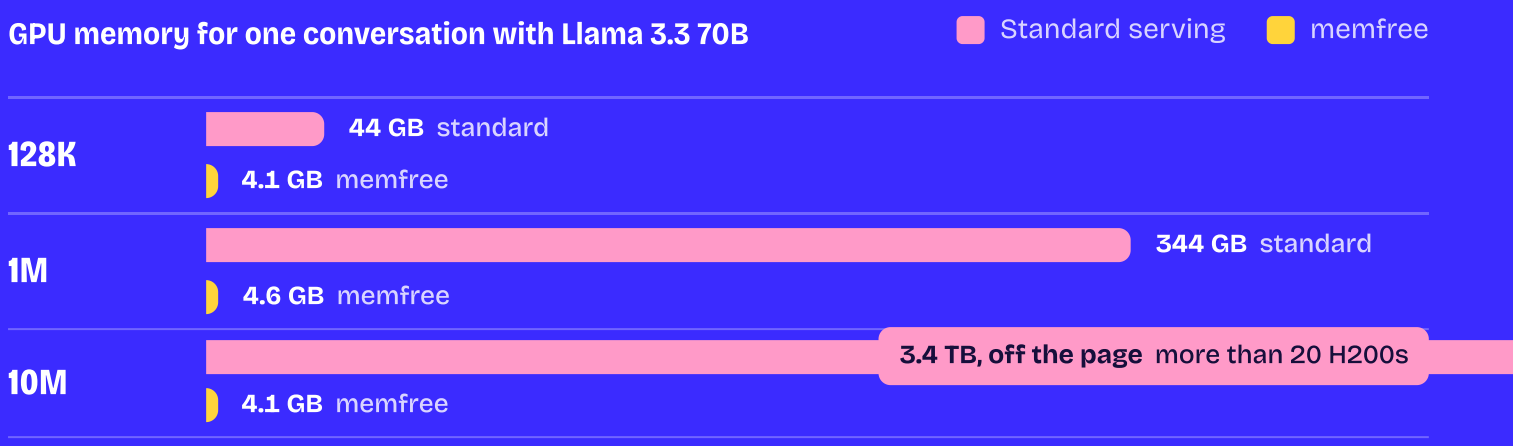
memfree

memfree.io

Ten million tokens. Four gigabytes.

memfree is an add-on for vLLM that serves conversations of a million tokens and more in about **4 GB of GPU memory each**. Your customers get long context. Your cards hold 13 times as many conversations.

GPU memory for one conversation with Llama 3.3 70B



Drawn to scale. memfree's memory and the standard figure at 128K measured on one NVIDIA H200, 7 November 2026; the standard figures at 1M and 10M are the model's own cache, worked out, because no single card can hold them.

Measured on one H200, side by side

	Standard serving	memfree
Right answers at 128K tokens	45 of 48	48 of 48
Right answers at 1M tokens	Does not fit on the card	48 of 48
Right answers at 10M tokens	Does not fit on the card	48 of 48
GPU memory per conversation	44 GB at 128K	About 4 GB, any length
Conversations per H200 at 128K	2	26
Card time to make a text answerable	207 s for 128K	17 s, up to 10M

NVIDIA H200, Llama 3.3 70B Instruct, the same 48 questions at every length, 7 November 2026.

Memory stays flat

About 4 GB per conversation from 128 thousand tokens to 10 million.

Answers stay right

48 of 48 at a million and at ten million tokens on one H200.

Your stack stays yours

Runs in front of vLLM on your machines. vLLM, your models and your data stay where they are.

Run a pilot on your own cards

omar@memfree.io memfree.io/pilot.html

Measurements use Llama 3.3 70B Instruct.
Built with Llama.