LLM Cheatsheet - May 2025

Vendor	Model	Parameters, B	Context, k tokens	API	OS	Reasoning	Public	Release
AI21 Labs	Jamba 1.6 Mini	? / 52	256	✅	✅	❌	✅	**Mar 2025**
AI21 Labs	Jamba 1.6 Large	94 / 398	256	✅	✅	❌	✅	**Mar 2025**
Anthropic	Claude 3.5 Sonnet	?	200	✅	❌	❌	✅	Jun 2024
Anthropic	Claude 3.7 Sonnet	?	200	✅	❌	✅	✅	**Feb 2025**
Cohere	Command A	? / 111	256	✅	❌	❌	✅	**Mar 2025**
Cohere	Command R+	? / 104	128	✅	❌	❌	✅	Aug 2024
DeepSeek	DeepSeek LLM	7 / 67	4	✅	✅	❌	✅	Nov 2023
DeepSeek	DeepSeek-V3	37 / 671	128	✅	❌	✅	✅	**Mar 2025**
Google	Gemini 2.5 Pro	?	1,000	✅	❌	✅	🚧	**Mar 2025**
Meta	Llama 3.1	? / 405	128	✅	✅	❌	✅	Jul 2024
Meta	Llama 4 Scout	? / 109	10,000	✅	✅	❌	✅	**Apr 2025**
Meta	Llama 4 Maverick	? / 400	1000	✅	✅	❌	✅	**Apr 2025**
Mistral	Mistral Large	?	128	✅	❌	❌	✅	Jul 2024
Mistral	Mixtral 8×22B	? / 141	65	✅	✅	❌	✅	Apr 2024
MosaicML	MPT-30B	30 / 30	8	✅	✅	❌	✅	Jun 2023
OpenAI	GPT-4.1	?	1000	✅	❌	❌	🚧	**Apr 2025**
OpenAI	GPT-4o	?	128	✅	❌	❌	✅	May 2024
OpenAI	o1	?	128	✅	❌	✅	🚧	Sep 2024
OpenAI	o3	?	128	✅	❌	✅	🚧	**Apr 2025**
OpenAI	o3-mini	?	200	✅	❌	✅	🚧	**Jan 2025**
OpenAI	o4-mini	?	128	✅	❌	✅	🚧	**Apr 2025**
Stability AI	StableLM-Alpha	7 / 7	4	✅	✅	❌	✅	Apr 2023