Google Gemma 4: Apache 2.0 Open Models with 256K Context That Run on Your Phone

Kabui, Charles

2026-04-08

Read at ToKnow.ai

Google released Gemma 4, a family of four open models under a fully permissive Apache 2.0 license. The lineup spans a 31B dense model, a 26B mixture-of-experts (MoE) that activates only 3.8B of its 25.2B parameters per forward pass, and two edge models: E4B at 4.5B effective parameters and E2B at 2.3B effective. The 31B scores 89.2% on AIME 2026 (a math competition benchmark), 85.2% on MMLU Pro, and reaches 2150 Codeforces ELO. The MoE

variant hits 82.3% on GPQA Diamond (a graduate-level science benchmark) while running nearly as fast as a 4B model. All models support up to 256K tokens of context (128K for edge variants), native function calling for agent workflows, and a configurable reasoning mode. The E2B and E4B add native audio input and process images at variable resolutions, all on-device with no internet required.

Previous Gemma releases had restrictions that complicated commercial use. Now anyone can build proprietary products, modify weights, and redistribute without limitations. The MoE model's 3.8B active parameters deliver close to 31B quality at a fraction of the compute cost. Edge models handle vision, audio, and agentic tasks on a phone offline. The 31B instruction-tuned variant alone crossed 1.1 million Hugging Face downloads within its first week.

Google now treats fully open licensing as competitive advantage rather than risk. With Meta's LLaMA, Alibaba's Qwen, and Gemma all under Apache 2.0 or equivalent terms, restrictive model licenses are becoming a liability. The race has shifted from who builds the best closed model to who builds the best open one. For a similar MoE efficiency approach, see NVIDIA's Nemotron Cascade 2.

Sources:

- Google Blog: Gemma 4 Launch
- Gemma 4 Model Card
- Gemma 4 on Hugging Face
- Google AI Studio: Gemma 4
