LLM

under 16GB

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

model	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
ministral-3:14b	15G	262k	Q4_K_M	23.67	3876.16