**Intel Core i7 12700H Laptop**

BUG on CPU with ollama 0.6.6 :

source=ggml.go:152 msg="key not found" key=general.alignment default=32

<https://github.com/ollama/ollama/issues/10503>

gemma3:4b - 3-6 tokens/sec (5.7 GB)

gemma3:1b - 10-28 tokens/sec (1.5 GB)

llama3.2:3b - 10-13 tokens/sec (3.5 GB)

llama3.2:1b - 10-25 tokens/sec (2.2 GB)

qwen3:1.7b - 8-9 tokens/sec unreliable (2.5 GB)

qwen3:0.6b - 8-15 tokens/sec unreliable (1.7 GB)

phi4-mini:3.8b - 8 tokens/sec (4.2 GB)

granite3.3:2b - 10 tokens/sec (2.7 GB)

**RTX 3070ti Laptop**

ollama 0.6.6

ctx-size 8192

Google

hf.co/unsloth/gemma-3-12b-it-GGUF:Q3\_K\_XL

gemma3:4b - 85 tokens/sec (5.7 GB) 3.3 GB

gemma3:1b - 160 tokens/sec (1.9 GB) 0.8 GB

(hf.co/unsloth/gemma-3-12b-it-GGUF:Q3\_K\_XL - 5 tokens/sec - 25%/75% CPU/GPU)

(gemma3:12b - too big- 40%/60% CPU/GPU)

Meta

llama3.1:8b - 60 tokens/sec (6.9 GB) 4.9 GB

llama3.2:3b - 105 tokens/sec (4.0 GB) 2.0 GB

llama3.2:1b - 160 tokens/sec (2.7 GB) 1.3 GB

Alibaba

qwen3:8b - 50 tokens /sec (6.0 GB) 5.2 GB

qwen3:4b - 85 tokens /sec (5.2 GB) 2.6 GB

qwen3:1.7b - 145 tokens /sec (2.0 GB) 1.4GB

qwen3:0.6b - 200 tokens /sec (2.2 GB) 0.5 GB

(hf.co/unsloth/Qwen3-14B-GGUF:Q3\_K\_M - 10 tokens/sec - 12%/88% CPU/GPU)

(qwen3:14b - too big- 28%/72% CPU/GPU)

Microsoft

phi4-mini:3.8b - 85 tokens/sec (4.7 GB) 2.5 GB

IBM

granite3.3:8b - 50 tokens/sec (6.1 GB) 4.9 GB

granite3.3:2b - 115 tokens/sec (3.2 GB) 1.5 GB

**RTX 4090 desktop**

Google

gemma3:27b - 40 tokens/sec (22 GB) 17 GB

gemma3:12b - 75 tokens/sec (12 GB) 8.1 GB

gemma3:4b - 150 tokens/sec (6.2 GB) 3.3 GB

gemma3:1b - 265 tokens/sec (1.9 GB) 0.8 GB

Mistral

mistral-small3.1:24b - 15 tokens/sec (26 GB 9%/91% CPU/GPU) 15 GB

+ num\_gpu 100 (open webui controls, on the top right of the screen)

mistral-small3.1:24b - 50 tokens/sec (26 GB 9%/91% CPU/GPU) 15 GB

Meta

llama3.1:8b - 125 tokens/sec (6.9 GB) 4.9 GB

llama3.2:3b - 205 tokens/sec (4.0 GB) 2.0 GB

llama3.2:1b - 300 tokens/sec (2.7 GB) 1.3 GB

Alibaba

hf.co/unsloth/Qwen3-32B-GGUF:Q4\_K\_M - 35 tokens/sec (21 GB) 19 GB

hf.co/unsloth/Qwen3-32B-GGUF:Q3\_K\_M - 40 tokens/sec (17 GB) 15 GB

qwen3:14b - 75 tokens/sec (12 GB) 9.3 GB

qwen3:8b - 115 tokens /sec (7.5 GB) 5.2 GB

qwen3:4b - 165 tokens /sec (5.2 GB) 2.6 GB

qwen3:1.7b - 265 tokens /sec (2.0 GB) 1.4GB

qwen3:0.6b - 340 tokens /sec (1.2 GB) 0.5 GB

qwen3:30b - 12 tokens/sec (21 GB) 18 GB

=> [Qwen3 MoE 30b-a3b, poor performance and Low GPU utilization issue · Issue #10458 · ollama/ollama](https://github.com/ollama/ollama/issues/10458)

=> with ollama v0.6.8 : 138 tokens/sec

Microsoft

phi4:14b - 75 tokens/sec (10 GB) 9.1 GB

phi4-mini:3.8b - 165 tokens/sec (3.4 GB) 2.5 GB

IBM

granite3.3:8b - 115 tokens/sec (6.1 GB) 4.9 GB

granite3.3:2b - 205 tokens/sec (2.3 GB) 1.5 GB

**RTX 5090 Runpod**

mistral-small3.1:24b - 86 tokens/sec

gemma3:27b - 61 tokens/sec

gemma3:12b - 107 tokens/sec

gemma3:4b - 205 tokens/sec

gemma3:1b - 322 tokens/sec

hf.co/unsloth/Qwen3-32B-GGUF:Q4\_K\_M - 50 tokens/sec

**B200 Runpod**

mistral-small3.1:24b - 92 tokens/sec

gemma3:27b - 47 tokens/sec