Qwen3 Complete Guide: Every Model from 0.6B to 235B

February 16, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Qwen3 is the strongest open-source model family for local AI right now. The 4B dense model rivals Qwen 2.5 72B on benchmarks. The 30B-A3B MoE fits on 8GB VRAM and outperforms QwQ-32B. Every model has a /think toggle that switches between chain-of-thought reasoning and fast chat mode. All Apache 2.0 licensed. For 8GB VRAM, run Qwen3-4B or the 30B-A3B MoE. For 12GB, the 14B is the sweet spot. For 24GB, the 32B dense model is hard to beat. Start with: ollama run qwen3:4b

📚 **More on this topic:** Qwen 2.5 Guide · VRAM Requirements · Best Models for Coding · Llama 4 Guide

Qwen3-4B matches Qwen 2.5-72B on benchmarks. Read that again.

A model that fits in 3GB of VRAM competes with one that needs 43GB. That's not marketing — it's the actual benchmark data from Alibaba's technical report, and it reflects a generational leap in what small models can do.

Qwen3 is the strongest open model family for local AI right now. Eight sizes from 0.6B to 235B, two MoE models that punch way above their weight, a /think toggle that no other family offers, and everything under Apache 2.0. This guide covers every model, what hardware you need, and which one to pick.

The Full Qwen3 Family

Dense Models

Model	Params	VRAM (Q4)	Ollama	Best For
Qwen3-0.6B	0.6B	~1GB	`ollama run qwen3:0.6b`	Phones, Raspberry Pi, edge devices
Qwen3-1.7B	1.7B	~1.5GB		Embedded, Pi, basic classification

Model	Params	VRAM (Q4)	Ollama	Best For
			`ollama run qwen3:1.7b`	
Qwen3-4B	4B	~3GB	`ollama run qwen3:4b`	Best quality/VRAM ratio in any model family
Qwen3-8B	8B	~6GB	`ollama run qwen3:8b`	General-purpose on 8-12GB GPUs
Qwen3-14B	14B	~9GB	`ollama run qwen3:14b`	Sweet spot for instruction-following
Qwen3-32B	32B	~20GB	`ollama run qwen3:32b`	Best dense model for 24GB GPUs

MoE Models

Model	Total / Active	VRAM (Q4)	Ollama	Best For
Qwen3-30B-A3B	30B / 3B active	~18GB*	`ollama run qwen3:30b-a3b`	Hidden gem — outperforms QwQ-32B
Qwen3-235B-A22B	235B / 22B active	~143GB	`ollama run qwen3:235b-a22b`	Flagship — competes with DeepSeek-R1, o1

*The 30B-A3B needs ~18GB at Q4, but Unsloth's aggressive quants can squeeze it into 8GB. At Q4_K_M you'll want 16-18GB.

Coding & Vision Variants

Model	Type	Notes
Qwen3-Coder 480B-A35B	MoE coding	Agentic coding monster, 256K context, needs 250GB+
Qwen3-Coder 30B-A3B	MoE coding	Practical coding model, fits in consumer hardware
Qwen3-VL 4B/8B/32B	Vision	Image understanding at each VRAM tier
Qwen3-VL 30B-A3B	Vision MoE	Multimodal with MoE efficiency

All models are **Apache 2.0** licensed — genuinely open for commercial and personal use.

The Killer Feature: /think and /no_think

Every Qwen3 model has a built-in reasoning toggle. No other model family does this as cleanly.

`/think` mode — chain-of-thought reasoning. The model works through the problem step by step before answering. Use for math, coding, logic puzzles, complex analysis.

`/no_think` mode — fast response. Skip the reasoning chain and answer directly. Use for chat, simple questions, creative writing, translation.

```
You: /think What's the time complexity of mergesort and why?

Qwen3: <think>
Mergesort divides the array in half recursively...
Each level of recursion does O(n) work for merging...
There are log(n) levels...
</think>

O(n log n). Here's why: mergesort splits the array in half
at each level (log n levels), and each level merges all n
elements. That gives n * log(n) total operations...
```

```
You: /no_think Translate "good morning" to French

Qwen3: Bonjour
```

The toggle works per-turn in multi-turn conversations. You can use `/think` for one question and `/no_think` for the next. The model follows the most recent instruction.

This matters for local AI because you can use one model for both quick chat and deep reasoning — no need to swap between a fast model and a thinking model.

Which Model for Which Hardware

This is the question that matters. Here's the honest breakdown:

4GB VRAM (What can you run?)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-0.6B	Q4_K_M	~1GB	40-60 tok/s	Fast but limited

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-1.7B	Q4_K_M	~1.5GB	30-45 tok/s	Better quality, still quick
Qwen3-4B	Q4_K_M	~3GB	20-35 tok/s	Best you can get at this tier

Pick: Qwen3-4B. It fits, it's fast, and its benchmark scores embarrass models 10x its size.

8GB VRAM (What can you run?)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-4B	Q8_0	~5GB	25-35 tok/s	Higher quality from same model
Qwen3-8B	Q4_K_M	~6GB	20-30 tok/s	Solid general-purpose
Qwen3-30B-A3B	1.78-bit	~8GB	15-25 tok/s	MoE gem, punches way above

Pick: Qwen3-8B for general use, or Qwen3-30B-A3B (Unsloth quant) if you want MoE magic. The 30B-A3B outperforms QwQ-32B despite activating only 3B parameters per token.

12GB VRAM (What can you run?)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-8B	Q6_K	~8GB	22-30 tok/s	Higher quant, better quality
Qwen3-14B	Q4_K_M	~9GB	18-25 tok/s	Best instruction-following

Pick: Qwen3-14B. This is the sweet spot. Strong instruction-following, great at structured output and tool calling, and it fits comfortably with room for context.

16GB VRAM (What can you run?)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-14B	Q6_K	~12GB	18-25 tok/s	Higher quant = better quality
Qwen3-30B-A3B	Q4_K_M	~18GB	Tight fit	MoE model, needs most of the VRAM
Qwen3-32B	Q4_K_M	~20GB	Won't fit	Over budget — need 24GB

Pick: Qwen3-14B at Q6_K. You get the best quality-per-VRAM at this tier. The 30B MoE at Q4 is tight — you'll sacrifice context window.

24GB VRAM (What can you run?)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-32B	Q4_K_M	~20GB	15-22 tok/s	Best dense model for 24GB
Qwen3-30B-A3B	Q4_K_M	~18GB	20-30 tok/s	Faster inference, headroom for context

Pick: Qwen3-32B for best quality, Qwen3-30B-A3B for faster inference. The 32B is a dense model — all 32B params work on every token. The 30B MoE is faster per token (only 3B active) but the 32B edges it on raw quality.

CPU Only

Model	Quant	RAM Needed	Speed	Verdict
Qwen3-4B	Q4_K_M	~4GB	5-8 tok/s	Usable for chat
Qwen3-8B	Q4_K_M	~8GB	3-5 tok/s	Slow but functional

CPU inference is always slow but Qwen3-4B at Q4 is genuinely usable at 5-8 tok/s. Acceptable for interactive chat on a decent CPU.

→ Check what fits your hardware with our Planning Tool.

Qwen3 vs Qwen 2.5

If you're on Qwen 2.5, here's what changed:

	Qwen 2.5	Qwen3
Architecture	Dense only	Dense + MoE
4B quality	Decent for size	Rivals Qwen 2.5-72B
Reasoning toggle	No	/think and /no_think
Training data	18T tokens	36T tokens
Languages	~29	119
MoE options	None	30B-A3B, 235B-A22B
License	Apache 2.0	Apache 2.0

	Qwen 2.5	Qwen3
Tool calling	Good	Better (agentic RL training)
Coding	Good	Qwen3-Coder is top-tier

The biggest upgrade is quality at every size. Qwen3-4B scores 83.7 on MMLU-Redux and 97.0 on MATH-500. Qwen 2.5-7B scored 74.2 on MMLU. The 4B model trained on 2x the data with better architecture now beats the previous-gen 7B comfortably.

The MoE models are entirely new. Qwen 2.5 had no MoE option. The 30B-A3B is particularly impressive — it gives you 30B-scale knowledge with 3B-scale inference speed.

Benchmarks

Dense Models — Quality by Size

Model	MMLU-Redux	MATH-500	Arena-Hard	LiveCodeBench
Qwen3-4B	83.7	97.0	—	—
Qwen3-8B	84.9	97.4	—	—
Qwen3-14B	86.7	97.4	—	—
Qwen3-32B	87.8	97.4	81.5	52.7

Flagship Comparisons

Model	Arena-Hard	AIME'24	AIME'25	LiveCodeBench
Qwen3-235B-A22B	95.6	85.7	81.4	70.7
DeepSeek-R1	92.3	79.8	70.0	65.9
o1	91.5	79.2	—	—
Grok-3	93.3	83.9	68.0	—

Qwen3-235B-A22B beats DeepSeek-R1 and o1 on Arena-Hard and both AIME math benchmarks. That's the flagship, though — you need ~143GB to run it.

Qwen3-Coder: The Coding Specialist

Two coding-specific models:

Qwen3-Coder 480B-A35B — the agentic coding monster. 480B total params, 35B active, 256K native context. Comparable to Claude Sonnet 4 on agentic coding benchmarks. Needs 250GB+ VRAM — datacenter territory.

Qwen3-Coder 30B-A3B — the practical option. Same MoE architecture as Qwen3-30B-A3B but trained on 7.5T tokens of code-heavy data. Fits on consumer hardware. Comes with an open-source CLI tool (Qwen Code) for agentic coding workflows.

```
ollama run qwen3-coder:30b-a3b   # Practical coding model
```

If you're using local LLMs for coding, the Coder 30B-A3B is the best option that fits on consumer hardware right now.

Getting Started

The fastest path from zero to running Qwen3:

```
# Install Ollama if you haven't
curl -fsSL https://ollama.com/install.sh | sh

# Pick your model based on VRAM
ollama run qwen3:4b        # 4-8GB VRAM — start here
ollama run qwen3:8b        # 8-12GB VRAM
ollama run qwen3:14b       # 12-16GB VRAM — sweet spot
ollama run qwen3:32b       # 24GB VRAM
ollama run qwen3:30b-a3b   # 16-24GB VRAM — MoE gem
```

Try the reasoning toggle:

```
>>> /think Explain why quicksort is O(n log n) average but O(n²) worst case
>>> /no_think What's the capital of France?
```

If you're coming from Qwen 2.5, the upgrade is worth it at every size. If you're coming from Llama 3, Qwen3 now beats it on most benchmarks at equivalent parameter counts. If you're on a budget GPU and want the absolute best model for your VRAM, Qwen3 is the answer.
