Bots

Llama-3.1-Nemotron-70B-Instruct-HF-IQ2_M

- Path: “mradermacher/Llama-3.3-70B-Instruct-i1-GGUF/Llama-3.3-70B-Instruct.i1-IQ2_M.gguf”
- Size: 24.12
- Param: 70
- Quant: “iq2m”
- GPU Layers: 80
- GPU Layers Used: 21
- Ctx: 131072
- Ctx Used: 32768

Llama-3.3-70B-Instruct.i1-IQ2_M

- Path: “bartowski/Llama-3.1-Nemotron-70B-Instruct-HF-GGUF/Llama-3.1-Nemotron-70B-Instruct-HF-IQ2_M.gguf”
- Size: 24.12
- Param: 70
- Quant: “iq2m”
- GPU Layers: 80
- GPU Layers Used: 21
- Ctx: 131072
- Ctx Used: 32768

Qwen2.5-32B-Instruct-Q5_K_L

- Path: “bartowski/Qwen2.5-32B-Instruct-GGUF/Qwen2.5-32B-Instruct-Q5_K_L.gguf”
- Size: 23.74
- Param: 32
- Quant: “q5kl”
- GPU Layers: 64
- GPU Layers Used: 17
- Ctx: 32768
- Ctx Used: 32768

Mistral-Small-Instruct-2409-Q8_0

- Path: “bartowski/Mistral-Small-Instruct-2409-GGUF/Mistral-Small-Instruct-2409-Q8_0.gguf”
- Size: 23.64
- Param: 22B
- Quant: “q80”
- GPU Layers: 56
- GPU Layers Used: 15
- Ctx: 131072
- Ctx Used: 32768

Mistral-Small-22B-ArliAI-RPMax-v1.1-q8_0

- Path: “ArliAI/Mistral-Small-22B-ArliAI-RPMax-v1.1-GGUF/Mistral-Small-22B-ArliAI-RPMax-v1.1-q8_0.gguf”
- Size: 23.64
- Param: 22
- Quant: “q80”
- GPU Layers: 56
- GPU Layers Used:15

- Ctx: 32768
- Ctx Used: 32768

Codestral-22B-v0.1-Q8_0

- Path: “bartowski/Codestral-22B-v0.1-GGUF/Codestral-22B-v0.1-Q8_0.gguf”
- Size: 23.64
- Param: 22
- Quant: “q80”
- GPU Layers: 56
- GPU Layers Used: 15
- Ctx: 32768
- Ctx Used: 32768

---

aya-expanse-32b-Q5_K_L (23.56 GB)

- Path: “bartowski/aya-expanse-32b-GGUF/aya-expanse-32b-Q5_K_L.gguf”
- Size: 23.56
- Param: 32
- Quant: “q5kl”
- GPU Layers: 40
- GPU Layers Used: 11
- Ctx: 131072
- Ctx Used: 32768

---

c4ai-command-r-08-2024-Q5_K_L

- Path: “bartowski/c4ai-command-r-08-2024-GGUF/c4ai-command-r-08-2024-Q5_K_L.gguf”
- Size: 23.56
- Param: 32
- Quant: “q5kl”
- GPU Layers: 40
- GPU Layers Used: 11
- Ctx: 131072
- Ctx Used: 32768

---

magnum-v4-27b-Q6_K_L

- Path: “bartowski/magnum-v4-27b-GGUF/magnum-v4-27b-Q6_K_L.gguf”
- Size: 22.63
- Param: 27
- Quant: “q6kl”
- GPU Layers: 46
- GPU Layers Used: 13
- Ctx: 8192
- Ctx Used: 8192

Mixtral-8x7B-Instruct-v0.1-exhaustive-LoRA.i1-IQ3_M

- Path: “mradermacher/Mixtral-8x7B-Instruct-v0.1-exhaustive-LoRA-i1-GGUF/Mixtral-8x7B-Instruct-v0.1-exhaustive-LoRA.i1-IQ3_M.gguf”
- Size: 21.48
- Param: 56
- Quant: “iq3m”

- GPU Layers: 32
- GPU Layers Used: 9
- Ctx: 32768
- Ctx Used: 32768

---

qwen2.5-coder-14b-instruct-q8_0

- Path: “Qwen/Qwen2.5-Coder-14B-Instruct-GGUF/qwen2.5-coder-14b-instruct-q8_0-00001-of-00002.gguf”
- Size: 15.7
- Param: 14
- Quant: “q80”
- GPU Layers: 48
- GPU Layers Used: 19
- Ctx: 131072
- Ctx Used: 32768

Virtuoso-Small-Q8_0

- Path: “arcee-ai/Virtuoso-Small-GGUF/Virtuoso-Small-Q8_0.gguf”
- Size: 15.7
- Param: 14
- Quant: “q80”
- GPU Layers: 48
- GPU Layers Used: 19
- Ctx: 131072
- Ctx Used: 32768

---

phi-4-Q8_0

- Path: “matteogeniaccio/phi-4/phi-4-Q8_0.gguf”
- Size: 15.58
- Param: 14.7
- Quant: “q80”
- GPU Layers: 40
- GPU Layers Used: 16
- Ctx: 16384
- Ctx Used: 16384

Mistral-Nemo-Instruct-2407-Q8_0

- Path: “lmstudio-community/Mistral-Nemo-Instruct-2407-GGUF/Mistral-Nemo-Instruct-2407-Q8_0.gguf”
- Size: 13.02
- Param: 12
- Quant: “q80”
- GPU Layers: 40
- GPU Layers Used: 20
- Ctx: 1024000
- Ctx Used: 32768

NemoMix-Unleashed-12B-Q8_0

- Path: “bartowski/NemoMix-Unleashed-12B-GGUF/NemoMix-Unleashed-12B-Q8_0.gguf”
- Size: 13.02
- Param: 12

- Quant: “q80”
- GPU Layers: 40
- GPU Layers Used: 20
- Ctx: 1024000
- Ctx Used: 32768

Rocinante-12B-v1.1-Q8_0

- Path: “TheDrummer/Rocinante-12B-v1.1-GGUF/Rocinante-12B-v1.1-Q8_0.gguf”
- Size: 13.02
- Param: 12
- Quant: “q80”
- GPU Layers: 40
- GPU Layers Used: 20
- Ctx: 1024000
- Ctx Used: 32768

Moistral-11B-v3-Q8_0

- Path: “TheDrummer/Moistral-11B-v3-GGUF/Moistral-11B-v3-Q8_0.gguf”
- Size: 11.4
- Param: 11
- Quant: “q80”
- GPU Layers: 48
- GPU Layers Used: 27
- Ctx: 8192
- Ctx Used: 8192

---

codegeex4-all-9b-Q8_0

- Path: “THUDM/codegeex4-all-9b-GGUF/codegeex4-all-9b-Q8_0.gguf”
- Size: 9.99
- Param: 9
- Quant: “q80”
- GPU Layers: 40
- GPU Layers Used: 26
- Ctx: 131072
- Ctx Used: 32768

Darkest-muse-v1-Q8_0

- Path: “bartowski/Darkest-muse-v1-GGUF/Darkest-muse-v1-Q8_0.gguf”
- Size: 9.83
- Param: 9
- Quant: “q80”
- GPU Layers: 42
- GPU Layers Used: 27
- Ctx: 8192
- Ctx Used: 8192

Tiger-Gemma-9B-v3-Q8_0

- Path: “TheDrummer/Tiger-Gemma-9B-v3-GGUF/Tiger-Gemma-9B-v3-Q8_0.gguf”
- Size: 9.83

- Param: 9
- Quant: “q80”
- GPU Layers: 42
- GPU Layers Used: 27
- Ctx: 8192
- Ctx Used: 8192

---

aya-expanse-8b-Q8_0

- Path: “bartowski/aya-expanse-8b-GGUF/aya-expanse-8b-Q8_0.gguf”
- Size: 8.54
- Param: 8
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 24
- Ctx: 131072
- Ctx Used: 32768

---

c4ai-command-r7b-12-2024-q8_0

- Path: “dranger003/c4ai-command-r7b-12-2024-GGUF/ggml-c4ai-command-r7b-12-2024-q8_0.gguf”
- Size: 8.54
- Param: 8
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 24
- Ctx: 131072
- Ctx Used: 32768

Hermes-3-Llama-3.1-8B-Q8_0

- Path: “bartowski/Hermes-3-Llama-3.1-8B-GGUF/Hermes-3-Llama-3.1-8B-Q8_0.gguf”
- Size: 8.54
- Param: 8
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 24
- Ctx: 131072
- Ctx Used: 32768

L3-8B-Stheno-v3.2-Q8_0-imat

- Path: “Lewdiculous/L3-8B-Stheno-v3.2-GGUF-IQ-Imatrix/L3-8B-Stheno-v3.2-Q8_0-imat.gguf”
- Size: 8.54
- Param: 8
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 24
- Ctx: 8192
- Ctx Used: 8192

Ministral-8B-Instruct-2410-Q8_0

- Path: “mradermacher/Ministral-8B-Instruct-2410-GGUF/Ministral-8B-Instruct-2410.Q8_0.gguf”

- Size: 8.53
- Param: 8
- Quant: “q80”
- GPU Layers: 36
- GPU Layers Used: 27
- Ctx: 32768
- Ctx Used: 32768

Nemotron-Mini-4B-Instruct-f16

- Path: “bartowski/Nemotron-Mini-4B-Instruct-GGUF/Nemotron-Mini-4B-Instruct-f16.gguf”
- Size: 8.39
- Param: 4
- Quant: “f16”
- GPU Layers: 32
- GPU Layers Used: 24
- Ctx: 4096
- Ctx Used: 4096

Qwen2.5-Coder-7B-Instruct-Q8_0

- Path: “lmstudio-community/Qwen2.5-Coder-7B-Instruct-GGUF/Qwen2.5-Coder-7B-Instruct-Q8_0.gguf”
- Size: 8.1
- Param: 7
- Quant: “q80”
- GPU Layers: 28
- GPU Layers Used: 22
- Ctx: 32768
- Ctx Used: 32768

SeaLLMs-v3-7B-Chat-Q8_0

- Path: “tensorblock/SeaLLMs-v3-7B-Chat-GGUF/SeaLLMs-v3-7B-Chat-Q8_0.gguf”
- Size: 8.1
- Param: 7
- Quant: “q80”
- GPU Layers: 28
- GPU Layers Used: 22
- Ctx: 131072
- Ctx Used: 32768

Llava-v1.5-7B-Q8_0

- Path: “second-state/Llava-v1.5-7B-GGUF/llava-v1.5-7b-Q8_0.gguf”
- Size: 7.79
- Param: 7
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 26
- Ctx: 4096
- Ctx Used: 4096

---

falcon-mamba-7b-instruct-Q8_0

- Path: “tensorblock/falcon-mamba-7b-instruct-GGUF/falcon-mamba-7b-instruct-Q8_0.gguf”
- Size: 7.77
- Param: 7
- Quant: “q80”
- GPU Layers: 64
- GPU Layers Used: 53
- Ctx: 1048576
- Ctx Used: 32768

---

codeqwen-1_5-7b-chat-q8_0

- Path: “Qwen/CodeQwen1.5-7B-Chat-GGUF/codeqwen-1_5-7b-chat-q8_0.gguf”
- Size: 7.71
- Param: 7
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 27
- Ctx: 65536
- Ctx Used: 32768

---

mathstral-7B-v0.1.Q8_0

- Path: “DevQuasar/mathstral-7B-v0.1-GGUF/mathstral-7B-v0.1.Q8_0.gguf”
- Size: 7.7
- Param: 7
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 27
- Ctx: 32768
- Ctx Used: 32768

---

rho-math-7b-v0.1-Q8_0

- Path: “tensorblock/rho-math-7b-v0.1-GGUF/rho-math-7b-v0.1-Q8_0.gguf”
- Size: 7.7
- Param: 7
- Quant: “q80”
- GPU Layers: 32
- GPU Layers Used: 27
- Ctx: 32768
- Ctx Used: 32768

Phi-3.5-mini-instruct.f16

- Path: “mradermacher/Phi-3.5-mini-instruct-GGUF/Phi-3.5-mini-instruct.f16.gguf”
- Size: 7.64
- Param: 3.8
- Quant: “f16”
- GPU Layers: 32
- GPU Layers Used: 27
- Ctx: 131072
- Ctx Used: 32768

Ministral-3b-instruct.f16

- Path: “mradermacher/Ministral-3b-instruct-GGUF/Ministral-3b-instruct.f16.gguf”
- Size: 6.63
- Param: 3
- Quant: “f16”
- GPU Layers: 14
- GPU Layers Used: 13
- Ctx: 131072
- Ctx Used: 32768

Llama-Doctor-3.2-3B-Instruct-f16

- Path: “bartowski/Llama-Doctor-3.2-3B-Instruct-GGUF/Llama-Doctor-3.2-3B-Instruct-f16.gguf”
- Size: 6.43
- Param: 3
- Quant: “f16”
- GPU Layers: 28
- GPU Layers Used: 28
- Ctx: 131072
- Ctx Used: 32768

Hermes-3-Llama-3.2-3B-f16

- Path: “bartowski/Hermes-3-Llama-3.2-3B-GGUF/Hermes-3-Llama-3.2-3B-f16.gguf”
- Size: 6.43
- Param: 3
- Quant: “f16”
- GPU Layers: 28
- GPU Layers Used: 28
- Ctx: 131072
- Ctx Used: 32768

Qwen2.5-Coder-3B-Instruct-f16

- Path: “bartowski/Qwen2.5-Coder-3B-Instruct-GGUF/Qwen2.5-Coder-3B-Instruct-f16.gguf”
- Size: 6.18
- Param: 3
- Quant: “f16”
- GPU Layers: 36
- GPU Layers Used: 36
- Ctx: 32768
- Ctx Used: 32768

SmolLM2-1.7B-Instruct-f16

- Path: “bartowski/SmolLM2-1.7B-Instruct-GGUF/SmolLM2-1.7B-Instruct-f16.gguf”
- Size: 3.42
- Param: 1.7
- Quant: “f16”
- GPU Layers: 24
- GPU Layers Used: 24
- Ctx: 8192
- Ctx Used: 8192

Llama-3.2-1B-Instruct-f16

- Path: bartowski/Llama-3.2-1B-Instruct-GGUF/Llama-3.2-1B-Instruct-f16.gguf
- Size: 2.48
- Param: 1
- Quant: “f16”
- GPU Layers: 16
- GPU Layers Used: 16
- Ctx: 131072
- Ctx Used: 32768

Qwen2.5-Coder-0.5B-Instruct-f16

- Path: “bartowski/Qwen2.5-Coder-0.5B-Instruct-GGUF/Qwen2.5-Coder-0.5B-Instruct-f16.gguf”
- Size: 0.99
- Param: 0.5
- Quant: “f16”
- GPU Layers: 24
- GPU Layers Used: 24
- Ctx: 32768
- Ctx Used: 32768

Qwen2.5-0.5B-Instruct-f16

- Path: “bartowski/Qwen2.5-0.5B-Instruct-GGUF/Qwen2.5-0.5B-Instruct-f16.gguf”
- Size: 0.99
- Param: 0.5
- Quant: “f16”
- GPU Layers: 24
- GPU Layers Used: 24
- Ctx: 32768
- Ctx Used: 32768

SmolLM2-360M-Instruct-f16

- Path: “bartowski/SmolLM2-360M-Instruct-GGUF/SmolLM2-360M-Instruct-f16.gguf”
- Size: 0.73
- Param: 0.36
- Quant: “f16”
- GPU Layers: 32
- GPU Layers Used: 32
- Ctx: 8192
- Ctx Used: 8192

SmolLM2-135M-Instruct-f16

- Path: “bartowski/SmolLM2-135M-Instruct-GGUF/SmolLM2-135M-Instruct-f16.gguf”
- Size: 0.27
- Param: 0.135
- Quant: “f16”
- GPU Layers: 30
- GPU Layers Used: 30
- Ctx: 8192
- Ctx Used: 8192
