DeepSeek Models Guide: R1, V3, and Coder

February 2, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For local use, the DeepSeek R1 distilled models are the ones that matter. R1-Distill-Qwen-14B is the sweet spot for 12 GB GPUs — it scores 69.7% on AIME 2024 and 93.9% on MATH-500, rivaling models 4x its size on reasoning tasks. On 24 GB, the 32B distill outperforms OpenAI's o1-mini. The full R1 and V3 (671B each) need 400+ GB of memory and aren't practical on consumer hardware. For general chat, Qwen3 is better. For pure math and logic, the R1 distills still hold their own.

📚 **More on this topic:** Best Models for Math & Reasoning · Qwen Models Guide · Llama 3 Guide · VRAM Requirements

DeepSeek made the biggest splash in local AI when R1 dropped in January 2025 — a reasoning model that matched OpenAI's o1 on math benchmarks, fully open-source, with distilled versions that run on a single consumer GPU.

But DeepSeek has a whole family of models now: R1, V3, V3.1, Coder V2, and more. It's confusing. This guide cuts through the noise: which ones actually matter for local use, what hardware they need, and when you should pick something else instead.

The DeepSeek Lineup at a Glance

Model	Type	Total Params	Active Params	Best For
R1 Distills	Reasoning (dense)	1.5B–70B	All (dense)	Math, logic, analysis, coding puzzles
R1 Full	Reasoning (MoE)	671B	37B	Same but better — needs datacenter hardware
V3 / V3.1	General-purpose (MoE)	671B–685B	37B	Everything — needs datacenter hardware
Coder V2	Code (MoE)	16B / 236B	2.4B / 21B	Programming — largely superseded

For consumer hardware, the R1 distills are the story. Everything else is either too big to run locally or has been outpaced by newer alternatives. The rest of this guide focuses on what you can actually use.

DeepSeek R1: The Reasoning Model

How It Works

R1 generates "thinking tokens" before answering. Ask it a math problem and it produces a chain-of-thought wrapped in `<think>...</think>` tags — working through the problem step by step, backtracking when it hits dead ends, verifying its own logic — before giving you a final answer.

This was trained via pure reinforcement learning. The model learned to reason by being rewarded for correct answers, not by imitating human reasoning traces. The result: it sometimes discovers solution paths that humans wouldn't take.

The tradeoff: Thinking tokens add overhead. A simple question might generate a few hundred thinking tokens. A hard math problem can produce tens of thousands. This makes R1 slower and more verbose than non-reasoning models — by design.

The Distilled Versions

DeepSeek distilled R1's reasoning ability into smaller, dense models. These are what you run locally:

Model	Base	VRAM (Q4_K_M)	AIME '24	MATH-500	GPQA	Ollama Command
R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	~2 GB	28.9%	83.9%	33.8%	`ollama pull deepseek-r1:1.5b`
R1-Distill-Qwen-7B	Qwen2.5-Math-7B	~6 GB	55.5%	92.8%	49.1%	`ollama pull deepseek-r1:7b`
R1-Distill-Llama-8B	Llama-3.1-8B	~6 GB	50.4%	89.1%	49.0%	`ollama pull deepseek-r1:8b`
R1-Distill-Qwen-14B	Qwen2.5-14B	**~11 GB**	**69.7%**	**93.9%**	**59.1%**	`ollama pull deepseek-r1:14b`
	Qwen2.5-32B	~22 GB	72.6%	94.3%	62.1%	

Model	Base	VRAM (Q4_K_M)	AIME '24	MATH-500	GPQA	Ollama Command
R1-Distill-Qwen-32B						`ollama pull deepseek-r1:32b`
R1-Distill-Llama-70B	Llama-3.3-70B	~43 GB	70.0%	94.5%	65.2%	`ollama pull deepseek-r1:70b`

The 7B Qwen distill outperforms GPT-4o on several math benchmarks. The 32B distill outperforms OpenAI's o1-mini. These are genuinely impressive numbers for models you can run on a single GPU.

Between the Qwen and Llama variants at similar sizes, **pick Qwen**. The 7B Qwen distill beats the 8B Llama distill on every benchmark (AIME: 55.5 vs 50.4, MATH-500: 92.8 vs 89.1).

What R1 Is Good At

- **Math competitions:** AIME, MATH-500, competition-level problems
- **Logic puzzles:** Multi-step deduction, constraint satisfaction
- **Code reasoning:** Algorithm design, debugging complex logic
- **Analysis:** Breaking down complex questions into structured reasoning

What R1 Is Bad At

- **Simple chat:** It overthinks. Ask "what's the capital of France?" and it might generate 200 thinking tokens first. Use a regular chat model for casual conversation.
- **Creative writing:** Not conversationally fluent. The writing is functional, not engaging.
- **Speed:** Thinking tokens mean more output per query. Expect 2-5x more total tokens generated compared to a non-reasoning model answering the same question.
- **Instruction following on non-reasoning tasks:** DeepSeek themselves say "R1 falls short of V3 in general-purpose tasks."

Recommended R1 Distill by GPU

GPU	Best R1 Distill	Quant	Speed (est.)
8 GB (RTX 3060 8GB, RTX 4060)	7B Qwen	Q4_K_M	~50-55 tok/s
12 GB (RTX 3060 12GB, RTX 4070)	**14B Qwen**	Q4_K_M	~28-35 tok/s

GPU	Best R1 Distill	Quant	Speed (est.)
16 GB (RTX 4060 Ti 16GB, RTX 5060 Ti)	14B Qwen	Q8_0	~25-30 tok/s
24 GB (RTX 3090, RTX 4090)	**32B Qwen**	Q4_K_M	~15-25 tok/s
48 GB+ (dual 3090, A6000)	70B Llama	Q4_K_M	~8-12 tok/s

The **14B on 12 GB** is the value sweet spot. You get 69.7% on AIME and 93.9% on MATH-500 on a $200 used GPU. That's competitive with models that need $1,600 GPUs.

→ Check what fits your hardware with our Planning Tool.

Important: Increase Context Length

Ollama defaults to 4096 tokens. That's not enough for a reasoning model — thinking chains can easily exceed that. Create a Modelfile:

```
FROM deepseek-r1:14b
PARAMETER num_ctx 16384
```

```
ollama create deepseek-r1-14b-16k -f Modelfile
```

DeepSeek V3: The Flagship (Reality Check)

V3 is DeepSeek's general-purpose model. It's a 671B MoE (37B active) that matches or beats GPT-4o and Claude 3.5 Sonnet on most benchmarks. It's genuinely one of the best open models ever released.

But you can't run it on consumer hardware. The math:

Quantization	Memory Needed
FP16	~1,400 GB
FP8	~700 GB
Q4_K_M	~400 GB
Dynamic 1-bit (Unsloth)	~170 GB

Even the most aggressive quantization (1-bit, significant quality loss) needs 170 GB of memory. You could load it into system RAM on a machine with 128+ GB, offloading from a single GPU, but expect about 3-5 tok/s. Not practical for interactive use.

V3.1 (August 2025) is the current version worth knowing about. It combines V3 and R1 into a hybrid that switches between thinking and non-thinking modes. It surpasses both V3 and R1 by 40%+ on coding agent benchmarks (SWE-bench). Available as `ollama pull deepseek-v3.1` — but same hardware requirements.

If you want V3-class intelligence, use the API. DeepSeek's API is extremely cheap ($0.27/M input tokens) compared to OpenAI or Anthropic. Or use it through OpenRouter from within Open WebUI.

DeepSeek Coder V2: Largely Superseded

Coder V2 was impressive when it launched — 90.2% on HumanEval, matching GPT-4-Turbo on coding benchmarks. The 16B Lite version runs on an 8 GB GPU:

```
ollama pull deepseek-coder-v2:16b
```

But it's been largely overtaken:

	DeepSeek Coder V2 (236B)	Qwen 2.5 Coder 32B
HumanEval	90.2%	Competitive
LiveCodeBench	43.4%	~65-70%
Architecture	MoE (21B active)	Dense 32B
VRAM (Q4)	~133 GB	~22 GB
Practical?	Multi-GPU only	Single 24 GB GPU

Qwen 2.5 Coder 32B is the better choice for coding now. It outperforms DeepSeek Coder V2 on real-world coding benchmarks, runs on a single GPU, and is easier to deploy. See our best models for coding guide for current recommendations.

The 16B Lite version is still fine if you have limited VRAM and want a dedicated coding model, but Qwen3-8B in thinking mode handles code well enough that there's less reason to run a separate coding model.

DeepSeek R1 vs the Competition

This is the question everyone asks: should you run DeepSeek R1 or Qwen3?

R1 Distills vs Qwen3 (Same Size)

Tier	R1 Distill	Qwen3	Reasoning Winner	General Winner
7-8B	R1-Distill-Qwen-7B (MATH: 92.8%)	Qwen3-8B (MATH: ~88.8%)	**R1**	**Qwen3**
14B	R1-Distill-Qwen-14B (MATH: 93.9%)	Qwen3-14B (MATH: ~92.6%)	**R1** (slight edge)	**Qwen3**
32B	R1-Distill-Qwen-32B (AIME: 72.6%)	Qwen3-32B (AIME: est. high 70s-80s)	**Qwen3**	**Qwen3**

At 7-8B and 14B: R1 distills win on pure math and reasoning. The gap is real — 92.8% vs 88.8% on MATH-500 at 7B is significant. But Qwen3 is better at everything else: chat, coding, multilingual, instruction following. And Qwen3 can toggle thinking on and off, while R1 distills always reason.

At 32B: Qwen3 wins on both reasoning and general tasks. The R1-Distill-Qwen-32B's advantage has been erased by Qwen3's training improvements.

R1 Distills vs Llama 3.3 (Same Size)

	R1-Distill-Llama-8B	Llama 3.1 8B
MATH-500	89.1%	~52%

	R1-Distill-Llama-70B	Llama 3.3 70B
MATH-500	94.5%	77.0%
GPQA	65.2%	50.5%

The R1 distills destroy the base Llama models on reasoning — it's not even close. But Llama 3.3 70B is still better for general chat, creative writing, and instruction following where deep reasoning isn't needed.

When to Choose DeepSeek R1

- You're solving math problems, logic puzzles, or competition-style questions
- You want to see the model's reasoning process (the `<think>` tags are useful for learning)
- You need a dedicated reasoning model for a specific workflow
- You have 12 GB VRAM and want the best possible reasoning (14B distill)

When to Choose Something Else

- **General chat:** Qwen3 (see our chat guide)
- **Coding:** Qwen 2.5 Coder 32B or Qwen3 with thinking mode
- **Creative writing:** Qwen3 or Llama 3.3 (see our writing guide)
- **Speed matters:** Any non-reasoning model — R1's thinking overhead makes it 2-5x slower per query

---

Running DeepSeek R1 in Ollama

Basic Setup

```
# Pull the model for your GPU
ollama pull deepseek-r1:7b      # 8 GB VRAM
ollama pull deepseek-r1:14b     # 12 GB VRAM
ollama pull deepseek-r1:32b     # 24 GB VRAM

# Run it
ollama run deepseek-r1:14b
```

Recommended Settings

Create a Modelfile for optimal R1 performance:

```
FROM deepseek-r1:14b
PARAMETER num_ctx 16384
```

```
PARAMETER temperature 0.6
PARAMETER top_p 0.95
```

```
ollama create my-r1 -f Modelfile
ollama run my-r1
```

Temperature 0.6 is the sweet spot. Lower causes repetitive/degenerate output. Higher causes incoherence. The official recommendation is 0.5–0.7.

Prompting Tips

R1 works differently than regular chat models:

Don't use system prompts. R1 performs worse with them. Put all instructions in the user message.

Don't use few-shot examples. They consistently degrade R1's reasoning. Use zero-shot, direct prompts.

Don't say "think step by step." R1 already thinks internally. Explicit chain-of-thought prompting actually hurts performance because it conflicts with the model's native reasoning.

Do give clear, direct questions:

```
Solve this: What is the largest integer n such that n^3 + 2n^2 - 5n - 6 < 0?
```

Common Problems

Thinking tokens cluttering the output. The `<think>...</think>` tags are part of how R1 works. Most frontends (Open WebUI, SillyTavern) can hide them. In Ollama's CLI, they show by default. If you're piping output to another tool, strip everything between `<think>` and `</think>`.

Empty thinking tags (`<think>\n\n</think>`). Sometimes R1 skips reasoning entirely, which usually degrades answer quality. This happens more on simple questions. No reliable fix — the model decides when to think.

Excessive verbosity. R1 can generate thousands of thinking tokens, then repeat the same content in the final answer. Try adding "Be concise" to your prompt. The R1-0528 update made this worse (~23K tokens per question vs ~12K for original R1).

Slow generation. Thinking overhead is inherent. To speed things up:

- Use Q4_K_M instead of Q8 quantization
- Enable flash attention in Ollama (`OLLAMA_FLASH_ATTENTION=1`)
- Don't set context length higher than you need — more context = more VRAM = slower
- Accept that reasoning models are slower. If you need speed, use Qwen3 in non-thinking mode.

Censorship on sensitive topics. R1 has Chinese political censorship baked into its weights — not just an API filter. It persists when run locally. Standard "uncensoring" techniques (abliteration) don't work on R1. If this is a dealbreaker, try Perplexity's R1-1776 (a post-trained uncensored version) or switch to Qwen3/Llama 3.3 for sensitive topics.

Ollama context too short. Default 4096 tokens is insufficient. Reasoning chains regularly exceed this. Set `num_ctx` to at least 8192, ideally 16384. See the Modelfile example above.

Bottom Line

The DeepSeek R1 distills are the best reasoning models you can run on consumer hardware. The 14B on a 12 GB GPU gives you competition-level math performance that was unthinkable a year ago.

But they're specialists. For everyday use — chat, coding, writing — Qwen3 is the better all-rounder at every size tier. The ideal local setup is both: Qwen3 for general tasks, R1 for when you need serious reasoning.

Start here:

- **12 GB GPU:** `ollama pull deepseek-r1:14b` — the value king
- **24 GB GPU:** `ollama pull deepseek-r1:32b` — outperforms o1-mini
- **Any GPU:** `ollama pull qwen3:8b` — for everything else

The full V3 is one of the best models ever made, but it needs datacenter hardware. Use the API if you want that level of intelligence, or wait for the next generation of distills.

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