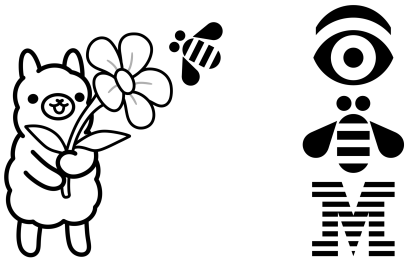
IBM Granite 3.0 models

October 21, 2024



A selection of IBM Granite 3.0 models are now available to run using Ollama. All models are offered under a standard **Apache 2.0 license**.

Performance on par with state-of-the-art open models

2B:

```
ollama run granite3-dense
```

8B:

```
ollama run granite3-dense:8b
```

Granite 2B and Granite 8B are text-only dense LLMs trained on over 12 trillion tokens of data, demonstrated significant improvements over their predecessors in performance and speed in IBM's initial testing. Granite 8B Instruct now rivals Llama 3.1 8B Instruct across both OpenLLM Leaderboard v1 and OpenLLM Leaderboard v2 benchmarks.

They are designed to support tool-based use cases and support for retrieval augmented generation (RAG), streamlining code generation, translation and bug fixing.

Mixture of Expert (MoE) models for low latency

1B:

```
ollama run granite3-moe
```

3B:

```
ollama run granite3-moe:3b
```

The **1B and 3B models** are the first mixture of experts (MoE) Granite models from IBM designed for low latency usage.

The models are trained on over 10 trillion tokens of data, the Granite MoE models are ideal for deployment in on-device applications or situations requiring instantaneous inference.

Capabilities

- Summarization
- Text classification
- Text extraction
- Question-answering
- Retrieval Augmented Generation (RAG)
- Code related
- Function-calling
- Multilingual dialog use cases

Get started

- Granite Dense 2B and 8B models
- Granite Mixture of Expert 1B and 3B models