechnical article.

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 a competitive edge in their industries. Thousands of governments 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.

Media contact:
Amy Angelini
IBM AI Communications
alangeli@us.ibm.com

1 Vision model benchmark results are available in IBM's technical article,*IBM Granite 3.2: Reasoning, Vision, Forecasting, and More*, published February 26, 2025.

2 Instruct model benchmark results are available in IBM's technical article,*IBM Granite 3.2: Reasoning, Vision, Forecasting, and More*, published February 26, 2025.

3 Inference scaling benchmark results are available in IBM's technical research blog,*Reasoning in Granite 3.2 Using Inference Scaling,* published February 26, 2025.

4 *Reasoning in Granite 3.2 Using Inference Scaling* IBM, published February 26, 2025.

SOURCE IBM

Additional assets available online: Photos

https://stage.mediaroom.com/ibmnewsroom/2025-02-26-ibm-expands-granite-model-family-with-new-multi-modal-and-reasoning-ai-built-for-the-enterprise