Llama 4 vs Qwen3 vs DeepSeek V3.2: Which to Run Locally in 2026

February 16, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For most local AI users in 2026, Qwen3 is the default. Widest range of sizes (0.6B to 235B), best quality per VRAM, Apache 2.0 license, and a /think toggle for reasoning. DeepSeek R1 distills win for pure reasoning — the 14B at 9GB and the 32B at 20GB produce chain-of-thought depth that's hard to match. GPT-OSS 20B is the dark horse at 16GB — OpenAI's first open model, Apache 2.0, strong at agentic coding. Llama 4 Scout is impressive but needs 55GB minimum. If you have one GPU under 24GB, start with Qwen3.

📚 **More on this topic:** Llama 4 Guide · Qwen3 Guide · DeepSeek V3.2 Guide · VRAM Requirements

Three model families are dominating local AI in 2026: Meta's Llama 4, Alibaba's Qwen3, and DeepSeek's V3.2/R1. Each has real strengths. Each has real limitations. And the answer to "which should I run?" depends almost entirely on how much VRAM you have.

The flagship models — Llama 4 Maverick (400B), DeepSeek V3.2 (685B), Qwen3-235B — are datacenter territory. You're not running any of them on a consumer GPU. But the models you can run locally? That's where the real competition is, and the answer isn't obvious.

The Contenders

Llama 4 (Meta) — MoE architecture with native multimodal. Scout has 109B total params but only 17B active per token. Native 10M token context window. The catch: even Scout needs ~55GB at Q4, putting it out of reach for most single-GPU setups.

Qwen3 (Alibaba) — The widest model range in any family. Eight dense sizes from 0.6B to 32B, two MoE models, and a `/think` toggle that switches between chain-of-thought reasoning and fast chat. Qwen3-4B rivals Qwen 2.5-72B on benchmarks. Apache 2.0 licensed.

DeepSeek V3.2/R1 (DeepSeek) — The flagship V3.2 competes with GPT-5 but needs 350GB+. The real story for local users is the R1-Distill lineup: dense reasoning models distilled from the

full R1, running on consumer hardware with chain-of-thought quality that punches way above their parameter count. MIT licensed.

GPT-OSS 20B (OpenAI) — The dark horse. OpenAI's first open-weight model, released August 2025. MoE with 21B total / 3.6B active, ships in MXFP4 at ~13GB. Apache 2.0 licensed. Doesn't get enough attention.

Head-to-Head Comparison

These are the models budget builders actually choose between:

	Llama 4 Scout	Qwen3-14B	Qwen3-30B-A3B	DeepSeek R1-14B	GPT-OSS 20B
Total params	109B	14B	30B	14B	21B
Active params	17B	14B	3B	14B	3.6B
Architecture	MoE (16 experts)	Dense	MoE (128 experts)	Dense (distilled)	MoE (32 experts)
VRAM (Q4)	~55GB	~9GB	~18GB*	~9GB	~13GB
Context	10M tokens	32K	32K	128K	131K
Multimodal	Yes (native)	No	No	No	No
License	Llama 4 Community	Apache 2.0	Apache 2.0	MIT	Apache 2.0
Best for	Vision + long docs	General purpose	Budget MoE	Reasoning	Agentic coding
Ollama	`llama4:scout`	`qwen3:14b`	`qwen3:30b-a3b`	`deepseek-r1:14b`	`gpt-oss:20b`

*Qwen3-30B-A3B needs ~18GB at Q4_K_M because MoE models load all expert weights. With Unsloth's 1.78-bit quant, it squeezes into ~8GB.

The table tells the story: Llama 4 Scout is the outlier at 55GB. Everything else fits on consumer hardware. GPT-OSS 20B packs 131K context and strong coding into 13GB — the kind of value you'd expect from someone trying to disrupt their own API business.

Winner by Hardware Tier

4GB VRAM (What can you run?)

Winner: Qwen3-4B — no contest.

Nothing else competitive fits. Qwen3-4B runs at Q4_K_M in ~3GB, leaving headroom for context. It scores 97.0 on MATH-500 in `/think` mode — a 4B model matching what took 72B parameters one generation ago. Llama 4 doesn't fit. DeepSeek R1-7B needs ~5GB. Qwen3 owns this tier.

```
ollama run qwen3:4b
```

8GB VRAM (What can you run?)

Winner: Qwen3-8B for general use. Qwen3-30B-A3B (Unsloth quant) for MoE magic.

Qwen3-8B at Q4_K_M uses ~6GB, runs at 20-30 tok/s, and handles chat, coding, and reasoning well. The 30B-A3B with Unsloth's 1.78-bit quant squeezes into 8GB and outperforms QwQ-32B on Arena-Hard (91.0 vs 89.5) despite activating only 3B parameters per token. The quality jump from 30B of expert knowledge is real — but you're running at aggressive quantization.

DeepSeek R1-7B (~5GB at Q4) is solid here for reasoning-focused work.

```
ollama run qwen3:8b          # Safe pick — 6GB
ollama run deepseek-r1:7b    # Reasoning focus — 5GB
```

12GB VRAM (What can you run?)

Winner: Qwen3-14B for general use. DeepSeek R1-14B for reasoning.

Both are 14B dense models at ~9GB. Both are excellent. The difference is specialization:

- **Qwen3-14B** — Better instruction following, better structured output, strong tool calling. The `/think` toggle means one model for both quick chat and deep reasoning. Scores 85.5 on Arena-Hard.

- **DeepSeek R1-14B** — Purpose-built reasoning from R1 distillation. Scores 69.7% on AIME 2024 and 93.9% on MATH-500. Produces deep, methodical chain-of-thought reasoning chains. When you need step-by-step problem solving, the depth of its thinking is hard to beat.

Qwen3-14B in `/think` mode actually matches or edges ahead on some reasoning benchmarks (76.3% AIME'24). But R1-14B's reasoning chains tend to be more thorough — fewer hallucinated logic steps on complex multi-hop problems. Pick based on whether you want a versatile generalist or a reasoning specialist.

```
ollama run qwen3:14b          # General purpose — sweet spot
ollama run deepseek-r1:14b    # Reasoning specialist
```

16GB VRAM (What can you run?)

Winner: GPT-OSS 20B for agentic coding. Qwen3-14B at Q6 for quality.

The dark horse tier. GPT-OSS 20B ships in MXFP4 at ~13GB, delivers fast inference with only 3.6B active params, and scores 60.7% on SWE-Bench Verified. OpenAI designed it for agentic workflows — strong instruction following, tool use, and 131K context. Apache 2.0 licensed. At 16GB, nothing else combines coding strength, context length, and speed like this.

Alternatively, Qwen3-14B at Q6_K uses ~12GB — higher quant means better quality with room for context.

```
ollama run gpt-oss:20b        # Agentic coding, fast inference
ollama run qwen3:14b          # Higher quant for better quality
```

24GB VRAM (What can you run?)

Winner: Qwen3-32B for general quality. DeepSeek R1-32B for reasoning.

The RTX 3090 tier. Both 32B models fit at Q4 (~20GB), and both are excellent:

- **Qwen3-32B** scores 91.0 on Arena-Hard. All 32B parameters work on every token. Best dense model at this VRAM tier for general use.
- **DeepSeek R1-32B** beats o1-mini on AIME (72.6% vs 63.6%), MATH-500 (94.3% vs 90.0%), and GPQA-Diamond (62.1% vs 60.0%). Former $20/month reasoning quality running on a $700 used GPU.

Llama 4 Scout barely fits at the lowest quants here. You'll sacrifice context window and quality. At 24GB, the dense 32B models are the better play.

```
ollama run qwen3:32b          # Best all-around at 24GB
ollama run deepseek-r1:32b    # Reasoning — beats o1-mini
```

48GB+ VRAM (Dual GPU / Mac)

Winner: Llama 4 Scout. Or Qwen3-235B-A22B if you're ambitious.

Scout at Q4 (~55GB) runs on dual GPUs or a Mac with 64GB+ unified memory. Native multimodal, 10M context, and 17B active params with 109B of expert knowledge behind them. This is where Llama 4 earns its place — the combination of vision, context, and quality isn't available at any lower tier.

Qwen3-235B-A22B (~143GB at Q4) beats DeepSeek-R1 and o1 on Arena-Hard (95.6) and AIME 2024 (85.7). Needs serious hardware, but the quality is flagship-tier.

```
ollama run llama4:scout           # ~55GB — vision + 10M context
ollama run qwen3:235b-a22b        # ~143GB — flagship quality
```

CPU Only (What actually works?)

Winner: Qwen3-4B (5-8 tok/s). DeepSeek R1-7B if you have 16GB+ RAM.

CPU inference is always slow. Qwen3-4B at ~4GB RAM is genuinely usable for interactive chat. DeepSeek R1-7B at ~8GB RAM gives you reasoning capability at 3-5 tok/s — slow but functional for problems worth thinking about.

→ Check what fits your hardware with our Planning Tool.

Winner by Use Case

Use Case	Winner	Why
Chat & conversation	Qwen3	Best instruction following, `/think` toggle, widest size range

Use Case	Winner	Why
Coding	GPT-OSS 20B / Qwen3-Coder 30B-A3B	SWE-Bench 60.7% (GPT-OSS) and 262K context (Qwen3-Coder)
Math & reasoning	DeepSeek R1 distills	Deep chain-of-thought, R1-32B beats o1-mini
Vision / multimodal	Llama 4 Scout or Qwen3-VL	Scout has native vision; Qwen3-VL at each VRAM tier
Long documents	Llama 4 Scout (10M) or GPT-OSS 20B (131K)	If you have the VRAM for Scout, nothing beats 10M context
Budget build	Qwen3 family	Something for every GPU from 4GB to 48GB+

For coding specifically: Qwen3-Coder 30B-A3B has 262K native context and is trained for agentic coding workflows. GPT-OSS 20B scores higher on SWE-Bench (60.7% vs ~50%) and fits in less VRAM. Both run on consumer hardware.

```
ollama run qwen3-coder:30b   # 262K context, agentic coding
ollama run gpt-oss:20b       # SWE-Bench leader at 13GB
```

The Bottom Line

For most budget local AI users in 2026, **Qwen3 is the default recommendation.** Widest range of sizes, best quality per VRAM dollar, Apache 2.0 license, `/think` toggle for reasoning. If you have one GPU and want one model family, Qwen3 at whatever size fits your VRAM.

DeepSeek R1 distills are the specialist pick for reasoning. The 14B at 12GB and 32B at 24GB produce chain-of-thought depth that's hard to match. The R1-32B beats o1-mini on four out of five major benchmarks — that's former $20/month quality running on a used 3090.

GPT-OSS 20B is the sleeper. OpenAI's first open model, Apache 2.0, fits in 16GB, 131K context, and 60.7% on SWE-Bench Verified. If you do agentic coding, give it a serious look.

Llama 4 Scout is impressive but needs hardware most hobbyists don't have. At 55GB minimum, it's a dual-GPU or Mac play. The native multimodal and 10M context are genuinely unique — no other local model touches that. But at 8GB, 12GB, or 24GB, the other families serve you better.

Start here:

```
ollama run qwen3:14b          # 12GB — the default pick
ollama run deepseek-r1:14b    # 12GB — reasoning specialist
ollama run gpt-oss:20b        # 16GB — coding dark horse
ollama run qwen3:32b          # 24GB — best all-around
```

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/llama-4-vs-qwen3-vs-deepseek-v3-2-local/

Free guides for running AI locally