LLM Models

Master Cheatsheet 2025

Open Source vs Closed Source | Complete Comparison Guide

■ Open Source vs Closed Source - Key Differences

Aspect	■ Open Source	■ Closed Source
Code Access	Full access to weights & code	API access only
Cost	Free (compute cost only)	Pay per token/API call
Customization	Fine-tune, modify, self-host	Limited to API parameters
Privacy	Data stays on your servers	Data sent to provider
Updates	Community-driven, forks	Provider-controlled
Support	Community forums	Enterprise support available
Deployment	Self-host anywhere	Cloud API dependent
Examples	Llama, Mistral, Qwen	GPT-4, Claude, Gemini

■ Closed Source LLMs

■ OpenAI Models

Model	Context	Best For	Pricing (per 1M tokens)
GPT-4o	128K	Multimodal, fast, smart	$2.50 in / $10 out
GPT-4o mini	128K	Cost-effective, everyday tasks	$0.15 in / $0.60 out
GPT-4 Turbo	128K	Complex reasoning	$10 in / $30 out
o1	200K	PhD-level reasoning, math, code	$15 in / $60 out
o1-mini	128K	Fast reasoning, coding	$3 in / $12 out
o3-mini	200K	Latest reasoning model	Variable pricing
GPT-4.1	1M	Long context, coding	$2 in / $8 out
DALL-E 3	-	Image generation	$0.04-0.12 per image
Whisper	-	Speech-to-text	$0.006 per minute

■ Anthropic Claude Models

Model	Context	Best For	Pricing (per 1M tokens)
Claude Opus 4	200K	Most intelligent, complex tasks	$15 in / $75 out
Claude Sonnet 4	200K	Balanced speed & intelligence	$3 in / $15 out
Claude Haiku 3.5	200K	Fast, cost-effective	$0.25 in / $1.25 out
Claude Sonnet 3.5	200K	Previous gen, still capable	$3 in / $15 out

■ Google Gemini Models

Model	Context	Best For	Pricing (per 1M tokens)
Gemini 2.5 Pro	1M	Best reasoning, thinking model	$1.25-2.50 in / $10-15 out
Gemini 2.0 Flash	1M	Fast, multimodal, agents	$0.10 in / $0.40 out
Gemini 1.5 Pro	2M	Longest context available	$1.25 in / $5 out
Gemini 1.5 Flash	1M	Speed optimized	$0.075 in / $0.30 out
Imagen 3	-	Image generation	Variable
Veo 2	-	Video generation	Variable

■ xAI Grok Models

Model	Context	Best For	Pricing (per 1M tokens)
Grok 3	128K	Flagship, real-time X data	$3 in / $15 out
Grok 3 mini	128K	Fast reasoning	$0.30 in / $0.50 out
Grok 2	128K	Previous gen, capable	$2 in / $10 out
Grok Vision	128K	Image understanding	$5 in / $15 out

■ Open Source LLMs

■ Meta Llama Models

Model	Parameters	Context	Best For
Llama 4 Scout	17B (109B MoE)	10M	Longest context, lightweight
Llama 4 Maverick	17B (400B MoE)	1M	Best multimodal open model
Llama 3.3	70B	128K	Strong all-rounder
Llama 3.2	1B, 3B, 11B, 90B	128K	Edge devices to vision
Llama 3.1	8B, 70B, 405B	128K	Previous gen, proven
Code Llama	7B, 13B, 34B	16K	Code generation

■■ Mistral AI Models

Model	Parameters	Context	Best For
Mistral Large 2	123B	128K	Flagship, multilingual
Mistral Medium	~70B	32K	Balanced performance
Mistral Small	22B	32K	Cost-effective
Codestral	22B	32K	Code generation (80+ langs)
Ministral 8B	8B	128K	Edge deployment
Ministral 3B	3B	128K	Mobile/IoT devices
Pixtral Large	124B	128K	Vision + language
Mixtral 8x22B	141B MoE	64K	Open weights, MoE

Mixtral 8x7B	46B MoE	32K	Efficient MoE

■ DeepSeek Models (China)

Model	Parameters	Context	Best For
DeepSeek V3	671B MoE	128K	GPT-4 level, very cheap
DeepSeek R1	671B MoE	128K	Reasoning (o1 competitor)
DeepSeek R1 Distill	1.5B-70B	128K	Distilled reasoning
DeepSeek Coder V3	~30B	128K	Code specialist
DeepSeek V2.5	236B MoE	128K	Previous gen

■ ***DeepSeek API: $0.14/M input, $0.28/M output - Cheapest in market!***

■ Qwen Models (Alibaba)

Model	Parameters	Context	Best For
Qwen 2.5	0.5B-72B	128K	Multilingual, coding
Qwen 2.5 Coder	1.5B-32B	128K	Code specialist
Qwen 2.5 Math	1.5B-72B	128K	Math reasoning
QwQ	32B	32K	Reasoning model
Qwen VL	2B-72B	32K	Vision-language
Qwen Audio	7B	-	Audio understanding

■ Other Notable Open Source Models

Model	Provider	Parameters	Best For
Gemma 2	Google	2B, 9B, 27B	Lightweight, efficient
Phi-4	Microsoft	14B	Small but powerful
Phi-3.5	Microsoft	3.8B-128K MoE	Edge devices
DBRX	Databricks	132B MoE	Enterprise, MoE
Falcon	TII	7B, 40B, 180B	Multilingual
Yi	01.AI	6B-34B	Bilingual EN/CN
InternLM	Shanghai AI Lab	7B-20B	Chinese focus
Command R+	Cohere	104B	RAG, enterprise
StarCoder 2	BigCode	3B-15B	Code generation
Stable LM	Stability AI	3B-12B	Conversational

■■ Where to Run Open Source LLMs

Platform	Type	Best For	Pricing
Ollama	Local	Run locally on Mac/Linux/Win	Free (your hardware)

LM Studio	Local	GUI for local models	Free
Hugging Face	Cloud/Local	Model hub, inference API	Free tier + paid
Together AI	Cloud API	Fast inference, many models	Pay per token
Groq	Cloud API	Ultra-fast inference (LPU)	Pay per token
Fireworks AI	Cloud API	Fast, fine-tuning support	Pay per token
Replicate	Cloud API	Easy deployment, pay per run	Pay per second
AWS Bedrock	Cloud	Enterprise, Llama/Mistral	Pay per token
Azure ML	Cloud	Enterprise, Meta models	Pay per token
Google Vertex	Cloud	Gemma, enterprise	Pay per token
RunPod	GPU Cloud	Rent GPUs, self-host	Per hour GPU
Vast.ai	GPU Cloud	Cheap GPU rental	Per hour GPU

■ Head-to-Head Comparison

Category	Best Closed Source	Best Open Source
Overall Intelligence	Claude Opus 4, GPT-4o	Llama 4, DeepSeek V3
Reasoning	o1, o3-mini	DeepSeek R1, QwQ
Coding	Claude Sonnet 4, GPT-4.1	DeepSeek Coder, Codestral
Long Context	Gemini 1.5 Pro (2M)	Llama 4 Scout (10M)
Multimodal	GPT-4o, Gemini 2.0	Llama 4 Maverick, Pixtral
Speed	GPT-4o mini, Gemini Flash	Groq + Llama, Mistral
Cost Effective	GPT-4o mini, Haiku	DeepSeek V3 (cheapest)
Privacy	N/A (API based)	Self-hosted Llama/Mistral
Enterprise	Claude, GPT Enterprise	Llama + AWS/Azure

■ Which LLM to Choose?

Use Case	Recommended Model	Why
Production App (budget)	GPT-4o mini / DeepSeek V3	Cheap, reliable, good quality
Production App (quality)	Claude Sonnet 4 / GPT-4o	Best balance of speed & quality
Complex Reasoning	o1 / DeepSeek R1	PhD-level thinking
Coding Assistant	Claude Sonnet 4 / Codestral	Best code generation
Data Privacy Required	Llama 4 / Mistral (self-host)	Data never leaves your server
Real-time Chat	GPT-4o mini / Gemini Flash	Fast response times
Document Analysis	Gemini 1.5 Pro / Llama 4 Scout	Longest context windows
Image Understanding	GPT-4o / Llama 4 Maverick	Best vision capabilities

Low Budget Startup	DeepSeek V3 / Qwen 2.5	Extremely cheap, good quality
Enterprise/Compliance	Claude / GPT Enterprise	SOC2, HIPAA compliance

■ API Endpoints Quick Reference

Provider	Base URL	Auth
OpenAI	https://api.openai.com/v1	Bearer token
Anthropic	https://api.anthropic.com/v1	x-api-key header
Google AI	https://generativelanguage.googleapis.com	API key
Groq	https://api.groq.com/openai/v1	Bearer token
Together	https://api.together.xyz/v1	Bearer token
DeepSeek	https://api.deepseek.com/v1	Bearer token
Mistral	https://api.mistral.ai/v1	Bearer token
Fireworks	https://api.fireworks.ai/inference/v1	Bearer token
Ollama (local)	http://localhost:11434/api	None

Created by Manoj | The AI Dude Tamil ■

YouTube: The AI Dude Tamil | Master AI, Automation & Prompt Engineering

■ Choose wisely: Open for privacy, Closed for convenience!