Gemma 4

Gemma 4

Released **March 31, 2026** by Google DeepMind. Apache 2.0 licensed. Multimodal (text + image, audio on small models).

Model Sizes

Model	Type	Effective Params	Context	Modalities
E2B	Dense	2.3B (5.1B w/ embeddings)	128K	Text, Image, Audio
E4B	Dense	4.5B (8B w/ embeddings)	128K	Text, Image, Audio
26B A4B	MoE	3.8B active / 25.2B total	256K	Text, Image
31B	Dense	30.7B	256K	Text, Image

The **26B A4B** is the standout — a MoE model that runs almost as fast as a 4B model despite 26B total params. The **E2B/E4B** use Per-Layer Embeddings for on-device efficiency.

Local Running Options

1. **Ollama** — `ollama run google/gemma-4` (all sizes). Easiest one-command setup.
2. **llama.cpp** — GGUF quantized versions available on Hugging Face. Good for CPU/GPU hybrid inference.
3. **vLLM** — For higher-throughput server deployment. Supports the native HF safetensor weights.
4. **LM Studio** — GUI-based, supports GGUF formats. Good for desktop use.
5. **Hugging Face Transformers** — Direct Python API. Full precision or QLoRA fine-tuning.

Hardware Requirements (rough)

- **E2B (2.3B eff.)** — Runs on phones, any modern laptop (4-8 GB RAM)
- **E4B (4.5B eff.)** — 8-16 GB RAM, most 2024+ MacBooks
- **26B A4B** — 16-24 GB VRAM (single GPU), or CPU with enough RAM
- **31B** — 24-48 GB VRAM (A100/H100 recommended), or multi-GPU

The **26B A4B** is generally considered the sweet spot for local use — frontier-level benchmarks (82.6 MMLU Pro, 88.3 AIME) with ~4B active parameter compute cost.

All models are on Hugging Face under `google/gemma-4-*`.

Revision #1
Created 2026-04-28 22:18:14 UTC by Clive
Updated 2026-07-12 21:42:37 UTC by Clive