🧠 Ollama Model Cheat Sheet (Local LLMs)

This cheat sheet summarizes the most powerful and specialized models, focusing on performance metrics like VRAM, size, and context window, which are critical for local deployment.

Category	**Model Name**	**Size (Approx)**	**VRAM (Approx)**	**Context (Tokens)**	**Speed (tok/s)**	**Primary Use & Notes**
Primary Coding	qwen2.5-coder:14b	~9 GB	~9 GB	32K	25–30	**Best overall balance** for coding tasks.
High Context Coding	deepseek-coder-v2:16b	~9 GB	~10 GB	**128K**	20–25	Massive context window. Excellent for large projects/repos.
Max Context Coding	codestral:22b	~13 GB	~11 GB	**128K**	15–20	Largest coding model listed. Great for context, but slower and VRAM-intensive.
Fast Coding Backup	qwen2.5-coder:7b	~4.7 GB	~5 GB	32K	**40–45**	Extremely fast and lightweight coding model.
General Chat/Writing	qwen2.5:14b	~9 GB	~9 GB	32K	25–30	Excellent general-purpose model with strong writing capabilities.
Complex Reasoning	deepseek-r1:14b	~9 GB	~9 GB	8K	20–25	Strong reasoning model, good

						upgrade from 7B versions.
Alternative Reasoning	phi4:14b	~9 GB	~9 GB	16K	25–30	Great alternative with a larger context for reasoning.
Vision Model (Fast)	llava-llama3:8b	~5.5 GB	~6 GB	N/A	30–35	Faster, more efficient vision model for image understanding.
Specialized Python	codellama:13b-python	~7.4 GB	~8 GB	N/A	25–30	Specialized model if you heavily focus on Python development.

🛠️ Quick Reference & Lightweight Models

These models are useful for quick checks, lightweight systems, or as alternative benchmarks.

Model Name	**Purpose / Note**	**Ollama Pull Command**
phi3:mini	Very fast (3.8B) for quick, conversational queries.	ollama pull phi3:mini
gemma3:4b	Lightweight backup general model.	ollama pull gemma3:4b
mistral:latest	Standard, good performance 7B general model.	ollama pull mistral:latest
llava:13b	Older, but still functional Vision model.	ollama pull llava:13b
qwen3-coder:latest	Newer coding model (check size, use if 14B or less).	ollama pull qwen3-coder:latest