Best Local LLMs for Chat & Conversation

January 31, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For everyday chat, Qwen3 is the current king at every size. Qwen3-8B on 8GB VRAM matches the previous generation's 14B models. Qwen3-14B on 16GB VRAM matches what used to require 32B. At 24GB, Gemma 3 27B (QAT int4, fits in 14GB VRAM) has the highest chat arena scores of any open model its size — it outperforms models 3x larger in blind human preference tests. For the best possible local chat experience, Qwen3-32B on 24GB VRAM is hard to beat. All of these run in Ollama with a single pull command.

📚 **More on this topic:** Best Models for Coding · Best Models for Writing · VRAM Requirements

You want a local model you can just talk to. Ask it questions, bounce ideas off it, get help thinking through problems — without sending every thought to OpenAI's servers.

The good news: chat is where local models have improved the most. The Qwen3 family (released April 2025) effectively doubled performance per parameter — an 8B model now matches what 14B models did six months ago. Gemma 3 27B scores higher in blind human preference tests than models three times its size. And all of them run on consumer hardware with a single Ollama command.

Here's what to download depending on your GPU.

Best Models by VRAM Tier

VRAM	Model	Quant	Speed (est.)	Best For
8 GB	**Qwen3-8B**	Q4_K_M	~38-45 tok/s	Best all-around at this size
8 GB	Llama 3.1 8B	Q4_K_M	~38 tok/s	Proven, widest tool support
8 GB	Gemma 2 9B	Q4_K_M	~35 tok/s	Most natural-sounding prose
12 GB	**Qwen3-14B**	Q4_K_M	~22-30 tok/s	Matches previous-gen 32B quality
12 GB	Gemma 3 12B	QAT int4	~25-35 tok/s	Multimodal (text + images), only 6.6GB
16 GB	Qwen3-14B	Q6_K	~20-25 tok/s	Higher quality, same model
16 GB	Mistral Nemo 12B	Q6_K	~25-30 tok/s	Clean output, 128K context

VRAM	Model	Quant	Speed (est.)	Best For
24 GB	**Gemma 3 27B**	QAT int4	~20-30 tok/s	Highest chat arena scores at this tier
24 GB	**Qwen3-32B**	Q4_K_M	~20-30 tok/s	Best overall quality, versatile
24 GB	Mistral Small 3.1 (24B)	Q4_K_M	~30-50 tok/s	Fastest at this tier, multimodal
48 GB+	Llama 3.3 70B	Q4_K_M	~8-12 tok/s	Best instruction following
48 GB+	Qwen 2.5 72B	Q4_K_M	~8-12 tok/s	Best multilingual, highest MT-Bench

The generational leap: Qwen3 models match Qwen 2.5 models roughly 2x their size. Qwen3-8B ≈ Qwen 2.5-14B. Qwen3-14B ≈ Qwen 2.5-32B. Qwen3-32B ≈ Qwen 2.5-72B. If you're still running Qwen 2.5, upgrading to Qwen3 at the same size is like getting a free GPU upgrade.

→ Check what fits your hardware with our Planning Tool.

What Makes a Good Chat Model?

Chat is different from coding or writing. A good conversational model needs to:

- **Follow instructions consistently** — do what you ask without drifting
- **Handle multi-turn conversation** — remember what you discussed earlier in the chat
- **Sound natural** — not robotic, not overly formal, not stuffed with filler phrases
- **Know when it doesn't know** — hallucinate less, admit uncertainty
- **Be responsive** — fast enough for interactive back-and-forth (15+ tok/s minimum)

Benchmarks that correlate with chat quality: **Chatbot Arena Elo** (blind human preference), **IFEval** (instruction following), and **MT-Bench** (multi-turn conversation). Raw MMLU scores don't tell you much about how a model feels to talk to.

The 8GB Tier

This is where most people start. An 8GB GPU runs 7-9B models at Q4 comfortably with room for context.

Qwen3-8B — The New Default

Qwen3-8B is the standout. It scores 74 on MMLU-Pro versus Qwen 2.5 7B's 45 — that's not an incremental improvement, it's a generational leap. It matches Qwen 2.5-14B across 15 benchmarks while running at the same speed as any 8B model.

The hybrid thinking mode is particularly useful for chat: it responds quickly in non-thinking mode for casual conversation, and you can trigger thinking mode (with `/think`) for harder questions that need reasoning. No other 8B model offers this.

```
ollama pull qwen3:8b
```

Llama 3.1 8B — The Safe Pick

If you want the model with the widest ecosystem support, most documentation, and the most third-party integrations, Llama 3.1 8B is it. The quality is slightly behind Qwen3-8B on benchmarks, but the tooling maturity makes up for it. IFEval score of 92.1% at the 70B level shows Meta's strength in instruction following — and that DNA carries down to the 8B version.

```
ollama pull llama3.1:8b
```

Gemma 2 9B — Best Prose Quality

If you care more about how the responses sound than raw benchmark scores, Gemma 2 9B produces the most natural, human-sounding prose at this tier. Google distilled knowledge from Gemini into this model, and it shows in the conversational quality. The 8K context limit is the main drawback — for longer conversations, you'll need to manage context more carefully.

```
ollama pull gemma2:9b
```

What to Expect at 8GB

Be realistic: 8B models are good for casual chat, quick Q&A, brainstorming, and simple tasks. They struggle with complex multi-step reasoning, nuanced analysis, and maintaining coherence over very long conversations. If you find yourself thinking “this is useful but not quite smart enough,” the jump to 14B is significant.

The 12-16GB Tier

This is the sweet spot for daily chat. 12GB and 16GB GPUs run 12-14B models that are noticeably smarter than 8B — the quality jump is immediately apparent in conversation.

Qwen3-14B — The Sweet Spot

Qwen3-14B matches Qwen 2.5-32B on most benchmarks. Let that sink in: a model that fits on a $200 RTX 3060 12GB performs like what previously required a 24GB card. ArenaHard score of 85.5, 128K context, hybrid thinking modes, Apache 2.0 license.

For everyday chat, this is where local AI goes from “useful tool” to “genuinely good conversational partner.”

```
ollama pull qwen3:14b
```

At Q4 on 12GB VRAM, expect ~22-30 tok/s — fast enough for interactive conversation. On 16GB VRAM, run it at Q6 for slightly better quality.

Gemma 3 12B — Multimodal on a Budget

Gemma 3 12B does something no other model at this size does: it handles both text and images. Upload a photo and ask about it. The QAT int4 version fits in just 6.6GB of VRAM, leaving plenty of room for context and even a second model.

The conversation quality is strong — Google’s knowledge distillation from Gemini gives it broader knowledge than you’d expect at 12B. 128K context window.

```
ollama pull gemma3:12b
```

Mistral Nemo 12B — Clean and Structured

Mistral Nemo was trained with quantization awareness (FP8 without quality loss), which means the quantized versions run better than typical 12B models. 128K context. The output tends to be cleaner and more structured than Qwen or Llama — good if you prefer concise, organized responses.

```
ollama pull mistral-nemo:12b
```

Phi-4 14B — Reasoning Specialist

Phi-4 punches well above its weight on math and reasoning — it outperforms Llama 3.3 70B on GPQA and MATH benchmarks. But it has a 16K context limit and weaker instruction following (lower IFEval scores). Use it when you need a thinking partner for analytical problems, not as a general chat model.

```
ollama pull phi4:14b
```

The 24GB Tier

This is where local chat gets genuinely excellent. A 24GB GPU runs 24-32B models that compete with cloud AI for most conversational tasks.

Gemma 3 27B — Highest Chat Arena Scores

Gemma 3 27B's Chatbot Arena Elo of 1338-1339 puts it ahead of DeepSeek-V3, Llama 3.1 405B, and Qwen 2.5-72B in blind human preference tests. It also ranked 2nd on EQ-Bench for creative writing. For a model you can run on a single GPU, those numbers are remarkable.

The QAT int4 version fits in 14.1GB of VRAM — well within 24GB with plenty of room for long conversations. Vision capability included (upload images and ask about them). 128K context.

```
ollama pull gemma3:27b
```

Qwen3-32B — Best All-Around

If you want one model that handles everything — chat, coding, analysis, writing — Qwen3-32B is the most versatile pick at this tier. MMLU-Pro 79.4 beats Qwen 2.5-72B (76.1). The hybrid thinking/non-thinking mode means quick responses for casual chat and deep reasoning when you need it.

Requires Q4 quantization to fit in 24GB with reasonable context. Expect ~20-30 tok/s on an RTX 3090.

```
ollama pull qwen3:32b
```

Mistral Small 3.1 (24B) — The Speed Pick

Mistral Small 3.1 is the fastest model at this tier — 30-50 tok/s quantized on an RTX 4090. MMLU 81%+, multimodal (vision), 128K context, Apache 2.0. If response speed matters more than squeezing out the last few percent of quality, this is your pick.

The June 2025 update (3.2) improved accuracy from 82.75% to 84.78% and halved the infinite generation rate.

```
ollama pull mistral-small3.1:24b
```

QwQ-32B — The Deep Thinker

QwQ-32B is the reasoning specialist at this tier. It leads on coding, math, and logic problems with scores on par with DeepSeek R1. The thinking traces are more concise than DeepSeek's (less verbose), but it's still slower for casual chat because it reasons through everything. Best used as a thinking partner for complex problems rather than a general chatbot.

```
ollama pull qwq:32b
```

The 48GB+ Tier

For dual-GPU setups, Mac Studio with 64GB+ unified memory, or dedicated AI machines.

Llama 3.3 70B — Best Instruction Following

IFEval score of 92.1% — the highest among open models. If you want a model that consistently does exactly what you ask, Llama 3.3 70B is the gold standard. Clean, controllable prose. 128K context. The widest ecosystem support of any 70B model.

Requires ~40GB at Q4_K_M. Runs at ~8-12 tok/s on an RTX 4090 (with partial offloading) or ~8.3 tok/s on an RTX 3090. On a Mac Studio with 64GB+ unified memory, performance is better.

```
ollama pull llama3.3:70b
```

Qwen 2.5 72B — Highest MT-Bench

MT-Bench score of 9.35 — the highest among open models for multi-turn conversation. Stronger on math (MATH 83.1%), multilingual support (29 languages), and structured data handling. Apache 2.0 license.

```
ollama pull qwen2.5:72b
```

A Note on MoE Models

Qwen3-30B-A3B is a Mixture of Experts model with 30B total parameters but only 3B active at a time. This means it fits in less VRAM than you'd expect while delivering 30B-level quality. Worth trying on CPU-only setups with 32GB RAM — it runs at 12-15 tok/s with only 3B active parameters doing the work.

```
ollama pull qwen3:30b-a3b
```

Chat-Specific Settings That Matter

Temperature

Temperature controls randomness. For chat, the sweet spot depends on what you want:

Use Case	Temperature	Why
Factual Q&A	0.2-0.3	Consistent, accurate, minimal variation
Everyday conversation	0.5-0.7	Natural-sounding, some personality
Creative brainstorming	0.8-1.0	More varied, surprising ideas

Set it in Ollama:

```
ollama run qwen3:8b /set parameter temperature 0.7
```

Temperature 0 doesn't make the model more accurate — it just makes it more consistent. If the model is wrong at temperature 0, it'll be confidently wrong every time.

System Prompt

A good system prompt makes a measurable difference in chat quality:

```
You are a helpful, knowledgeable conversational partner. Keep responses
concise — 2-3 sentences for simple questions, longer only when the topic
needs depth. Use natural language with contractions. Never start with
"Great question!" or filler phrases. If you don't know something, say so
directly.
```

Set it in a Modelfile:

```
FROM qwen3:14b
PARAMETER temperature 0.7
PARAMETER num_ctx 8192
SYSTEM """You are a helpful, knowledgeable conversational partner. Keep responses concise — 2-3 s
```

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

Context Length for Long Chats

Chat conversations accumulate context fast. Each message (yours and the model's) adds to the token count. When you hit the context limit, the oldest messages get dropped silently.

Set `num_ctx` based on your VRAM and how long you want conversations to last:

num_ctx	Approx. Messages	VRAM Impact (14B model)
2048	~10-15 back-and-forth	Minimal
4096	~20-30 back-and-forth	~0.5 GB
8192	~40-60 back-and-forth	~1 GB
16384	~80-120 back-and-forth	~2 GB
32768	Very long sessions	~4 GB

If you're on tight VRAM, enable KV cache quantization (`OLLAMA_KV_CACHE_TYPE=q8_0`) to roughly double the context you can fit.

Ollama Quick Start Commands

Copy-paste these to start chatting immediately:

```
# 8GB VRAM — Best overall
ollama run qwen3:8b

# 8GB VRAM — Safest pick
ollama run llama3.1:8b

# 12GB VRAM — Sweet spot for daily chat
ollama run qwen3:14b

# 12GB VRAM — Multimodal (text + images)
ollama run gemma3:12b
```

```
# 24GB VRAM — Highest chat quality
ollama run gemma3:27b

# 24GB VRAM — Best all-around
ollama run qwen3:32b

# 24GB VRAM — Fastest
ollama run mistral-small3.1:24b

# 48GB+ — Best instruction following
ollama run llama3.3:70b
```

Don't have Ollama yet? Our getting started guide walks you through installation in under 15 minutes.

The Bottom Line

VRAM	Get This	Why
8 GB	Qwen3-8B	Matches previous-gen 14B models. Fast, versatile.
12 GB	Qwen3-14B	The daily driver. Matches previous-gen 32B quality.
16 GB	Qwen3-14B at Q6	Same model, higher quality quantization.
24 GB	Gemma 3 27B or Qwen3-32B	Gemma for natural conversation, Qwen3 for versatility.
48 GB+	Llama 3.3 70B	Best instruction following, cleanest output.
CPU only	Qwen3-30B-A3B	MoE model — 30B quality with only 3B active.

The quality gap between a local chat model and ChatGPT has never been smaller. Qwen3-32B on a 24GB GPU handles everyday conversation, Q&A, brainstorming, and analysis at a level that would have required a $20/month subscription a year ago. Download one, start chatting, and see for yourself.

Source: https://insiderllm.com/guides/best-local-llms-chat-conversation/

Free guides for running AI locally