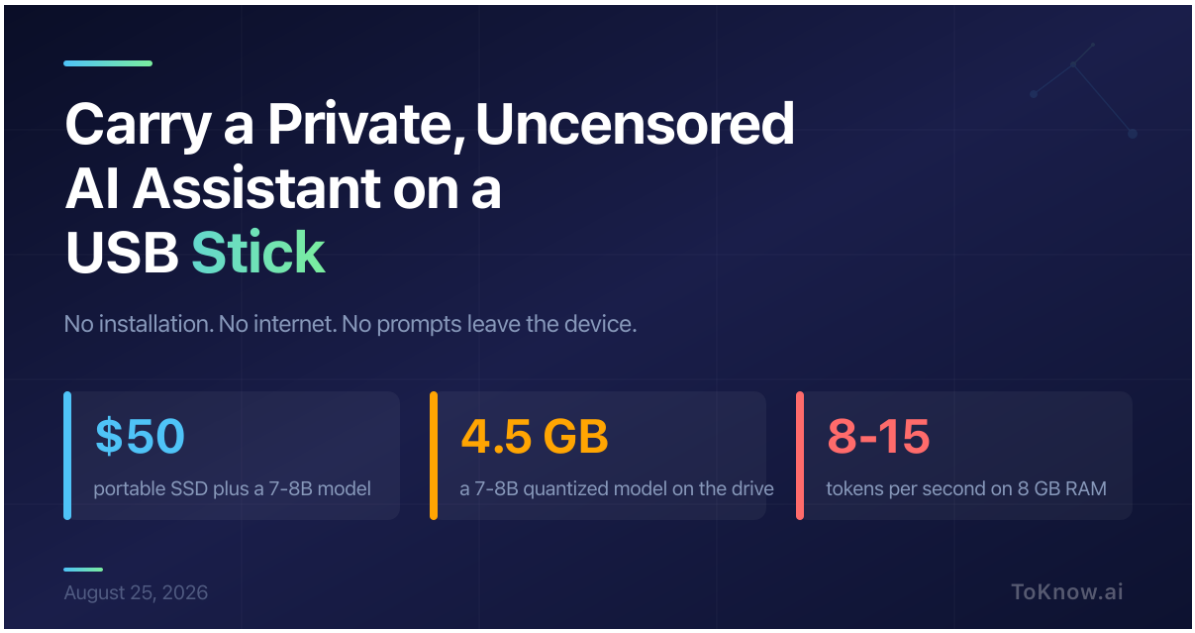
Carry a Private, Uncensored AI Assistant on a USB Stick

Kabui, Charles

2026-08-25

Read at ToKnow.ai



USB-Uncensored-LLM is a portable, air-gapped local AI environment that runs large language models straight from a USB drive or external SSD, with no installation and no internet. It ships its own portable Python and a custom Ollama engine, so nothing is installed on the host machine: plug in the drive, run one script, and a chat window opens in the browser. A shared volume keeps model weights in one place across Windows, macOS, and Linux, and the

launcher picks the right acceleration, CUDA, Metal, or CPU. The uncensored label means the bundled models are abliterated fine-tunes, such as a 1.6 GB Gemma 2 2B or a 5.3 GB Qwen 3.5 9B, with safety-refusal vectors removed. The original repository was archived in July 2026, and development moved to the MIT-licensed Uncensored-Local-Studio, which adds image generation, speech-to-text, and text-to-speech.

It is for anyone handling sensitive documents, working behind a firewall, or in a place where cloud AI is blocked. The model never phones home: no accounts, no telemetry, no prompt leaves the device. A portable SSD with a 7 to 8 billion parameter model costs under $50, and that 4.5 GB model runs at 8 to 15 tokens per second on a laptop with 8 GB of RAM. The trade-off is trust: abliterated weights are edited to stop refusing prompts, so you rely on whoever built them, and there is no sandbox beyond the code you run.

The point is ownership. AI on a stick cannot be switched off, filtered, or watched by anyone else, the same argument as running local models as a hedge against remote shutdown in its most portable form.

Sources:

- USB-Uncensored-LLM repository and README (GitHub)
- Uncensored-Local-Studio, the successor project (GitHub)
- Portable LLM on a USB Stick: Offline AI Setup (kunalganglani.com)
- USB-Uncensored-LLM overview (DeepWiki)
