Gemma Models Guide: Google's Lightweight Local LLMs

February 8, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Gemma 3 is the current generation — skip Gemma 2 for new setups. The sweet spots: Gemma 3 4B for 8GB cards (outperforms Gemma 2 27B on benchmarks), Gemma 3 12B for 12-16GB cards, and Gemma 3 27B for 24GB cards where it competes with models twice its size. All variants support 128K context. Gemma excels at instruction following, summarization, and structured output but loses to Qwen 2.5 for coding and creative writing. Google's license allows commercial use but has restrictions worth reading before deploying in production.

📚 **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 3 is the current generation** (released March 2025). It replaces Gemma 2 across the board. The 4B model outperforms the previous-gen 27B. The 27B model beats Google's own Gemini 1.5 Pro on benchmarks while running on a single consumer GPU.

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

The Gemma 3 Lineup

Model	Parameters	VRAM (Q4)	VRAM (Q8)	Context	Best For
Gemma 3 1B	1B	~1 GB	~1.5 GB	32K	Edge devices, Raspberry Pi
Gemma 3 4B	4B	~3 GB	~5 GB	128K	8GB cards, best tiny model
Gemma 3 12B	12B	~8 GB	~14 GB	128K	12-16GB cards, daily driver
Gemma 3 27B	27B	~16 GB	~28 GB	128K	24GB cards, best quality

All Gemma 3 models are available in base and instruction-tuned (-IT) variants. For local AI use, you almost always want the instruction-tuned version.

The big upgrade from Gemma 2: 128K context on all models except 1B (32K). Gemma 2 was stuck at 8K, which made it impractical for document analysis or long conversations. That limitation is gone.

How to Run

```
# Via Ollama
ollama run gemma3:4b
ollama run gemma3:12b
ollama run gemma3:27b

# Specific quantization
ollama run gemma3:27b-q4_K_M
```

Models are available on HuggingFace in safetensors format and on Ollama in GGUF. Both llama.cpp and vLLM support Gemma 3 natively.

What Each Size Gets You

Gemma 3 1B: The Edge Model

The smallest model in the lineup. Google added a 270M variant in August 2025 for task-specific fine-tuning, but for general use, 1B is the floor.

Runs on: Literally anything. Raspberry Pi 5, old laptops, phones. Under 1.5GB even at Q8.

Good for: Simple classification, basic Q&A, running as a background service with minimal resource usage. Not good for open-ended conversation or anything requiring reasoning.

Skip if: You have any GPU with 4GB+ VRAM. The 4B model is dramatically better and still lightweight.

Gemma 3 4B: The Surprise Performer

This is the model that turned heads. Gemma 3 4B-IT outperforms Gemma 2 27B-IT on key benchmarks — a model 7x smaller beating its predecessor's flagship. Google's architectural improvements (likely inherited from Gemini 2.0 research) pack more capability into fewer parameters than anyone expected.

Runs on: Any 8GB GPU comfortably. Even 4GB cards at aggressive quantization.

VRAM: ~3GB at Q4, leaving room for other processes on an 8GB card.

Good for: Summarization, instruction following, structured output, basic reasoning. A legitimate daily driver if you have 8GB or less.

Benchmark context: Competes with Llama 3.1 8B and Mistral 7B at half the parameter count. Faster inference because less data to move through the pipeline.

Gemma 3 12B: The Sweet Spot

If you have a 12GB card (RTX 3060, 4060) or 16GB card, the 12B is the model to default to. It fits comfortably at Q4 on 12GB with room for context, and at Q6 on 16GB.

VRAM: ~8GB at Q4. Comfortable on 12GB cards with overhead for KV cache.

Good for: General assistant tasks, document analysis (128K context fits substantial documents), summarization, structured data extraction.

vs Qwen 2.5 14B: Qwen wins on coding and multilingual tasks. Gemma 12B is faster and better at following specific formatting instructions. If you need JSON output or structured responses, Gemma tends to be more reliable.

Gemma 3 27B: Punching Up

The flagship runs on a single 24GB GPU at Q4 and beats Gemini 1.5 Pro on Google's own benchmarks. On Chatbot Arena, it ranks alongside models that need 4-8x more hardware.

VRAM: ~16GB at Q4 (fits on 24GB with room for long context). ~28GB at Q8 (needs multi-GPU or quantization trade-off).

Key benchmarks:

Benchmark	Gemma 3 27B	What It Measures
MMLU-Pro	67.5	Academic knowledge
MATH	69.0	Mathematical reasoning
LiveCodeBench	29.7	Real-world coding
GPQA Diamond	42.4	Graduate-level science
FACTS Grounding	74.9	Factual accuracy

Good for: Complex reasoning, analysis, detailed instruction following. The best Gemma model for tasks where quality matters more than speed.

vs Qwen 2.5 32B: Qwen wins on most benchmarks, especially coding and multilingual. Gemma 27B is 5B parameters smaller (faster) and better at structured output compliance. If you're choosing between them on a 24GB card, Qwen 2.5 32B is the stronger general model — but Gemma 27B is worth testing if you value response format consistency.

Where Gemma Shines

Instruction Following

Gemma's strongest trait. When you say "output JSON with these exact fields" or "respond in exactly three bullet points," Gemma complies more consistently than most open models. This makes it excellent for:

- **Structured data extraction:** "Parse this email and return sender, date, subject, action items as JSON"
- **Template-based generation:** Fill-in responses that match a specific format
- **API backends:** Where consistent response structure matters more than creative flair

Summarization

Google trained Gemma with strong compression abilities. It produces concise, accurate summaries without the padding that plagues some models. Combined with 128K context, Gemma 3 can summarize entire documents in a single pass.

Speed

Gemma's architecture is optimized for inference efficiency. At the same parameter count, Gemma typically generates tokens faster than Llama or Qwen. The 4B model is particularly fast — on an RTX 3060, expect 60-80 tok/s at Q4, fast enough that responses feel instant.

Multimodal (Vision)

Gemma 3 4B and larger include native vision capabilities — they can process images alongside text. Upload a screenshot, diagram, or photo and ask questions about it. No separate vision model needed.

PaliGemma 2 is the dedicated vision model (3B, 10B, 28B) for tasks like OCR, object detection, and image captioning. If vision is your primary use case, PaliGemma offers more precision. If you occasionally need image understanding alongside text, Gemma 3's built-in vision is convenient.

Where Gemma Struggles

Creative Writing

Gemma produces technically correct but often dry prose. It follows instructions precisely — which means it writes exactly what you ask for, but without the unexpected turns or stylistic personality that make creative writing interesting. For fiction, poetry, or marketing copy, Llama 3 and Qwen 2.5 produce more engaging output.

Coding

LiveCodeBench score of 29.7 for the 27B tells the story. It can write basic functions and explain code, but for serious development work — complex algorithms, debugging, multi-file refactoring — dedicated coding models like DeepSeek Coder and Qwen 2.5 Coder are substantially better.

CodeGemma exists (7B, based on Gemma 1) but hasn't been updated to the Gemma 3 architecture and falls behind current coding models. Skip it.

Multilingual

Gemma's training data skews English. It handles major European languages adequately but can't match Qwen's multilingual breadth (29 languages with strong coverage) or NLLB's translation-specific capabilities. If multilingual is important, Qwen 2.5 is the better choice.

---

vs The Competition

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

Aspect	Gemma 3 4B	Llama 3.2 3B	Phi-4 Mini 3.8B
VRAM (Q4)	~3 GB	~2.5 GB	~2.5 GB
Context	128K	128K	128K

Aspect	Gemma 3 4B	Llama 3.2 3B	Phi-4 Mini 3.8B
Instruction following	Best	Good	Good
Math/reasoning	Good	Below average	Best
Creative writing	Dry	Better	Dry
Vision	Built-in	Built-in	No

Pick Gemma 3 4B for structured tasks and vision. **Pick Phi-4 Mini** for math and reasoning. **Pick Llama 3.2 3B** for natural conversation.

Gemma 3 12B vs Qwen 2.5 14B vs Llama 3.1 8B

Aspect	Gemma 3 12B	Qwen 2.5 14B	Llama 3.1 8B
VRAM (Q4)	~8 GB	~9 GB	~5 GB
Context	128K	128K	128K
Coding	Moderate	Strong	Moderate
Instruction following	Best	Good	Good
Multilingual	Weak	Best	Moderate
Speed (same hardware)	Fastest	Slower	Fast

Pick Gemma 3 12B for structured output and speed. **Pick Qwen 2.5 14B** if it fits your VRAM and you need coding or multilingual. **Pick Llama 3.1 8B** if VRAM is tight.

Gemma 3 27B vs Qwen 2.5 32B

Aspect	Gemma 3 27B	Qwen 2.5 32B
VRAM (Q4)	~16 GB	~20 GB
Context	128K	128K
Overall benchmarks	Very good	Better
Coding	Moderate	Strong
Structured output	Best	Good
Speed (24GB card)	Faster (more headroom)	Slower (tighter fit)

Pick Gemma 3 27B if structured output and speed matter, or if you want more VRAM headroom for long context. **Pick Qwen 2.5 32B** for the strongest overall capabilities, especially coding.

The honest take: on a 24GB card, Qwen 2.5 32B is the better general-purpose model. Gemma 27B is the better specialist for tasks that need format compliance and don't need strong coding.

VRAM Requirements

Model	Q4_K_M	Q6_K	Q8_0	FP16
Gemma 3 1B	~1 GB	~1 GB	~1.5 GB	~2 GB
Gemma 3 4B	~3 GB	~4 GB	~5 GB	~8 GB
Gemma 3 12B	~8 GB	~10 GB	~14 GB	~24 GB
Gemma 3 27B	~16 GB	~20 GB	~28 GB	~54 GB

Recommended GPU pairings:

Your GPU	Best Gemma Model	Quantization
4GB (GTX 1650, RX 6500)	Gemma 3 4B	Q4 (tight)
8GB (RTX 3070, 4060)	Gemma 3 4B	Q8 (best quality)
12GB (RTX 3060, 4060 Ti)	Gemma 3 12B	Q4
16GB (RTX 4060 Ti 16GB)	Gemma 3 12B	Q6 or Gemma 3 27B at Q4 (tight)
24GB (RTX 3090, 4090)	Gemma 3 27B	Q4 with room for 128K 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 2?

Skip it for new setups. Gemma 3 is better at every size and adds 128K context (vs 8K). The only reason to run Gemma 2 is if you've already fine-tuned a Gemma 2 model for your specific use case and don't want to redo the training.

If you see benchmarks comparing "Gemma" to other models without specifying the generation, check whether they mean Gemma 2 or Gemma 3 — the difference is substantial.

Gemma 2 Quick Reference (Legacy)

Model	VRAM (Q4)	Context	Notes
Gemma 2 2B	~2 GB	8K	Replaced by Gemma 3 4B
Gemma 2 9B	~6 GB	8K	Replaced by Gemma 3 12B
Gemma 2 27B	~16 GB	8K	Replaced by Gemma 3 27B

Gemma 1: Completely obsolete. No reason to run any Gemma 1 model.

The License Situation

Gemma uses Google's Gemma Terms of Use — not Apache 2.0, not MIT, not a standard open-source license.

What's allowed:

- Commercial use (with conditions)
- Redistribution
- Fine-tuning and modification
- Research and academic use

What's restricted:

- Must comply with Google's Acceptable Use Policy
- Prohibited use for certain applications (weapons, surveillance, etc.)
- Some developers consider the terms ambiguous for commercial deployment

The practical reality: For hobbyist and personal use, the license is fine. For commercial products, read the full terms carefully and consider consulting a lawyer — the restrictions are broader than Llama's license and less clear-cut than Apache 2.0.

If license clarity matters for your use case, Qwen 2.5 (Apache 2.0) or Llama 3 (Meta's permissive license) have more straightforward terms.

Recommendations

Tightest budget (4-8GB VRAM): Gemma 3 4B. Legitimately useful at 3GB, faster than alternatives at the same quality tier, and the built-in vision is a bonus. On 8GB, run it at Q8 for the best quality this size class offers.

Mid-range (12-16GB VRAM): Gemma 3 12B for structured tasks and speed. But honestly, if your card fits Qwen 2.5 14B (~9GB at Q4), test both — Qwen is stronger overall for general use.

High-end (24GB VRAM): Start with Qwen 2.5 32B as your default, and keep Gemma 3 27B available for tasks where format compliance matters. Having both loaded in Ollama and switching based on the task is the power-user move.

Edge/embedded: Gemma 3 1B or the 270M fine-tuned variant for classification and simple extraction tasks.
