Best Uncensored Local LLMs (And Why You Might Want Them)

February 10, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The best uncensored local LLM for most people is Dolphin 3.0 on Mistral 24B — strong reasoning, truly unrestricted, and fits on 24GB VRAM at Q4. For 8GB VRAM, Dolphin 3.0 Llama 3.1 8B or an abliterated Llama 3.1 8B works well. For 70B-class quality, huihui_ai's abliterated Llama 3.3 70B on Ollama is the pick. 'Uncensored' means the model won't refuse topics — it's not jailbroken, it's properly fine-tuned to remove alignment refusals. Legitimate uses include fiction with mature themes, academic research, red teaming, and creative projects where guardrails get in the way. The quality tradeoff is real but small: abliterated models lose a few points on benchmarks, and Dolphin finetunes occasionally drift from the base model's style. For most creative and research work, that tradeoff is worth it.

📚 **Related:** Best LLMs for Writing · Mistral & Mixtral Guide · Qwen Models Guide · Running AI Offline

Ask a standard instruct model to write a scene where a character gets hurt and you'll get a polite refusal. Ask it to discuss the chemistry of explosives for a novel and it'll lecture you about safety. Ask it to roleplay a villain and it'll break character to remind you that violence is wrong.

These refusals exist because instruct-tuned models are trained to refuse broad categories of content. That training is useful for customer-facing chatbots. It's maddening for fiction writers, researchers, and anyone doing creative work that involves the full range of human experience.

Uncensored models fix this. Here's what that actually means, which ones to run, and the tradeoffs you're making.

What "Uncensored" Actually Means

Standard instruct models go through alignment training: reinforcement learning from human feedback (RLHF) or similar techniques that teach the model to refuse certain categories of requests. This is where "I can't help with that" comes from.

Uncensored models have that alignment layer removed. There are two approaches:

Dataset filtering (Dolphin method). Eric Hartford, the creator of the Dolphin series, filters the training dataset to remove all refusal, avoidance, and bias examples before fine-tuning. The model never learns to refuse because the refusal examples aren't in the training data. This produces a model that's helpful on any topic without the hedging.

Abliteration. A newer technique that removes alignment from an already-trained instruct model. Researchers found that refusal behavior lives in a single direction in the model's activation space — a specific mathematical vector. Abliteration identifies that vector and removes it, producing an uncensored model without retraining. This is faster and cheaper than fine-tuning from scratch.

What uncensored does NOT mean:

- **Not "jailbroken."** Jailbreaking tricks a model into ignoring its alignment through prompt engineering. That's fragile and inconsistent. Uncensored models have the alignment genuinely removed — they're not fighting against their training.
- **Not "will help you do crimes."** The model doesn't gain new knowledge by being uncensored. It still has the same training data and the same knowledge cutoff. It won't teach you things it doesn't know.
- **Not "no limitations."** The model still hallucinates. It still has a knowledge cutoff. It still struggles with the same things any LLM struggles with. It just won't refuse to try.

Why You'd Want One

If standard models never refused your requests, uncensored models wouldn't need to exist. But alignment training is a blunt instrument. It blocks legitimate use cases alongside genuinely harmful ones.

Fiction writing. You're writing a thriller and need a character to describe a violent scene. You're writing literary fiction with sexual content. You're writing a war novel. Standard models refuse or sanitize these scenes into something unusable. Uncensored models write the scene as directed.

Academic research. Studying radicalization, analyzing propaganda techniques, researching drug chemistry for a paper — standard models treat the request as the act. Uncensored models provide information without moral commentary.

Red teaming and security testing. If you're testing AI systems for vulnerabilities, you need a model that generates adversarial content. You can't test defenses with a model that refuses to attack.

Roleplay and interactive fiction. Tabletop RPG scenarios, interactive stories, character-driven roleplay — standard models break character to insert safety disclaimers. Uncensored models stay in character.

Philosophy and ethics discussions. Try getting a standard model to steelman a morally uncomfortable position without hedging. Uncensored models engage with difficult ideas directly, which is what philosophy actually requires.

Creative freedom. Your hardware, your model, your rules. Some people just don't want a corporation's content policy deciding what their local AI will discuss.

The Best Uncensored Models

Dolphin Series (Eric Hartford) — The Standard

Eric Hartford's Dolphin models are the most established uncensored finetunes. They've been running since the Llama 2 era and cover most major base model families. Available on HuggingFace and directly through Ollama.

Model	Base	Size	VRAM (Q4)	Best For
Dolphin 3.0 Mistral 24B	Mistral 24B	24B	~16GB	All-around best, coding + chat
Dolphin 3.0 R1 Mistral 24B	Mistral 24B	24B	~16GB	Reasoning-heavy tasks
Dolphin 3.0 Llama 3.1 8B	Llama 3.1	8B	~5GB	Budget hardware, 128K context
Dolphin 3.0 Qwen 2.5 3B	Qwen 2.5	3B	~2.5GB	Minimal hardware, fast responses
Dolphin 2.9.2 Qwen2 72B	Qwen 2	72B	~42GB	Maximum quality
Dolphin 2.5 Mixtral 8x7B	Mixtral	8x7B	~26GB	MoE, good at coding

The Dolphin 3.0 Mistral 24B is the current pick if you have 24GB VRAM. It handles coding, reasoning, creative writing, and general chat well, and it's truly unrestricted. The "Venice Edition" co-release is specifically tuned for minimum filtering.

The R1 variant adds reasoning traces — trained on 800K reasoning examples from the Dolphin-R1 dataset. Use it when you need step-by-step problem solving without refusals.

```
# Run Dolphin on Ollama
ollama run dolphin-llama3
```

```
ollama run dolphin-mixtral

# Dolphin 3.0 (check Ollama library for latest tags)
ollama run dolphin3:8b-llama3.1-q4_K_M
```

Abliterated Models — Uncensored Without Retraining

Abliteration takes an existing instruct model and removes the refusal direction from its weights. The advantage: you get the exact same model with the exact same capabilities, minus the refusals. No new training data, no capability drift from fine-tuning.

Key creators:

- **failspy** — Pioneered abliteration tooling, created abliterated versions of Llama 3 8B and 70B
- **mlabonne (Maxime Labonne)** — Wrote the definitive abliteration tutorial, maintains a collection of abliterated models on HuggingFace
- **huihui_ai** — Maintains abliterated versions on Ollama, including Llama 3.3 70B and DeepSeek R1

Model	Base	Size	VRAM (Q4)	Source
Llama 3.1 8B Instruct abliterated	Llama 3.1	8B	~5GB	mlabonne, failspy
Llama 3.3 70B abliterated	Llama 3.3	70B	~40GB	huihui_ai (Ollama)
Lexi-Llama-3.1 8B Uncensored	Llama 3.1	8B	~5GB	Orenguteng
DeepSeek R1 abliterated	DeepSeek R1	Various	Varies	huihui_ai (Ollama)
QwQ 32B abliterated	QwQ	32B	~20GB	bartowski (GGUF)

```
# Abliterated models on Ollama (community uploads)
ollama run huihui_ai/llama3.3-abliterated
ollama run huihui_ai/deepseek-r1-abliterated

# Or download GGUF from HuggingFace and load in Ollama/LM Studio
```

Mistral — Less Filtered by Default

Mistral models ship with lighter safety restrictions than Llama. The base Mistral 7B and Nemo 12B are less likely to refuse creative writing requests, controversial topics, or edgy roleplay. You may not need an uncensored finetune at all if your needs are moderate.

For fully unrestricted Mistral, Dolphin 3.0 Mistral 24B is the go-to. For the 7B tier, Dolphin 2.8 Mistral 7B works on 8GB VRAM.

Qwen — It's Complicated

Qwen has a reputation for being "less censored" than Llama, but the reality is more nuanced. Qwen models have sophisticated political alignment — they give curated answers on China-related topics (Taiwan, Tiananmen, Xinjiang) and vary their responses by language, with fewer refusals in Chinese than English.

For general creative content, Qwen is indeed more permissive than Llama out of the box. For political or historical topics involving China, it's more restricted. A Berkeley study documented this pattern in detail.

Uncensored Qwen finetunes exist — Dolphin 2.9.2 Qwen2 72B and Orion-zhen's Qwen2.5-7B-Instruct-Uncensored — but they remove the creative refusals, not necessarily the political ones embedded deeper in the base weights.

WizardLM Uncensored — The Legacy Pick

WizardLM Uncensored dates back to the Llama 2 era. It's available in 7B and 13B on Ollama and still gets pulled regularly. For creative writing on low VRAM, the 13B variant is adequate.

But it's showing its age. Newer uncensored models (Dolphin 3.0, abliterated Llama 3.x) are significantly better at following instructions, maintaining coherence, and producing quality output. Use WizardLM Uncensored if you're on very constrained hardware and need something that just works. Otherwise, go with Dolphin or an abliterated model.

```
ollama run wizardlm-uncensored       # 7B
ollama run wizardlm-uncensored:13b   # 13B
```

VRAM Requirements

Uncensored models have identical VRAM requirements to their base models — removing alignment doesn't change the parameter count or architecture. If you can run Llama 3.1 8B, you can run the abliterated version.

Size	Q4_K_M	Q8_0	FP16	Min GPU
3B	~2.5GB	~4GB	~6GB	Any 4GB+ card
7-8B	~5GB	~9GB	~16GB	RTX 3060 (8GB)
12-14B	~9GB	~15GB	~28GB	RTX 3060 12GB
24B	~16GB	~26GB	~48GB	RTX 3090/4090
70-72B	~40GB	~72GB	~140GB	2x RTX 3090

The VRAM requirements guide covers this in detail. For uncensored specifically:

- **8GB VRAM:** Dolphin 3.0 Llama 3.1 8B (Q4) or abliterated Llama 3.1 8B
- **12GB VRAM:** Dolphin 2.8 Mistral 7B (Q8 for better quality) or abliterated Nemo 12B (Q4)
- **16GB VRAM:** Dolphin 3.0 Mistral 24B (Q4) — the sweet spot
- **24GB VRAM:** Dolphin 3.0 Mistral 24B (Q6 or Q8) or abliterated QwQ 32B (Q4)
- **48GB+ VRAM:** Dolphin 2.9.2 Qwen2 72B or abliterated Llama 3.3 70B

Where to Find Them

Ollama (Easiest)

Search the Ollama library for "uncensored" or browse community uploads for "abliterated":

```
# Official library models
ollama run dolphin-llama3        # Dolphin 2.9, Llama 3 8B
ollama run dolphin-mixtral       # Dolphin, Mixtral 8x7B
ollama run dolphin3              # Dolphin 3.0, Llama 3.1 8B
ollama run wizardlm-uncensored   # WizardLM 7B
ollama run nous-hermes3          # Less filtered, good for creative work

# Community models (huihui_ai)
ollama run huihui_ai/llama3.3-abliterated
```

```
ollama run huihui_ai/deepseek-r1-abliterated
ollama run huihui_ai/dolphin3-abliterated
```

HuggingFace (More Options)

Search for "uncensored" or "abliterated" and filter by GGUF format. The key quantizers to look for:

- **bartowski** — The current go-to for high-quality GGUF quants (successor to TheBloke). Uses imatrix quantization for better quality at low bit rates.
- **mradermacher** — Another reliable quantizer with extensive coverage. Offers both static and imatrix quants.
- **TheBloke** — No longer actively uploading, but older models are still available and widely used.

Download a GGUF file, then load it in LM Studio or import it into Ollama:

```
# Create a Modelfile for Ollama
echo 'FROM ./model-name.Q4_K_M.gguf' > Modelfile
ollama create my-uncensored-model -f Modelfile
ollama run my-uncensored-model
```

The Quality Tradeoff

Uncensored models aren't free lunch. Removing alignment has a cost.

Abliteration preserves most capability because it targets a narrow mathematical direction in the model's activation space. On normal, non-refused prompts, the model behaves identically. But on edge cases — prompts that are near the refusal boundary — you may see slightly degraded coherence. Maxime Labonne's tutorial demonstrates a "healing" step that recovers some of this loss.

Dataset-filtered finetunes (Dolphin) can drift from the base model's behavior. The training data is different, so the model's style, instruction following, and knowledge can shift. Dolphin models are generally strong, but they're not identical to running the base instruct model with refusals removed. They're a different model.

Practical impact:

- For creative writing: the tradeoff is almost always worth it. The quality loss is minor, and the freedom to write without refusals is major.
- For coding: use the base instruct model unless it's refusing something specific. Dolphin finetunes are decent at code but not as polished as purpose-built coding models.
- For reasoning: abliterated versions of reasoning models (DeepSeek R1, QwQ) preserve reasoning quality well. The Dolphin R1 Mistral 24B variant adds reasoning without sacrificing the uncensored behavior.

Test before committing. Run the same prompts through the base instruct model and the uncensored variant. If the uncensored version handles your actual workflow well, use it. If you notice quality drops on the tasks you care about, try a different uncensored variant or a different technique (Dolphin vs. abliterated).

The Ethics of Uncensored Models

This is straightforward: you own your hardware, and you control what runs on it.

When Meta or Mistral release an open-weight model, they're giving you the weights. What you do with those weights on your own machine is your decision. Fine-tuning to remove alignment is no different from fine-tuning for any other purpose — it's using the model as the creators licensed you to use it.

The alignment in standard models reflects one company's content policy, designed for their liability concerns when serving millions of users through a web API. Those concerns don't apply to you running a model on your own hardware for your own work. A fiction writer doesn't need OpenAI's content policy. A security researcher doesn't need Meta's refusal training. A philosopher doesn't need Google's hedging.

With that freedom comes responsibility. An uncensored model will do what you ask. It won't remind you about ethics, suggest you reconsider, or refuse to engage. That means the judgment call is entirely yours. For most people running local AI for creative and research work, that's exactly how it should be.

What Uncensored Won't Do

A few misconceptions worth clearing up:

It won't teach you things the model doesn't know. Removing alignment doesn't add knowledge. If the model's training data didn't include specific synthesis procedures, classified information, or working exploit code, the uncensored version doesn't have it either. It just won't refuse to discuss the topic in general terms.

It won't access the real world. An uncensored model can't hack systems, access databases, or interact with anything outside its context window. It generates text. That's all it does.

It won't be consistently "better." On most everyday tasks — summarization, Q&A, coding, analysis — the standard instruct model performs the same or slightly better because its training is more polished. You only benefit from uncensored when standard models refuse what you need.

It won't bypass the model's actual limits. Hallucination, knowledge cutoff, context window, reasoning failures — these are all still present. Uncensored models are honest about not knowing things instead of refusing to engage, which can actually make it easier to spot when they're making things up.

Recommended Setups

8GB VRAM — Budget creative writing:

```
ollama run dolphin3:8b-llama3.1-q4_K_M
```

Dolphin 3.0 on Llama 3.1 8B. Handles fiction, roleplay, and general creative work. 128K context window for longer documents.

16GB VRAM — The sweet spot:

```
# Download GGUF from bartowski on HuggingFace
# Load in LM Studio or create Ollama Modelfile
```

Dolphin 3.0 Mistral 24B at Q4_K_M. Best all-around uncensored model. Good at coding, reasoning, and creative work.

24GB VRAM — High quality: Run the Mistral 24B at Q6 or Q8 for better output quality, or step up to abliterated QwQ 32B for reasoning tasks.

48GB+ — No compromises:

```
ollama run huihui_ai/llama3.3-abliterated
```

Abliterated Llama 3.3 70B. Near-frontier quality with zero refusals.

The Bottom Line

Uncensored local models exist because alignment training blocks legitimate creative and research work alongside genuinely harmful content. If standard models never refuse your requests, you don't need these. If they do, Dolphin 3.0 Mistral 24B is the current best option for most VRAM configurations, and abliterated versions of Llama 3.3 and DeepSeek R1 give you frontier-class models without restrictions.

The privacy advantages of local AI get even stronger with uncensored models — there's no content policy log somewhere tracking what you generated. Your creative work stays on your machine, unrestricted and unmonitored.

Start with `ollama run dolphin-llama3` to see how it feels. If you want more capability, step up to Dolphin 3.0 Mistral 24B or grab an abliterated model from huihui_ai on Ollama.

Related Guides

- Best LLMs for Writing & Creative Work
- Mistral & Mixtral Guide
- Qwen Models Guide
- VRAM Requirements for Local LLMs
- Running AI Offline
- Local AI Privacy Guide
- Quantization Explained
- Model Formats: GGUF vs GPTQ vs AWQ
- Local AI Planning Tool — VRAM Calculator

Get notified when we publish new guides.
