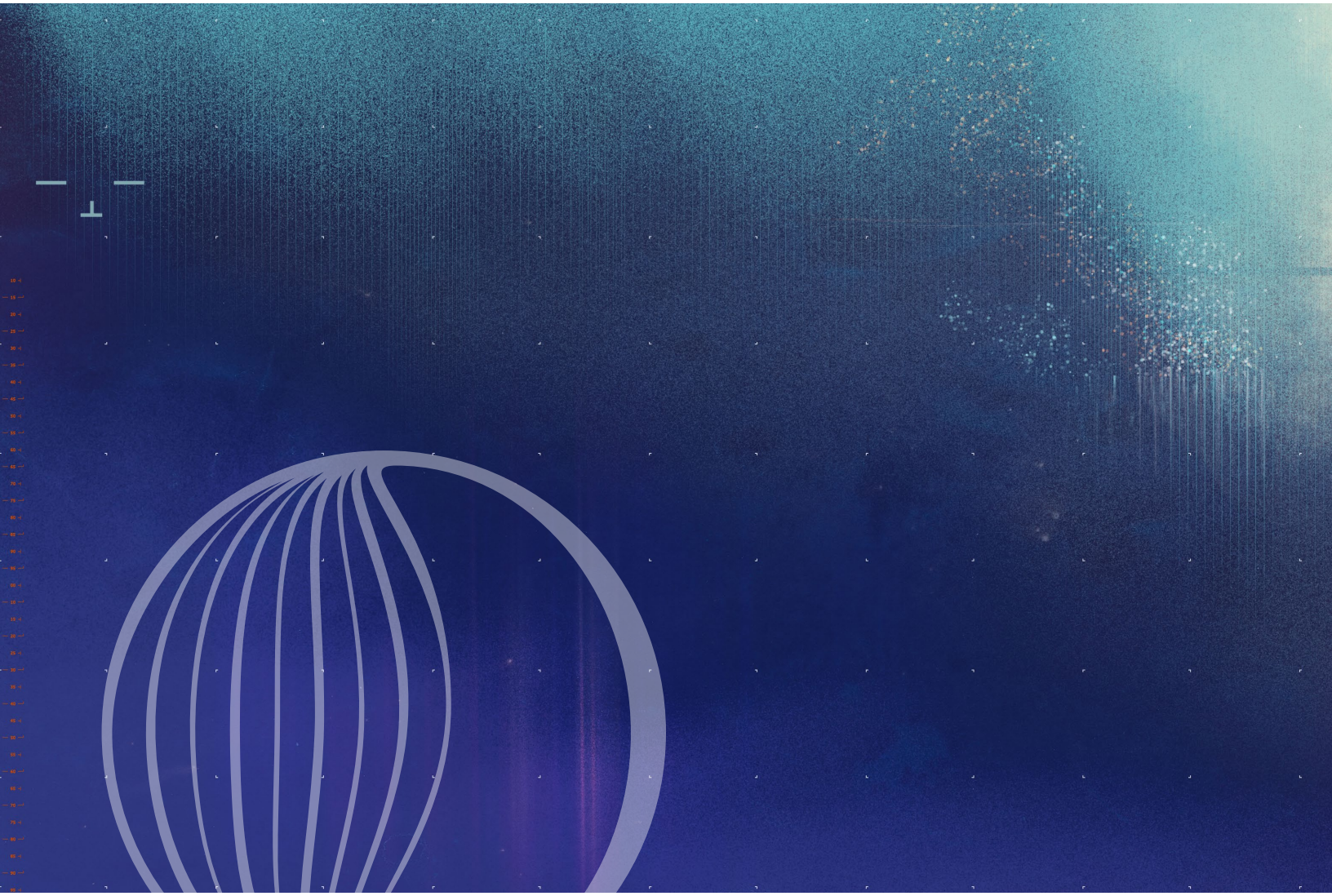
signal65



SIGNAL65 VALIDATION:

IBM Granite Benchmarking and Enterprise Readiness

AUTHOR

Mitch Lewis
Performance Analyst | Signal65

IN PARTNERSHIP WITH

IBM

FEBRUARY 2025

Executive Summary

The field of AI has experienced rapid innovation within the past few years, with generative AI and Large Language Models (LLMs) in particular garnering significant attention from enterprise organizations. The quick pace of innovation has led to a plethora of new and intriguing models available in the market, while the relatively new nature of the technology presents a challenge for enterprise organizations to evaluate their options.

LLMs can be evaluated by a wide range of metrics including logical reasoning abilities, math and coding capabilities, safety, and more. Different models may exceed in various areas due to different model sizes, architectures, and mixtures of training data. Enterprise organizations choosing LLMs should be aware how different models perform in the areas that are most beneficial to their intended use cases. In addition, factors such as size, security, and continuous development are core considerations that may impact the practicality of model usage in an enterprise environment.

This paper discusses an overview of LLM evaluation criteria and reviews the performance of IBM Granite models in several key areas. This paper additionally evaluates how IBM Granite models are positioned as competitive solutions for enterprise AI requirements.

LLM Evaluation

The renewed interest in AI technology experienced over the past few years, led primarily by generative AI applications, has resulted in rapid development of new and impressive LLMs. The current AI landscape includes various LLMs capable of delivering increasingly impressive results across a wide range of tasks. For enterprise organizations focused on leveraging AI technology for their business use cases, it can be challenging to select the most appropriate model. Organizations should be aware of various LLM evaluation criteria and be intentional in selecting a model that is well suited for their needs.

The evaluation of LLMs is a complex task in which there is often not a definitive best solution. The various characteristics of each LLM, such as reading comprehension or coding capabilities, may hold different weight for different organizations, depending on their specific priorities and intended use cases. In addition, model selection may be constrained by parameters such as ease of deployment or hardware limitations.

Measuring the capabilities of LLMs is additionally difficult as it often involves evaluation of vague or ambiguous criteria, such as reasoning and common sense. To evaluate LLMs, there exists a large, and evolving, array of benchmarks that measure various capabilities of LLMs. While no single benchmark is capable of showcasing all aspects of an LLM, evaluating models across a wide range of benchmarks can help organizations understand their strengths and weaknesses.

IBM Granite Overview

IBM Granite is a family of generative AI models developed by IBM and targeted to meet the varying needs of enterprise AI. IBM developed its Granite models with business use cases in mind and has released several variations to meet different enterprise needs, all open source under Apache 2.0 licensing, encouraging transparency and enabling users to customize models according to their needs. The Granite 3.0 family includes both pre-trained and post-trained models with 2 Billion and 8 Billion parameter dense models as well as smaller Mixture-of-Experts (MoE) sparse models with 400 Million and 800 Million activated parameters. All models in the Granite 3.0 family are relatively small, created with the intention of meeting practical enterprise deployment requirements and enabling customization using enterprise data to achieve state-of-the-art performance at low computing costs. The various model sizes provide additional flexibility to meet a wide array of both use cases and infrastructure capabilities.

The Granite 2B and 8B dense models are decoder-only transformer models, with a similar architecture to many other state-of-the art language models. Granite 3.0 dense models utilize Grouped Query Attention (GQA) for its attention mechanism, Rotary Position Embedding (RoPE) for positional encoding, and RMSNorm for normalization before each multi-layer perceptron (MLP) layer. Granite's MLP layers utilize the SwiGLU activation function. To reduce the size of the models, parameter sharing of the input and output layers is utilized. The MoE models share a similar architecture with the larger dense Granite models, with MoE layers substituted for the standard MLP layers. Other key design choices for the MoE models include Dropless Token Routing, Fine Grained Experts, and Load Balancing Loss.

A key differentiator of IBM Granite 3.0 models is the data used to train the models. IBM has trained its Granite models on a combination of publicly available text and code data sets as well as synthetic data sets created by IBM. Data used in training Granite models is highly curated and filtered, ensuring data is high quality and meets all data governance and licensing requirements. Training data includes multilingual data across 12 languages, as well as numerous academic, technical, math, and code sources. IBM utilizes a two-stage approach to model pre-training with different data mixtures at each stage. The first stage uses large data sets to maximize knowledge across a diverse range of domains, while the second stage utilizes smaller, higher quality data sets to improve model performance on specific tasks. For the post-trained versions of Granite models, IBM additionally employs a wide range of post-training techniques to further improve instruction following capabilities and align the models with human values.

Granite 3.0 Benchmarking

To evaluate the performance of Granite 3.0 models, IBM has tested its models, as well as several competitive models, against a comprehensive suite of LLM benchmarks., Several leading open-source models of similar parameter sizes were tested as competitors for a fair comparison.

IBM Granite 2B and 8B dense models were compared against the following competitors:

- Gemma-2-2B
- Llama-3.2=2B
- Mistral-7B
- Llama-3.1-8B

IBM Granite MoE 400M and 800M models were compared against the following competitors:

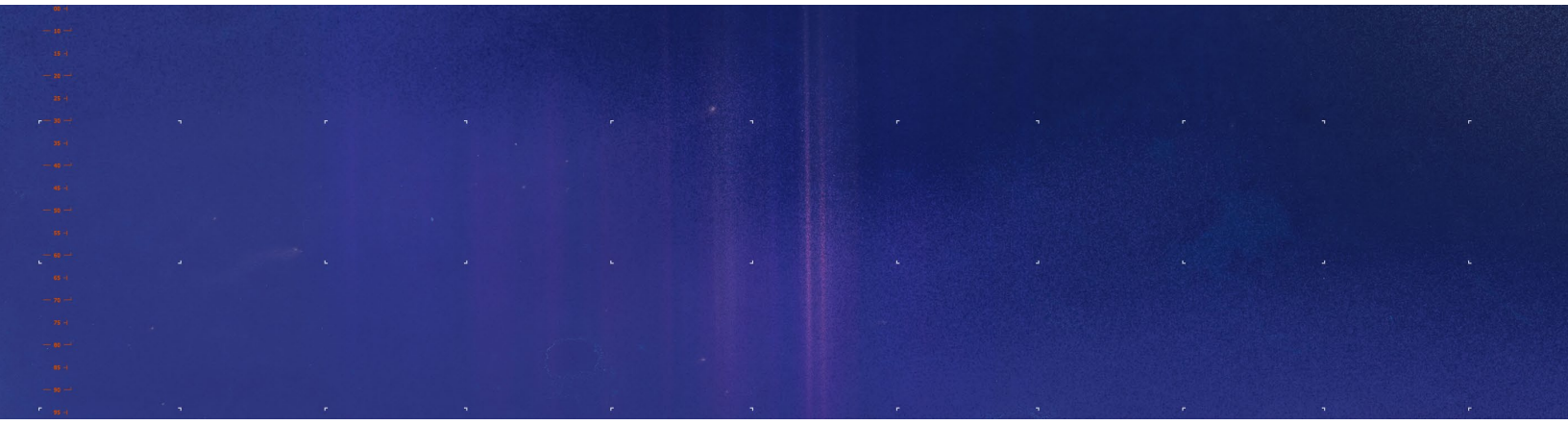
- SmolLM-360M
- Llama-3.2 1B
- SmolLM-1.7B

Models were evaluated using a suite of standard LLM benchmarks to test various tasks for both base models and instruct models. Models were evaluated across the following areas:

Base Models	Instruct Models
• Human Exams • Common Sense • Reading Comprehension • Reasoning • Code • Math	• Instruction Following • Reasoning • Multilingual • RAG • Code • Cybersecurity • Function Calling • Safety

Figure 1: *Base and Instruct Model Evaluation Metrics*

To conduct an independent third-party evaluation of IBM's testing, Signal65 has both reviewed the published results and observed the methodology of IBM's benchmarking process.



Results Overview

In order to process input and submit data to AI accelerators efficiently, the inferencing software creates tokens from input data and then sends those tokens in large groups (called batching) in order to help increase overall token processing rates.

Figure 2: *Average Performance of Base Models*

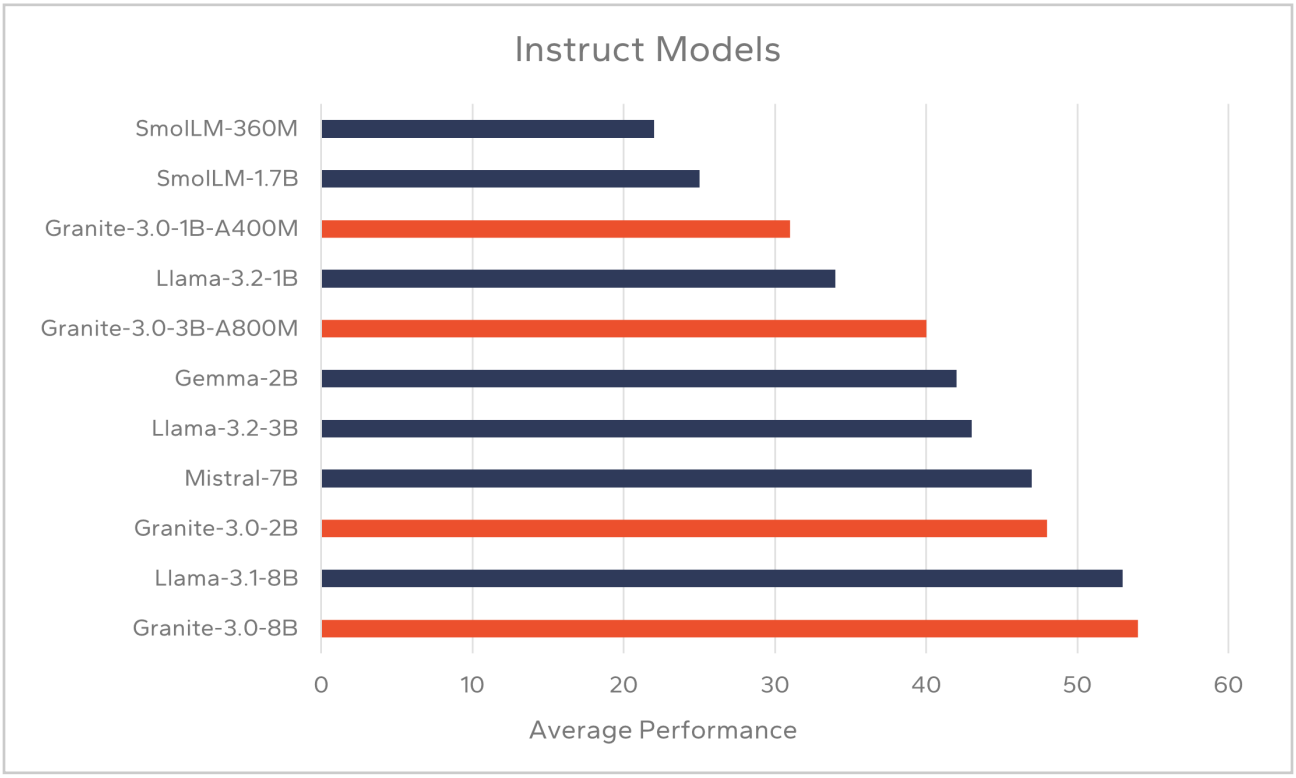
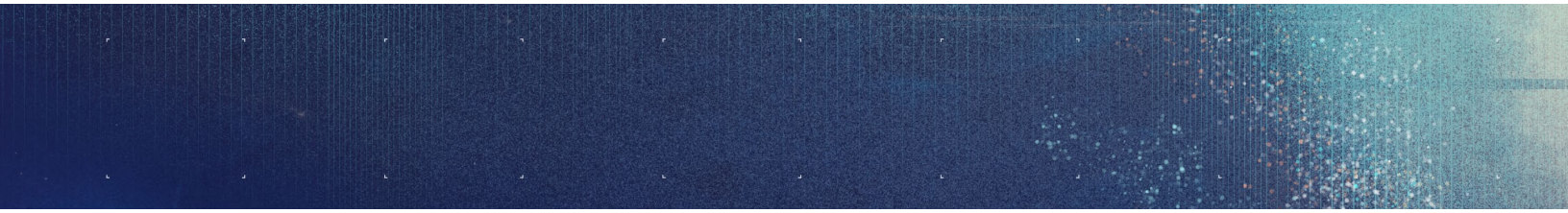


Figure 3: *Average Performance of Instruct Models*

Notably, the Granite-3.0-8B models were found to have the highest average performance, both when evaluating base models and instruct models. Smaller Granite 3.0 models not only outperform competitive models of the same size but were additionally found to outperform significantly larger models. The Granite-3.0-2B model, for example, not only outperforms the similarly sized Gemma-2B and Llama-3.2-3B models, but also outperforms Mistral-7B. Similarly, both Granite MoE models, which utilize less than 1B active parameters, were found to outperform larger models. These results highlight IBM's capability to develop highly capable models with manageable parameter sizes.

Additionally, the Granite-3.0-8B models, both base and instruct versions, lead their nearest competitors in each individual category tested. This demonstrates that IBM Granite not only provides impressive performance on average, but additionally provides flexibility to excel in a wide range of possible tasks.



Performance Comparison

To measure a range of general LLM capabilities, IBM Granite 3.0 instruct models were tested against its competitors with a suite of standard benchmarks that fit several core categories. The benchmarks ran can be seen in Figure 4.

Category	Benchmarks
Instruction Following	IFEval MT-Bench
Human Exams	AGI-Eval MMLU MMLU-Pro
Commonsense	OBQA SIQA Hellaswag WinoGrande TruthfulQA
Reading Comprehension	BoolQ SQuAD 2.0
Reasoning	ARC-C GPQA BBH
Code	HumanEvalSynthesis HumanEvalExplain HumanEvalFix MBPP
Math	GSM8K MATH
Multilingual	PAWS-X MGSM

Figure 4: *General Knowledge and Instruction Benchmarks*

In general, Granite 3.0 models of all sizes performed were found to perform highly across all categories. Due to the dynamic nature of LLMs and the benchmarks used to measure them, model performance often varies between individual benchmarks within a single category. Running several benchmarks, however, can create a stronger indicator of its overall abilities. IBM Granite models were found to achieve leading scores in a majority of the categories tested, and on average, each granite model outperformed its closest sized competitors.

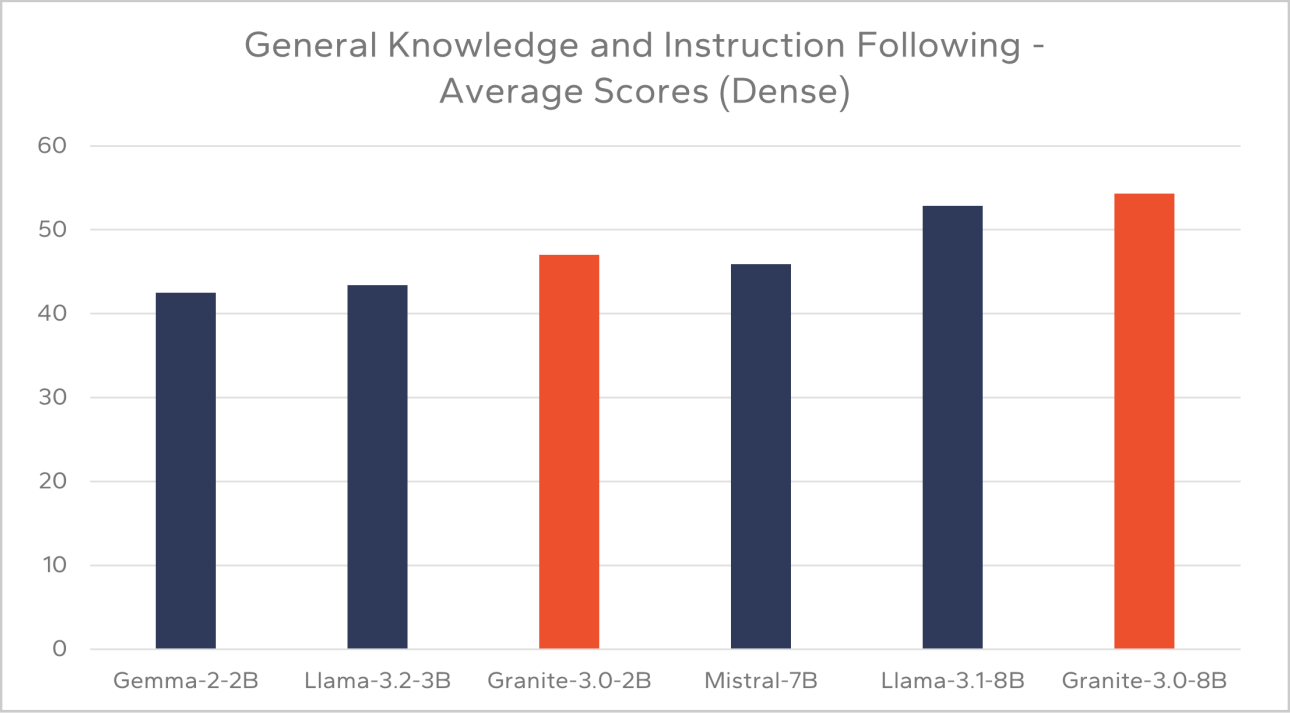


Figure 5: *General Knowledge and Instruction (Dense Models)*

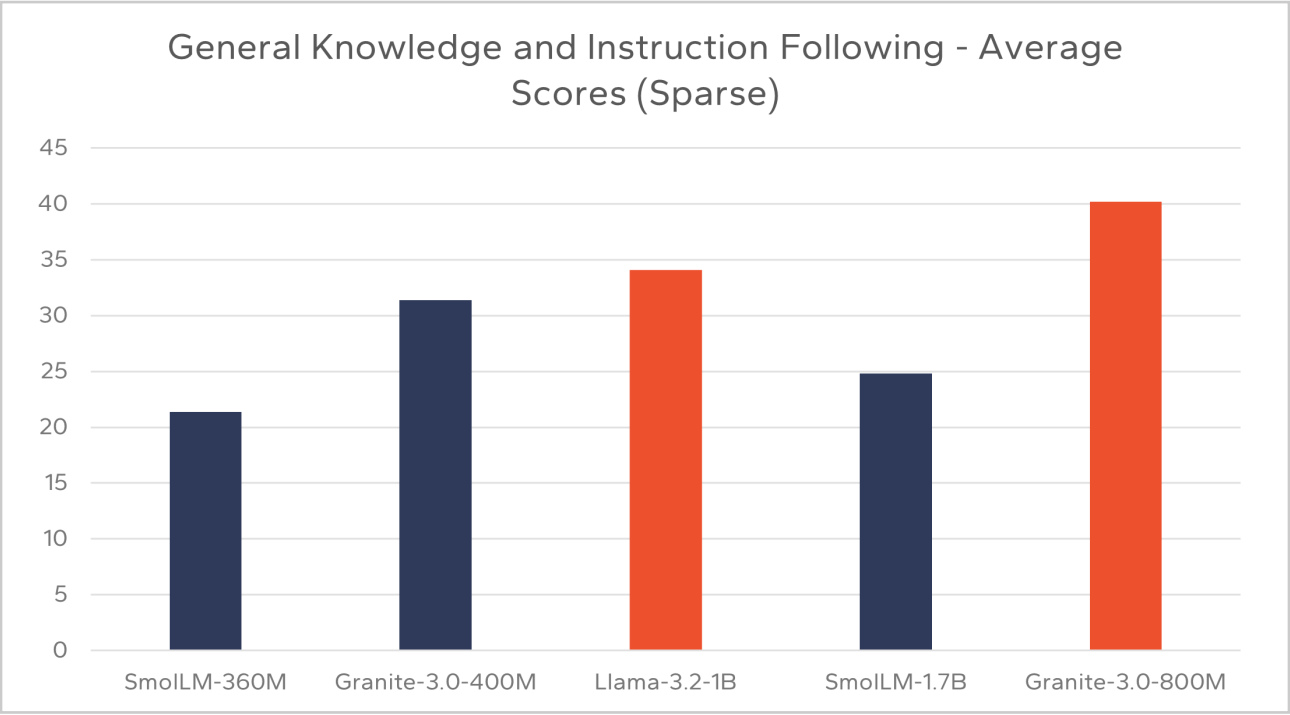


Figure 6: *General Knowledge and Instruction (Sparse Models)*

Function Calling

In many cases, enterprise AI use cases may require models to leverage APIs or other external tool calls. This function calling capability of each model tested was evaluated utilizing the following benchmarks:

Category	Benchmarks
Function Calling	• BFCL V2 • ToolAlpaca • Nexus • API Bank • SealTools • API Bench

Figure 7: *Function Calling Benchmarks*

When averaging the scores across the function calling benchmarks, IBM Granite 3.0 models once again led all similar sized competitors, with Granite-3.0-8B achieving the highest overall results.

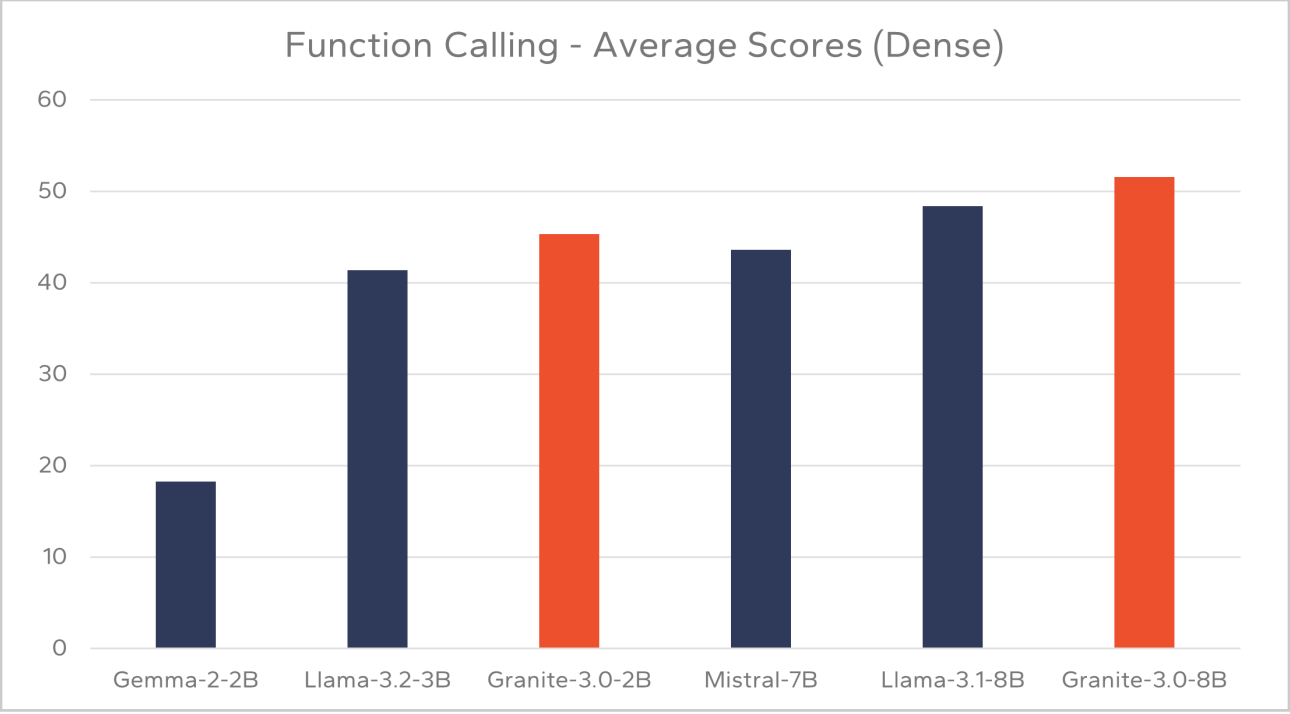


Figure 8: *Function Calling (Dense Models)*

When evaluating the average performance of the sparse Granite-3.0-400M and Granite-3.0-800M models, it should be noted that the SmolLM models were excluded, as results were not achievable from all benchmarks. For the benchmarks that were achievable, however, both sparse Granite 3.0 models outperformed the SmolLM models. It is also notable that both the Granite-3.0-400M and Granite-3.0-800M outperformed the Llama-3.2-1B model across all function calling benchmarks.

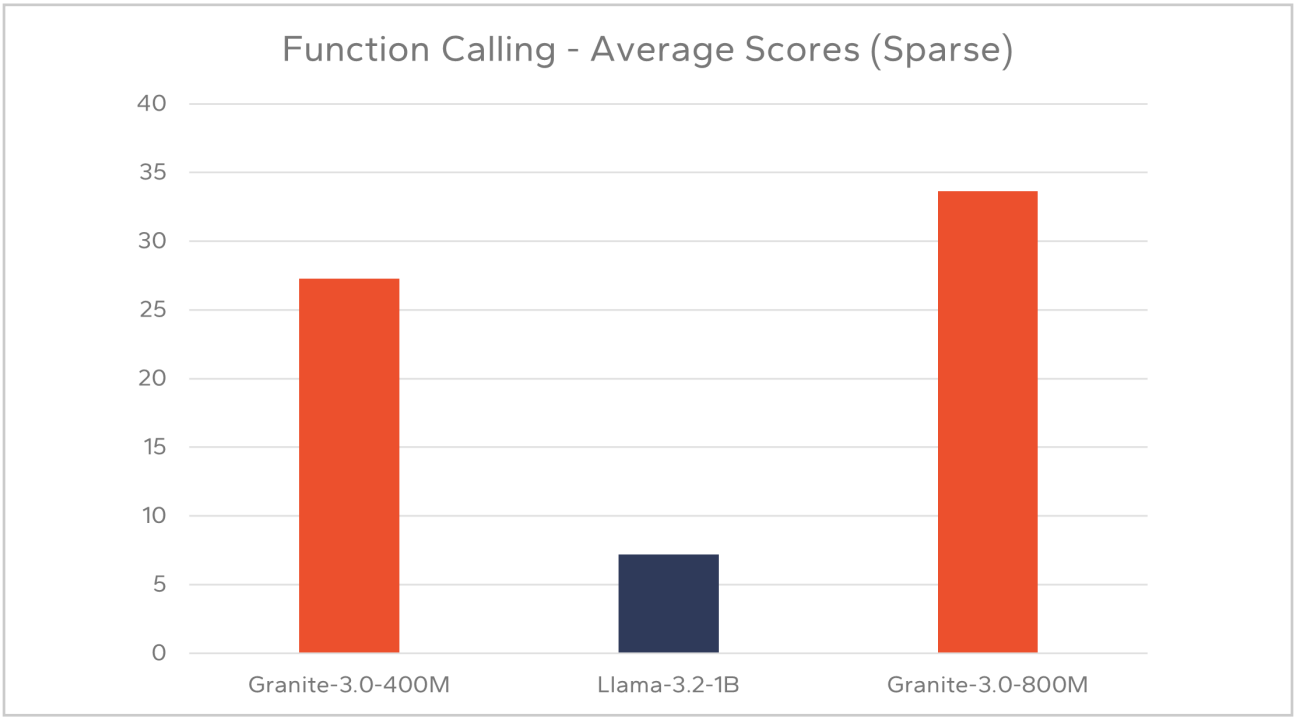
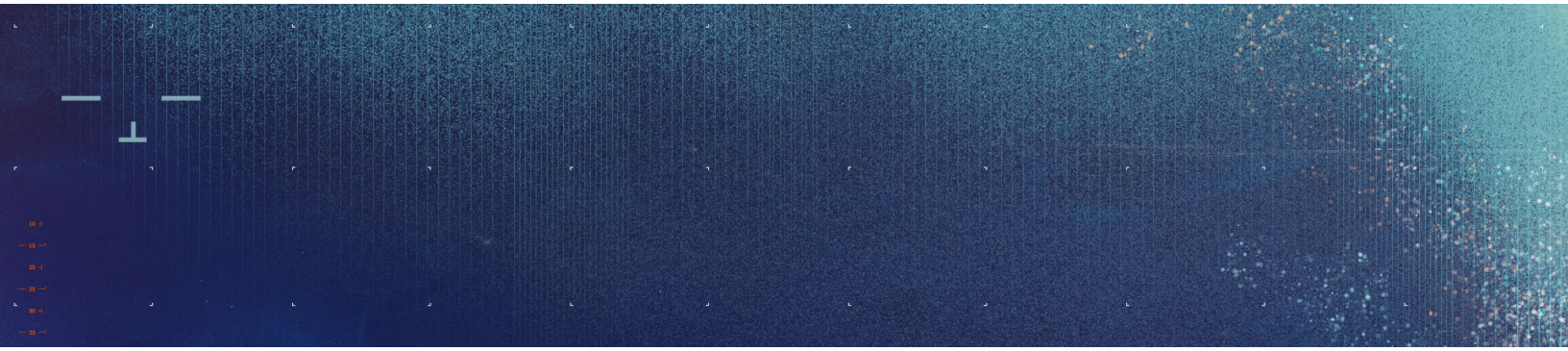


Figure 9: *Function Calling (Sparse Models)*



Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) has become a key methodology for enabling enterprise AI workloads. RAG provides models with additional context, enabling more accurate results and reducing additional training requirements. For enterprise workloads, RAG is a key strategy enabling foundational LLMs to function in industry specific use cases.

Testing of RAG capabilities utilized the RAGBench dataset and RAGAS evaluation framework. Benchmarks used to evaluate the RAG capabilities of Granite 3.0 dense models and competitors include the following:

Category	Benchmarks
RAG	• CovidQA • DelucionQA • Emanual • ExpertQA • HAGRID • HotpotQA • MS Marco • PubMedQA • TAT-QA • TechQA • FinQA

Figure 10: *RAG Benchmarks*

Models were evaluated on Faithfulness and Correctness metrics, using GPT-4 as an LLM judge. Faithfulness uses both the generated answer and given context to measure the factual consistency of an answer with respect to the given context. Correctness is measured as a combination of factuality and semantic similarity by comparing answers to a ground truth response. Both Granite-3.0-2B and Granite-3.0-8B models outperformed similarly sized models, leading in both Faithfulness and Correctness metrics.

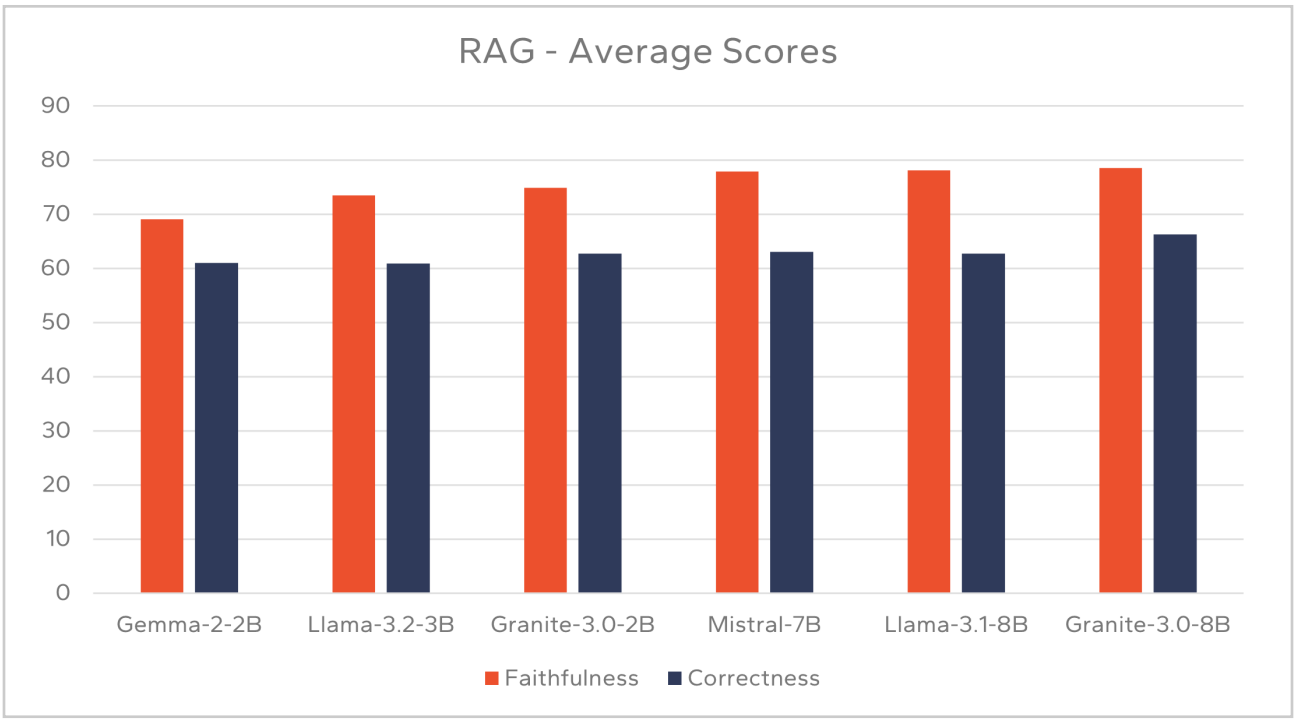


Figure 11: *Average RAG Scores*

Safety and Cybersecurity

Enterprises deploying AI solutions must ensure that their models are both secured from outside threats, as well as safe from providing harmful, inappropriate, or sensitive information. These metrics are crucial to evaluating a model's viability in an enterprise environment, especially when sensitive information may be concerned.

To evaluate cybersecurity capabilities, each model was tested across 15 total tasks, 8 internal IBM benchmarks and 7 publicly available security benchmarks. Benchmarks tested include the following:

Category	Benchmarks
Cybersecurity (Internal)	• Adversarial MITRE ATT&CK • SIEM Rule TTP Mapping • CTI Detection and Mitigation Mapping • CWE Technical Impact Mapping • CTI Relationship Prediction • CTI Entity Classification • MITRE ATTT&CK Entity Classification • CWE Description Summarization
Cybersecurity (Public)	• SecEval • CISSP Assessment Questions • Cybersecurity Skill Assessment • CyberMetric • Cyber Threat Intelligence Multiple Choice Questions • Cyber Threat Intelligence Root Cause Mapping • MMLU Computer Security (SecMMLU)

Figure 12: *Cybersecurity Benchmarks*

The results of the performance benchmarking show that the IBM Granite 3.0 models, at all sizes, outperformed their closest sized competitors. The Granite 3.0 models, both dense and MoE, were found to achieve a higher score on average for both the internal IBM benchmarks as well as the publicly available benchmarks. These results highlight strong cybersecurity characteristics across the entire IBM Granite 3.0 family.

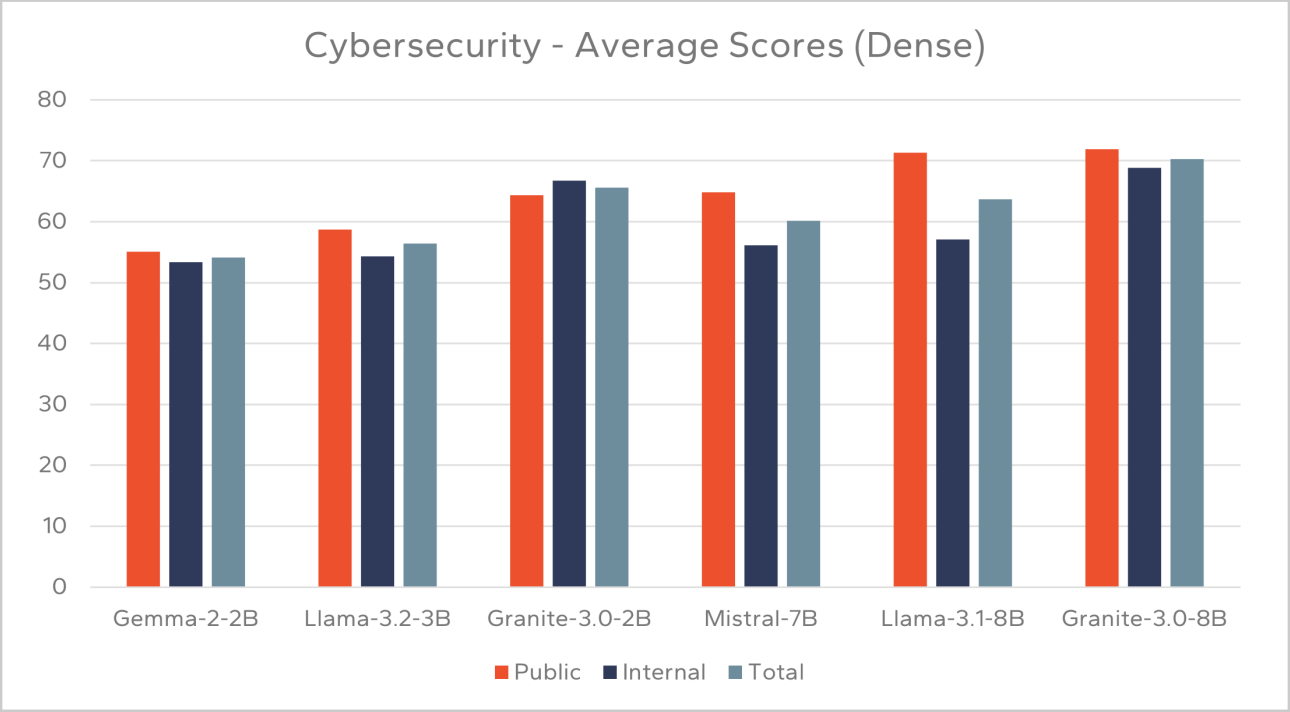


Figure 13: *Cybersecurity Average Scores (Dense Models)*

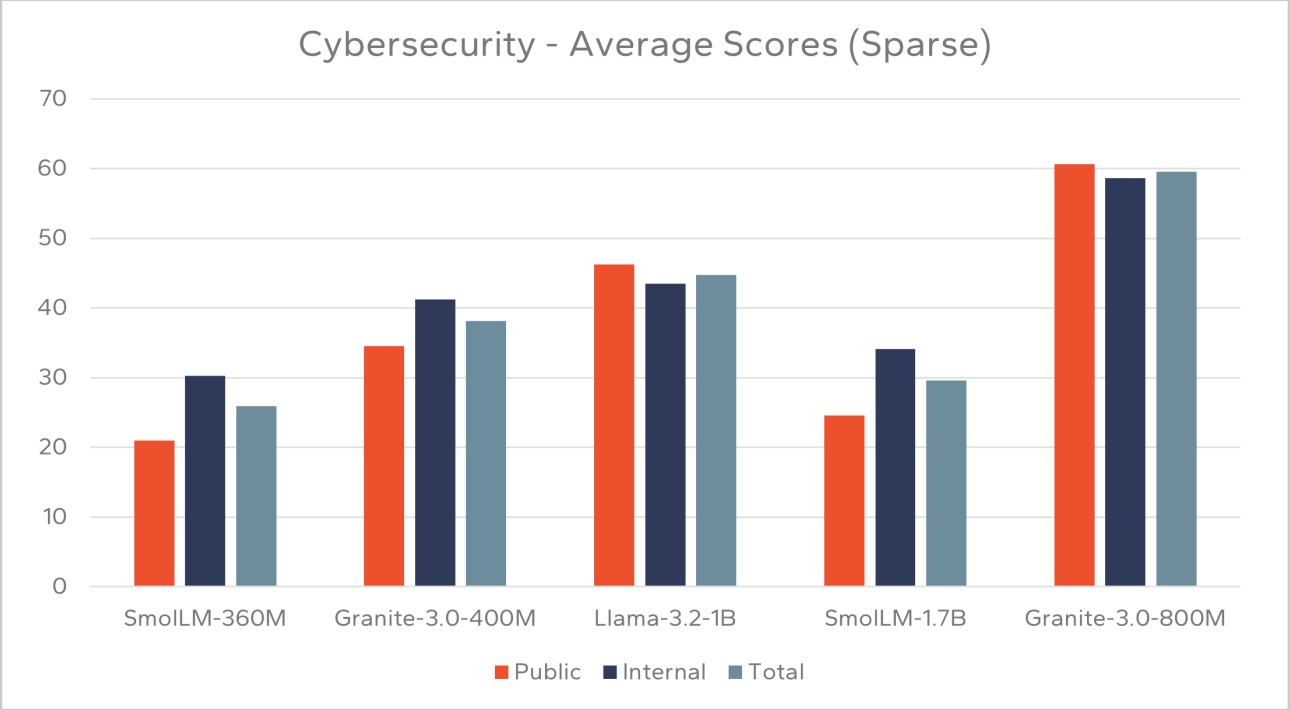


Figure 14: *Cybersecurity Average Scores (Sparse Models)*

Granite 3.0 models and their competitors were additionally tested on an array of benchmarks to measure model safety. These benchmarks measure a model's ability to avoid generation of harmful, inappropriate, or illegal content. Safety benchmarks evaluated include the following:

Category	Benchmarks
Safety	• BOLD • AttaQ • Crows-Pairs • ALERT • SALAD-Bench

Figure 15: *Safety Benchmarks*

The AttaQ benchmark measures a model's resistance to generating content related to 7 categories: Harmful Info, PII, Substance Abuse, Explicit Content, Violence, Discrimination, and Deception. Granite 3.0 models, at all sizes, were found to outperform similar sized models for all 7 categories. In addition to AttaQ, the Granite-3.0-8B model, along with Llama-3.1-8B and Mistral-7B, were tested with the BOLD, Crows-Pairs, ALERT, and SALAD-Bench safety benchmarks. Granite-3.0-8B was found to achieve the best scores for both BOLD and SALAD-Bench benchmarks. While Granite was outperformed in the Crows-Pairs and ALERT benchmarks, the scores were found to be highly competitive with the other two models.

Signal65 Analysis

Signal65 believes the results of IBM's benchmarking showcase that IBM Granite 3.0 models are well suited to address the various needs of enterprise AI workloads. Granite models, at all sizes, were found to consistently excel when compared to similar sized open-source models.

It should be noted that Granite models did not outperform competitive models in all benchmarks. In addition, the competitive models tested do not represent an exhaustive list of all LLMs available in the market. It is possible that other models, especially much larger models, may outperform the Granite models at various tasks. The results, however, do demonstrate leading performance from Granite models as compared to several leading competitors, across a wide range of tasks. Additionally, by leveraging the customizability of IBM Granite models, organizations may be able to further increase performance for their specific workloads, achieving state-of-the-art results while avoiding the high infrastructure costs required of larger, more resource intensive models.

The strong performance of the Granite models becomes especially apparent when evaluating the average scores across several benchmarks. On average, Granite showcased leading performance for general LLM tasks, function calling, RAG, cybersecurity, and safety. The areas tested are key to both general LLM performance, as well as viability in enterprise settings.

The leading performance across the wide range of general knowledge and instruction tasks demonstrate Granite's general usefulness and its broad range of capabilities. The benchmarks in this area included a broad range of tasks including reading comprehension, math, coding, and multi-lingual capabilities. Different organizations, as well as different industries, will require various combinations of these tasks for their AI applications. IBM Granite's leading scores across several of these areas demonstrate its flexibility to meet the needs of various enterprise use cases, spanning from technical to linguistic.

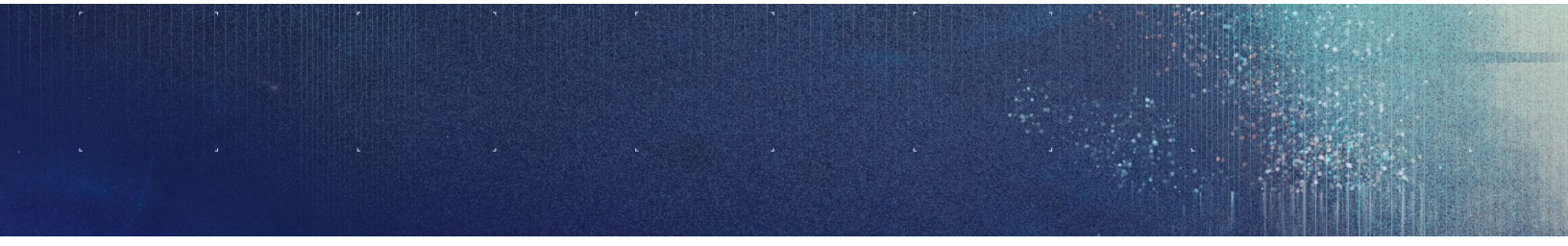
The benchmark scores for both RAG and function calling tasks further demonstrate Granite's ability to meet the needs of enterprise AI development. RAG is crucial to many enterprise AI workloads, helping align the model with industry specific information. Function calling is similarly important, enabling models to integrate into enterprise workflows and complete key business tasks. The high performance achieved by Granite across both RAG and function calling benchmarks demonstrates its ability to be used in highly dynamic, accurate AI solutions for real business use cases.

Perhaps most important to determining an LLMs viability in an enterprise environment are safety and cybersecurity characteristics. LLMs, like all other IT workloads, present a possible vector for attack from malicious actors, and security is a non-negotiable characteristic for enterprise organizations. In addition, the variable nature of AI can present safety concerns in model outputs. Ensuring that models are both safe and secure is critical for enterprise use. IBM Granite showed impressive results in both cybersecurity and safety benchmarks, indicating its enterprise readiness.

Beyond the specific benchmark results, Signal65 also believes that the IBM Granite family of models present several characteristics that may make them well suited for enterprise AI needs. The range of models, both sparse MoE models and the larger dense models, offer significant flexibility to meet different enterprise environments and needs. In general, all of the models within the IBM Granite family are relatively small, when compared to the growing parameter sizes of many other state-of-the-art models. While these larger models may provide powerful results, they require significant infrastructure, introducing additional costs and resources. Despite their relatively small size, each Granite model has demonstrated significant AI capabilities throughout the various benchmarks. This offers enterprise organizations lightweight, manageable AI solutions that can be quickly deployed in a variety of environments, including both the edge and core datacenters. In addition, the customizability of Granite models provides further flexibility for organizations to achieve their specific AI use cases, without introducing additional infrastructure or cost.

Signal65 additionally believes that AI models, such as Granite, developed by IBM should provide enterprise organizations with an added layer of trust. IBM, as a leading technology vendor with a long and reputable history, is aware of enterprise-specific needs and the potential risks involved with new technologies such as AI. IBM previously established its commitment to responsible AI by launching the AI Alliance alongside several other leading technology and research organizations. The AI Alliance is dedicated to the advancement of open, safe, and reliable AI, tenets in which IBM has maintained in its development of Granite models.

IBM Granite models have been designed with a commitment to ethical and responsible AI practices, suitable for enterprise requirements. This can be noted by IBM's attention to data collection and governance practices, as well as the clear focus on safety and security demonstrated from the benchmark results. IBM has additionally developed companion models focused specifically on AI safety and security, known as IBM Granite Guardian, which are designed to detect risk in prompts and responses. When tested, the IBM Granite Guardian models achieved top performance in over 15 safety benchmarks. For enterprise AI, safety and security are critical, and Signal65 believes that organizations should look to deploy models, such as Granite, which heavily prioritize such metrics.



IBM Granite 3.1 and Future Outlook

In addition to the impressive benchmark results published for Granite 3.0 models, IBM has demonstrated their commitment to continuous development of enterprise ready AI models with Granite 3.1. Signal65 has reviewed the Hugging Face Open LLM Leaderboard v1 and v2 results published for both Granite 3.0 and 3.1 models. The Hugging Face Open LLM Leaderboards consist of an average of several benchmarks. While the set of benchmarks included in the two leaderboards are not extensive as the full set of testing done on Granite 3.0 models, they do demonstrate a high level view of model performance. Benchmarks included in the two leaderboards can be seen below:

Category	Benchmarks
Open LLM v1	• ARC-Challenge • Hellaswag • MMLU • TruthfulQA • Winogrande • GSM8K
Open LLM v2	• IFEval • BBH • MATH • GPQA • MUSR • MMLU-Pro

Figure 16: *Open LLM Benchmarks*

The Open LLM Leaderboard v1 results demonstrate minor generation over generation improvements for the Granite 8 B dense models and the Granite 800M MoE models. The Granite 400M MoE models achieved approximately the same score while the Granite-3.0-2B model slightly outperformed the 3.1 version.

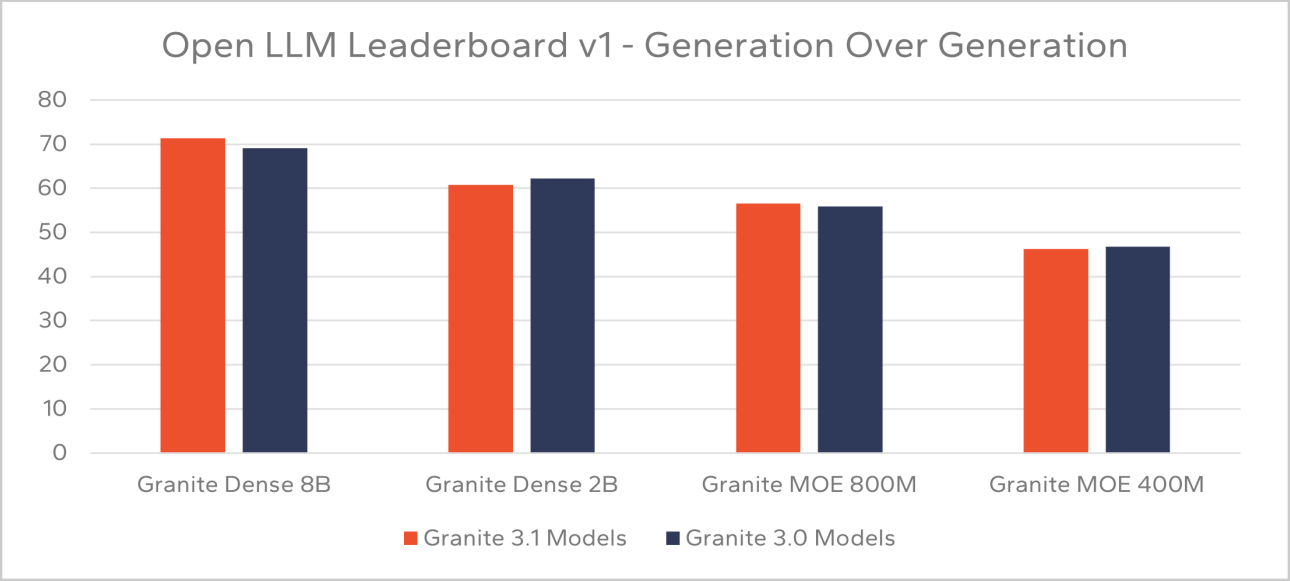


Figure 17: *Open LLM Leaderboard v1 Generation over Generation*

When examining the Open LLM Leaderboard v2 results, the Granite 3.1 models show significant improvement over the previous generation, at all model sizes. This highlights the benefits of continuous model development and the improvements that can be achieved. While the Granite 3.0 models were found to be competitive with other leading models, the advances achieved with Granite 3.1 models demonstrate IBM's ability to further improve the Granite family of models.

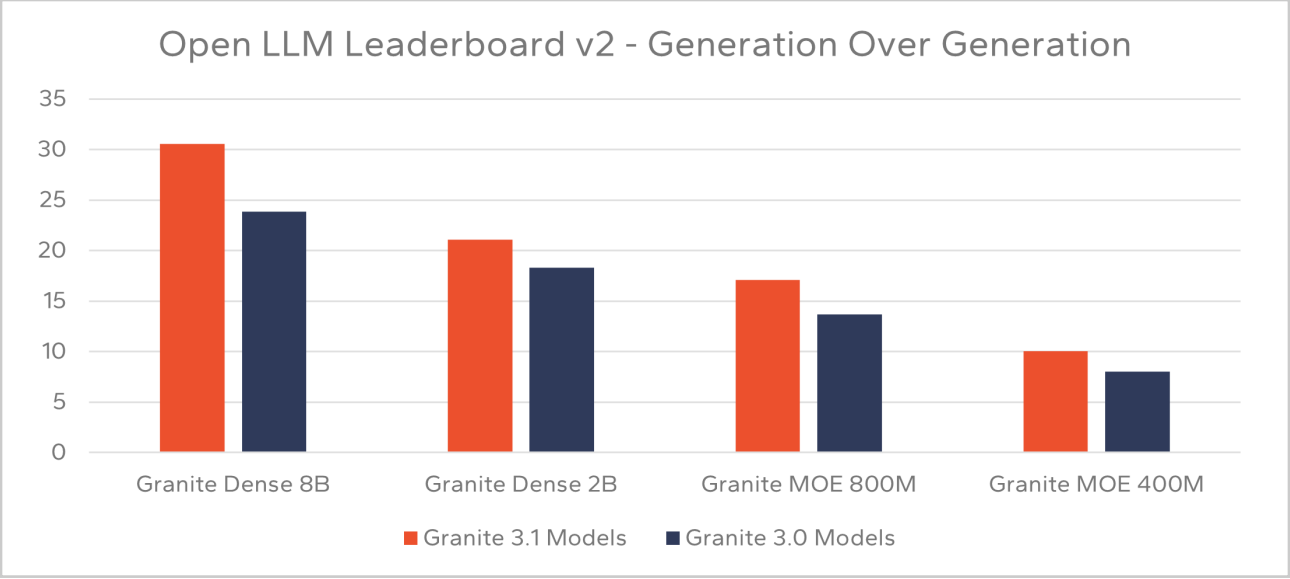


Figure 18: *Open LLM Leaderboard v2 Generation over Generation*

When examining the Open LLM Leaderboard v2 results, the Granite 3.1 models show significant improvement over the previous generation, at all model sizes. This highlights the benefits of continuous model development and the improvements that can be achieved. While the Granite 3.0 models were found to be competitive with other leading models, the advances achieved with Granite 3.1 models demonstrate IBM's ability to further improve the Granite family of models.

Final Thoughts

Generative AI technology holds significant potential to transform and enhance many enterprise workflows. To do so, however, enterprise organizations must evaluate and select AI models that meet their specific criteria. The needs of organizations vary, and evaluating LLMs can be a complex process. Evaluation of various benchmark results can help organizations understand how different AI models may be well suited for their needs.

The array of benchmark results published by IBM, and verified by Signal65, highlight IBM Granite models as highly competitive LLMs that are suitable for enterprise needs. Granite models provide lightweight, flexible AI models that excel in a wide range of tasks, while easily integrating into enterprise workflows, and addressing key licensing, data governance, safety, and security concerns.

While there exists a plethora of new and evolving LLMs available in the market, Signal65 believes that IBM Granite models are uniquely suited for the enterprise and should be considered alongside other leading models. Signal65 has additionally noted the commitment to ongoing AI development from IBM, and believes that IBM Granite models will continue to advance, offering increasingly impressive future performance and evolving to meet the unique requirements of enterprise AI over time.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

IBM

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION
Signal65 I signal65.com