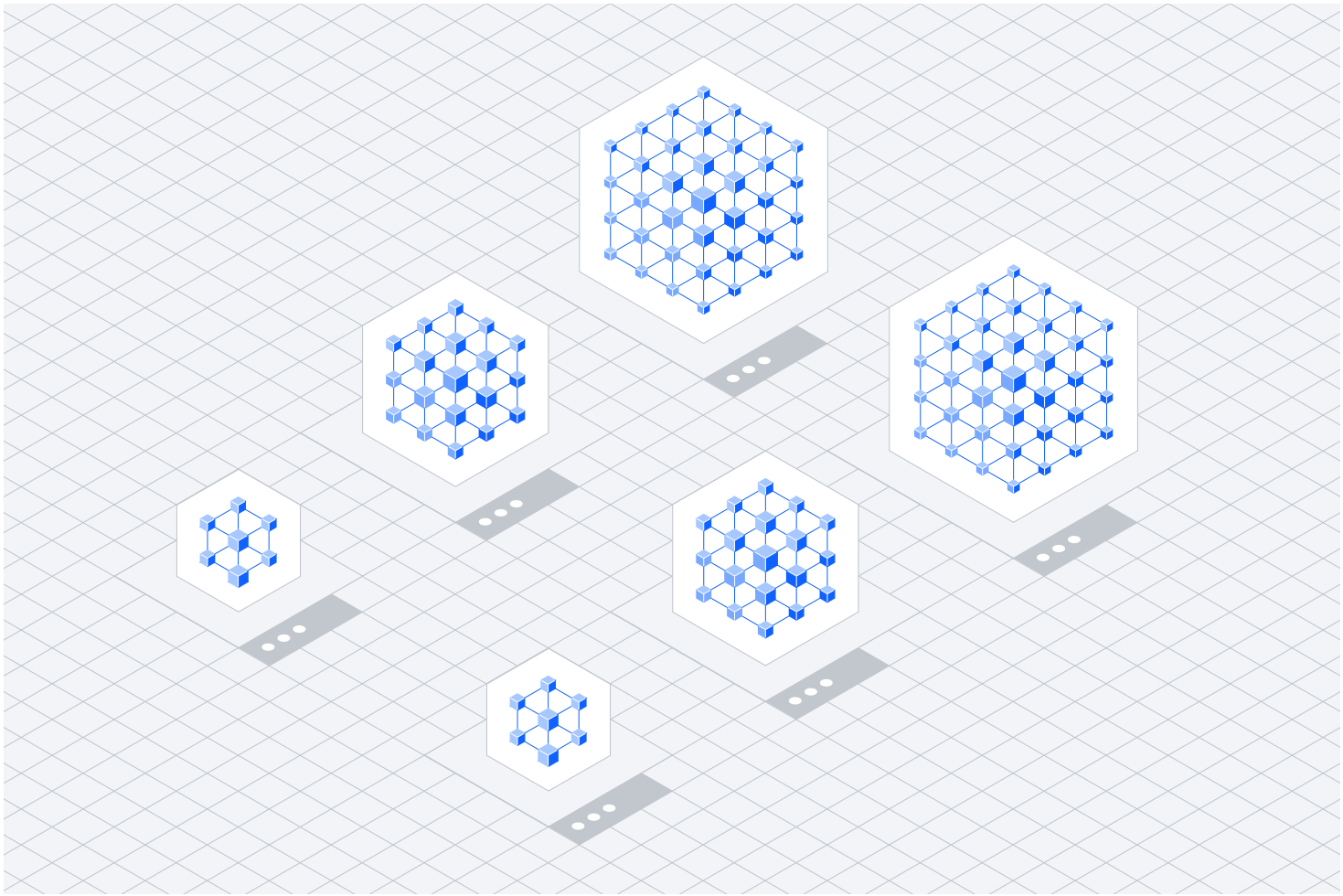


A differentiated approach to AI foundation models

Scale generative AI with enterprise-grade models

IBM

Table of contents

Introduction

In 2023, organizational departments such as HR, IT and customer care have focused on the generative AI use cases, such as summarization, code generation and Q&A to take out costs and drive productivity. According to a recent study from the IBM Institute for Business Value (IBV), 75% of CEOs believe competitive advantage will depend on who has the most advanced generative AI and 50% are now integrating generative AI into their products and services.[1]

The IBV study also deduces that the experimentation phase for generative AI leaders will be short and intense, with 74% of executives reporting that generative AI will be ready for general rollout in the next three years. Multimodal, multilingual foundation models and automation agents will expand the gamut of generative AI use cases and adoption across business workflows. Additionally, vertical and domain-specific foundation models with a smaller number of parameters that can match the performance of larger models at lower cost of inferencing are predicted to gain more market traction. Furthermore, the techniques for training foundation models will continue to evolve, unlocking advanced capabilities and new orders of efficiencies to make generative AI even more attractive for businesses.

61%

of CEOs identify concerns about data lineage and provenance as a barrier to adopting generative AI.

A mandate to embrace generative AI with eyes wide open

Business and technology executives are handed a clear mandate from their leadership and boards to transform their business models, offerings and operations with generative AI in 2024. An IBM study on responsible AI and ethics finds that CEOs say they feel over six times more pressure from their boards and investors to accelerate the adoption of generative AI rather than slow it down.[2]

According to the IBM CEO study, 58% of business executives believe that major ethical risks abound with the adoption of generative AI, and 79% say AI ethics is important to their enterprise-wide AI approach.[2] 61% of CEOs identify concerns about data lineage and provenance as a barrier to adopting generative AI,[1] even as 85% of executives anticipate direct interactions between generative AI and customers within the next two years.[1]

Policymakers and regulators are waking up to the potential risks and social harms of generative AI and are drafting policies and laws to ensure sustainable innovation and diffusion. Both the EU AI Act and the US White House executive order on AI announced in 2023 signal government commitment to the scrutiny of AI. In addition, business leaders are concerned about exposing their company's proprietary data to foundation models hosted in the cloud, yet they see the value in fine-tuning AI with its business context.

Taking steps to ensure the value exceeds the cost

Cost is another major consideration of policymakers and regulators. According to an article in the scientific journal *Nature*, a search driven by generative AI uses 4–5 times the energy needed to run a conventional web search, and large models in market require volumes of expensive GPU compute resources.[3]

Enterprise decision-makers leading generative AI initiatives within their organizations have a lot to think about. As organizations move from exploration to investigation and production with generative AI, they need foundation models to power key tasks in various departments. In this white paper, we share the value of enterprise-grade foundation models that provide trust, performance and cost-effective benefits to all industries.

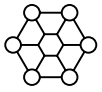
The needs of enterprises as they scale generative AI

The major challenge facing enterprise decision-makers is striking a balance between scaling generative AI faster across a global organization that leverages its proprietary data and minimizing foundational model-related risk exposure and total cost of ownership (TCO).

Moving from exploration to investigation and production with generative AI will require enterprises to adopt the right model choices for the right use cases and a robust platform to customize models and infuse AI into their applications. In addition, they'll need hybrid cloud to deploy AI in their infrastructure of choice and a reliable partner who can help scale and operationalize AI with minimal risks.

A strategic approach is key to successful adoption
As decision-makers evaluate options to choose the right foundation models for scaling generative AI, it's crucial to consider a strategic point of view that incorporates both enterprise-grade foundation models and a robust platform to operationalize them.

Trust in foundation models, gauged through metrics like transparency indexes and hallucination scores, hinges on transparency in data management, training and evaluation processes.



Performance measures can then be used to determine which attributes, such as versatility, accuracy, latency and throughput, are critical for the enterprise use case.

Cost-effectiveness measures can help narrow down model choices that deliver the necessary performance at lower inferencing costs and with fewer compute resources.

Customization, specialization and flexibility
Users should be able to customize the foundation models for their use cases, company and industry domain using fine-tuning and prompt-tuning with an easy-to-use toolkit. They would also need specialized database capabilities for storing, managing and retrieving high-dimensional vectors that fuel generative AI applications. Depending on the use case, the nature of the content used to inference the model and operational considerations, clients should have the flexibility to deploy the model in the infrastructure of their choice. AI guardrails and continuous monitoring make model deployments secure and reliable as organizations scale-up generative AI applications.

Enterprise decision-makers also seek a reliable technology partner that grasps opportunities and risks in enterprise AI adoption, understands the key model dimensions and bakes in AI ethics and regulatory preparedness across the generative AI lifecycle, starting with foundation model development.

IBM's differentiated approach to delivering enterprise-grade foundation models

Enterprise-grade foundation models represent a class of models characterized by their trustworthiness, performance and cost-effectiveness, which makes them well-suited for deployment in enterprise settings where reliability, efficiency and value are paramount considerations. While most model providers obsess about building and bringing the most capable models to the market, enterprise clients should step back and asses what optimal mix of model attributes they need to succeed.

IBM's approach to delivering enterprise-grade models encompasses four key principles:

1. **Open:** Bring best-in-class open models from IBM and open source to our IBM watsonx™ foundation models library.
2. **Trusted:** Train models on trusted and governed data for applications that require enterprise-level transparency, governance and performance in accordance with IBM AI Ethics principles.
3. **Targeted:** Design for the enterprise and optimize for targeted business domains and use cases with open-sourced or insourced skills and knowledge, using large-scale alignment techniques.
4. **Empowering:** Provide clients with competitively priced model choices to build AI that best suits their unique business needs and risk profiles.

Open models and third-party models offer choice
IBM pursues a multimodel strategy with enterprise-grade multimodal (text, code audio, image, geospatial) and multilingual model choices that best suits clients' unique business needs, regional interests and risk profiles. In adherence to its open-source and ecosystem strategy, IBM gives clients the freedom to choose from a diverse library that includes proprietary, open-source models and third-party models hosted on IBM® watsonx.ai™, our next-generation enterprise studio for AI builders.

Clients can harness the IBM foundation model library directly on the watsonx.ai platform. IBM watsonx.ai is a next-generation enterprise studio. Part of the watsonx™ AI and data platform, it brings together new generative AI capabilities powered by foundation models and traditional machine learning (ML) into a powerful studio spanning the AI lifecycle. It allows you to tune and guide models with your enterprise data to meet your needs with easy-to-use tools for building and refining performant prompts.

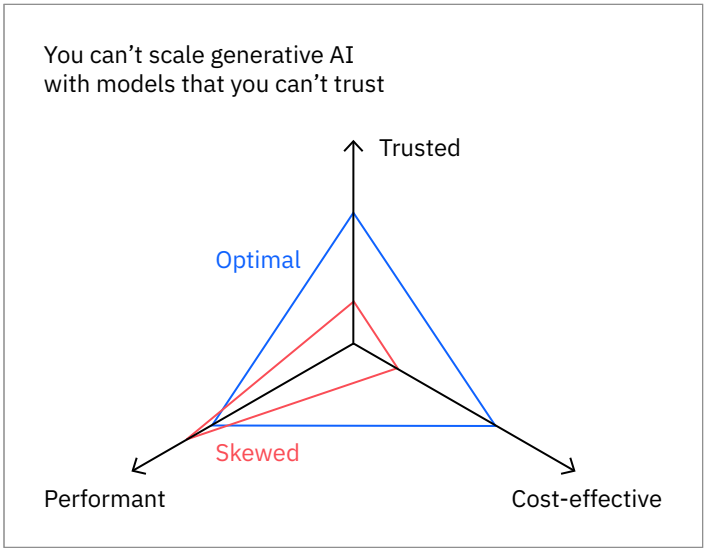


Figure 1. IBM's point of view on enterprise-grade foundation models—trusted, performant, cost-effective

Key benefits of the watsonx AI and data platform

- **Model customization and choice:** Tailor models with proprietary data and expertise to unique use cases and company and industry domains for easy integrations with ready-to-use AI applications. The watsonx platform brings complementary capabilities, such as vector database and embedding models, to support popular use cases like retrieval-augmented generation (RAG). With the Bring-Your-Own-Model (BYOM) capability, clients can enjoy much more flexibility and freedom over their model choices on watsonx.
- **Robust governance:** Make AI safe and secure at scale with AI guardrails, continuous risk monitoring, integrated governance and third-party large language model (LLM) support.
- **Flexible deployment:** Work with your infrastructure of choice with hybrid multicloud and on-prem options to avoid vendor lock-in and reduce TCO.

IBM Granite was trained on enterprise-relevant content that meets rigorous data governance, regulatory and risk criteria defined and enforced by IBM's AI Ethics principles and its Office of Privacy & Responsible Technology.

IBM Granite, developed by IBM Research: An example of IBM's differentiated approach to models

IBM Granite™ is IBM's flagship series of enterprise-grade foundation models based on decoder-only transformer architecture. Granite models are designed to possess an optimal mix of model attributes to meet the trusted, performant and cost-effective requirements of the enterprise. Granite language models are trained on trusted enterprise data spanning internet, academic, code, legal and finance data sources.

Currently, we have the following models in the Granite series:

1. **Granite-13b-chat-v2.1:** A chat model optimized for dialogue use cases that works well with virtual agents and chat applications
2. **Granite-13b-instruct-v2.1:** An instruct model trained on high-quality finance data to perform well in finance domain tasks
3. **Granite-20b-multilingual:** Trained to understand and generate text in English, German, Spanish, French and Portuguese
4. **Granite-8b-japanese:** Designed to perform language tasks on Japanese text
5. **Granite-7b-lab:** Supports general-purpose tasks and is tuned using the IBM's large-scale alignment of chatbots (LAB) methodology to incorporate new skills
6. **Granite code models:** A family of models trained in 116 programming languages and ranging in size from 3 to 34 billion parameters—with base model and instruction-following model variants

A foundation model with trust built in

IBM's approach to developing the Granite series of foundation models is guided by our AI ethics policies, with a strong emphasis on trust and transparency in training data and the model-training process. IBM stands behind its models by offering clients intellectual property (IP) indemnity protection so they can focus on the business impact of AI, not on costly courtroom litigation. Granite models were built for the enterprise and governed by rules and safeguards to minimize hateful and profane content, or "HAP." The models are trained on datasets that meet IBM's rigorous data governance, risk and compliance criteria.

IBM helps ensure the integrity of its Granite models through comprehensive data management across the training phases to build trusted models.

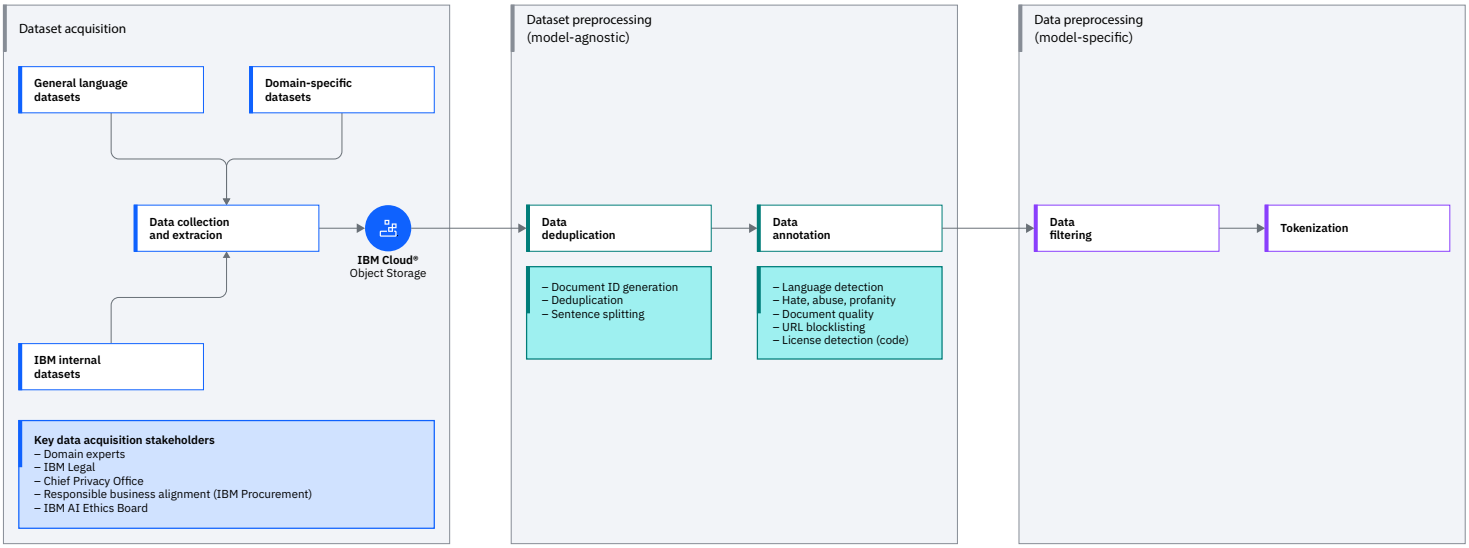


Figure 2. The training phases of the Granite models used to build trusted models

Improved accuracy for targeted enterprise business domains like finance and use cases like RAG is achieved through chat fine-tuning and model alignment techniques.

Chat fine-tuning techniques were designed to prevent hallucinations and misalignments in the model outputs.

The Granite-13b-v2.1 foundation model demonstrates some of the least biased behavior of studied models, according to a measures of bias in open-ended language generation dataset (BOLD) evaluation. It also consistently scores high on harmlessness across attack domains, according to AttaQ dataset evaluation.[4]

With the IBM Granite series, pick a competitively priced model with less infrastructure requirement, IP indemnification and an easy-to-use toolkit for model customization and application integration.

Granite models were built for the enterprise and governed by rules and safeguards to minimize hateful and profane content, or “HAP.”

A foundation model that’s performant
Bigger isn’t always better—smaller and more targeted models can perform on par with larger general-purpose models.

- According to IBM internal benchmark testing in February 2024, the Granite-13b-chat-v2.1 model showed almost comparable performance across several key enterprise use cases—summarization and entity extraction—when compared to llama-13b-chat.[5]
- According to IBM internal benchmark testing in March 2024, the Granite-20b-multilingual model approaches the quality of three popular open-source models—llama-2-13b-chat, llama 2-70b-chat and mixtral 8x7b—across three languages—German, Portuguese and Spanish—within translation tasks when compared to llama-2-13b-chat.[5]
- In just three months, the new Granite-13b-chat-v2.1 has achieved a 45% improved performance in RAG compared to the earlier v1 model, based on internal IBM performance benchmark testing for enterprise tasks that include RAG.[5]
- According to IBM internal benchmark testing in Feb 2024, on ConFinQA, a standard benchmark measuring accuracy of multiturn numeric reasoning in financial services Q&A, the Granite.13b.chat.v2.1 model demonstrated 18% better performance than the second best model llama2.70. chat model.[5]
- According to IBM internal benchmark testing in 2024, in JCommonsenseQA, a multiple-choice question answering dataset in Japanese that tests commonsense knowledge, Granite-8b-japanese outperformed open-source elyza-7b-llama-japanese by more than 2 times for zero-shot and 1.2 times for few shot.[5]
- Evaluation on a comprehensive set of tasks has shown that Granite code models consistently match state-of-the-art performance among open-source code LLMs currently available.[5]

With lower latencies, IBM Granite models have been shown to generate faster responses for summarization tasks.

A foundation model that’s cost-effective
Granite models require only a fraction of GPU capacity and compute, which can result in a lower carbon footprint and reduce TCO.

- Granite models are competitively priced in the market and can run on less than 1 GPU.
- IBM Watson large speech model, which is 5 times smaller than OpenAI’s Whisper, processes audio 10 times faster on the same hardware, which can translate to lower costs.[6]

Conclusion: The outcomes of successfully scaling and operationalizing generative AI

As we reflect on the transformative impact of generative AI and foundation models across industries, business domains and geographies, it's evident that the integration of generative AI isn't just imminent but already in motion. More than half of organizations are actively exploring its capabilities. With the forthcoming expansion into multimodal and smaller specialty models, businesses will further accelerate the impact of AI. The journey ahead calls for executives to balance rapid implementation with ethical and risk considerations, while navigating the fast-paced evolution of generative AI.

A leader in enterprise AI and hybrid cloud, IBM is well-positioned in the marketplace to consistently bring trusted, performant and cost-effective generative AI products and solutions to our clients. IBM Research® will continue to evolve Granite models through open-source engagement to deliver more performant and trusted models at competitive pricing. Through the open-source project, InstructLab, initiated by IBM and Red Hat®, we democratize model development and alignment with open-sourced skills and knowledge.

Enterprises already leveraging IBM foundation models on watsonx.ai

- The Championships, Wimbledon used watsonx.ai foundation models to train their AI to create tennis commentary and generate informative and engaging video clip narrations for fans with varied sentence structures and vocabulary.[7]
- Bradesco, a prominent financial institution in Latin America, is excited about using watsonx.ai foundation models to enhance its AI strategy within its cutting-edge corporate infrastructure.
- NASA is using foundation models to make it easier to build AI applications for text and sensor data, enabling models that detect natural hazards and track changes to vegetation and wildlife habitat.
- Quantum Street AI is applying IBM foundation models to provide investment managers with AI and ML tools to identify market opportunities and manage risk.

Take the next steps
Experience our interactive demo and sign up for a trial of watsonx.ai.

Additional resources:

- Visit our website to learn more about IBM's model PoV and offerings.
- Learn more about IBM Granite models—data sources, training steps and performance evaluations—by reading the latest research paper.
- Use our model evaluation guide for simple decision heuristics you can apply to refine your model choices.

1. The CEO's guide to generative AI: What you need to know and do to win with transformative technology, IBM Institute for Business Value, January 2024.
2. The CEO's guide to generative AI: Responsible AI & ethics, IBM Institute for Business Value, 23 October 2023.
3. Generative AI's environmental costs are soaring—and mostly secret, Springer Nature Limited, 20 February 2024.
4. Watsonx foundation model evaluation summary, IBM Research AI and watsonx.ai teams, 23 January 2023.
5. "Open sourcing IBM's Granite code models," IBM Research blog, 6 May 2024.
6. IBM Research benchmarking study on customer care use cases where each turn in the conversation lasted less than 30 seconds.
7. IBM, Wimbledon and the power of watsonx, IBM Corp., January 2024.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
May 2024