DeepSeek V3.2 Guide: What Changed and How to Run It Locally

February 16, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** DeepSeek V3.2 is a 685B MoE model (~37B active) that competes with GPT-5 and Claude on reasoning benchmarks. You can't run it locally on consumer hardware — Q4 still needs 350GB+. But the R1-Distill models are the real story for budget builders: the 7B fits in 5GB VRAM with strong reasoning, the 14B is the best reasoning model at 12GB, and the 32B at 20GB rivals o1-mini. Use the V3.2 API for the full model ($0.25/M input tokens) or run the distills locally with Ollama.

📚 **More on this topic:** DeepSeek Models Guide (V3/R1) · Best LLMs for Math & Reasoning · VRAM Requirements · Running 70B Models Locally

DeepSeek V3.2 matches GPT-5 on benchmarks at a fraction of the API cost. On GPQA Diamond (PhD-level science), it jumped from 59.1 to ~80. On AIME math, from 39.6 to 59.4. On LiveCodeBench, from 39.2 to 49.2. These aren't incremental improvements — they're generational.

But here's the local AI reality: the full V3.2 model is a 685B-parameter MoE beast that needs 350GB+ even at Q4. You're not running it on a 3090.

The good news: the R1-Distill models — dense reasoning specialists distilled from DeepSeek-R1 — are genuinely excellent and fit on consumer hardware. The 32B distill rivals o1-mini. The 14B is arguably the best reasoning model you can run on 12GB. This guide covers both the flagship and the models you can actually use.

What Changed: V3 vs V3.2

Benchmark	V3	V3.2	Improvement
MMLU-Pro	75.9	~85	+9.1
GPQA Diamond (PhD-level science)	59.1	~80	+20.9
AIME (math competition)	39.6	59.4	+19.8

Benchmark	V3	V3.2	Improvement
LiveCodeBench	39.2	49.2	+10.0

The architecture stayed the same — 685B total parameters, ~37B active per token, MoE with 256 routed experts plus 1 shared expert. What changed is training: more data, better RL alignment, and DeepSeek Sparse Attention (DSA) that drops context processing from O(L^2) to roughly O(L*k).

V3.2 now competes directly with GPT-5 on general benchmarks and with Claude on reasoning tasks. The high-compute variant (V3.2-Speciale) surpasses GPT-5 on several benchmarks.

The DeepSeek Model Lineup

Model	Type	Params	Active	Context	Best For
V3.2	MoE	685B	~37B	128K	Everything — flagship
R1	MoE + reasoning	685B	~37B	128K	Deep chain-of-thought reasoning
R1-Distill-7B	Dense	7B	7B	128K	Budget reasoning
R1-Distill-14B	Dense	14B	14B	128K	Sweet spot reasoning
R1-Distill-32B	Dense	32B	32B	128K	Near-full R1 quality
R1-Distill-70B	Dense	70B	70B	128K	Best distill, needs serious VRAM
Coder-V2	MoE	236B	21B	128K	Coding specialist

The R1-Distill models were created by generating 800,000 reasoning samples from the full R1 model and fine-tuning smaller open-source models (Qwen 2.5 and Llama 3 bases) on that data. They're dense — not MoE — so VRAM requirements scale linearly with parameter count. All are MIT licensed.

VRAM Requirements

Full V3.2 (685B MoE)

Precision	VRAM	Hardware	Practical?
FP16	~1,543GB	Multi-node H200 cluster	Datacenter
INT8	~700GB	8-10x H100	Enterprise
Q4	~350-400GB	5-8x H100	Enterprise

Bottom line for the full model: Not happening on consumer hardware. Use the API.

R1-Distill Models (Dense — The Practical Ones)

Model	VRAM (Q4)	VRAM (Q8)	Ollama Command
R1-Distill-1.5B	~1.5GB	~2.5GB	`ollama run deepseek-r1:1.5b`
R1-Distill-7B	~5GB	~8GB	`ollama run deepseek-r1:7b`
R1-Distill-8B	~5.5GB	~9GB	`ollama run deepseek-r1:8b`
R1-Distill-14B	~9GB	~15GB	`ollama run deepseek-r1:14b`
R1-Distill-32B	~20GB	~34GB	`ollama run deepseek-r1:32b`
R1-Distill-70B	~40GB	~75GB	`ollama run deepseek-r1:70b`

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

These are the models that matter for local AI users. Dense architecture means predictable VRAM usage — no MoE surprises.

Which Distill for Which Hardware

8GB VRAM (What can you run?)

Pick: R1-Distill-7B (~5GB at Q4)

Strong reasoning for its size. The chain-of-thought output from R1 distillation means it works through problems step-by-step, catching errors that standard 7B models miss. Leaves headroom for a decent context window.

Compared to Qwen3-8B: Qwen3 is more versatile (chat, coding, translation). The R1-Distill-7B is better specifically at reasoning and math. If reasoning is your priority, go DeepSeek. For general use, Qwen3.

12GB VRAM (What can you run?)

Pick: R1-Distill-14B (~9GB at Q4)

Probably the best reasoning model at this VRAM tier. The jump from 7B to 14B is significant for complex reasoning — fewer hallucinated steps, better at multi-hop logic, stronger at math.

Compared to Qwen3-14B: Similar tradeoff to the 7B tier. Qwen3-14B has the /think toggle and better tool calling. R1-Distill-14B has deeper reasoning from the R1 distillation. Both are excellent — pick based on your primary use case.

24GB VRAM (What can you run?)

Pick: R1-Distill-32B (~20GB at Q4)

This model rivals o1-mini on reasoning benchmarks. It outperforms OpenAI-o1-mini across multiple evaluations. At 20GB, it fits on a single 3090 with room for ~8K context.

The updated R1-0528 version (distilled from Qwen3-8B architecture) is even stronger — it surpasses both Qwen3-8B and Qwen3-32B on AIME benchmarks.

48GB+ VRAM (Dual GPU)

Pick: R1-Distill-70B (~40GB at Q4)

Needs dual GPUs or a Mac with 64GB+ unified memory. The quality jump from 32B to 70B is real but the hardware requirement doubles. Worth it if you have the setup, but the 32B is the better value.

Using the V3.2 API

If you want the full V3.2 model (not the distills), the API is the realistic option:

	Price
Input tokens	$0.25/M
Output tokens	$0.38/M
Cache hit input	$0.028/M

That's roughly 10x cheaper than GPT-4o and 5-6x cheaper than Claude Sonnet for comparable quality. If you're building an application that needs the full model's capability, this is the most cost-effective flagship API on the market.

```
from openai import OpenAI

client = OpenAI(
    api_key="your-deepseek-key",
    base_url="https://api.deepseek.com"
)

response = client.chat.completions.create(
    model="deepseek-chat",  # Points to V3.2
    messages=[{"role": "user", "content": "Explain group theory"}],
)
```

The API is OpenAI-compatible — swap your base URL and API key and most existing code works unchanged.

V3.2 vs the Competition

Flagship Tier

Model	MMLU-Pro	GPQA Diamond	AIME	LiveCodeBench
DeepSeek V3.2	~85	~80	59.4	49.2
GPT-5	~85	~78	—	—
Claude 3.5 Sonnet	~84	~65	—	—
Qwen3-235B-A22B	—	—	85.7	70.7

V3.2 competes with GPT-5 on general knowledge and beats Claude on science reasoning. Qwen3-235B leads on math (AIME) and coding (LiveCodeBench), but needs ~143GB to run.

Budget Local Tier (What You Can Actually Run)

Model	VRAM (Q4)	Reasoning	Coding	Chat
R1-Distill-14B	~9GB	Excellent	Good	Good
Qwen3-14B	~9GB	Very good	Very good	Excellent
Llama 3.1 8B	~5GB	Decent	Decent	Good

At 12GB VRAM, R1-Distill-14B and Qwen3-14B are both top-tier. The DeepSeek model wins on pure reasoning depth. Qwen3 wins on versatility, tool calling, and the /think toggle.

The China Factor

DeepSeek is a Chinese company. Some users have concerns about data privacy. Here's what matters:

If you run locally: Your data never leaves your machine. The model weights are open, the community has audited them, and inference happens entirely on your hardware. It doesn't matter where the company is headquartered if the computation happens on your GPU.

If you use the API: Your prompts go to DeepSeek's servers in China. If that's a concern for your use case, run locally or use a different API provider. Several third-party providers serve DeepSeek models from US/EU infrastructure.

The license is real: MIT for the R1-Distill models. The weights are genuinely open. You can inspect, modify, and redistribute them.

Practical take: For local AI, the origin of the model doesn't affect your privacy. That's the whole point of running locally. If you're paranoid, run the R1-Distills on your own hardware and your data goes nowhere.

Getting Started

Local (R1-Distill Models)

```
# Pick based on your VRAM
ollama run deepseek-r1:7b     # 8GB VRAM
ollama run deepseek-r1:14b    # 12GB VRAM — sweet spot
ollama run deepseek-r1:32b    # 24GB VRAM — rivals o1-mini
ollama run deepseek-r1:70b    # 48GB+ VRAM
```

Full V3.2 (API)

Sign up at platform.deepseek.com, grab an API key, and use the OpenAI-compatible endpoint. $0.25/M input tokens makes it the cheapest flagship API available.

Which Path?

- **Budget builder, single GPU:** Run the R1-Distill that fits your VRAM. The 14B at 12GB is the reasoning sweet spot.
- **Need the full flagship:** Use the API. It's dirt cheap.
- **Privacy-first:** Run distills locally. Your data stays on your machine.
- **Best overall quality at 24GB:** Compare R1-Distill-32B vs Qwen3-32B — both are excellent, different strengths.

The R1-Distill models prove that distillation from a frontier model can produce genuinely good small models. At $200 for a used 12GB GPU running the 14B distill, you get reasoning quality that didn't exist at this price point a year ago.