Best Local LLMs for Writing & Creative Work

January 30, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For fiction and creative writing, Qwen 2.5 32B (Q4 on 24GB VRAM) is the sweet spot — coherent prose, good pacing, follows style instructions well. If you have 8GB VRAM, Nous Hermes 3 8B punches above its weight for creative work. For blog posts and structured content, Qwen 2.5 14B on 16GB handles it well. The real quality jump happens at 32B+ parameters — that's where models shift from 'usable for writing' to 'genuinely good at it.' If you need uncensored output for fiction with dark themes, look for abliterated model variants.

📚 **More on this topic:** Best Models for Coding · Best Models Under 3B · VRAM Requirements

Cloud AI writes well, but it reads everything you write. Your novel drafts, journal entries, client work, half-formed ideas — all stored on someone else's servers. Local models let you write, brainstorm, edit, and experiment without sending a single word to the cloud.

The catch: not every local model writes well. Some produce generic, stilted prose. Others refuse to write conflict, romance, or anything remotely dark. And the difference between a 7B and a 32B model for writing quality is enormous — far bigger than for coding or Q&A tasks.

This guide covers which models actually produce good writing, organized by what you want to write and what hardware you have.

What Makes a Good Writing Model?

Writing is one of the hardest tasks for local LLMs. Unlike coding (where output is correct or isn't) or Q&A (where facts are verifiable), good writing requires:

- **Coherence over long passages** — maintaining tone, character, and narrative threads
- **Stylistic range** — matching different voices, genres, and registers
- **Instruction following** — doing what you ask without drifting
- **Not being boring** — avoiding the same safe, generic, corporate-sounding prose

Bigger models are significantly better at all of these. The quality jump from 14B to 32B for writing is more dramatic than for almost any other task. If you can run 32B, do it.

Best Models by VRAM Tier

VRAM	Model	Quant	Best For	Quality
8 GB	Nous Hermes 3 8B	Q4_K_M	Fiction, creative RP	Good for size
8 GB	Llama 3.1 8B Instruct	Q4_K_M	Blog posts, structured content	Solid all-around
8 GB	Mistral 7B Instruct	Q4_K_M	Quick drafts, brainstorming	Fast, serviceable
12 GB	MN Violet Lotus 12B	Q4_K_M	Character-driven fiction	Strong emotional intelligence
12 GB	Mistral Nemo 12B	Q4_K_M	Blog posts, editing	Clean, structured output
16 GB	Qwen 2.5 14B	Q6_K	Articles, SEO content, editing	Very good
16 GB	Qwen3-14B	Q4_K_M	Balanced creative + factual	Best value mid-range
24 GB	**Qwen 2.5 32B**	**Q4_K_M**	**Fiction, long-form, editing**	**Excellent — the sweet spot**
24 GB	DeepSeek-R1-Distill-Qwen-32B	Q4_K_M	Plotted fiction, complex narrative	Great reasoning + prose
24 GB	Mistral Small 24B	Q4_K_M	Nonfiction, editing, rewriting	Reliable, structured
48 GB+	Midnight Miqu 70B v1.5	Q4_K_M	Literary fiction, prose quality	Best local prose available
48 GB+	Llama 3.3 70B Euryale v2.3	Q4_K_M	Immersive storytelling	Vivid, descriptive

The jump that matters: 32B is where writing quality shifts from “useful assistant” to “genuinely good collaborator.” If you’re serious about using local AI for writing, a 24GB GPU running Qwen 2.5 32B is the target.

→ Check what fits your hardware with our Planning Tool.

Best for Fiction & Creative Writing

Fiction is the hardest test for a language model. It needs to maintain character voice, pace a scene, build tension, and produce prose that doesn’t sound like a corporate memo.

Top picks:

Tier	Model	Why
Best overall (if hardware allows)	Midnight Miqu 70B v1.5	Community's top pick for prose quality. "Writes like a novelist." Understands subtext, pacing, and tone in ways other models don't.
Best on 24GB	Qwen 2.5 32B	Strong coherence, follows style instructions, good at sustained narrative. Detailed and contextually aware.
Best on 24GB (plotted work)	DeepSeek-R1-Distill-Qwen-32B	The reasoning capability helps with complex plotting and maintaining story logic. Can over-think simple scenes though.
Best on 12GB	MN Violet Lotus 12B	A merge of Violet Twilight and Lumimaid. High emotional intelligence — maintains character motivations and feelings across long conversations.
Best on 8GB	Nous Hermes 3 8B	Coherent long-form output, maintains character consistency. Best fiction model at this size.

What to expect by size:

- **7-8B:** Generates readable prose but tends to rush scenes, repeat phrases, and lose track of story details after a few thousand tokens.
- **14B:** Noticeably better coherence. ~30% quality improvement over 8B for long-form text per community testing. Can maintain a scene but may struggle with complex multi-character interactions.
- **32B:** Where fiction gets genuinely good. Models understand subtext, can maintain narrative threads across chapters, and produce prose with actual stylistic variety.
- **70B:** The premium tier. Natural pacing, subtlety, sustained coherence. This is where the top community models (Midnight Miqu, Euryale) live.

Best for Blog Posts & Articles

Blog writing is more structured than fiction — you need clear sections, factual tone, and consistent formatting. The good news: smaller models handle this well because the structure does a lot of the heavy lifting.

Top picks:

VRAM	Model	Why
8 GB	Llama 3.1 8B	Clean, controllable prose. Good at following outline structures.
16 GB	Qwen 2.5 14B	Detailed, contextually complete answers. Strong instruction following.
24 GB	Qwen 2.5 32B	Best local option for researchy, detailed articles.
24 GB	Mistral Small 24B	Reliable, well-structured nonfiction. Fast.

Tips for blog writing with local models:

- Provide an outline in the prompt — models follow structure much better than they invent it
- Generate section by section, not the entire article at once
- Use a system prompt that specifies tone and audience: "Write in a direct, practical tone for technically literate readers. No filler phrases."

Best for Editing & Rewriting

Editing is harder than generating. The model needs to understand your intent, preserve what's good, and improve what isn't — without rewriting everything in its own voice.

Top picks:

VRAM	Model	Why
8 GB	Phi-4 (14B)	Excels at text tasks — rewriting, summarization, rephrasing. Fits at Q4.
16 GB	Qwen 2.5 14B	Strong instruction following. Does what you ask without going rogue.
24 GB	Mistral Small 24B	Good at targeted edits. Fast iteration.
24 GB	Qwen 2.5 32B	Best at complex editing tasks that require understanding context.

Key: For editing, instruction-following matters more than raw creativity. You want a model that can execute "rewrite this paragraph to be more concise while keeping the technical details" without deciding to restructure your entire article. Qwen models excel at this.

Best for Brainstorming & Outlining

Speed matters more than quality here. You want fast idea generation, not polished prose.

Any 7-8B model works well for brainstorming. Run it at Q4_K_M on 8GB VRAM and you'll get 30-40+ tok/s — fast enough for real-time conversation.

Good picks: Llama 3.1 8B, Mistral 7B Instruct, Qwen 2.5 7B. Don't waste 24GB of VRAM on brainstorming.

The Censorship Problem

You're writing a thriller. A character picks up a knife. The model refuses to continue because "violence."

This is the biggest frustration with local writing models. Default instruct models have safety filters that trigger on violence, romance, dark themes, morally complex characters, and sometimes even mild conflict. For fiction writing, this is crippling.

The Solution: Abliterated Models

Abliteration is the current standard for uncensored local models. Instead of retraining on "edgy" data (which degrades model intelligence), abliteration surgically removes the refusal mechanism from the model's weights. The result: same intelligence, no safety refusals.

Recommended uncensored models for writing:

Model	Size	VRAM	Notes
Eva Qwen 2.5	7B-72B	8-48GB	Uncensored Qwen variants. Good across sizes.
Dolphin 3.0	8B-70B	8-48GB	Strong conversational flow and instruction following.
Nous Hermes 3	8B	~8 GB	Creative writing focused. Coherent long-form.
Mistral Small 3.1 (abliterated)	24B	~12 GB	Budget uncensored option.
Llama 3.3 70B (abliterated)	70B	~40 GB	Full power, no filters.
Midnight Miqu 70B v1.5	70B	~40 GB	Already uncensored. Best prose quality.

Search HuggingFace for "abliterated" + your preferred model name. Most popular models have community-made abliterated variants.

Tradeoff: Abliterated models occasionally produce lower-quality output on non-creative tasks compared to their filtered counterparts. Use the filtered version for factual work, abliterated for fiction.

System Prompts That Actually Help

The right system prompt makes a measurable difference in writing quality. Here are three that work:

For fiction writing:

```
You are an experienced literary fiction author. Write vivid, emotionally
engaging prose with natural dialogue. Show, don't tell. Focus on sensory
details and character psychology. Avoid these words: tapestry, delve,
testament, beacon, journey, realm. Avoid single-sentence paragraphs.
Do not summarize emotions — show them through action and dialogue.
Do not rush to resolution. Build scenes gradually.
```

For blog/article writing:

```
You are a technical writer who explains complex topics clearly. Write
in a direct, practical tone. Lead with the useful information, not
background context. Use short paragraphs. No filler phrases like
"it's worth noting" or "in today's landscape." Be specific — use
numbers, examples, and concrete details instead of vague claims.
```

For editing:

```
You are a careful editor. When asked to edit text, preserve the
author's voice and intent. Only change what is specifically requested.
Do not add new content unless asked. Do not restructure unless asked.
Explain each change you make and why.
```

Critical tip: Repeat your most important instructions. Models deprioritize instructions over long conversations. If “show, don’t tell” matters, say it in the system prompt and again in your scene prompt.

Context Length and Long-Form Writing

Models advertise 128K context windows, but writing quality degrades well before that. Research consistently shows performance drops start at 8K-16K tokens, even when models can technically handle more.

Practical limits for writing:

Context	Pages	What Works	What Breaks
2-4K tokens	3-6 pages	Single scenes, dialogue	—
8-16K tokens	12-24 pages	Chapters, short stories	Character details start drifting
16-32K tokens	24-49 pages	Multi-chapter with summaries	Tone inconsistency, repetition
32K+ tokens	49+ pages	Possible but quality suffers	Narrative coherence degrades

The practical approach: Write chapter by chapter. Keep a running summary of characters, plot points, and tone in the system prompt. Feed the summary + current chapter into context rather than trying to keep the entire manuscript loaded. Most professional AI-assisted writers cap generation at 800-1200 words per turn for best quality.

Set `num_ctx` appropriately — don't rely on defaults. On constrained VRAM, KV cache quantization (Q8) can double your usable context with minimal quality loss.

Hardware Quick Reference

Writing Goal	Minimum VRAM	Recommended GPU	Model to Run
Brainstorming, outlining	8 GB	RTX 3060 8GB, RX 6600	Any 7-8B
Blog posts, articles	12-16 GB	RTX 3060 12GB	Qwen 2.5 14B
Serious fiction	24 GB	RTX 3090 (used, ~$750)	Qwen 2.5 32B
Best local prose	40-48 GB	Dual 3090 or Mac Studio	Midnight Miqu 70B
CPU-only option	16-32 GB RAM	—	7-14B at Q4

The Bottom Line

Local models are genuinely useful for writing now — not just as gimmicks, but as real tools. The key is matching the right model to the right task:

- **Brainstorming:** Any 7-8B model. Speed over quality.
- **Blog posts:** Qwen 2.5 14B on 16GB VRAM. Structured, reliable.

- **Fiction:** Qwen 2.5 32B on 24GB VRAM. The sweet spot where prose gets good.
- **Best prose:** Midnight Miqu 70B if you have the hardware. Nothing local comes close.
- **Uncensored:** Look for abliterated variants of whatever model you're already using.

Don't fight safety filters with clever prompting — switch to an abliterated model. Don't try to generate entire novels in one shot — work chapter by chapter. And if your writing feels generic, go bigger. The 14B-to-32B jump is where local AI stops sounding like an AI.
