Gemma Models Guide: Google's Lightweight Local LLMs

February 8, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Gemma 4 is the current generation (April 2026) and it's a massive leap. The 31B dense model scores 80% on LiveCodeBench and 89% on AIME — turning Gemma from a structured-output specialist into a genuine all-rounder. The 26B-A4B MoE variant activates only 4B parameters per token, so it runs fast on modest hardware while punching way above its weight. Edge models (E2B, E4B) handle vision and audio. Biggest practical change: Gemma 4 ships under Apache 2.0 — no more ambiguous custom license. For new setups, skip everything before Gemma 4.

📚 **Related:** Qwen Models Guide · Llama 3 Guide · VRAM Requirements · Best LLMs for Chat

Google has a reputation problem with open models. They release things, rename them, deprecate them, and release something else. Keeping track of the Gemma lineup requires more effort than it should.

Here's the short version: **Gemma 4 is the current generation** (released April 2, 2026). It doesn't just iterate on Gemma 3 — it rewrites the competitive picture. The 31B model scores 80% on LiveCodeBench and 89% on AIME, turning Gemma from "good at structured output, mediocre at everything else" into a genuine contender across the board. And it ships under **Apache 2.0** — no more custom license headaches.

If you're starting fresh, skip Gemma 1, 2, and 3. This guide covers what's worth running today.

The Gemma 4 Lineup

Gemma 4 splits into two tiers: **Edge** models (E2B, E4B) for lightweight and mobile use, and **Workstation** models (26B-A4B MoE, 31B dense) for serious local inference.

Model	Total Params	Active Params	Type	Context	Best For
Gemma 4 E2B	5.1B	2.3B	Dense (edge)	128K	Phones, Raspberry Pi, background tasks

Model	Total Params	Active Params	Type	Context	Best For
Gemma 4 E4B	8B	4.5B	Dense (edge)	128K	8GB cards, fast daily driver
Gemma 4 26B-A4B	~25B	~4B	MoE (workstation)	256K	16-24GB cards, best efficiency
Gemma 4 31B	31B	31B	Dense (workstation)	256K	24GB cards, maximum quality

All models include instruction-tuned variants. For local use, grab the `-it` versions.

What changed from Gemma 3: Everything. Context jumps to 256K on workstation models. The MoE architecture means the 26B model only activates ~4B parameters per token — so it runs at small-model speeds with large-model quality. Benchmarks aren't even close (more on that below).

How to Run

```
# Via Ollama (available day one)
ollama run gemma4          # defaults to E4B
ollama run gemma4:e2b
ollama run gemma4:e4b
ollama run gemma4:26b      # 26B-A4B MoE
ollama run gemma4:31b
```

Models are available on HuggingFace in safetensors and GGUF formats. Both llama.cpp and vLLM support Gemma 4 natively from launch.

```
# Via llama.cpp
llama-server -hf ggml-org/gemma-4-E2B-it-GGUF
```

What Each Size Gets You

Gemma 4 E2B: The Edge Model

The smallest useful model. 5.1B total parameters but only 2.3B active — designed for phones, edge devices, and background services.

Runs on: Anything with 4GB+ RAM. Raspberry Pi 5, old laptops, phones.

Good for: Classification, basic Q&A, simple summarization. Surprisingly capable for its size — 60% on MMLU-Pro and 44% on LiveCodeBench. It also handles vision and audio natively.

Skip if: You have an 8GB+ GPU. The E4B is substantially better and still lightweight.

Gemma 4 E4B: The Daily Driver

The sweet spot for 8GB cards. 8B total parameters, 4.5B active. This replaces Gemma 3 4B as the go-to lightweight model, and the jump is dramatic.

Runs on: Any 8GB GPU comfortably at Q4. Even some 6GB cards with aggressive quantization.

Benchmarks that matter:

Benchmark	Gemma 4 E4B	Gemma 3 4B	What It Measures
MMLU-Pro	69.4%	~52%	Academic knowledge
LiveCodeBench v6	52.0%	~15%	Real-world coding
AIME 2026	42.5%	~8%	Math competition
GPQA Diamond	58.6%	~30%	Graduate-level science

Good for: General assistant tasks, coding help, summarization, vision tasks (screenshots, diagrams), and audio processing. A legitimate daily driver on budget hardware.

The upgrade from Gemma 3 4B: Night and day. Gemma 3 4B was good at structured output but weak at reasoning and coding. Gemma 4 E4B is competent across the board.

Gemma 4 26B-A4B: The Efficiency King

This is the model that makes Gemma 4 special. 25B total parameters organized as a Mixture-of-Experts, but only ~4B parameters activate per token. You get large-model quality at small-model speeds and VRAM costs.

VRAM: ~16-18GB at Q4. Fits on a 24GB card with room for long context. Tight on 16GB cards but possible.

Key benchmarks:

Benchmark	26B-A4B	Gemma 3 27B	Improvement
MMLU-Pro	82.6%	67.5%	+15 points
LiveCodeBench v6	77.1%	29.7%	+47 points
AIME 2026	88.3%	20.8%	+68 points
GPQA Diamond	82.3%	42.4%	+40 points
LMArena	~1441	~1200	Massive jump

Good for: Complex reasoning, coding, analysis, agentic workflows with native function calling. The 256K context handles entire codebases or long documents in a single pass. Video understanding is supported on this tier.

The MoE advantage: Because only ~4B parameters are active per token, inference is fast. You get benchmark scores that compete with the 31B dense model at a fraction of the compute cost. If you have a 24GB card, this is the model to default to.

Gemma 4 31B: Maximum Quality

The dense flagship. All 31B parameters active on every token — slower than the MoE variant but marginally better on the hardest benchmarks.

VRAM: ~18-20GB at Q4. Fits on 24GB with some room for context. Q8 needs ~34-38GB (multi-GPU territory).

Key benchmarks:

Benchmark	Gemma 4 31B	26B-A4B	What It Measures
MMLU-Pro	85.2%	82.6%	Academic knowledge
AIME 2026	89.2%	88.3%	Math competition
LiveCodeBench v6	80.0%	77.1%	Real-world coding
GPQA Diamond	84.3%	82.3%	Graduate-level science
Codeforces ELO	2150	1718	Competitive programming
tau2-bench (agentic)	86.4%	—	Tool use / agents

Good for: When you need the absolute best quality Gemma can offer. Complex multi-step reasoning, competitive programming, agentic tasks with tool use.

vs the 26B-A4B: The 31B is 2-3 points better on most benchmarks but noticeably slower because all parameters are active. For most tasks, the 26B-A4B gives you 95% of the quality at much better speed. The 31B is worth it for competitive programming and the hardest reasoning tasks.

What's New in Gemma 4

Apache 2.0 License

The biggest practical change. Previous Gemma models used Google's custom "Gemma Terms of Use" — technically allowing commercial use but with ambiguous restrictions that made lawyers nervous. Gemma 4 ships under **Apache 2.0**, the same license as Qwen and Mistral.

What this means for you:

- **Commercial use:** No restrictions, no ambiguity
- **Redistribution:** Standard open-source terms
- **Fine-tuning:** Train and deploy without worrying about Google's acceptable use policy
- **No more license comparison headaches:** Gemma is now on equal legal footing with the competition

If the old Gemma license was the reason you picked Qwen or Llama instead, that reason is gone.

Native Function Calling

Gemma 4 is trained for multi-turn agentic workflows with native function calling. Previous Gemma models could follow instructions well but weren't designed for tool use. The 31B scores 86.4% on tau2-bench (agentic task completion) — this is a model you can build agent pipelines around.

Vision, Audio, and Video

All Gemma 4 models support vision (image understanding). The split:

- **E2B & E4B (edge):** Image + text + audio
- **26B-A4B & 31B (workstation):** Image + text + video

The vision encoder handles variable aspect ratios with configurable token budgets (70 to 1,120 tokens per image), so you can balance detail vs speed. Document OCR, screenshot analysis, diagram understanding — all built in.

Architecture Improvements

Under the hood: Per-Layer Embeddings (PLE) feed a second embedding signal into every decoder layer. Shared KV cache lets later layers reuse earlier layers' key/value states, cutting memory usage. Dual RoPE handles both sliding-window attention (local context) and global attention (full context) efficiently.

The practical result: better quality per parameter and lower memory overhead at long context lengths.

Where Gemma 4 Shines

Coding (Finally)

Gemma 3's coding story was "use something else." Gemma 4 flips that. LiveCodeBench jumps from 29.7% to 80.0% on the 31B, and the 26B-A4B hits 77.1%. The Codeforces ELO of 2150 on the 31B puts it in competitive territory. You no longer need a separate coding model if you're running Gemma 4.

Reasoning and Math

AIME scores went from 20.8% (Gemma 3 27B) to 89.2% (Gemma 4 31B). This isn't incremental improvement — it's a generational leap. Graduate-level science (GPQA Diamond) nearly doubled. Gemma 4 can handle complex multi-step reasoning that Gemma 3 simply couldn't.

Instruction Following (Still)

Gemma's traditional strength carries forward. When you say "output JSON with these exact fields" or "respond in exactly three bullet points," Gemma 4 still complies more consistently than most open models. The difference is that now it can also think while following your format.

Agentic Workflows

Native function calling plus 256K context plus strong reasoning makes Gemma 4 viable for agent pipelines. The 31B's 86.4% on tau2-bench means it can reliably use tools, maintain conversation state, and execute multi-step plans.

Multimodal Everything

Vision across all sizes. Audio on edge models. Video on workstation models. Upload a screenshot, ask about a diagram, process a voice note, analyze a video clip — depending on which model you're running. No separate models needed for most multimodal tasks.

Where Gemma 4 Still Struggles

Creative Writing

The structured-output DNA that makes Gemma excellent at following instructions still produces relatively dry creative prose. It's better than Gemma 3, but if you need personality in fiction, poetry, or marketing copy, Llama and Qwen still produce more engaging output.

Multilingual

Gemma's training data still skews English. MMMLU scores are solid (88.4% on the 31B), but Qwen's multilingual breadth across 29 languages with strong coverage remains unmatched. If non-English is your primary use case, Qwen is still the better choice.

Long-Context Retrieval

The 128K needle-in-haystack scores (MRCR v2) show some weakness: 66.4% for the 31B, 44.1% for the 26B-A4B. The context window is large, but retrieval accuracy at extreme lengths isn't perfect. For RAG-style tasks with very long documents, chunk strategically rather than dumping everything in.

---

vs The Competition

Gemma 4 E4B vs Llama 3.2 3B vs Phi-4 Mini 3.8B

Aspect	Gemma 4 E4B	Llama 3.2 3B	Phi-4 Mini 3.8B
VRAM (Q4)	~5-6 GB	~2.5 GB	~2.5 GB
Context	128K	128K	128K
Coding	Strong (52% LCB)	Below average	Good
Math/reasoning	Strong (42.5% AIME)	Below average	Good
Vision	Built-in + audio	Built-in	No
License	Apache 2.0	Meta license	MIT

Pick Gemma 4 E4B for the best overall capability at the 8GB tier, especially if you need vision or audio. **Pick Phi-4 Mini** if VRAM is extremely tight. **Pick Llama 3.2 3B** for natural conversation tone.

Gemma 4 26B-A4B vs Qwen 2.5 14B vs Llama 3.1 8B

Aspect	Gemma 4 26B-A4B	Qwen 2.5 14B	Llama 3.1 8B
VRAM (Q4)	~16-18 GB	~9 GB	~5 GB
Active params	~4B	14B	8B
Context	256K	128K	128K
Coding	Strong (77% LCB)	Strong	Moderate
Reasoning	Excellent (88% AIME)	Good	Moderate
Multilingual	Moderate	Best	Moderate
Speed (tok/s)	Fast (MoE)	Moderate	Fast

Pick Gemma 4 26B-A4B if you have the VRAM — the MoE efficiency means it runs fast despite the large total parameter count, and it dominates on reasoning and coding. **Pick Qwen 2.5 14B** for multilingual work or if VRAM is limited to 12-16GB. **Pick Llama 3.1 8B** if you're on 8GB.

Gemma 4 31B vs Qwen 2.5 32B

Aspect	Gemma 4 31B	Qwen 2.5 32B
VRAM (Q4)	~18-20 GB	~20 GB
Context	256K	128K
Coding (LCB)	80.0%	~65%
Math (AIME)	89.2%	~55%
Structured output	Excellent	Good
Multilingual	Moderate	Strong
License	Apache 2.0	Apache 2.0

The calculus has changed. On a 24GB card, Gemma 4 31B now beats Qwen 2.5 32B on most benchmarks — especially coding and reasoning. Qwen still wins for multilingual tasks, but Gemma 4 is the stronger general-purpose model in this size class. This is a reversal from Gemma 3, where Qwen was the clear overall winner.

VRAM Requirements

Gemma 4 Models

Model	Q4_K_M	Q8_0	BF16
Gemma 4 E2B	~4 GB	~5-8 GB	~10 GB
Gemma 4 E4B	~5-6 GB	~9-12 GB	~16 GB
Gemma 4 26B-A4B	~16-18 GB	~28-30 GB	~52 GB
Gemma 4 31B	~18-20 GB	~34-38 GB	~62 GB

Recommended GPU pairings:

Your GPU	Best Gemma 4 Model	Quantization
4GB (GTX 1650, RX 6500)	Gemma 4 E2B	Q4
8GB (RTX 3070, 4060)	Gemma 4 E4B	Q4 (comfortable)

Your GPU	Best Gemma 4 Model	Quantization
12GB (RTX 3060, 4060 Ti)	Gemma 4 E4B	Q8 (best quality at this tier)
16GB (RTX 4060 Ti 16GB, Arc A770)	Gemma 4 26B-A4B	Q4 (tight but works)
24GB (RTX 3090, 4090)	Gemma 4 26B-A4B or 31B	Q4 with room for 256K context

→ Use our Planning Tool to check exact VRAM for your setup.

For full VRAM breakdowns across all models, see the VRAM requirements guide.

What About Gemma 3? (Previous Generation)

Gemma 3 was the generation that put Gemma on the map — the 27B beat Gemini 1.5 Pro on benchmarks and the 4B outperformed Gemma 2 27B. But Gemma 4 is so far ahead that Gemma 3 is now only worth running if you've already fine-tuned a Gemma 3 model and don't want to redo the training.

Gemma 3 Quick Reference (Legacy)

Model	VRAM (Q4)	Context	Status
Gemma 3 1B	~1 GB	32K	Replaced by Gemma 4 E2B
Gemma 3 4B	~3 GB	128K	Replaced by Gemma 4 E4B
Gemma 3 12B	~8 GB	128K	Replaced by Gemma 4 26B-A4B (MoE)
Gemma 3 27B	~16 GB	128K	Replaced by Gemma 4 31B / 26B-A4B

```
# Gemma 3 is still available on Ollama
ollama run gemma3:4b
ollama run gemma3:12b
ollama run gemma3:27b
```

Gemma 2 and Gemma 1: Completely obsolete. No reason to run either.

The License Situation (Resolved)

Gemma 4 ships under **Apache 2.0** — the same permissive open-source license used by Qwen, Mistral, and most of the open-weight ecosystem.

Previous Gemma models used Google's custom "Gemma Terms of Use" which technically allowed commercial use but had restrictions around acceptable use policy that made deployment decisions murky. That ambiguity is gone.

What Apache 2.0 means:

- Commercial use with no restrictions
- Redistribution, modification, fine-tuning — all standard open-source terms
- No need to comply with Google's acceptable use policy
- Same legal clarity as Qwen 2.5 or Mistral

If you avoided Gemma in the past because of the license, that concern no longer applies.

Note: Gemma 3 and earlier still use the old custom license. If you're running legacy Gemma models commercially, the original terms still apply.

Recommendations

Tightest budget (4-8GB VRAM): Gemma 4 E4B. At ~5-6GB for Q4, it fits comfortably on 8GB cards and delivers remarkably strong coding and reasoning for its size. Built-in vision and audio are bonuses you won't find on competing models at this tier.

Mid-range (12-16GB VRAM): This is where it gets interesting. On 12GB, run E4B at Q8 for best quality. On 16GB, the 26B-A4B MoE fits at Q4 — and because only ~4B parameters are active per token, it runs fast while delivering workstation-class benchmarks. Test both and see which fits your workflow.

High-end (24GB VRAM): Gemma 4 26B-A4B as your daily driver, with the 31B available for the hardest tasks. The MoE model gives you the best speed-to-quality ratio at this VRAM tier. The 31B is 2-3 points better on benchmarks but noticeably slower — use it when you need peak quality on complex reasoning or competitive programming.

Edge/embedded: Gemma 4 E2B for devices with limited resources. Vision and audio support built in.

Coming from Gemma 3: Upgrade. The benchmark improvements are too large to ignore — AIME went from 20.8% to 89.2%, LiveCodeBench from 29.7% to 80.0%. Unless you have a fine-tuned Gemma 3 model, switch to Gemma 4.

📚 **Model comparisons:** Qwen Models Guide · Llama 3 Guide · Mistral & Mixtral Guide · DeepSeek Guide

📚 **Hardware pairing:** VRAM Requirements · 8GB VRAM Guide · 12GB VRAM Guide · GPU Buying Guide

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/gemma-models-guide/

Free guides for running AI locally