gemma3

```
ollama run gemma3
```

12.3M Downloads Updated 2 days ago

The current, most capable model that runs on a single GPU.

vision 270m 1b 4b 12b 27b

Models

View all →

Name	Size	Context	Input
gemma3:latest	3.3GB	128K	Text, Image
gemma3:270m	292MB	32K	Text
gemma3:1b	815MB	32K	Text
gemma3:4b latest	3.3GB	128K	Text, Image
gemma3:12b	8.1GB	128K	Text, Image
gemma3:27b	17GB	128K	Text, Image

Readme

> This model requires Ollama 0.6 or later. **Download Ollama**

Gemma is a lightweight, family of models from Google built on Gemini technology. The Gemma 3 models are multimodal—processing text and images—and feature a 128K context window with support for over 140 languages. Available in 270M, 1B, 4B, 12B, and 27B parameter sizes, they excel in tasks like question answering, summarization, and reasoning, while their compact design allows deployment on resource-limited devices.

Models

Text

270M parameter model (32k context window)

```
ollama run gemma3:270m
```

1B parameter model (32k context window)

```
ollama run gemma3:1b
```

Multimodal (Vision)

4B parameter model (128k context window)

```
ollama run gemma3:4b
```

12B parameter model (128k context window)

```
ollama run gemma3:12b
```

27B parameter model (128k context window)

```
ollama run gemma3:27b
```

Quantization aware trained models (QAT)

The quantization aware trained Gemma 3 models preserves similar quality as half precision models (BF16) while maintaining a lower memory footprint (3x less compared to non-quantized models).

1B parameter model

```
ollama run gemma3:1b-it-qat
```

4B parameter model

```
ollama run gemma3:4b-it-qat
```

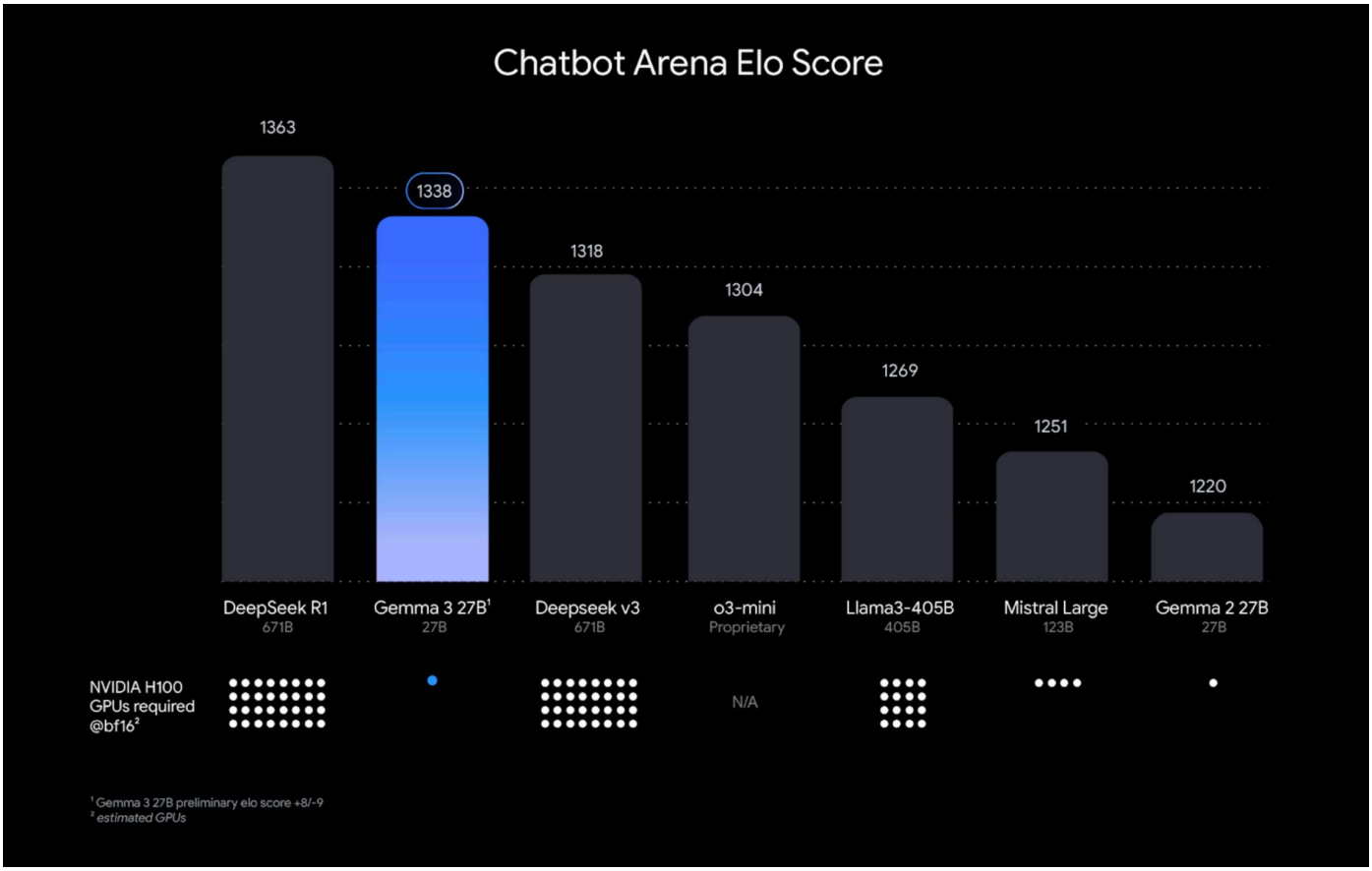
12B parameter model

```
ollama run gemma3:12b-it-qat
```

27B parameter model

```
ollama run gemma3:27b-it-qat
```

Evaluation



Benchmark Results

Gemma 3 270M

Benchmark	n-shot	Gemma 3 270m instruction tuned
HellaSwag	0-shot	37.7
PIQA	0-shot	66.2
ARC-c	0-shot	28.2
WinoGrande	0-shot	52.3
BIG-Bench Hard	few-shot	26.7
IF Eval	0-shot	51.2

These models were evaluated against a large collection of different datasets and metrics to cover different aspects of text generation:

Reasoning, logic and code capabilities

Benchmark	Metric	Gemma 3 PT 1B	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
HellaSwag	10-shot	62.3	77.2	84.2	85.6
BoolQ	0-shot	63.2	72.3	78.8	82.4
PIQA	0-shot	73.8	79.6	81.8	83.3
SocialIQA	0-shot	48.9	51.9	53.4	54.9
TriviaQA	5-shot	39.8	65.8	78.2	85.5
Natural Questions	5-shot	9.48	20.0	31.4	36.1
ARC-c	25-shot	38.4	56.2	68.9	70.6
ARC-e	0-shot	73.0	82.4	88.3	89.0
WinoGrande	5-shot	58.2	64.7	74.3	78.8
BIG-Bench Hard		28.4	50.9	72.6	77.7
DROP	3-shot, F1	42.4	60.1	72.2	77.2
AGIEval	3-5-shot	22.2	42.1	57.4	66.2
MMLU	5-shot, top-1	26.5	59.6	74.5	78.6
MATH	4-shot	–	24.2	43.3	50.0
GSM8K	5-shot, maj@1	1.36	38.4	71.0	82.6
GPQA		9.38	15.0	25.4	24.3
MMLU (Pro)	5-shot	11.2	23.7	40.8	43.9
MBPP	3-shot	9.80	46.0	60.4	65.6
HumanEval	pass@1	6.10	36.0	45.7	48.8
MMLU (Pro COT)	5-shot	9.7	NaN	NaN	NaN

Multilingual capabilities

Benchmark	Gemma 3 PT 1B	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
MGSM	2.04	34.7	64.3	74.3
Global-MMLU-Lite	24.9	57.0	69.4	75.7
Belebele	26.6	59.4	78.0	–
WMT24++ (ChrF)	36.7	48.4	53.9	55.7
FloRes	29.5	39.2	46.0	48.8
XL-Sum	4.82	8.55	12.2	14.9
XQuAD (all)	43.9	68.0	74.5	76.8

Multimodal capabilities

Benchmark	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
COCOcap	102	111	116
DocVQA (val)	72.8	82.3	85.6
InfoVQA (val)	44.1	54.8	59.4
MMMU (pt)	39.2	50.3	56.1
TextVQA (val)	58.9	66.5	68.6
RealWorldQA	45.5	52.2	53.9
ReMI	27.3	38.5	44.8
AI2D	63.2	75.2	79.0
ChartQA	45.4	60.9	63.8
ChartQA (augmented)	81.8	88.5	88.7
VQAv2	–	–	–
BLINK	38.0	35.9	39.6

SpatialSense VQA	50.9	60.0	59.4
CountBenchQA	26.1	17.8	68.0

Reference

Gemma Terms of Use