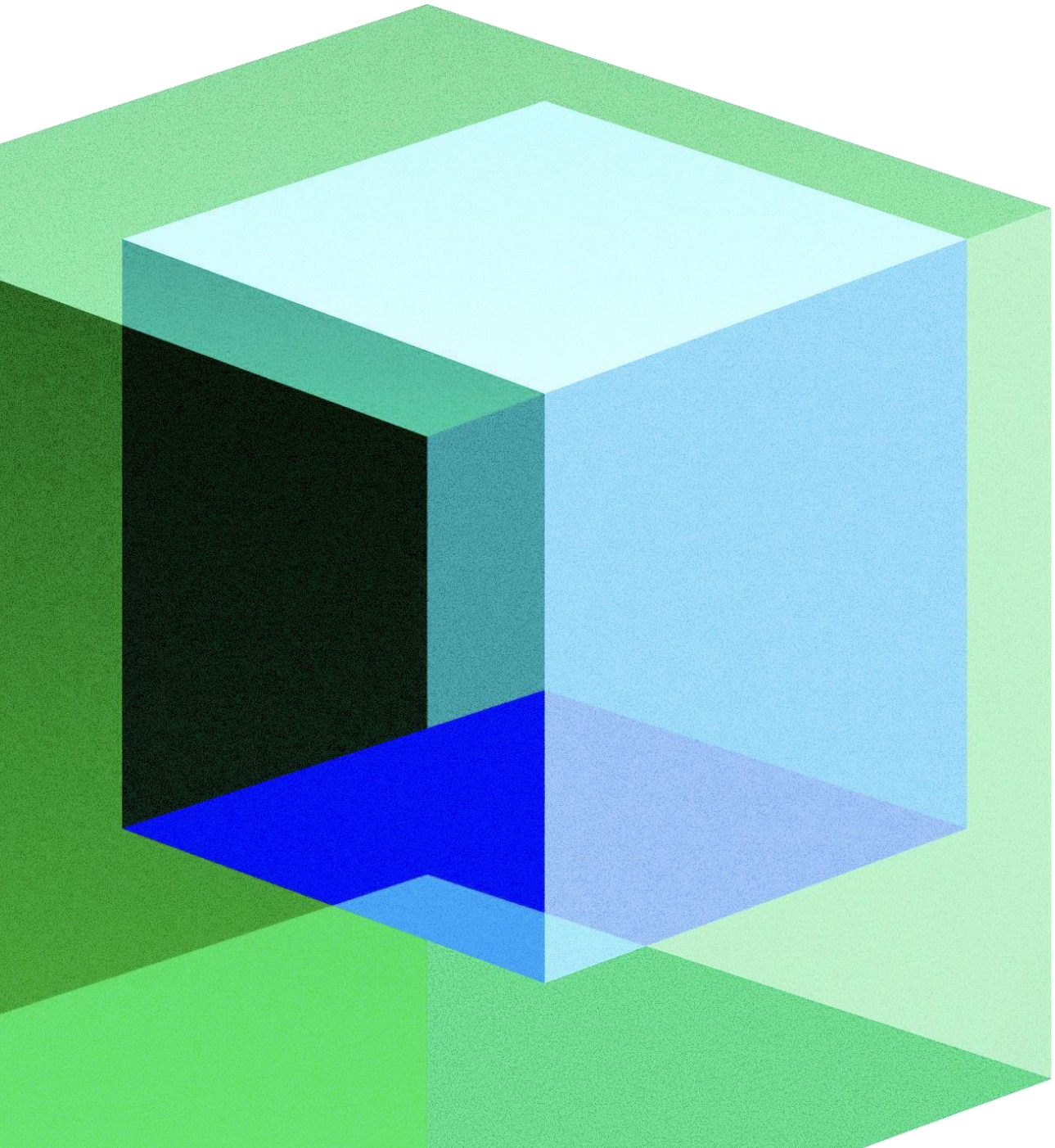
IBM Granite 3

Powerful performance, longer context,
new embedding models, and more.



Contents

Success Stories

High-impact results

IBM Granite delivers real results for enterprises

AI is driving impact across many areas, including customer service, employee productivity and IT operations. With Granite, you can match the right model to the right use case in these high-impact areas. With that focus in these strategic areas, you can avoid the pitfalls of stalled projects and pilots that never took flight.

As you build that foundation of ROI & show the value of AI to your board, you can build on your success so that your organization can keep investing and keep innovating.

This isn't hypothetical. As you can see, our customers have achieved real results - in productivity gains with AI and automation since January 2023.

Customer Experience	*Digital Labor*	*IT Operations*
Customer Service Improved customer service containment rates by 20% vodafone	Human Resources Reduced time from resume to offer by 60% blue pearl	Application Modernization 62% reduction in Ansible Playbook creation time citi
Customer Experience Reduced commentary editing from hours to minutes for 250+ matches us open	Financial Services Reduced bank workload by 50% COMPARUS	Application Modernization 1,500 hours of manual labor saved yearly WATER CORPORATION
E-Commerce A truly personalized shopping experience for millions of shoppers yappyBuy	Digital Assistant Streamlined proposal gathering for 160,000 consultants IBM	Insight Engine Bringing a new level of understanding to UFC fans (700M+) around the world UFC

Success Stories

Blue Pearl | IBM

Boosting efficiency, quality matches and talent retainment with gen AI

Blue Pearl had been encountering significant inefficiencies associated with traditional HR practices, particularly job-description creation and the matching of suitable candidates to hiring companies' needs. To address these challenges, Blue Pearl partnered with IBM® Client Engineering to develop an automated solution.

The proof of concept used large language models from the IBM watsonx.ai™ enterprise studio for AI builders to analyze structured and unstructured HR data, including job descriptions and résumés. This led to the creation of an automated job-description generator, which is capable of producing highly specific and relevant job postings tailored to the unique needs of various organizations.

Leveraging IBM Granite™, the team developed a job-matching engine. This engine aligns applicant credentials with job criteria by understanding the nuanced meanings within unstructured data, surpassing simple keyword matching.

The new system reduced the average time to fill vacancies by 60%, and the advanced semantic capabilities of watsonx.ai increased the fit of skilled professionals to job roles by 85%. Workplace cohesiveness and synergy saw a 35% improvement, attributed to the better alignment of candidate skills with organizational needs.

60%
reduced average time to fill vacancies

85%
increased fit of professionals to job roles

35%
increased workplace cohesiveness

Success Stories

US Open | IBM

AI models built with watsonx transform data into insight

For two weeks at the end of summer, nearly one million people make the journey to Flushing, New York, to watch the best tennis players in the world compete in the US Open Tennis Championships.

To help the US Open stay on the cutting edge of customer experience, IBM Consulting worked closely with the United States Tennis Association (USTA) to develop generative AI models that transform tennis data into insights and original content on the US Open app and website.

To do this, the USTA used IBM® watsonx™, a portfolio of AI products, and powerful AI models, including IBM Granite™ foundation models, to help develop key app features, such as Match Reports and AI Commentary for US Open highlight reels.

Reduced commentary editing from hours to minutes for 250+ matches

15M

World-class digital experiences for fans

7M

Data points captured and analyzed throughout the tournament

Success Stories

Citi | IBM

Driving developer productivity and code quality improvements

In a pilot with Citi, WCA for Red Hat Ansible Lightspeed has already led to 62% less time to create playbooks, 2X less critical failures, and an expected $80M+ savings in 3 years.

No developer required external support while using watsonx Code Assistant (WCA) (i.e. documentation, Stack Overflow), as compared to 24X external outreach without WCA.

Our AI for Code research is also powering IBM watsonx Code Assistant for Z, a new capability to help enterprises accelerate the transformation of their legacy applications, translate services from COBOL to Java, and save developers time and reduce human error.

IBM will continue to introduce new modalities, like WCA for Enterprise Java, and a code assistant base that solves unique enterprise challenges. These allow companies to bring their own domains or languages, and customize the state-of-the-art WCA code model, and deploy anywhere.

62%
reduced time to create automation Playbooks

2X
fewer critical failures

$80M
expected savings in 3 years

Overview

Granite

Granite is IBM's suite of generative AI models that provides enterprise control and customization through accessible model weights and architectures. It enables businesses to adapt models to specific needs, enhance security through transparency, and maintain version control across deployments. Built for production environments, Granite allows organizations to host specific versions in their systems, providing greater control over infrastructure while ensuring long-term stability and flexibility.

Granite Family

LLMs for enterprise

- Granite-7B-Base
- Granite-3.0/3.1/3.2-8B-Instruct
- Granite-3.0/3.1/3.2-2B-Instruct

Inference-efficient Mixture-of-experts (MoE):

- Granite-3.0/3.1-3B-A800M
- Granite-3.0/3.1-1B-A400M

Guardrail models

- Granite-Guardian-HAP-38M
- Granite-Guardian-HAP-125M
- Granite-Guardian-3.0/3.1-2B
- Granite-Guradian-3.0/3.1-8B

Time Series models

- Granite-TimeSeries-TTM-r1
- Granite-TimeSeries-TTM-r2
- Granite-TimeSeries-PatchTST
- Granite-TimeSeries-PatchTSMixer

Code models

- Granite-3B-Code
- Granite-8B-Code
- Granite-20B-Code
- Granite-34B-Code

Speculative decoding models

- Granite-3B-Code-Instruct-Accelerator
- Granite-8B-Code-Instruct-Accelerator
- Granite-20B-Code-Instruct-Accelerator
- Granite-34B-Code-Instruct-Accelerator
- Granite-7B-Instruct-Accelerator
- Granite-3.0/3.1-2B-Instruct-Accelerator

Geospatial model – Earth

- Granite-EarthObservation-HLS-Biomass
- Granite-EarthObservation-HLS-CanopyHeight
- Granite-EarthObservation-HLS-Landslide

Geospatial model – Weather and Climate

- Granite-WeatherClimate-Precip-Downscaling
- Granite-WeatherClimate-WindForecasting

Granite 3.2

The new **Granite 3.2 8B Instruct** and **Granite 3.2 2B Instruct** offer experimental chain-of-thought reasoning capabilities that significantly improve their ability to follow complex instructions with no sacrifice to general performance. The reasoning process can be toggled on and off, allowing for efficient use of computing resources.

When combined with IBM's inference scaling techniques, Granite 3.2 8B Instruct's extended thought process enables it to meet or exceed the reasoning performance of much larger models, including GPT-4o and Claude 3.5 Sonnet.

Multi-modal	Transparent with indemnification	Easily and quickly customizable	Granite models offered with Apache 2.0

Parameters	**2B, 8B Dense**
Training Data	**Web data + Synthetic data + Publicly available datasets with permissible licenses**
Input Modalities	**Multilingual Text and Image**
Output Modalities	**Multilingual Text and Code**
Context Length	**128k**
Training Tokens	**Up to 12T tokens**
Knowledge Cutoff	**April 2024**

Granite 3.2

Features of Granite 3.2 models

Lightweight

Our largest dense model has 8 billion parameters, and our smallest model has 2 billion, enabling hosting, or even fine-tuning, on more limited compute resources.

Chain-of-thought Reasoning

The new Granite 3.2 8B Instruct and Granite 3.2 2B Instruct offer experimental chain-of-thought reasoning capabilities that significantly improve their ability to follow complex instructions with no sacrifice to general performance. The reasoning process can be toggled on and off, allowing for efficient use of computing resources.

Time Series Models

The latest additions to the Granite Timeseries model family, Granite-Timeseries-TTM-R2.1, expand TTM's forecasting capabilities to include daily and weekly predictions in addition to the minutely and hourly forecasting tasks already supported by prior TTM models.

Trustworthy

We're introducing new model sizes for Granite Guardian 3.2, including a variant derived from our 3B-A800M mixture of experts (MoE) language model. The new models offer increased efficiency with minimal loss in performance.

Vision

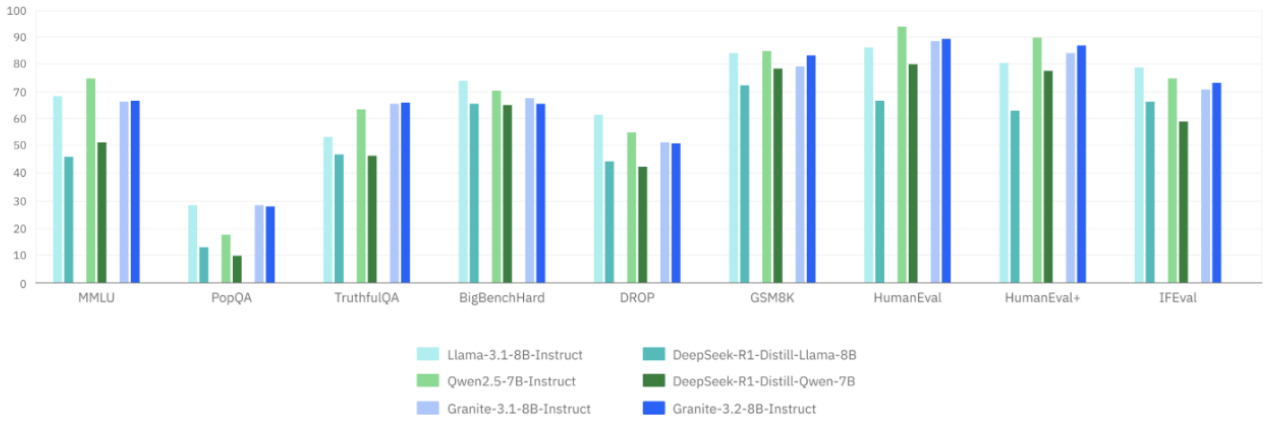
Our new multimodal model, Granite Vision 3.2 2B, was developed with a particular focus on document understanding, on which it matches the performance prominent open models 5 times its size.

Embedding Models

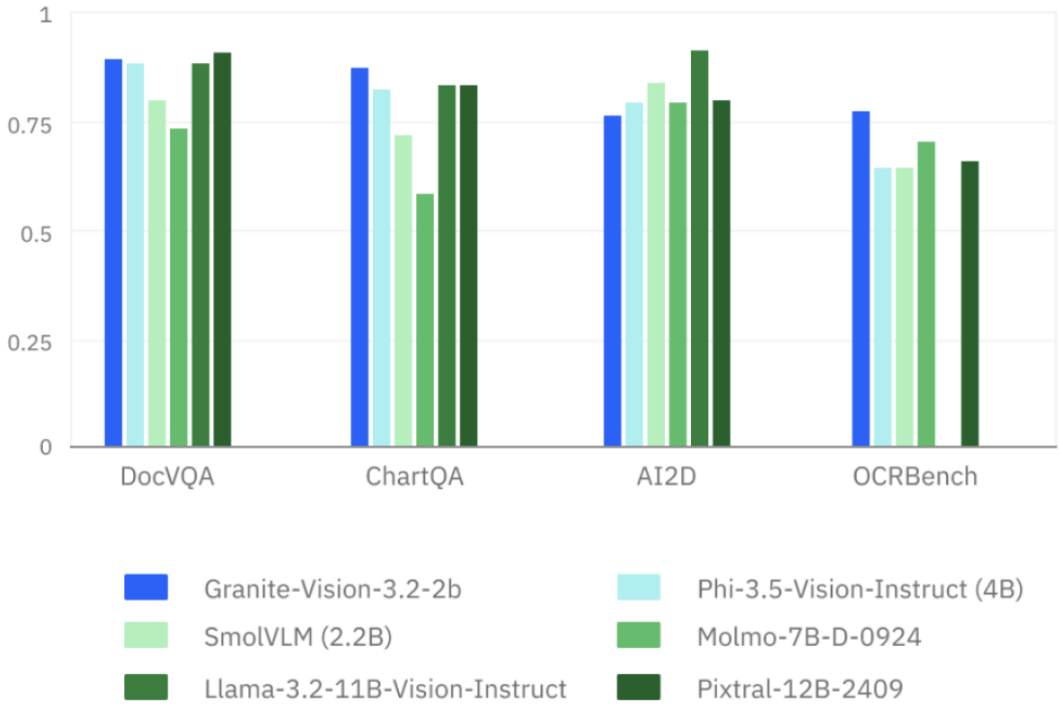
We're introducing new model sizes for Granite Guardian 3.2, including a variant derived from our 3B-A800M mixture of experts (MoE) language model. The new models offer increased efficiency with minimal loss in performance.

Granite 3.2

Competitive performance



Comparison of pre- and post-reasoning performance on general academic performance benchmarks



Comparison on benchmarks that measure performance on document understanding tasks

Granite 3.1

Granite 3.1 language models are lightweight, state-of-the-art, open foundation models that natively support **multilingual experience**, **coding**, **reasoning**, and **tool calling**, including the potential to be run on constrained compute resources. All the models are publicly released under an Apache 2.0 license for both research and commercial use. The models' data curation and training procedure were designed for enterprise usage and customization, with a process that evaluates datasets for governance, risk and compliance (GRC) criteria, in addition to IBM's standard data clearance process and document quality checks.

Fit-for-purpose	Transparent with indemnification	Easily and quickly customizable	Granite models offered with Apache 2.0

Parameters	**2B, 8B Dense** **1B, 3B MoE**
Training Data	**Web data + Synthetic data + Publicly available datasets with permissible licenses**
Input Modalities	**Multilingual Text**
Output Modalities	**Multilingual Text and Code**
Context Length	**128k**
Training Tokens	**Up to 12T tokens**
Knowledge Cutoff	**April 2024**

Granite 3.1

Key advantages of Granite 3.1 models

Lightweight

Our largest dense model has 8 billion parameters, and our smallest MoE model has an activated parameter count of 400 million, enabling hosting, or even fine-tuning, on more limited compute resources.

Competitive performance

All our models demonstrate competitive performance on par with leading foundation models, evaluated on multiple benchmark datasets.

Trustworthy Enterprise-Grade LLM

All our models are trained on license-permissible data collected following IBM's Al Ethics principles for trustworthy enterprise usage. We describe in great detail the sources of our data, data processing pipeline, and data mixture search to strengthen trust in our models for mission-critical and regulated applications.

Robust Models with Permissive License

Combined with excellent performance across various benchmarks, our Granite 3.1 models provide a great foundation for enterprise customization. All our models, including instruct variants, use an Apache 2.0 license, allowing for more consumer and enterprise usage flexibility over the more restrictive licenses of other available models in the same class.

Reduced operational costs

Runs training and inference tasks at a fraction of the cost of leading closed models, significantly reducing operational costs.

Granite 3.1

Architecture

Granite 3.1 models come in 4 varying sizes and 2 architectures:

Dense Models
2B and 8B parameter models

- Trained on 12 trillion tokens in total.
- State-of-the-art training and data recipes 12T+ tokens training data
- Designed for enterprise tasks:
- Language (RAG, summarization, entity extraction, classification, etc.),
- Code (generation, translation, bug fixing),
- Agents (tool use, advanced reasoning),
- Multilingual support (en, de, es, fr, ja, pt, ar, cs, it, ko, nl, zh)

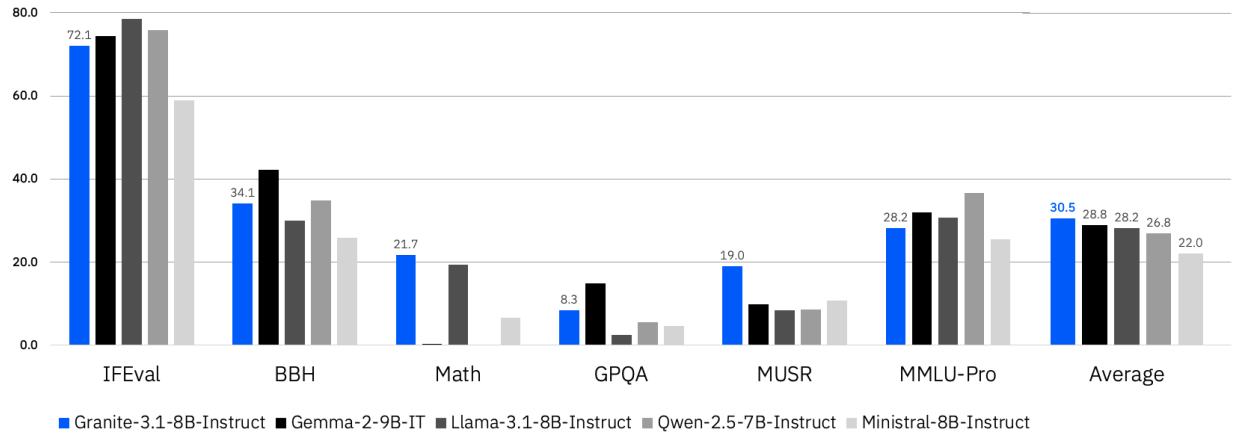
Mixture-of-Expert (MoE) Models
Sparse 1B and 3B MoE models

- 400M and 800M activated parameters respectively, trained on 10 trillion tokens in total.
- 10T+ tokens training data
- Runs with <1B parameters at inference time with minimal performance trade-off
- Ideal for on-device applications or runtimes requiring extremely low latency

Accordingly, these options provide a range of models with different compute requirements to choose from, with appropriate trade-offs with their performance on downstream tasks. At each scale, we release base model — checkpoints of models after pretraining, as well as instruct checkpoints — models finetuned for dialogue, instruction-following, helpfulness, and safety.

Granite 3.1

Competitive performance



Granite 3.1 compared against Benchmarks Genmma-2, Llama-3.1, Qwen-2.5, and Ministral

Benchmark Datasets:

1. **Instruction-Following Evaluation (IFEval):** Tests model's ability to follow explicit formatting instructions - Instruction following, Formatting, Generation
2. **Big Bench Hard (BBH):** Collection of challenging for LLM tasks across domains, for example Language understanding, Mathematical reasoning, Common sense and world knowledge
3. **Mathematics Aptitude Test of Heuristics (MATH):** High school level competitions mathematical problems - Complex algebra, Geometry problems, Advanced calculus
4. **Graduate-Level Google-Proof Q&A (GPQA):** PhD-level knowledge multiple choice questions in science – Chemistry, Biology, Physics
5. **Multistep Soft Reasoning (MuSR):** Reasoning and understanding on/of long texts - Language understanding, Reasoning capabilities, Long context reasoning
6. **Massive Multitask Language Understanding - Professional (MMLU-Pro):** Expertly reviewed multichoice questions across domains, for example - Medicine and healthcare, Law and ethics, Engineering, Mathematics

Granite 3.1

Use-cases

Use-cases	Recommendation	Model Size
• Personal information management • Multilingual knowledge retrieval • Rewriting tasks running locally on edge	Retrieval, Summarization, faster inference	**1B** (MoE - A400M) (base, Instruct)
• Mobile AI-powered writing assistant	Retrieval, Summarization, faster inference	**2B** (base, Instruct)
• Mobile AI-powered writing assistant	Query and prompt rewriting, mobile AI-powered writing assistant, edge devices	**3B** (MoE - A800M) (base, Instruct)
• Text summarization • Text classification • Sentiment analysis • Language translation • Entity recognition	Ideal for limited computational power and resources, faster training times	**8B** (base, Instruct)
• Text summarization • Text classification • Sentiment analysis • Language translation • Entity recognition	Ideal for limited computational power and resources, faster training times, faster inference	**8B** (Instruct-Accelerator) [Unique selling point]

Granite Embedding Models

The Granite Embedding collection delivers innovative sentence-transformer models purpose-built for retrieval-based applications. Featuring a bi-encoder architecture, these models generate high-quality embeddings for textual inputs such as queries, passages, and documents, enabling seamless comparison through cosine similarity. Built using retrieval oriented pretraining, contrastive finetuning, knowledge distillation, and model merging, the Granite Embedding lineup is optimized to ensure strong alignment between query and passage embeddings.

Built on a foundation of carefully curated, permissibly licensed public datasets, the Granite Embedding models set a high standard for performance, maintaining competitive scores not only on academic benchmarks such as BEIR, but also out-performing models of the same size on many enterprise use cases. Developed to meet enterprise-grade expectations, they are crafted transparently in accordance with IBM's AI Ethics principles and offered under the Apache 2.0 license for both research and commercial innovation.

Model	granite-embedding-30m-english	granite-embedding-125m-english	granite-embedding-107m-multilingual	granite-embedding-278m-multilingual
Embedding Size	**384**	**768**	**384**	**768**
Vocabulary Size	**50265**	**50265**	**250002**	**250002**
Max Sequence Length	**512**	**512**	**512**	**512**
No. of Parameters	**30M**	**125M**	**107M**	**278M**

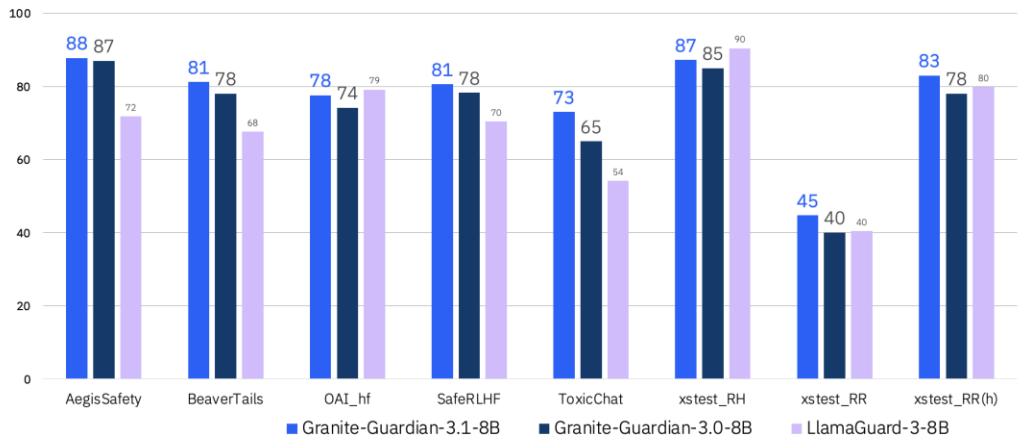
Granite Guardian Models

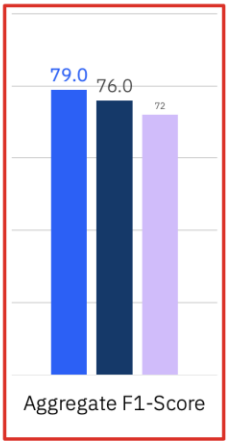
The Granite Guardian models are a collection of models designed to detect risks in prompts and responses. Trained on instruction fine-tuned Granite languages models, these models can help with risk detection along many key dimensions catalogued in the IBM AI Risk Atlas. These models are trained on unique data comprising human annotations from socioeconomically diverse people and synthetic data informed by internal red-teaming.

Granite Guardian provides comprehensive coverage of Risks

Breadth – Spans Social risks, Security risks and risks specific to RAG use-cases

Depth – Enables explicit detection of social risks such as unethical behavior, social-bias, violence, profanity, and sexual content; security risks like jailbreaks; and RAG-specific hallucination risks





Guardian Comparision

Granite Guardian

Models and Use-cases

Granite Guardian models

- Granite-Guardian-3.0/3.1-2B
- Granite-Guardian-3.0/3.1-8B
- Granite-Guardian-3.2-5B
- Granite-Guardian-3.2-3B-A800M

Use-cases	Recommendation	Model Size
• Detecting harm-related risks within prompt text or model response (as guardrails). These present two fundamentally different use cases as the former assesses user-supplied text while the latter evaluates model-generated text. • RAG (retrieval-augmented generation) use-case where the guardian model assesses three key issues: context relevance (whether the retrieved context is relevant to the query), groundedness (whether the response is accurate and faithful to the provided context),and answer relevance (whether the response directly addresses the user's query). • Function calling risk detection within agentic workflows, where Granite Guardian evaluates intermediate steps for syntactic and semantic hallucinations. This includes assessing the validity of function calls and detecting fabricated information, particularly during query translation.	Ideal for edge devices	2B/3B
	Ideal for limited computational power and resources, faster training times	8B/5B

Model Selection

Comparing Models

Granite models

- Granite-base
- Granite-3.0 (Oct 2024)
- Granite-3.1 (Dec 2024)
- Granite-3.2 (Feb 2025)

Feature	**Granite-base and code**	**Granite 3.0** *(Oct 2024)*	**Granite 3.1** *(Dec 2024)*	**Granite 3.2** *(Feb 2025)*
Context length	4k	4k	128k	128k
Input modalities	English text and code	Multilingual text (12 languages)	Multilingual text (12 languages)	Multilingual text and images (12 languages)
Output modalities	Multilingual text and code	Multilingual text and code	Multilingual text and code	Multilingual text and code
Training Tokens	2.5T	12T	12T	12T
Languages	English	English, German, Spanish, French, Japanese, Portuguese, Arabic, Czech, Italian, Korean, Dutch, and Chinese	English, German, Spanish, French, Japanese, Portuguese, Arabic, Czech, Italian, Korean, Dutch, and Chinese	English, German, Spanish, French, Japanese, Portuguese, Arabic, Czech, Italian, Korean, Dutch, and Chinese

Accuracy, Cost and Latency based model selection

When selecting the appropriate model for deployment, target metrics must be carefully evaluated.

An example: To enhance customer support efficiency, the e-commerce company intends to implement an AI-driven customer service bot to streamline issue resolution.

Define Accuracy Requirements

Ensure at least 90% accuracy in classifying and escalating customer service requests.

Prioritize Economy

Ensure that operational costs do not exceed $10 for every 1000 requests handled.

Latency

Ensure each request is processed in under one minute to improve customer satisfaction.

The e-commerce company has done following three experimentations.

	Method	Accuracy	Cost	Avg. Latency
Granite-3.1-2b	Fine-tuned (1000 examples)	96%	$7.0	<1s
Granite-3.1-3b	Few-shot (n=5)	91%	$8.0	<1s
Granite-3.1-8b	Zero-shot	86%	$15.0	<1s

✅ **Granite-3.1-2b**
(Fine-tuned, 1000 examples)

This model offers the best balance of high accuracy, low cost, and fast response time. Since it has been fine-tuned with 1000 examples, it is highly optimized for the task, making it the most reliable choice for deployment.

❌ **Granite-3.1-3b**
(Few-shot, n=5)

While this model performs well, its accuracy is lower than the fine-tuned model, and the cost is slightly higher. Given that the fine-tuned model provides better performance at a lower cost, this model is not the optimal choice.

❌ **Granite-3.1-8b**
(Zero-shot):

Despite meeting latency requirements, this model has the lowest accuracy and the highest cost. Since it operates in a zero-shot setting, it lacks optimization that fine-tuned and few-shot models offer. The cost per request is double that of the fine-tuned model, making it economically unviable for deployment.

Final Recommendation

The **Granite-3.1-2b** model is the best choice for deployment due to its superior accuracy, lowest cost, and fast response time. The other models either fail to meet cost-effectiveness or accuracy benchmarks, making them less suitable for implementation.

Use-case based model selection

The process of selecting an appropriate model from the Granite family begin with a careful categorization of the specific use case at hand. Examples of such use cases include code generation, time series analysis, and geospatial applications.

Language-based tasks

Retrieval augmented generation (RAG), summarization, content generation, insight extraction, and classification based on documents or dynamic content

IBM Granite Language models

- Granite-7B-Base
- Granite-3.1-1B-A400M
- Granite-3.1-3B-A800M
- Granite-3.1/3.2-2B
- Granite-3.1/3.2-8B

Example: Building a Q&A resource from a broad knowledge base, providing customer service assistance

Code

Optimize the software development lifecycle with code generative tasks, including code generation, code explanation, and code editing

IBM Granite Language models

- Granite-34B-Code
- Granite-20B-Code
- Granite-8B-Code
- Granite-3B-Code

Example: AI-generated code recommendations, IT application modernization from COBOL to Java

Safety

Safeguard AI with models ensuring enterprise data security and mitigate risks across a variety of user prompts and LLM responses

IBM Granite Language models

- Granite-Guardian-HAP-125M
- Granite-Guardian-HAP-38M
- Granite-Guardian-3.1-2B
- Granite-Guardian-3.1-8B
- Granite-Guardian-3.2-5B/3B-A800M

Example: AI compliance with regulatory requirements in financial services, healthcare, and government

Time series

Time-series forecasting to easily analyze current data to make predictions and informed decisions

IBM Granite Language models

- Granite-TimeSeries-TTM-r1
- Granite-TimeSeries-PatchTSMixer
- Granite-TimeSeries-PatchTST
- Granite-TimeSeries-TTM-r2

Example: Predicting future customer demand for a given product and period, using historical sales and other data sources

Geospatial

Uncover patterns and trends in geospatial data

IBM Granite Language models

- Granite-EarthObservation-HLS-Biomass
- Granite-EarthObservation-HLS-CanopyHeight
- Granite-EarthObservation-HLS-Landslide
- Granite-WeatherClimate-Precip-Downscaling
- Granite-WeatherClimate-WindForecasting

Example: NASA and IBM teamed up to create an AI foundation model for Earth observations using large-scale satellite and remote sensing data

Accelerators

Accelerates token generation when using the accelerator with the base model

IBM Granite Language models

- Granite-3B-Code-Instruct-Accelerator
- Granite-8B-Code-Instruct-Accelerator
- Granite-20B-Code-Instruct-Accelerator
- Granite-34B-Code-Instruct-Accelerator
- Granite-7B-Instruct-Accelerator

Example: conversational AI, Code generation, Machine translation which needs speeding up inference, reducing latency.

Model Deployment

Model Availability & Adoption

How and where to deploy

IBM's acquirement of RedHat has been a game-changer in terms of where our large language models can be deployed. Depending on the client's wants, any model of Granite can be deployed:

- On Prem
- On IBM servers
- Hybrid; on both
- On Cloud provider of the Client's choosing

Granite availabilities

watsonx
Granite Language
Granite Guardian

Nvidia NIM
Granite Language
Granite Guardian
Granite MoE

Google Vertex
Granite Language
Granite MoE

HuggingFace
Granite Language
Granite Guardian
Granite Accelerators
Granite MoE

Replicate
Granite Language
Granite MoE

RHEL AI
Granite Language

OpenShift AI
Granite Language

Ollama
Granite Language
Granite MoE

Model Deployment

Stay in control of how you deploy and run Granite

- Choose your infrastructure
- Avoid vendor lock-in
- Maximize system efficiency and performance
- Enhance system reliability
- Deliver exceptional global system performance
- Decide what options to choose and when

Deployment options	Benefits	Ideal For
Cloud-based Managed Services (IBM Cloud (watsonx), Google Vertex, NVidia NIM, HuggingFace)	• Ease of use • Rapid deployment, minimal setup • Overhead management • Flexible pay-as-you-go pricing	Businesses looking to quickly validate AI use cases and reach product market fit
VPC self-hosting (CP4D, RHEL, OpenShift)	• Balance of control and convenience • Enhanced security and customization within familiar cloud environment • Cloud scalability	Businesses requiring data privacy and moderate customization, with some in-house technical expertise
On-premises self-hosting (On customers env)	• Ultimate control, customization and security • Data and infrastructure sovereignty • Cost savings for high-volume use cases • Ability to leverage existing infrastructure investments	Businesses operating in highly regulated industries with significant technical resources, or those with use cases that require high data privacy and security

Model Deployment

Select the right model adoption strategy to fit your unique use-case

Granite out-of-the-box

Utilizing a pre-trained Granite model without any modifications or additional training is suitable where customization is not critical, and ideal for internal applications and proofs-of-concept. This approach is generally the most cost-effective, as it leverages preexisting models and does not require additional training or customization. Applications and use-cases include:

- IBM's AskIBM
- Internal request and ticket handling

Prompt Engineering (Granite prompting guide)

Prompt engineering refines the AI's responses, making it suitable for more complex applications. Additionally, incorporating retrieval-augmented generation (RAG) can enhance the use of Granite. This approach requires prompt engineering tailored to the specific domain, ensuring the prompts are optimally designed to retrieve and generate relevant information. Applications and use-cases include:

- Enhanced fraud detection for financial services
- Creative writing

Fine-tuning

Fine-tuning a pre-trained model is more costly than out-of-the-box use or prompt engineering because it requires additional training resources and expertise. However, the benefits of improved performance may outweigh these costs for certain applications or use-cases including:

- Request for Proposal (RFP)
- Personalized recommendations
- Legal and medical document analysis

Model Deployment

Fine-tuning: InstructLAB

InstructLAB allows anyone to improve an existing LLM by fine-tuning it with additional data sources. This allows LLMs to continuously gain new knowledge, supplementing gaps in their initial training, even about current events that happened since their pre-training phase.

InstructLAB is compute-efficient, and catastrophic-forgetting averse.

InstructLAB is a model-agnostic technology.

InstructLAB can augment models through **skill recipes** used to generate synthetic data for tuning. Experiments can be run locally on quantized version of these models

Skill recipes take the form of example inputs/outputs for a desired skill. These skills are organized in structured **taxonomy** and anyone can contribute to it.

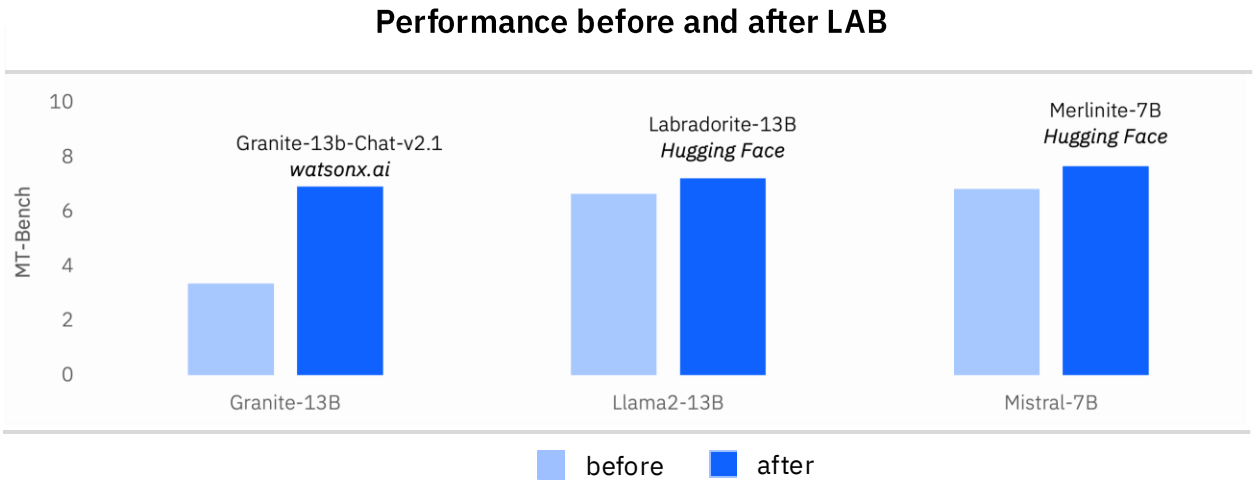
InstructLAB uses the skill recipes to systematically generate new **synthetic data.**

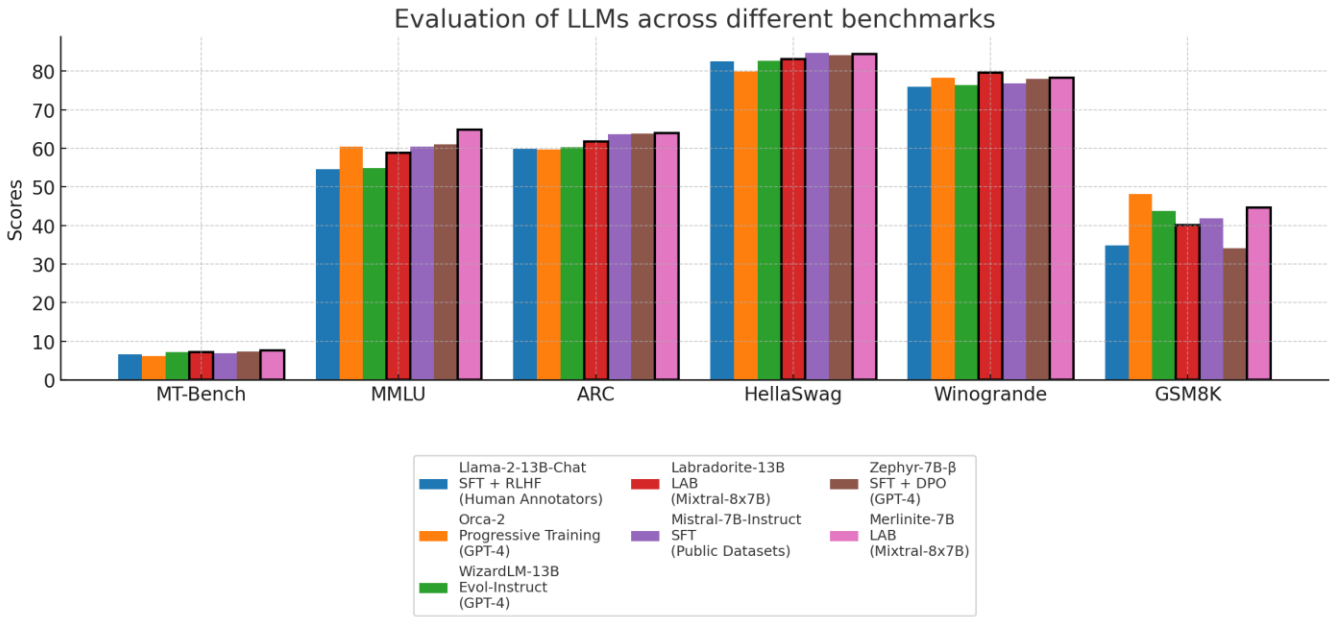
The InstructLAB base model is re-tuned using all synthetic data generated to date. This includes any new contributions, which introduces **new skill.**

Clients who would like to fine-tune Granite models for their use-cases can benefit from regularly reported tests on various fields from the InstructLAB Community.

Model Deployment

InstructLAB Benchmarks





Model Deployment

Cost considerations

The cost of using Granite is based on number of tokens generated.

Model name	Provider	Deployed by IBM Pay per token	Price Per 1 million tokens in USD*	Deploy on demand Pay by the hour	Price Per hour in USD*
llama-3-2-1b-instruct	Meta	✓	USD 0.10*		
llama-3-2-3b-instruct	Meta	✓	USD 0.15*		
llama-3-1-8b-instruct	Meta	✓	USD 0.60*	✓	USD 5.22*
llama-guard-11b-vision	Meta	✓	USD 0.35*		

Model name	Provider	Deployed by IBM Pay per token	Price Per 1 million tokens in USD*	Deploy on demand Pay by the hour	Price Per hour in USD*
granite-3-2b-instruct (v3.1)	IBM	✓	USD 0.10*		
granite-3-8b-instruct (v3.1)	IBM	✓	USD 0.20*		
granite-guardian-3-2b (v3.1)	IBM	✓	USD 0.10*		
granite-guardian-3-8b (v3.1)	IBM	✓	USD 0.20*		
granite-vision-3-2-2b	IBM	✓	USD 0.10*		
Granite-3-2-8b-instruct	IBM	✓	USD 0.20*		

Appendix

Additional Resources

Official Website:
IBM Granite

Github:
IBM Granite Github

Huggingface:
IBM Granite Huggingface

Youtube:
IBM Granite Youtube

Playground:
IBM Granite Playground

Recipes for Agents:
IBM Granite Recipes for agents