GRANITE 3.0 LANGUAGE MODELS

Granite Team, IBM[1]
[1]See Contributions and Acknowledgments section for full author list.
Please send correspondence to granite-inquiries@ibm.com.

ABSTRACT

This report presents Granite 3.0, a new set of lightweight, state-of-the-art, open foundation models ranging in scale from 400 million to 8 billion active parameters. Equipped with native support of multilingual, coding, function calling, and strong safety performance, these models target enterprise use cases, including on-premise and on-device settings. Evaluations on a comprehensive set of tasks demonstrate that our models consistently reach state-of-the-art performance for their size (as shown in Figure 1 and 2). This report also discloses technical details of pre-training and post-training that may help the research community accelerate the collective efforts to develop open foundation models. We publicly release pre-trained and post-trained versions of all our Granite 3.0 models under a standard permissive Apache 2.0 license allowing both research and commercial use. With support from the open source community, the Granite 3.0 models have been integrated with a range of existing tools for quantization, fine-tuning, and deployment.

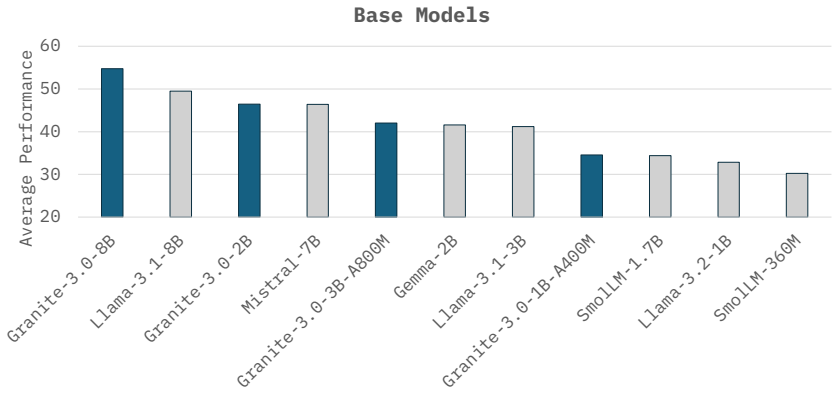


Figure 1: Average performance of base models across 19 tasks from 6 domains.

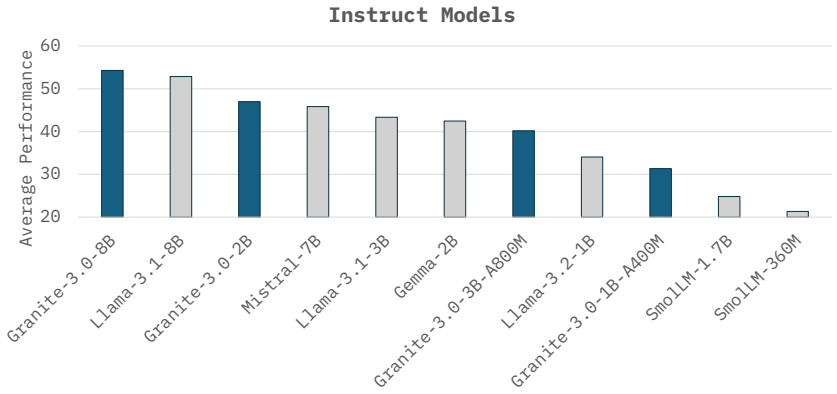
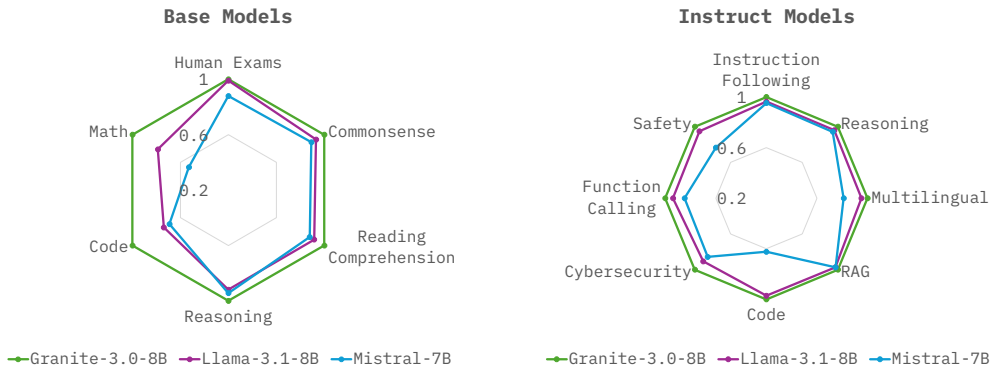


Figure 2: Average performance of instruct models across 23 tasks from 8 domains.



(a) (b)

Figure 3: The relative performance of Granite-3.0-8B and baseline models across different domains. See Table 8 and Table 9 for details of benchmarks included in each category.

1 Introduction

The adoption of large language models (LLMs) across different applications has spread quickly. While commercial options that are consumer-facing via a web interface or API call are widely available, there is a demand for on-premise models. For accessibility, being able to fine-tune a pretrained LLM for on-premise use requires models with lower hardware requirements.

There are many lightweight models like Gemma (Team et al., 2024) and Llama (Dubey et al., 2024) that perform well and fit the bill. However, in an enterprise setting, the adoption of LLMs can have further constraints. The provenance and transparency around data usage and processing can have legal and compliance implications. In particular, the license that an LLM is released under can also restrict companies from using a model on their specific use cases.

In this report, we present the **Granite 3.0** family of language models natively supporting multilinguality, coding, reasoning, and tool usage, 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 procedures were designed for enterprise usage and customization in mind, 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. Specifically, Granite 3.0 includes 4 different models of varying sizes:

- **Dense Models:** 2B and 8B parameter models, trained on 12 trillion tokens in total.
- **Mixture-of-Expert (MoE) Models:** Sparse 1B and 3B MoE models, with 400M and 800M activated parameters respectively, trained on 10 trillion tokens in total.

Accordingly, these models provide a range of options with different compute requirements to choose from, with appropriate trade-offs with their performance on downstream tasks. At each scale, we release a base model — checkpoints of models after pretraining, as well as instruct checkpoints — models finetuned for dialogue, instruction-following, helpfulness, and safety. The base models are trained from scratch with a two-stage training procedure. In stage 1, our dense and MoE models are trained on 10 trillion and 8 trillion tokens, respectively. Stage 1 training data consists of unstructured multilingual language data from diverse sources across academia, the internet, enterprise (e.g., financial, legal), and code, including publicly available datasets with permissive licenses. In stage 2, we train on a mixture of 2 trillion tokens of data. Some of the data sources for stage 2 are the same as the stage 1 data sources, mixed with a small amount of high-quality open-source and synthetic corpora with permissive licenses. The data mixtures are derived through a data mixture search focusing on robustness across different domains and tasks. The instruct models are derived by supervised fine-tuning (SFT) of the pre-trained checkpoints, followed by model alignment using reinforcement learning (PPO, BRAIn (Pandey et al., 2024)). We find that both SFT and PPO/BRAIn are important for improved performance on downstream automatic evaluations, including better chat capabilities.

Additionally, the models were trained with techniques that leverage different methods found in the existing literature: μP (Yang & Hu, 2020; Yang et al., 2022; 2023) allowed for hyperparameter transfer after a hyperparameter search on smaller models, and Power scheduler (Shen et al., 2024c) allowed for learning rate transfer across batch size and total number of training tokens. For our MoE models, we used a dropless MoE (Gale et al., 2023) approach for better model performance using the ScatterMoE (Tan et al., 2024) implementation.

Experiment results show that our Granite 3.0 models outperform models of similar parameter sizes on many benchmarks, demonstrating strong performance in knowledge, reasoning, function calling, multilingual, code support, as well as enterprise tasks like cybersecurity and retrieval augmented generation (RAG). Figure 3 shows that our Granite-3.0-8B models consistently outperform Llama-3.1-8B and Mistral-7B on various domains. The key advantages of Granite 3.0 models are:

- **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.
- **Robust Models with Permissive License:** Combined with excellent performance across various benchmarks, our Granite 3.0 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.
- **Trustworthy Enterprise-Grade LLM**: All our models are trained on license-permissible data collected following IBM's AI Ethics principles[1] 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.

We describe the model architecture and background on MoE models in Section 2. Then, we describe our data collection, filtering, and preprocessing pipeline in Section 3. We then go into detail about our data mixture and hyperparameter search for pretraining in Section 4, followed by our post-training methodology in Section 5, and our compute infrastructure in Section 6. Section 7 describes the results of our comprehensive evaluation of the trained models, including a comparison with other open-source LLMs. Finally, Section 8 discusses the social harms and risks of this project.

2 Model Architecture

The Granite 3.0 language models are based on two architectures: a decoder-only dense transformer and a decoder-only sparse Mixture-of-Expert (MoE) transformer.

Table 1: Hyperparameters for Granite 3.0 models.

Model	2B	8B	1B-A400M	3B-A800M
Embedding size	2048	4096	1024	1536
Number of layers	40	40	24	32
Attention head size	64	128	64	64
Number of attention heads	32	32	16	24
Number of KV heads	8	8	8	8
MLP hidden size	8192	12800	512	512
MLP activation	SwiGLU	SwiGLU	SwiGLU	SwiGLU
Number of Experts	–	–	32	40
MoE TopK	–	–	8	8
Initialization std	0.1	0.1	0.1	0.1
Sequence Length	4096	4096	4096	4096
Position Embedding	RoPE	RoPE	RoPE	RoPE
#Parameters	2.5B	8.1B	1.3B	3.3B
#Active Parameters	2.5B	8.1B	400M	800M
#Training tokens	12T	12T	10T	10T

[1] https://www.ibm.com/impact/ai-ethics

2.1 DENSE MODELS

Granite 3.0 2B and 8B dense models share a similar architecture as popular language models like Llama and our previous Granite Code models Mishra et al. (2024), ensuring strong compatibility with open-source inference and fine-tuning pipelines. We use Grouped Query Attention (GQA; Ainslie et al. 2023) with 8 key-value heads to get a good balance between memory cost and model performance, and Rotary Position Embedding (RoPE; Su et al. 2024) to model the relative position between tokens. For the MLP layers, Granite 3.0 Dense models use SwiGLU as the activation function. Before each MLP and attention layer, we use RMSNorm to normalize the layer's input. We also share parameters between the input embedding and the output linear transform. This reduces the size of the model, and we have observed that the tying of these embeddings have zero, or even a positive impact on model performance.

2.2 MIXTURE-OF-EXPERT MODELS

Granite 3.0 1B and 3B MoE models use similar architecture as Granite Dense models, with the MLP layers substituted with MoE layers. A Mixture of Experts (MoE) layer comprises N modules $f_1, \ldots, f_N$ and a router $g(e \mid \mathbf{x})$. Given an input $\mathbf{x}$ to the MoE layer, the router predicts a probability distribution over the N modules. Of these, we select the top k experts. When $k < N$, we are using a Sparse Mixture of Experts (SMoE; Shazeer et al. 2017). For this series of Granite MoE models, we use a linear layer to model the router:

$$\mathbf{s} = \mathbf{W}_{\text{router}}\mathbf{x}, \tag{1}$$

$$g(e \mid \mathbf{x}) = \begin{cases} \text{softmax}\left(\text{Top}k\left(\mathbf{s}\right)\right)_i, & \mathbf{s}_i \in \text{Top}k\left(\mathbf{s}\right) \\ 0, & \mathbf{s}_i \notin \text{Top}k\left(\mathbf{s}\right) \end{cases} \tag{2}$$

where $\mathbf{W}_{\text{router}}$ is the expert embedding matrix of shape (N, D_{emb}), and Top$k$ is the operator that selects the top k logits from $\mathbf{s}$. The final output of the SMoE is then given by

$$y = \sum_{e=1}^{N} g(e \mid \mathbf{x}) \cdot f_e(\mathbf{x}) \tag{3}$$

When $g(e \mid \mathbf{x}) = 0$, $f_e(\mathbf{x})$ will not need to be evaluated, thus reducing computation cost during training and inference. The key designs of the Granite MoE models are summarized below:

Dropless Token Routing. Since each token selects experts independently, some experts could receive more tokens than others. In previous MoE models, like Switch Transformer (Fedus et al., 2022) and Deepseek-V2 (Liu et al., 2024a), a capacity cap is set for each expert or device, and the extra tokens that exceed the cap are dropped. As observed in Gale et al. (2023), this cap negatively affects the model training stability and loss. In our training, we use ScatterMoE (Tan et al., 2024), a dropless MoE implementation, to avoid token dropping and improve training efficiency.

Fine-grained Experts. Recent studies (Krajewski et al., 2024; Dai et al., 2024) suggest that setting the size of experts in MoE to mirror the feed-forward layer is not optimal. Instead, increasing the expert granularity, number of experts, and number of activated experts could increase the possible combinations of experts and result in better model performance. Following these observations, we use fine-grained experts and a larger number of activated experts in Granite 3.0 MoE models. Specifically, we use a top-k of 8 out of 32 and 40 experts respectively for the 1B and 3B MoE models.

Load Balancing Loss. To avoid routing tokens repeatedly to the same expert and wasting the extra capacity in other experts, we use the frequency-based auxiliary loss introduced in Fedus et al. (2022)

$$\mathcal{L}_b = N \sum_{i=1}^{N} f_i P_i \tag{4}$$

where N is the number of experts, f_i is the fraction of tokens dispatched to expert i, and P_i is the fraction of the router probability allocated for expert i. Intuitively, this loss penalises over-usage of