Best Local LLMs for Math & Reasoning: What Actually Works

February 2, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For pure math on 24 GB VRAM, DeepSeek R1-Distill-Qwen-32B (72.6% AIME, 94.3% MATH-500) is the proven choice. But the real story is Phi-4-reasoning-plus — a 14.7B model that scores 81.3% on AIME and fits on a 12 GB GPU at Q4. It's a math-only model, not a general assistant, but nothing else comes close per-parameter. For an all-rounder that also handles math, Qwen 3 32B with /think mode scores 95.2% on MATH-500 and does everything else too. At 8 GB, DeepSeek R1-Distill-Qwen-7B (55.5% AIME, 92.8% MATH-500) beats every general-purpose model its size on reasoning. The key insight: thinking/reasoning models that show their chain of thought dramatically outperform standard models on math — Qwen 3 32B drops from 95.2% to 43.6% on MATH-500 when thinking is disabled. If you care about math, you need a model that thinks.

📚 **More on this topic:** DeepSeek Models Guide · Qwen Models Guide · Llama 3 Guide · VRAM Requirements

Standard LLMs are bad at math. Ask Llama 3.1 8B to solve a competition-level problem and it'll confidently produce wrong answers. The model doesn't reason — it pattern-matches, and math requires actual step-by-step logic.

That changed in 2025 with reasoning models. DeepSeek R1 proved that chain-of-thought training could make open-source models competitive with OpenAI's o1 on math benchmarks. Now there's a whole class of models — R1 distills, Qwen 3's thinking mode, Phi-4-reasoning — that genuinely think through problems before answering.

The difference is dramatic. Qwen 3 32B scores 95.2% on MATH-500 with thinking enabled. Turn thinking off: 43.6%. Same model, same weights, same hardware. The reasoning process is the entire game.

This guide covers which reasoning model to run on your hardware, what the benchmarks actually mean, and when local models fall short.

Why Math Is Different

Regular LLMs generate text left-to-right, picking the most likely next token. This works for chat, summarization, and creative writing. It fails for math because math requires working through intermediate steps — and a wrong step early means everything after it is wrong too.

Reasoning models fix this by generating a chain of thought before the final answer. The model literally writes out its reasoning — "let me try substituting x = 3… that gives 27 + 9 = 36, not 42, so let me try x = 4…" — then produces the answer based on that work.

This costs more tokens (2-5x more per response) and takes longer, but the accuracy improvement is massive. It's the difference between a calculator and a student who shows their work.

The benchmarks that matter

GSM8K (grade school math) is saturated — top models score 95%+. It no longer separates good from great. The benchmarks that actually differentiate reasoning models in 2026:

Benchmark	What It Tests	Difficulty
MATH-500	Competition-level math (500 problems from MATH dataset)	Hard
AIME 2024	American Invitational Math Exam — genuinely difficult	Very hard
GPQA Diamond	Graduate-level science questions	Expert
LiveCodeBench	Real coding problems, regularly updated	Hard

If a model scores well on AIME and MATH-500, it can handle anything you'll throw at it in practice.

The Contenders at a Glance

Every model worth considering for local math and reasoning, with the numbers that matter:

Model	Size	AIME '24	MATH-500	GPQA-D	Q4 VRAM	Type
Phi-4-reasoning-plus	14.7B	81.3%	~97.7%	69.3%	~8-10 GB	Math specialist

Model	Size	AIME '24	MATH-500	GPQA-D	Q4 VRAM	Type
DS R1-Distill-Qwen-32B	32B	72.6%	94.3%	62.1%	~18 GB	Always-think reasoning
Qwen 3 32B (thinking)	32B	70.0%	95.2%	60.0%	~20 GB	All-rounder + thinking
DS R1-Distill-Qwen-14B	14B	69.7%	93.9%	59.1%	~6.5 GB	Always-think reasoning
Qwen 3 4B (thinking)	4B	73.8%	91.4%	—	~2.5 GB	Tiny reasoning model
DS R1-Distill-Qwen-7B	7B	55.5%	92.8%	49.1%	~3.3 GB	Lightweight reasoning
Gemma 3 27B	27B	—	89.0%*	42.4%	~14 GB	General + multimodal
Llama 3.3 70B	70B	—	—	—	~43 GB	General purpose

*Gemma 3's MATH score is from the full MATH benchmark, not MATH-500 specifically. Direct comparison with MATH-500 scores should be made cautiously.

The standout: **Phi-4-reasoning-plus at 14.7B scores higher on AIME than the DeepSeek R1-Distill-32B**. A model that fits on a 12 GB GPU outperforming one that needs 24 GB. That's the kind of efficiency jump that matters for local use.

Best Model by VRAM Tier

Skip the analysis if you just want a recommendation for your GPU:

4 GB VRAM

Pick: Qwen 3 4B (thinking)

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

This is absurd. A 4B model scoring 73.8% on AIME 2024 — higher than the 32B DeepSeek R1 distill. Qwen 3's thinking mode at 4B is genuinely useful for math on hardware that can barely run a chatbot. Use `/think` for math problems, `/no_think` for everything else.

VRAM: ~2.5 GB at Q4_K_M

8 GB VRAM

Pick: DeepSeek R1-Distill-Qwen-7B for pure math, **Qwen 3 8B** for math + general use

```
ollama run deepseek-r1:7b      # math specialist
ollama run qwen3:8b             # all-rounder with /think
```

The R1 7B distill scores 92.8% on MATH-500 and 55.5% on AIME — strong for a model under 4 GB. But it always thinks and isn't great for general chat. Qwen 3 8B at ~5 GB gives you the thinking toggle: `/think` for math, `/no_think` for quick questions.

12-16 GB VRAM

Pick: Phi-4-reasoning-plus for math, **Qwen 3 14B** for all-rounder

```
ollama run phi4-reasoning-plus  # math beast
ollama run qwen3:14b            # all-rounder with /think
```

Phi-4-reasoning-plus is the best math model you can run on a 12-16 GB card. 81.3% AIME, ~97.7% MATH-500 — at 14.7B parameters. Those numbers beat models 2-4x its size. The catch: it's a math-only model, not a general assistant. Microsoft's own model card says "designed and tested for math reasoning only."

For a model that handles math well and everything else too, Qwen 3 14B with thinking mode is the pick.

If you have 16 GB and want the best of both: keep both pulled in Ollama and switch based on the task.

24 GB VRAM

Pick: DeepSeek R1-Distill-Qwen-32B for math, **Qwen 3 32B** for all-rounder

```
ollama run deepseek-r1:32b      # reasoning specialist
ollama run qwen3:32b            # all-rounder with /think
```

At 24 GB, you have two excellent options. The R1 32B distill (72.6% AIME, 94.3% MATH-500) is a dedicated reasoning model — it always shows chain of thought and excels at math, logic, and complex analysis. Qwen 3 32B with thinking (70.0% AIME, 95.2% MATH-500) is slightly behind on AIME but slightly ahead on MATH-500, and it toggles between thinking and fast chat.

For a dedicated math/reasoning setup, the R1 32B. For one model that does everything, Qwen 3 32B.

Both fit at Q4 on an RTX 3090 or RTX 4090, though context length will be limited (~4-8K tokens).

→ Check what fits your hardware with our Planning Tool.

DeepSeek R1 Distills: The Math Specialists

The DeepSeek R1 distills are purpose-built for reasoning. They were trained by distilling the full 671B R1 model's reasoning behavior into smaller models using supervised fine-tuning. They always generate chain-of-thought tokens — there's no way to turn it off.

Distill	AIME '24	MATH-500	GPQA-D	Q4 VRAM	Minimum GPU
1.5B	28.9%	83.9%	33.8%	~2 GB	Any
7B	55.5%	92.8%	49.1%	~3.3 GB	RTX 3060 8GB
14B	69.7%	93.9%	59.1%	~6.5 GB	RTX 3060 12GB
32B	72.6%	94.3%	62.1%	~18 GB	RTX 3090/4090
70B (Llama)	70.0%	94.5%	65.2%	~40 GB	Dual GPU

Key observations:

The 14B is the value king. Scoring 69.7% on AIME at only ~6.5 GB VRAM, it's the most reasoning per gigabyte in the R1 lineup. Fits comfortably on a 12 GB card.

The 32B beats OpenAI's o1-mini on AIME and MATH-500. On a single consumer GPU.

The 70B isn't worth it. Going from 32B to 70B (Llama base) only gains 0.2% on MATH-500 and drops slightly on AIME. Double the VRAM for negligible improvement. The Qwen-based 32B is the sweet spot.

The always-think tradeoff. Every response includes reasoning tokens — even for "what's 2+2." This makes them slower for simple tasks and annoying for general chat. They're specialists, not daily drivers.

```
# Pull the one that fits your GPU
ollama run deepseek-r1:7b
ollama run deepseek-r1:14b
ollama run deepseek-r1:32b
```

Phi-4-Reasoning: The Dark Horse

Microsoft's Phi-4-reasoning-plus is the most underrated math model for local use. At 14.7B parameters, it scores:

- **81.3% on AIME 2024** — higher than DS R1-Distill-32B (72.6%) and Qwen 3 32B (70.0%)
- **~97.7% on MATH-500** — higher than full DeepSeek R1 671B (97.3%)
- **69.3% on GPQA Diamond** — competitive with models 4x its size

This is a 14B model outperforming 32B reasoning specialists. At Q4 quantization, it needs ~8-10 GB VRAM — meaning it runs on an RTX 3060 12GB or RTX 4060 Ti 16GB.

The catch

Phi-4-reasoning-plus is a **math-only model**. Microsoft's model card explicitly states it's "designed and tested for math reasoning only." Don't expect good creative writing, chat, or code generation. It does one thing and does it extraordinarily well.

The base Phi-4 (non-reasoning) is a general-purpose model, but it scores dramatically lower on math benchmarks. The reasoning version was fine-tuned on o3-mini demonstrations plus 90 steps of reinforcement learning on ~6,000 math problems.

There's also **Phi-4-mini-reasoning** at 3.8B — a tiny reasoning model for edge devices. Useful if you're truly VRAM-constrained but still need reasoning capabilities.

Setup

```
ollama run phi4-reasoning-plus   # 14.7B, math specialist
ollama run phi4-mini-reasoning   # 3.8B, lightweight
```

When to pick Phi-4 over R1

- You have 12-16 GB VRAM and want the absolute best math performance
- Your use case is specifically math, physics, or formal logic
- You don't need the model for chat, code, or creative tasks

When to pick R1 instead

- You want stronger performance on code (LiveCodeBench, Codeforces)
- You want a model that handles reasoning across domains, not just math
- You have 24 GB and the R1 32B distill fits

Qwen 3 Thinking Mode: The All-Rounder

Qwen 3 isn't a dedicated reasoning model, but its hybrid thinking mode makes it competitive on math while being useful for everything else.

The numbers with thinking enabled:

Qwen 3	AIME '24	MATH-500	Thinking Off MATH-500
4B	73.8%	91.4%	—
8B	—	88.8%	—
14B	—	92.6%	—
32B	70.0%	95.2%	43.6%

That last column is the key stat. **Qwen 3 32B drops from 95.2% to 43.6% on MATH-500 when thinking is disabled.** The chain of thought more than doubles the model's math accuracy. Same weights, same quantization, same hardware — the only difference is whether it reasons through the problem first.

The toggle advantage

Unlike R1 distills that always think, Qwen 3 lets you switch per-message:

```
>>> /think Prove that the square root of 2 is irrational
[detailed chain-of-thought reasoning...]

>>> /no_think What's the capital of France?
Paris.
```

This means one model handles your math homework and your quick questions without wasting tokens on reasoning when it's not needed.

Qwen 3 4B: The surprise

Qwen 3 4B with thinking scores **73.8% on AIME 2024**. That's higher than the DeepSeek R1-Distill-Qwen-32B (72.6%). A model that fits in 2.5 GB of VRAM outperforming a model that needs 18 GB on competition math. The 4B won't match the 32B on complex multi-step reasoning or non-math tasks, but for the AIME benchmark specifically, the result speaks for itself.

Gemma 3: The Multimodal Option

Google's Gemma 3 27B is a strong general model with decent math capabilities, but it's not a reasoning specialist. Scores:

- **89.0% on MATH** (full benchmark, not MATH-500)
- **95.9% on GSM8K**
- **42.4% on GPQA Diamond**

That GPQA score (42.4%) tells the story — it's 20-30 points behind the reasoning models. Gemma 3 doesn't do chain-of-thought reasoning. It's a standard autoregressive model that happens to be well-trained.

When Gemma 3 makes sense for math: If you need a multimodal model that can read equations from images (photos of textbooks, handwritten problems) and solve them, Gemma 3 27B's vision capabilities plus decent math performance is a unique combination. No reasoning model currently handles image input as well.

At ~14 GB with Google's QAT quantization, the 27B fits on a 24 GB GPU with room for context.

```
ollama run gemma3:27b
```

Setup Guide

Quick start: Best model for your GPU

```
# 4 GB VRAM — tiny reasoning model
ollama run qwen3:4b

# 8 GB VRAM — lightweight math specialist
ollama run deepseek-r1:7b

# 12 GB VRAM — math beast
ollama run phi4-reasoning-plus

# 16 GB VRAM — all-rounder with thinking
ollama run qwen3:14b

# 24 GB VRAM — full reasoning power
ollama run deepseek-r1:32b
```

System prompts for better math output

Reasoning models benefit from system prompts that set expectations:

```
# Modelfile for math-focused Qwen 3
FROM qwen3:14b

PARAMETER num_ctx 8192
PARAMETER temperature 0.1

SYSTEM """You are a precise math and reasoning assistant. When solving problems:
1. Show your complete work step by step
2. Verify your answer by checking it against the original problem
3. If you're uncertain, state your confidence level"""
```

```
ollama create math-assistant -f Modelfile
ollama run math-assistant
```

Low temperature (0.1-0.3) helps for math — you want deterministic reasoning, not creative variation.

Thinking mode with Qwen 3

Force thinking on for all math sessions:

```
# Interactive toggle
>>> /set think

# Or per-message
>>> /think Solve: find all real solutions to x^4 - 5x^2 + 4 = 0
```

For a permanently thinking version:

```
FROM qwen3:14b
PARAMETER think true
PARAMETER temperature 0.1
```

Managing multiple reasoning models

Keep both a specialist and an all-rounder pulled:

```
# Pull both
ollama pull deepseek-r1:14b
ollama pull qwen3:14b

# Use R1 for heavy math
ollama run deepseek-r1:14b "Prove that there are infinitely many primes"

# Use Qwen for math + everything else
ollama run qwen3:14b
```

Ollama keeps models on disk when not in use and loads them into VRAM on demand. Having multiple models pulled doesn't cost VRAM — only the running model uses GPU memory.

When Local Falls Short

Honesty time. Local reasoning models are impressive, but they have limits.

Frontier math. On the hardest problems — FrontierMath, ARC-AGI, novel research-level math — cloud models (o1-pro, o3, Claude with extended thinking) still lead by a wide margin. The full DeepSeek R1 at 671B is competitive, but you can't run that locally.

Multi-step proofs. Long chain-of-thought reasoning (20+ steps) degrades in smaller models. The 7B and 14B models sometimes lose track of earlier steps or contradict themselves mid-proof. The 32B models are significantly more reliable here.

Context length under load. Reasoning tokens eat into your context window. A problem that needs 2,000 tokens of thinking plus 500 tokens of answer needs 2,500 tokens of KV cache space. On VRAM-constrained setups, this limits how complex your problems can be.

Verification. Reasoning models can produce confident, well-structured proofs that are wrong. The chain of thought looks convincing even when it contains logical errors. Always verify important results independently.

For day-to-day math help — homework, data analysis, engineering calculations, statistics — local reasoning models are more than capable. For publishing mathematical proofs or tackling unsolved problems, cloud APIs are still the safer bet.

Bottom Line

The math reasoning landscape for local AI has changed dramatically. A year ago, you needed cloud APIs for anything beyond basic arithmetic. Now:

- **Phi-4-reasoning-plus** scores 81.3% on AIME on a 12 GB GPU
- **DeepSeek R1-Distill-32B** beats o1-mini on a single RTX 3090
- **Qwen 3 4B** with thinking outperforms 32B models on competition math at 2.5 GB VRAM
- **Qwen 3 32B** gives you 95.2% MATH-500 with a thinking toggle for everyday use

The key insight: **thinking is everything**. The same Qwen 3 32B that scores 95.2% with `/think` drops to 43.6% without it. If you care about math accuracy, you need a model that reasons, not just predicts.

For most people: **Qwen 3 at your VRAM tier with thinking enabled** handles math and everything else. For math-only workflows: **Phi-4-reasoning-plus** if you have 12-16 GB, **R1-Distill-32B** if you have 24 GB. For the tightest budgets: **Qwen 3 4B** with `/think` is genuinely useful at 2.5 GB.

```
# Just try it — the difference is immediately obvious
ollama run qwen3:8b
# Then type: /think Solve: x^3 - 6x^2 + 11x - 6 = 0
```
