

LLM

model	capabilities	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
llama3.2	completion tools	"3.2B"	131072	"Q4KM"	88.14	715.43
ministral-3:14b	completion vision tools	"13.9B"	262144	"Q4KM"	23.78	302.07
qwen3-coder:30b	completion tools	"30.5B"	262144	"Q4KM"	73.75	72.41
llama3:70b	completion	"70.6B"	8192	"Q4"	5.55	9.72
llava	completion vision	"7B"	32768	"Q4"	49.92	207.27
deepseek-coder-v2:16b	completion insert	"15.7B"	163840	"Q4"	84.44	111.71
bjoernb/qwen3-coder-30b-1m:latest	completion tools	"30.5B"	1048576	"Q4KM"	74.23	94.84
freehuntx/qwen3-coder:8b	completion tools	"8.2B"	40960	"Q4KM"	37.97	565.68
networkjohnny/deepseek-coder-v2-lite-base-q4km-gguf:latest	completion tools	"3.2B"	131072	"Q4_K_M"	86.02	1124.53

From:

<https://wiki.csgalileo.org/> - **Galileo Labs**

Permanent link:

<https://wiki.csgalileo.org/tips/llm?rev=1765730960>

Last update: **2025/12/14 17:49**

