Models by RAM

I will be using the new QAT variants of Gemma 4. These are aledgedly the bee's knee's and I have been meaning to give them a go. They are designed to reduce memory requirements while preserving model quality.

More info can be found here: QAT

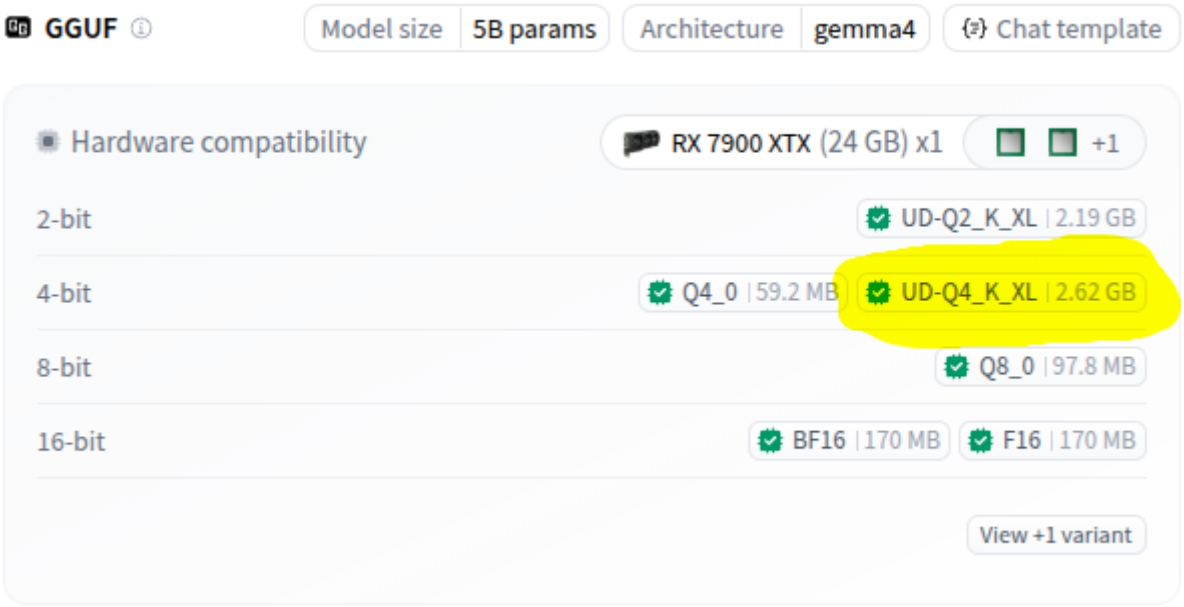
Property	E2B	E4B	12B Unified	31B Dense
Total Parameters	2.3B effective (5.1B with embeddings)	4.5B effective (8B with embeddings)	11.95B	30.7B
Layers	35	42	48	60
Sliding Window	512 tokens	512 tokens	1024 tokens	1024 tokens
Context Length	128K tokens	128K tokens	256K tokens	256K tokens
Vocabulary Size	262K	262K	262K	262K
Supported Modalities	Text, Image, Audio	Text, Image, Audio	Text, Image, Audio	Text, Image
Vision Encoder Parameters	*~150M*	*~150M*	-	*~550M*
Audio Encoder Parameters	*~300M*	*~300M*	-	No Audio

4GB

E2B

E2B is the model designed for mobile devices. "2.62GB"

https://huggingface.co/unsloth/gemma-4-E2B-it-qat-GGUF



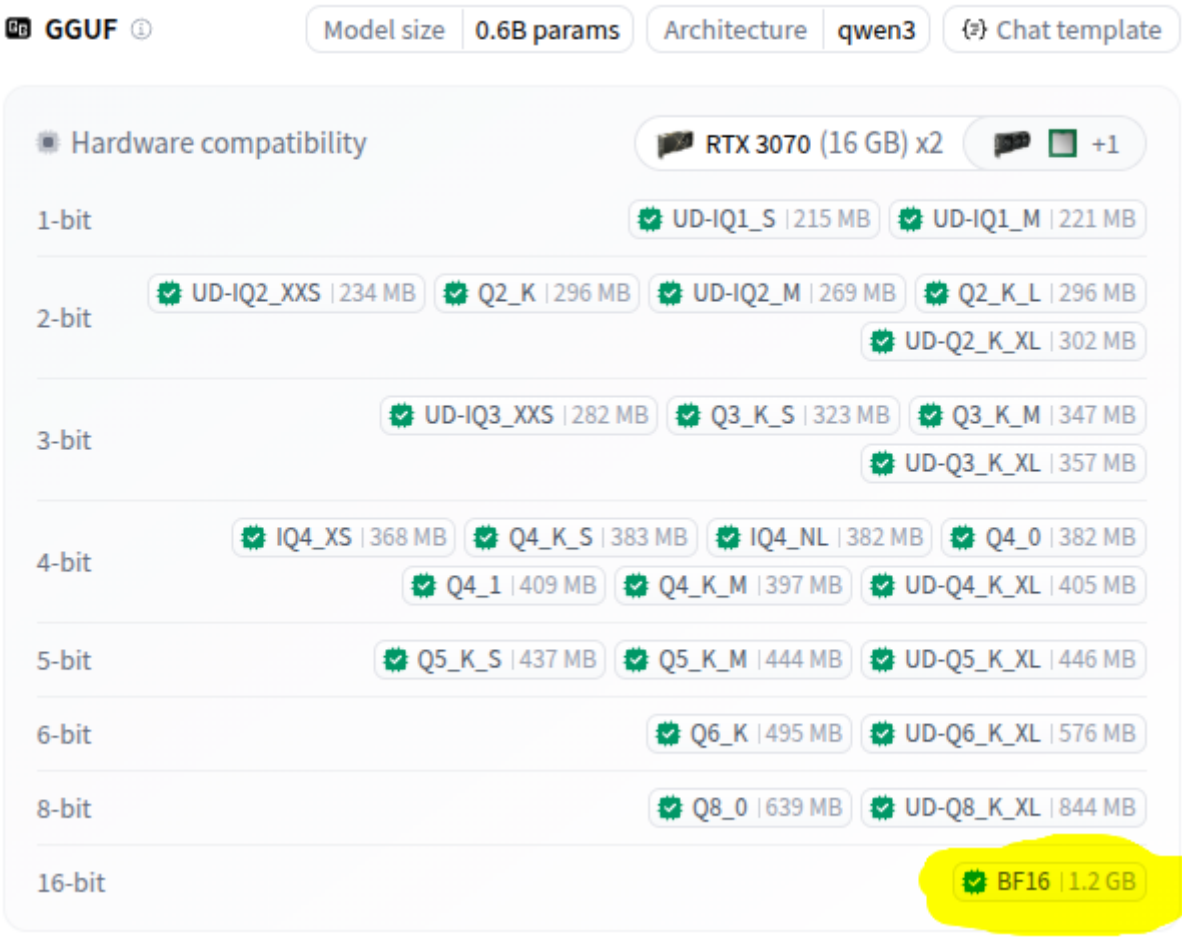
If 4GB is not enough to run E2B I have not tested! Qwen3-0.6B is considered the stepping stone model into local AI.

Qwen3-0.6B

The Intro to AI special, this model has 1.3 million downloads in the past month and can run on a toaster.

This is the only model outside of the QAT family I would suggest for this as I know it will work.

https://huggingface.co/unsloth/Qwen3-0.6B-GGUF

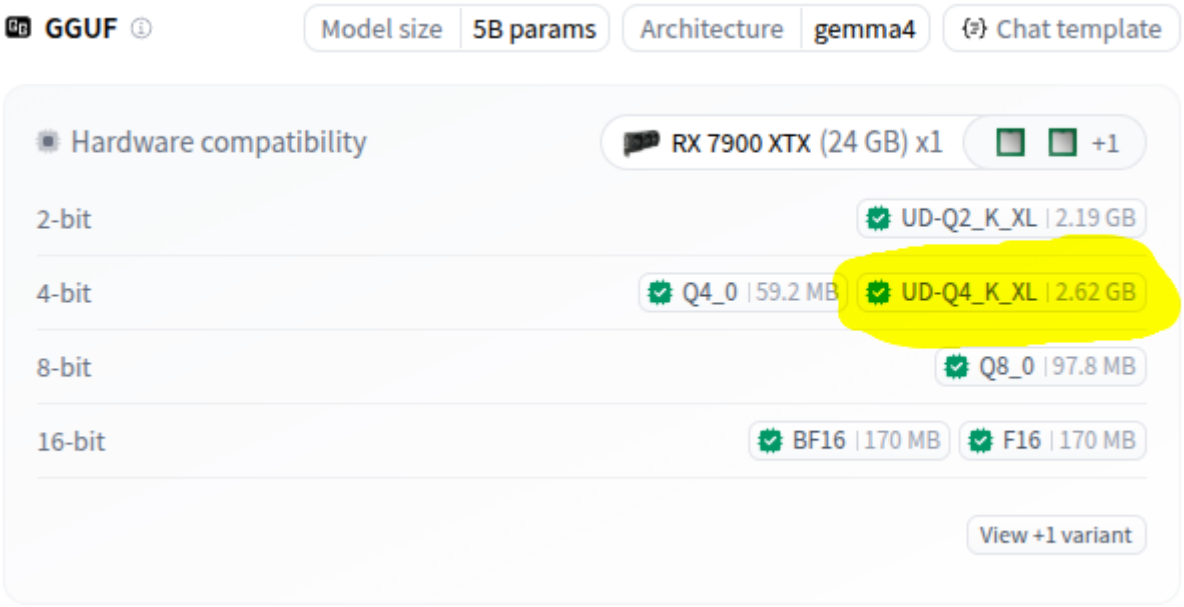


8GB

E2B Again...

Probably still a good choice if you have other services running.

https://huggingface.co/unsloth/gemma-4-E2B-it-qat-GGUF

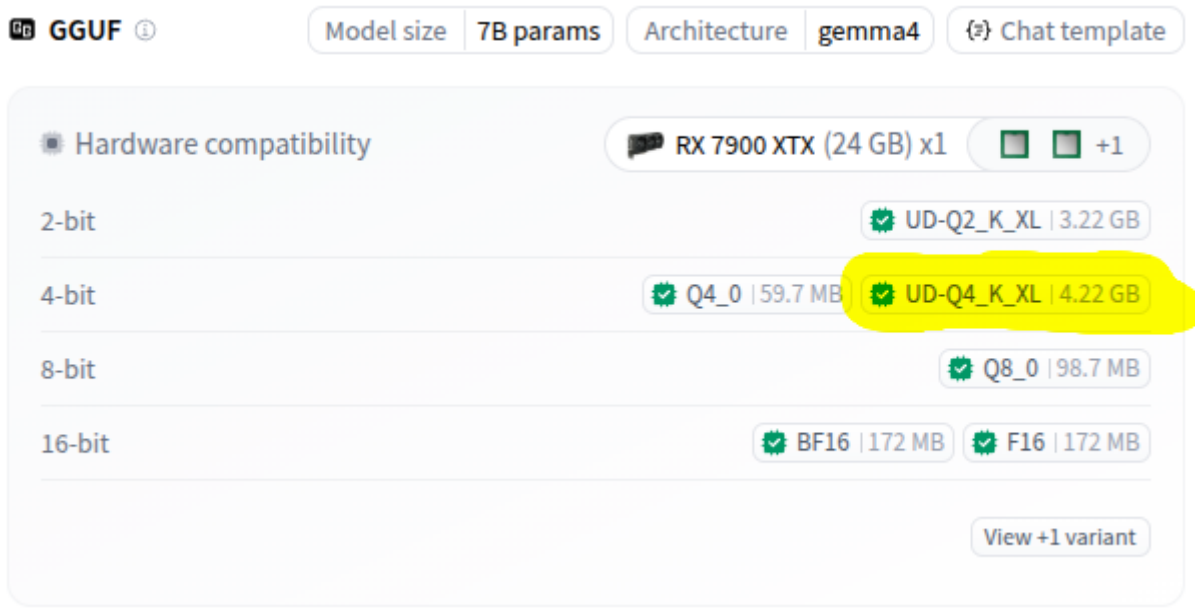


I would suggest the E4B if you have the spare resources.

E4B

I think this one could be very useful. I may even use it so I can run other stuff in parallel.

https://huggingface.co/unsloth/gemma-4-E4B-it-qat-GGUF



16GB ram

Gemma-4 12B

This is the unified version of the above models and what I will be using.

https://huggingface.co/unsloth/gemma-4-12B-it-qat-GGUF