

# Models by RAM

I will be using the new QAT variants of Gemma 4. These are alledgedly 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
<b>Total Parameters</b>	2.3B effective (5.1B with embeddings)	4.5B effective (8B with embeddings)	11.95B	30.7B
<b>Layers</b>	35	42	48	60
<b>Sliding Window</b>	512 tokens	512 tokens	1024 tokens	1024 tokens
<b>Context Length</b>	128K tokens	128K tokens	256K tokens	256K tokens
<b>Vocabulary Size</b>	262K	262K	262K	262K
<b>Supported Modalities</b>	Text, Image, Audio	Text, Image, Audio	Text, Image, Audio	Text, Image
<b>Vision Encoder Parameters</b>	~150M	~150M	-	~550M
<b>Audio Encoder Parameters</b>	~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>

**GGUF** ⓘ

Model size | 5B params | Architecture | gemma4 | Chat template

Hardware compatibility

RX 7900 XTX (24 GB) x1 +1

2-bit	UD-Q2_K_XL   2.19 GB
4-bit	Q4_0   59.2 MB   <b>UD-Q4_K_XL   2.62 GB</b>
8-bit	Q8_0   97.8 MB
16-bit	BF16   170 MB   F16   170 MB

View +1 variant



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>

 GGUF 


Model size

0.6B params

Architecture

qwen3


 Chat template


 Hardware compatibility

 RTX 3070 (16 GB) x2


  +1


1-bit


 UD-IQ1\_S | 215 MB


 UD-IQ1\_M | 221 MB


2-bit

 UD-IQ2\_XXS | 234 MB

 Q2\_K | 296 MB


 UD-IQ2\_M | 269 MB


 Q2\_K\_L | 296 MB


 UD-Q2\_K\_XL | 302 MB

3-bit


 UD-IQ3\_XXS | 282 MB


 Q3\_K\_S | 323 MB

 Q3\_K\_M | 347 MB

 UD-Q3\_K\_XL | 357 MB


4-bit


 IQ4\_XS | 368 MB


 Q4\_K\_S | 383 MB

 IQ4\_NL | 382 MB


 Q4\_0 | 382 MB


 Q4\_1 | 409 MB


 Q4\_K\_M | 397 MB

 UD-Q4\_K\_XL | 405 MB


5-bit

 Q5\_K\_S | 437 MB

 Q5\_K\_M | 444 MB

 UD-Q5\_K\_XL | 446 MB

6-bit

 Q6\_K | 495 MB


 UD-Q6\_K\_XL | 576 MB

8-bit

 Q8\_0 | 639 MB

 UD-Q8\_K\_XL | 844 MB

16-bit

 BF16 | 1.2 GB

# 8GB

## E2B Again...

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

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

**GGUF** ⓘ

Model size | 5B params | Architecture | gemma4 | Chat template

Hardware compatibility

RX 7900 XTX (24 GB) x1 +1

2-bit	UD-Q2_K_XL   2.19 GB	
4-bit	Q4_0   59.2 MB	UD-Q4_K_XL   2.62 GB
8-bit	Q8_0   97.8 MB	
16-bit	BF16   170 MB	F16   170 MB

View +1 variant

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>

**GGUF** ⓘ

Model size 7B params Architecture gemma4 (#) Chat template

Hardware compatibility  RX 7900 XTX (24 GB) x1   +1

2-bit	 UD-Q2_K_XL   3.22 GB
4-bit	 Q4_0   59.7 MB  UD-Q4_K_XL   4.22 GB
8-bit	 Q8_0   98.7 MB
16-bit	 BF16   172 MB  F16   172 MB

View +1 variant

# 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>

**GGUF** ⓘ

Model size | 12B params | Architecture | gemma4 | Chat template

Hardware compatibility

RX 7900 XTX (24 GB) x1 +1

4-bit

Q4\_0 | 254 MB | **UD-Q4\_K\_XL | 6.72 GB**

8-bit

Q8\_0 | 465 MB

16-bit

BF16 | 862 MB | F16 | 862 MB

View +1 variant