

Will it run? The formulas.

1

WEIGHTS

$$\text{GB} = \text{B params} \times \text{bytes per param}$$

Params are in the name: 8B = 8 billion. It's a floor, so check the real file size.

2

KV CACHE

$$2 \times \text{layers} \times \text{KV heads} \times \text{head dim} \times \text{tokens} \times \text{bytes}$$

The conversation's memory, growing with every token. Values from config.json; bytes = **2** at 16-bit, **1** at 8-bit. Skip layers that reuse KV.

3

FIT

$$\text{weights} + \text{KV} + 1.5 \text{ GB} \leq \text{usable memory}$$

1.5 GB covers runtime and buffers. Usable = GPU **VRAM**, system **RAM** minus OS and apps, or **~70–75%** of a Mac's unified memory.

4

SPEED

$$\text{tok/s} \approx 0.6 \times \text{bandwidth} \div \text{GB read per token}$$

Each token reads all weights once. Memory bandwidth, not compute, sets the pace.

5

MOE

$$\text{memory} \leftarrow \text{total params} \cdot \text{speed} \leftarrow \text{active params}$$

Mixture-of-Experts: every expert sits in memory, but each token reads only the active ones. A **30B-A3B** needs 30B of memory, runs at ~3B speed.

6

PROMPT

$$\text{seconds} \approx 2 \times \text{params} \times \text{prompt tokens} \div \text{FLOP/s}$$

Prompts are processed in parallel, so compute is the limit. Use achievable FLOP/s, ~30–50% of spec. Why GPUs read long documents **10×+** faster.

BYTES PER PARAM

16-bit (FP16)	2
8-bit	~1
6-bit	~0.8
4-bit (default choice)	~0.6

MODEL SIZES AT 4-BIT

MODEL	SIZE · WITH 8K CTX
3B	~2 · ~4 GB
8B	~5 · ~7.5 GB
14B	~9 · ~11 GB
32B	~20 · ~23 GB
70B	~42 · ~46 GB

Context column includes the 1.5 GB headroom.

KV CACHE · ~8B MODEL

8K tokens	~1 GB
32K tokens	~4 GB
128K tokens	~16 GB

Bigger models need more. 8-bit KV halves it.

USABLE MEMORY

Graphics card	VRAM - headroom
PC without GPU	RAM - OS/apps
Mac	~70–75% unified

BANDWIDTH (GB/S)

DDR4 RAM	~40–50
DDR5 RAM	~60–90
Apple base	~100–120
Apple Pro	~200–270
Apple Max	~400–550
RTX 4060 Ti	~290
RTX 3090 / 4090	~950–1,000

SPEED FEEL (TOK/S)

< 3 · painful
5–10 · reading speed
20+ · comfortable
50+ · instant

Worked example: 8B at 4-bit ≈ 5 GB; on 100 GB/s, $0.6 \times 100 \div 5 \approx 12$ tok/s.

DOESN'T FIT?

1 Lower precision

2 Smaller model

3 Shorter context or 8-bit KV cache

4 Partial offload to RAM (PC only; slower)