LLM

under 16GB

- vision: **llama3.2-vision**
- coding and agentic: **deepseek-coder-v2:lite**
- general reasoning: **llama3.1:8b**

model	**capabilities**	**size**	**context**	**quantization**	**eval rate [token/s]**	**prompt eval rate [token/s]**
llama3.2	completion tools	“3.2B”	131072	“Q4*K*M”	88.14	715.43
ministral-3:14b	completion vision tools	“13.9B”	262144	“Q4*K*M”	23.78	302.07
qwen3-coder:30b	completion tools	“30.5B”	262144	“Q4*K*M”	73.75	72.41
llama3:70b	completion	“70.6B”	8192	“Q4*0” | 5.55 | 9.72 | | llava | completion vision | “7B” | 32768 | “Q40”*	49.92	207.27
deepseek-coder-v2:16b	completion insert	“15.7B”	163840	“Q4_0”	84.44	111.71