Large Language Models (LLMs) Overview

1. Open-Source LLMs

Model	Developer	Architecture	License	Max Tokens
LLaMA 2	Meta	Transformer	Open (custom)	4K-32K
Mistral 7B	Mistral AI	Transformer	Apache 2.0	8K
Mixtral 8x7B	Mistral AI	MoE (Sparse)	Apache 2.0	32K
Gemma	Google	Transformer	Apache 2.0	8K
Phi-2	Microsoft	Transformer	MIT	4K
Yi 34B	01.AI	Transformer	Open	32K
GPT-J / GPT-NeoX	EleutherAI	Transformer	Apache/MIT	2.7B-20B
OpenChat	Community	LLaMA-based	Apache 2.0	8K-32K

2. Proprietary LLMs

Model	Developer	Access Type	Max Tokens	Use Cases
GPT-4	OpenAI	API, ChatGPT+	128K	General AI tasks, coding
Claude 3	Anthropic	API, Claude.ai	200K+	Document processing, chat
Gemini 1.5	Google DeepMind	API, Gemini App	1M+	Advanced reasoning
Command R+	Cohere	API	128K+	RAG, search, summarization
Jurassic-2	AI21 Labs	API	8K+	Text generation
ERNIE Bot 4.0	Baidu	API (China)	?	Multilingual, Chinese NLP

3. Special-Purpose & Lightweight LLMs

Model	Developer	Specialization
LLamaGuard	Meta	Content moderation
PrivateGPT	Open-source	Local, offline use
TinyLLaMA	Community	Ultra-lightweight
Dolly	Databricks	Instruction tuning
StableLM	Stability AI	General-purpose
RedPajama	Together	Dataset replication