Gemma 4 Just Dropped: What Local AI Builders Need to Know

April 2, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Gemma 4 is Google's strongest open model yet, now under Apache 2.0. Four variants: E2B and E4B for edge devices (with audio input), 26B-A4B MoE (3.8B active, ~17GB at Q4, 150 tok/s on RTX 4090), and 31B dense (~20GB at Q4, #3 open model on LMArena at ELO 1452). All support vision and video. The 26B-A4B MoE is the standout for local users -- competitive with Qwen 3.5 35B-A3B at similar VRAM. The Apache 2.0 license switch removes the biggest reason people avoided Gemma for production. Run it today: `ollama run gemma4:26b`.

More on this topic: Gemma Models Guide | Qwen 3.5 Local Guide | VRAM Requirements | GPU Buying Guide

Google just shipped Gemma 4, and two things matter more than the benchmarks: it's Apache 2.0, and it does vision, video, and audio in a single model that fits on consumer hardware.

Gemma 3 had a restrictive license that scared off anyone building commercial products. Qwen and Llama ate its lunch. Gemma 4 fixes that with a clean Apache 2.0 license – no custom clauses, no "Harmful Use" carve-outs, no legal overhead. That alone makes this worth paying attention to.

The model quality is strong too. The 31B dense variant ranks #3 among open models on LMArena (ELO 1452), and the 26B-A4B MoE punches at nearly the same level (#6, ELO 1441) while running at 150 tok/s on an RTX 4090. Here's what you need to know to run it locally.

The lineup

Gemma 4 ships in four sizes, all multimodal:

Model	Total Params	Active Params	Context	Modalities	Ollama Size
E2B	5.1B	2.3B	128K	Text + Image + Video + Audio	7.2 GB
E4B	8B	4.5B	128K	Text + Image + Video + Audio	9.6 GB

Model	Total Params	Active Params	Context	Modalities	Ollama Size
26B-A4B	26B	3.8B (MoE)	256K	Text + Image + Video	18 GB
31B	31B	31B (Dense)	256K	Text + Image + Video	20 GB

The naming is a bit confusing. E2B and E4B are “edge” models – the numbers reflect effective compute, not raw parameters. The 26B-A4B is a Mixture-of-Experts model with 128 experts where 8 are active per token, giving you big-model quality at small-model inference cost.

All four models support image and video input out of the box. The E2B and E4B edge models also handle audio input (speech recognition and audio scene understanding), with audio for the larger models coming later.

Architecture highlights: alternating sliding-window and full-context attention layers, Per-Layer Embeddings (PLE) for better representation, and shared KV cache where later layers reuse K/V from earlier ones to save VRAM.

Benchmarks: the actual numbers

Reasoning and knowledge

Benchmark	E2B	E4B	26B-A4B	31B	Gemma 3 27B
MMLU Pro	60.0%	69.4%	82.6%	85.2%	67.6%
AIME 2026	37.5%	42.5%	88.3%	89.2%	20.8%
GPQA Diamond	43.4%	58.6%	82.3%	84.3%	42.4%
MMMLU	67.4%	76.6%	86.3%	88.4%	70.7%

The jump from Gemma 3 to Gemma 4 is massive. AIME goes from 20.8% to 89.2% on the 31B. That’s not an incremental improvement – it’s a generational leap.

Coding

Benchmark	E4B	26B-A4B	31B	Gemma 3 27B
LiveCodeBench v6	52.0%	77.1%	80.0%	29.1%
Codeforces ELO	940	1718	2150	110

The 31B's Codeforces ELO of 2150 puts it in competitive programmer territory. The 26B-A4B at 1718 is strong too, especially considering it only uses 3.8B active params per token.

Vision

Benchmark	E4B	26B-A4B	31B	Gemma 3 27B
MMMU Pro	52.6%	73.8%	76.9%	49.7%
MATH-Vision	59.5%	82.4%	85.6%	46.0%

Multimodal isn't an afterthought. Gemma 4 handles document parsing, chart understanding, OCR, GUI element detection, and generates HTML from screenshots. The vision token budget is configurable (70 to 1120 tokens per image).

LMArena rankings

- **31B:** ELO 1452, #3 open model globally
- **26B-A4B:** ELO 1441, #6 open model globally

For context, the Chinese models (Qwen 3.5, GLM-5, Kimi K2.5) still hold the top spots, but not by much.

VRAM requirements

Model	Q4 VRAM	Q8 VRAM	BF16 VRAM	Best GPU Fit
E2B	~4 GB	5-8 GB	10 GB	Any 8GB GPU, Apple Silicon 8GB
E4B	5-6 GB	9-12 GB	16 GB	RTX 3060 12GB, M-series 8-16GB
26B-A4B	16-18 GB	28-30 GB	52 GB	RTX 4090, RTX 5060 Ti 16GB (tight), M-series 32GB+
31B	17-20 GB	34-38 GB	62 GB	RTX 3090/4090, M-series 32GB+

The 26B-A4B at Q4_K_M is about 17 GB on disk. That's a squeeze on 16GB cards (RTX 5060 Ti, RTX 4070 Ti Super) but fits comfortably on 24GB cards. On a Mac with 32GB unified memory, it runs fine.

The E4B is the 8GB GPU play. At Q4, it needs about 6 GB – leaving room for context on a 12GB RTX 3060 or an 8GB M-series Mac.

GGUF quant sizes (26B-A4B, Unsloth Dynamic)

Quant	File Size
UD-Q2_K_XL	10.5 GB
UD-Q3_K_M	12.5 GB
UD-Q4_K_M	16.9 GB
UD-Q5_K_M	21.2 GB
Q8_0	26.9 GB
BF16	50.5 GB

How to run it

Ollama (easiest)

Day-0 support. Pull and run:

```
ollama run gemma4          # Default: E4B
ollama run gemma4:e2b      # Edge 2B
ollama run gemma4:26b      # MoE 26B-A4B
ollama run gemma4:31b      # Dense 31B
```

LM Studio

Available at launch. All four variants in GGUF format through the model browser. Search “gemma-4” and pick your quant level.

llama.cpp

Day-0 support with multimodal. Download the GGUF and the vision projector file:

```
./llama-cli -m gemma-4-26b-a4b-it-Q4_K_M.gguf \
  --mmproj gemma-4-26b-a4b-it-mmproj.gguf \
  -p "Describe this image" --image photo.jpg
```

Note: some chat template issues were reported at launch (PR #21326). Check for updates if you hit formatting problems.

Unsloth GGUFs

Already on HuggingFace with Dynamic quantization (UD-) for all variants. These use per-layer variable precision for better quality at the same file size. Look for `unsloth/gemma-4-26B-A4B-it-GGUF` and similar.

Gemma 4 vs Qwen 3.5: should you switch?

This is the question everyone's asking. Qwen 3.5 has been the local AI default for weeks. Does Gemma 4 change that?

26B-A4B MoE vs Qwen 3.5 35B-A3B MoE

	Gemma 4 26B-A4B	**Qwen 3.5 35B-A3B**
Active params	3.8B	3B
VRAM (Q4)	~17 GB	~17 GB
LMArena ELO	1441	Higher (ranked above)
Context	256K	131K
Vision	Yes	Yes
Audio	No (coming)	No
License	Apache 2.0	Apache 2.0

Similar VRAM, similar speed. Qwen 3.5 still edges ahead on community coding tests and multilingual (201 languages, 250K token vocabulary). Gemma 4 wins on context length (256K vs 131K) and has stronger vision benchmarks. The community consensus: competitive, but Qwen 3.5 still leads on coding quality and overall Arena rankings.

31B dense vs Qwen 3.5 27B dense

Both fit on a 24GB card at Q4. Gemma 4 31B scores higher on AIME (89.2% vs Qwen's strong but lower numbers) and GPQA Diamond. Qwen 3.5 27B wins on SWE-bench coding (72.4%) and multilingual tasks. Extended testing shows Qwen producing "more architecturally sound solutions" in complex coding scenarios.

E4B vs Qwen 3.5 9B at 8GB

The E4B has fewer effective parameters (4.5B vs 9B) but adds audio and video input that Qwen doesn't have. For pure text quality at the 8GB tier, Qwen 3.5 9B likely wins. For multimodal use cases on constrained hardware, Gemma 4 E4B is more capable.

The Apache 2.0 switch

This might matter more than the benchmarks.

Gemma 1 through 3 used Google's custom "Gemma Terms of Use" license. It had vague "Harmful Use" restrictions, limits on redistribution, and Google could update the terms whenever they wanted. Legal teams at companies considering Gemma for products would look at that license, look at Qwen's Apache 2.0, and pick Qwen.

Gemma 4 under Apache 2.0 removes all of that. No custom clauses. Full commercial use. Modify, redistribute, deploy however you want. VentureBeat called the license change "the most consequential commercial signal in the launch."

For hobbyists, this doesn't change much – you were running Gemma anyway. For anyone building a product, this puts Gemma 4 back on the table.

Who should care

If you're happy with Qwen 3.5: Don't switch yet. Community benchmarks are still rolling in, and Qwen 3.5 remains ahead on coding and multilingual. Monitor the next two weeks of testing.

If you need multimodal on consumer hardware: Gemma 4 E4B with vision and audio on 8GB VRAM is genuinely new. No other open model at this size does text + image + video + audio.

If you avoided Gemma because of the license: Apache 2.0 changes everything. Gemma 4 is now a first-class option for commercial deployment.

If you have 16-24GB VRAM: The 26B-A4B MoE at 150 tok/s on an RTX 4090 is fast. At 17GB Q4, it fits on most 24GB cards with room for context. Worth testing alongside Qwen 3.5 35B-A3B on your specific tasks.

If you're on edge hardware: The E2B runs on a Raspberry Pi 5 at ~3 tok/s. The E4B hits 54 tok/s on an M4 Pro. These are real options for embedded and mobile AI.

Bottom line

Gemma 4 is Google finally getting serious about open models. The quality is competitive with Qwen 3.5, the multimodal story is stronger, the license is fixed, and framework support was ready at launch.

It's not a clear winner over Qwen 3.5 – the Chinese models still lead on Arena rankings and coding. But it's close enough that the choice now comes down to your specific use case, not "Qwen wins by default."

The 26B-A4B MoE is the model to try first. It fits on a 24GB card, runs fast, handles vision, and has 256K context. Start there:

```
ollama run gemma4:26b
```

Related Guides

- Gemma Models Guide (Gemma 1-3)
- Qwen 3.5 Local AI Guide
- How Much VRAM Do You Need?
- GPU Buying Guide
- Quantization Explained

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/gemma-4-local-ai-guide/

Free guides for running AI locally