Phi Models Guide: Microsoft's Small but Mighty LLMs

February 8, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Phi-4 14B is the headliner — 84.8% MMLU and 82.6% HumanEval while fitting on a 12GB GPU at Q4. It's the best model under 14B for math and coding. Phi-4-mini (3.8B) is the new small option with 128K context, replacing Phi-3.5 Mini. Both are MIT licensed with no commercial restrictions. The catch: Phi models are trained heavily on synthetic data, making them excellent at benchmarks but weaker on creative writing, factual breadth, and multilingual tasks. For general-purpose use, Qwen 2.5 14B is still the better all-rounder — but for math, code, and reasoning on limited hardware, nothing at this size beats Phi-4.

📚 **Related:** Gemma Models Guide · Qwen Models Guide · VRAM Requirements · Best LLMs for Math

Microsoft's thesis with Phi is simple: small models trained on high-quality data can match models several times their size. And they've mostly proven it right.

Phi-4 14B scores 84.8% on MMLU — in the same range as Llama 3.3 70B and Qwen 2.5 72B, models that need 4-5x more VRAM. It hits 82.6% on HumanEval for coding. It outscores GPT-4o on GPQA and MATH benchmarks.

All on a model that fits on a 12GB GPU.

The trick: synthetic training data. Microsoft generates massive amounts of carefully crafted training examples using larger models, then distills that capability into smaller architectures. It works brilliantly for structured tasks. It falls flat for creative writing and broad knowledge. Understanding that trade-off is key to using Phi effectively.

The Current Lineup

Model	Parameters	Context	VRAM (Q4)	Released	Best For
Phi-4	14B	16K	~10 GB	Dec 2024	Math, coding, reasoning
Phi-4-mini	3.8B	128K	~3 GB	Feb 2025	Tight hardware, long context

Model	Parameters	Context	VRAM (Q4)	Released	Best For
Phi-4-multimodal	5.6B	128K	~4 GB	Feb 2025	Speech + vision + text
Phi-3.5 Mini	3.8B	128K	~3 GB	Aug 2024	Legacy (use Phi-4-mini instead)
Phi-3.5 MoE	42B (6.6B active)	128K	~10 GB	Aug 2024	Niche MoE use cases
Phi-3.5 Vision	4.2B	128K	~3 GB	Aug 2024	Legacy (use Phi-4-multimodal instead)

Skip: Phi-3, Phi-2, Phi-1.5 — all obsolete. If you see benchmarks referencing these, they're not relevant to what you'd run today.

How to Run

```
# Via Ollama
ollama run phi4                    # Phi-4 14B
ollama run phi4-mini               # Phi-4-mini 3.8B

# Specific quantization
ollama run phi4:q4_K_M             # Q4 for 12GB cards
ollama run phi4:q8_0               # Q8 for 24GB cards
```

Models are also available on HuggingFace under MIT license. GGUF quantizations are available from the usual community quantizers.

Phi-4 14B: The Main Event

Benchmarks

Benchmark	Phi-4 14B	What It Measures
MMLU	84.8%	Academic knowledge breadth
HumanEval	82.6%	Code generation (Python)

Benchmark	Phi-4 14B	What It Measures
MATH	80.4%	Mathematical reasoning
GPQA	Beats GPT-4o	Graduate-level science

These numbers are remarkable for a 14B model. For context, Llama 3.1 70B scores ~86% on MMLU — only 1.2 points higher with 5x the parameters and 5x the VRAM requirement.

VRAM Requirements

Quantization	VRAM	Fits On
Q4_K_M	~10 GB	RTX 3060 12GB, RTX 4060 Ti 16GB
Q6_K	~12 GB	RTX 3060 12GB (tight), RTX 3090
Q8_0	~16 GB	RTX 4060 Ti 16GB, RTX 3090
FP16	~30 GB	Dual GPU or A100

The sweet spot is **Q4 on a 12GB card**. You lose minimal quality on reasoning and coding tasks at Q4 — Phi's strengths are structural (logic, patterns) rather than knowledge-dependent, so quantization hurts less than it would on a general knowledge model.

Inference Speed

On consumer hardware, expect roughly:

- **RTX 3060 12GB (Q4):** ~25-30 tok/s generation
- **RTX 3090 24GB (Q4):** ~50-60 tok/s generation
- **RTX 4090 24GB (Q4):** ~70-80 tok/s generation

Prompt processing is fast at ~260 tok/s on capable hardware. The 14B size keeps inference snappy.

The 16K Context Limitation

Phi-4's biggest practical limitation: 16K context window. Started at 4K and was extended to 16K during training, but that's still short compared to Qwen 2.5 (128K), Llama 3.1 (128K), and Gemma 3 (128K).

For chat and coding, 16K is usually enough. For document analysis, summarization of long texts, or RAG with many retrieved chunks, it's a real constraint. If long context matters, Phi-4-mini (128K) or a different model family is the better choice.

Phi-4-mini 3.8B: Long Context on Tiny Hardware

Phi-4-mini is the replacement for Phi-3.5 Mini. Same 3.8B parameter count, but improved reasoning, math, multilingual support, and function calling.

The big upgrade: 128K context window. On 3GB of VRAM, you get a model that can process substantial documents — something the original Phi-4 14B can't do with its 16K limit.

Where it fits: Ultra-constrained hardware. 4GB VRAM cards, old laptops, Raspberry Pi setups. If your total VRAM is under 6GB, Phi-4-mini is one of the best options available.

Function calling: Phi-4-mini supports structured function calling out of the box — useful for agent-style workflows where the model needs to invoke tools. This is unusual for a 3.8B model.

```
ollama run phi4-mini
```

Phi-4-multimodal 5.6B

A unified model that handles text, vision, and speech. Topped the OpenASR leaderboard with 6.14% word error rate — competitive with dedicated speech models.

If you need one model to handle transcription, image understanding, and text chat on limited hardware, this is the most efficient option. At ~4GB VRAM (Q4), it replaces the need for separate Whisper + LLaVA + chat model setups.

Niche but genuinely useful for the right workflow.

Where Phi Shines

Math and Reasoning

This is Phi's core selling point. Microsoft's synthetic data approach generates millions of step-by-step math solutions, logic problems, and reasoning chains. The result: Phi-4 14B matches or beats 70B models on math benchmarks.

In practice, this means:

- Correct algebra, calculus, statistics at the 14B tier
- Logical deduction that doesn't fall apart on multi-step problems
- Word problems handled more reliably than Llama or Mistral at comparable sizes

If you're a student, researcher, or anyone who regularly asks AI to work through quantitative problems, Phi-4 is the best model under 30B parameters for this.

Coding

82.6% HumanEval isn't a fluke. Phi-4 writes good Python — clean, correct, and usually following best practices. The training data includes heavy Python representation (typing, math, random, and standard library modules in particular).

For other languages (JavaScript, Rust, Go), quality is still solid but less consistent. Qwen 2.5 Coder has broader language coverage if you need more than Python.

Instruction Following

Phi-4 follows format instructions well. "Output as JSON," "respond in bullet points," "give me exactly three examples" — these directives are followed more consistently than Llama 3.1 8B at comparable sizes. Useful for structured workflows and API backends.

Where Phi Struggles

Creative Writing

Phi models produce technically correct but lifeless prose. The synthetic training data optimizes for accuracy, not voice. If you ask Phi-4 to write a short story, you'll get something that reads like a textbook example of a short story — all the elements present, none of the spark.

For creative work, Qwen 2.5 and Llama 3 produce noticeably more engaging output.

Factual Knowledge

This is the hidden cost of being 14B. Phi-4 scores well on benchmarks by being good at reasoning through problems, but its factual knowledge base is narrow. Ask it about obscure history, niche technical topics, or current events and you'll hit gaps that a 70B model wouldn't have.

The community has noted that Microsoft may have over-optimized for benchmark performance at the expense of general knowledge. Phi-4's SimpleQA score (measuring factual accuracy) dropped from 7.6 to 3.0 compared to earlier versions — a sign that training focused on structured problem-solving over broad knowledge retention.

Multilingual

Phi-4 is English-first. It handles other languages but with significantly less capability than Qwen 2.5 (29 languages) or Llama 3 (strong European coverage). For multilingual work, look elsewhere.

JSON Output Quirks

Some users report inconsistent JSON formatting from Phi-4 — well-structured prompts sometimes produce malformed output. This appears related to a tokenizer quirk where the model auto-adds assistant prompts in certain serving configurations. If you're using Phi-4 as an API backend, test your specific prompt format carefully.

---

vs The Competition

Phi-4 14B vs Qwen 2.5 14B

Aspect	Phi-4 14B	Qwen 2.5 14B
VRAM (Q4)	~10 GB	~9 GB
Context	16K	128K
Math/reasoning	**Best in class**	Very good
Coding	Strong (Python focus)	**Stronger (broader languages)**

Aspect	Phi-4 14B	Qwen 2.5 14B
Creative writing	Weak	Good
Multilingual	Weak	**29 languages**
Knowledge breadth	Narrow	**Broader**
License	MIT	Apache 2.0

Pick Phi-4 for math-heavy, reasoning-focused work. **Pick Qwen 2.5 14B** for everything else. If you can only keep one 14B model on your system, Qwen is the more versatile choice.

Phi-4 14B vs Llama 3.1 8B

Aspect	Phi-4 14B	Llama 3.1 8B
VRAM (Q4)	~10 GB	~5 GB
Context	16K	128K
Reasoning	**Much stronger**	Average
Coding	**Stronger**	Moderate
Creative writing	Weak	**Better**
Speed (same GPU)	Slower (larger)	Faster (smaller)

Phi-4 is the better model if it fits your hardware. The question is whether the 5GB VRAM difference matters for your setup. On an 8GB card, Llama 3.1 8B is your only option. On a 12GB card, Phi-4 is the upgrade worth making.

Phi-4-mini 3.8B vs Gemma 3 4B

Aspect	Phi-4-mini 3.8B	Gemma 3 4B
VRAM (Q4)	~3 GB	~3 GB
Context	128K	128K
Math/reasoning	**Better**	Good
Vision	Via multimodal variant	**Built-in**
Instruction following	Good	**Better**
Creative writing	Weak	Slightly better

Aspect	Phi-4-mini 3.8B	Gemma 3 4B
Function calling	**Built-in**	Limited

At the 3-4B tier, both are excellent. **Phi-4-mini** for math, function calling, and agent workflows. **Gemma 3 4B** for general tasks, vision, and structured output.

VRAM Cheat Sheet

Your GPU	Best Phi Model	Quantization	What Else Fits
4GB	Phi-4-mini 3.8B	Q4	Tight but works
8GB	Phi-4-mini 3.8B	Q8 (best quality)	Room for embedding model too
12GB	Phi-4 14B	Q4	The sweet spot
16GB	Phi-4 14B	Q6	Better quality
24GB	Phi-4 14B	Q8	But consider Qwen 32B instead

On a 24GB card, you have better options than Phi-4. Qwen 2.5 32B at Q4 (~20GB) is a stronger general model. Phi-4 at Q8 on 24GB is overkill for a 14B model — spend that VRAM on a bigger model instead.

→ Check what fits your hardware with our Planning Tool.

The Phi-3.5 Lineup (Legacy)

Still available, but superseded by Phi-4:

Phi-3.5 Mini (3.8B)

Replaced by Phi-4-mini. Same size, less capable. No reason to run this on new setups.

Phi-3.5 MoE (16x3.8B, 6.6B active)

A mixture-of-experts model with 42B total parameters but only 6.6B active per token. Runs on ~10GB VRAM at Q4. Interesting architecture, but in practice Phi-4 14B on the same VRAM

budget is faster and often better. The MoE approach adds complexity without a clear win at this scale.

Phi-3.5 Vision (4.2B)

Replaced by Phi-4-multimodal, which adds speech on top of vision. Use the newer model.

Recommendations

Math and reasoning on limited hardware: Phi-4 14B. Nothing else at 10GB VRAM comes close for quantitative work. If you're a student or researcher who needs reliable math capability without cloud APIs, this is the model.

Tight VRAM budget (4-8GB): Phi-4-mini 3.8B. The 128K context and function calling support make it more than just a tiny chat model — it's a capable small agent.

General-purpose use: Skip Phi. Qwen 2.5 at equivalent sizes is more versatile, better at creative tasks, stronger multilingual, and has longer context (on the 14B). Phi is a specialist, not a generalist.

Coding: Phi-4 14B is genuinely good for Python. For broader language support, pair it with Qwen 2.5 Coder or use Qwen 2.5 14B as your single model.

The honest take: Phi-4 is the best model you should only use for specific things. It dominates math and reasoning benchmarks at its size class, but the narrow training shows everywhere else. Keep it alongside a general-purpose model, not instead of one.

📚 **Model comparisons:** Gemma Models Guide · Qwen Models Guide · Llama 3 Guide · Best LLMs for Math

📚 **Hardware pairing:** VRAM Requirements · 12GB VRAM Guide · Best GPU Under $300
