Open LLMs
for Commercial Use

A list of open LLMs available for commercial use.

Himanshu Ramchandani
https://www.linkedin.com/in/hemansnation/

Open LLMs

These LLMs (Large Language Models) are all licensed for commercial use (e.g., Apache 2.0, MIT, OpenRAIL-M). Contributions welcome!

Language Model	Release Date	Checkpoints	Paper/Blog	Params (B)	Context Length	Licence	Try it
T5	2019/10	T5 & Flan-T5, Flan-T5-xxl (HF)	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	0.06 - 11	512	Apache 2.0	T5-Large

RWKV 4	2021/08	RWKV, ChatRWKV	The RWKV Language Model (and my LM tricks)	0.1 - 14	infinity (RNN)	Apache 2.0	
GPT-NeoX-20B	2022/04	GPT-NEOX-20B	GPT-NeoX-20B: An Open-Source Autoregressive Language Model	20	2048	Apache 2.0	
YaLM-100B	2022/06	yalm-100b	Yandex publishes YaLM 100B, the largest GPT-like neural network in open source	100	1024	Apache 2.0	

UL2	2022/10	UL2 & Flan-UL2, Flan-UL2 (HF)	UL2 20B: An Open Source Unified Language Learner	20	512, 2048	Apache 2.0	
Bloom	2022/11	Bloom	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model	176	2048	OpenRAIL-M v1	
ChatGLM	2023/03	chatglm-6b	ChatGLM, Github	6	2048	Custom Free with some usage restriction (might require registration)	

Cerebras-GPT	2023/03	Cerebras-GPT	Cerebras-GPT: A Family of Open, Compute-efficient, Large Language Models (Paper)	0.111 - 13	2048	Apache 2.0	Cerebras-GPT-1.3B
Open Assistant (Pythia family)	2023/03	OA-Pythia-12B-SFT-8, OA-Pythia-12B-SFT-4, OA-Pythia-12B-SFT-1	Democratizing Large Language Model Alignment	12	2048	Apache 2.0	Pythia-2.8B
Pythia	2023/04	pythia 70M - 12B	Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling	0.07 - 12	2048	Apache 2.0	

Dolly	2023/04	dolly-v2-12b	Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM	3, 7, 12	2048	MIT	
StableLM-Alpha	2023/04	StableLM-Alpha	Stability AI Launches the First of its StableLM Suite of Language Models	3 - 65	4096	CC BY-SA-4.0	
FastChat-T5	2023/04	fastchat-t5-3b-v1.0	We are excited to release FastChat-T5: our compact and commercial-friendly chatbot!	3	512	Apache 2.0	

DLite	2023/05	dlite-v2-1_5b	Announcing DLite V2: Lightweight, Open LLMs That Can Run Anywhere	0.124 - 1.5	1024	Apache 2.0	DLite-v2-1.5B
h2oGPT	2023/05	h2oGPT	Building the World's Best Open-Source Large Language Model: H2O.ai's Journey	12 - 20	256 - 2048	Apache 2.0	
MPT-7B	2023/05	MPT-7B, MPT-7B-Instruct	Introducing MPT-7B: A New Standard for Open-Source, Commerci	7	84k (ALiBi)	Apache 2.0, CC BY-SA-3.0	

			ally Usable LLMs				
RedPajama-INCITE	2023/05	RedPajama-INCITE	Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models	3 - 7	2048	Apache 2.0	RedPajama-INCITE-Instruct-3B-v1
OpenLLaMA	2023/05	open_llama_3b, open_llama_7b, open_llama_13b	OpenLLaMA: An Open Reproduction of LLaMA	3, 7	2048	Apache 2.0	OpenLLaMA-7B-Preview_200bt

Falcon	2023/05	Falcon-180B, Falcon-40B, Falcon-7B	The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only	180, 40, 7	2048	Apache 2.0	
GPT-J-6B	2023/06	GPT-J-6B, GPT4All-J	GPT-J-6B: 6B JAX-Based Transformer	6	2048	Apache 2.0	
MPT-30B	2023/06	MPT-30B, MPT-30B-instruct	MPT-30B: Raising the bar for open-source foundation models	30	8192	Apache 2.0, CC BY-SA-3.0	MPT 30B inference code using CPU

| LLaMA 2 | 2023/06 | LLaMA 2 Weights | Llama 2: Open Foundation and Fine-Tuned Chat Models | 7 - 70 | 4096 | Custom Free if you have under 700M users and you cannot use LLaMA outputs to train other LLMs besides LLaMA and its derivatives | Hugging Chat |
| ChatGLM2 | 2023/06 | chatglm2-6b | ChatGLM2-6B, Github | 6 | 32k | Custom Free with some usage restriction (might require registration) | |

XGen-7B	2023/06	xgen-7b-4k-base, xgen-7b-8k-base	Long Sequence Modeling with XGen	7	4096, 8192	Apache 2.0	
Jais-13b	2023/08	jais-13b, jais-13b-chat	Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models	13	2048	Apache 2.0	
OpenHermes	2023/09	OpenHermes-7B, OpenHermes-13B	Nous Research	7, 13	4096	MIT	OpenHermes-V2 Finetuned on Mistral 7B

OpenLM	2023/09	OpenLM 1B, OpenLM 7B	Open LM: a minimal but performative language modeling (LM) repository	1, 7	2048	MIT	
Mistral 7B	2023/09	Mistral-7B-v0.1, Mistral-7B-Instruct-v0.1	Mistral 7B	7	4096-16K with Sliding Windows	Apache 2.0	Mistral Transformer
ChatGLM3	2023/10	chatglm3-6b, chatglm3-6b-base, chatglm3-6b-32k, chatglm3-6b-128k	ChatGLM3	6	8192, 32k, 128k	Custom Free with some usage restriction (might require registration)	

Skywork	2023/10	Skywork-13B-Base, Skywork-13B-Math	Skywork	13	4096	Custom Free with usage restriction and models trained on Skywork outputs become Skywork derivatives, subject to this license.	
Jais-30b	2023/11	jais-30b-v1, jais-30b-chat-v1	Jais-30B: Expanding the Horizon in Open-Source Arabic NLP	30	2048	Apache 2.0	
Zephyr	2023/11	Zephyr 7B	Website	7	8192	Apache 2.0	

DeepSeek	2023/11	deepseek-llm-7b-base, deepseek-llm-7b-chat, deepseek-llm-67b-base, deepseek-llm-67b-chat	Introducing DeepSeek LLM,	7, 67	4096	Custom Free with usage restriction and models trained on DeepSeek outputs become DeepSeek derivatives, subject to this license.	
Mistral 7B v0.2	2023/12	Mistral-7B-v0.2, Mistral-7B-Instruct-v0.2	La Plateforme	7	32k	Apache 2.0	
Mixtral 8x7B v0.1	2023/12	Mixtral-8x7B-v0.1, Mixtral-8x7B-Instruct-v0.1	Mixtral of experts	46.7	32k	Apache 2.0	
LLM360 Amber	2023/12	Amber, AmberChat, AmberSafe	Introducing LLM360: Fully Transparent	6.7	2048	Apache 2.0	

			Open-Source LLMs				
SOLAR	2023/12	Solar-10.7B	Upstage	10.7	4096	apache-2.0	
phi-2	2023/12	phi-2 2.7B	Microsoft	2.7	2048	MIT	
FLOR	2023/12	FLOR-760M, FLOR-1.3B, FLOR-1.3B-Instructed, FLOR-6.3B, FLOR-6.3B-Instructed	FLOR-6.3B: a chinchilla-compliant model for Catalan, Spanish and English	0.76, 1.3, 6.3	2048	Apache 2.0 with usage restriction inherited from BLOOM	
RWKV 5 v2	2024/01	rwkv-5-world-0.4b-2, rwkv-5-world-1.5b-2, rwkv-5-world-3b-2, rwkv-5-world-3b-2(16k), rwkv-5-world-7b-2	RWKV 5	0.4, 1.5, 3, 7	unlimited(RNN), trained on 4096 (and 16k for 3b)	Apache 2.0	
OLMo	2024/02	OLMo 1B, OLMo 7B, OLMo 7B Twin 2T	AI2	1,7	2048	Apache 2.0	

Qwen1.5	2024/02	Qwen1.5-7B, Qwen1.5-7B-Chat, Qwen1.5-14B, Qwen1.5-14B-Chat, Qwen1.5-72B, Qwen1.5-72B-Chat	Introducing Qwen1.5	7, 14, 72	32k	Custom Free if you have under 100M users and you cannot use Qwen outputs to train other LLMs besides Qwen and its derivatives	
LWM	2024/02	LWM-Text-Chat-128K, LWM-Text-Chat-256K, LWM-Text-Chat-512K, LWM-Text-Chat-1M, LWM-Text-128K, LWM-Text-256K, LWM-Text-512K, LWM-Text-1M	Large World Model (LWM)	7	128k, 256k, 512k, 1M	LLaMA 2 license	
Jais-30b v3	2024/03	jais-30b-v3, jais-30b-chat-v3	Jais 30b v3	30	8192	Apache 2.0	

Gemma	2024/02	Gemma 7B, Gemma 7B it, Gemma 2B, Gemma 2B it	Technical report	2-7	8192	Gemma Terms of Use Free with usage restriction and models trained on Gemma outputs become Gemma derivatives, subject to this license.	
Grok-1	2024/03	Grok-1	Open Release of Grok-1	314	8192	Apache 2.0	

| Qwen1.5 MoE | 2024/03 | Qwen1.5-MoE-A2.7B, Qwen1.5-MoE-A2.7B-Chat | Qwen1.5-MoE: Matching 7B Model Performance with 1/3 Activated Parameters | 14.3 | 8192 | Custom Free if you have under 100M users and you cannot use Qwen outputs to train other LLMs besides Qwen and its derivatives | |
| Jamba 0.1 | 2024/03 | Jamba-v0.1 | Introducing Jamba: AI21's Groundbreaking SSM-Transformer Model | 52 | 256k | Apache 2.0 | |

Qwen1.5 32B	2024/04	Qwen1.5-32B, Qwen1.5-32B-Chat	Qwen1.5-32B: Fitting the Capstone of the Qwen1.5 Language Model Series	32	32k	Custom Free if you have under 100M users and you cannot use Qwen outputs to train other LLMs besides Qwen and its derivatives	
Mamba-7B	2024/04	mamba-7b-rw	Toyota Research Institute	7	unlimited(RNN), trained on 2048	Apache 2.0	
Mixtral8x 22B v0.1	2024/04	Mixtral-8x22B-v0.1, Mixtral-8x22B-Instruct-v0.1	Cheaper, Better, Faster, Stronger	141	64k	Apache 2.0	

Llama 3	2024/04	Llama-3-8B, Llama-3-8B-Instruct, Llama-3-70B, Llama-3-70B-Instruct, Llama-Guard-2-8B	Introducing Meta Llama 3, Meta Llama 3	8, 70	8192	Meta Llama 3 Community License Agreement Free if you have under 700M users and you cannot use LLaMA 3 outputs to train other LLMs besides LLaMA 3 and its derivatives	
Phi-3 Mini	2024/04	Phi-3-mini-4k-instruct, Phi-3-mini-128k-instruct	Introducing Phi-3, Technical Report	3.8	4096, 128k	MIT	

OpenELM	2024/04	OpenELM-270M, OpenELM-270M-Instruct, OpenELM-450M, OpenELM-450M-Instruct, OpenELM-1_1B, OpenELM-1_1B-Instruct, OpenELM-3B, OpenELM-3B-Instruct	OpenELM: An Efficient Language Model Family with Open Training and Inference Framework	0.27, 0.45, 1.1, 3	2048	Custom open license No usage or training restrictions	
Snowflake Arctic	2024/04	snowflake-arctic-base, snowflake-arctic-instruct	Snowflake Arctic: The Best LLM for Enterprise AI — Efficiently Intelligent, Truly Open	480	4096	Apache 2.0	

Qwen1.5 110B	2024/04	Qwen1.5-110B, Qwen1.5-110B-Chat	Qwen1.5-110B: The First 100B+ Model of the Qwen1.5 Series	110	32k	Custom Free if you have under 100M users and you cannot use Qwen outputs to train other LLMs besides Qwen and its derivatives	
RWKV 6 v2.1	2024/05	rwkv-6-world-1.6b-2.1, rwkv-6-world-3b-2.1, rwkv-6-world-7b-2.1	RWKV 6	1.6, 3, 7	unlimited(RNN), trained on 4096	Apache 2.0	

DeepSeek-V2	2024/05	DeepSeek-V2, DeepSeek-V2-Chat	DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model	236	128k	Custom Free with usage restriction and models trained on DeepSeek outputs become DeepSeek derivatives, subject to this license.	
Fugaku-LLM	2024/05	Fugaku-LLM-13B, Fugaku-LLM-13B-instruct	Release of "Fugaku-LLM" – a large language model trained on the supercomputer "Fugaku"	13	2048	Custom Free with usage restrictions	

Falcon 2	2024/05	falcon2-11B	Meet Falcon 2: TII Releases New AI Model Series, Outperforming Meta's New Llama 3	11	8192	Custom Apache 2.0 with mild acceptable use policy	
Yi-1.5	2024/05	Yi-1.5-6B, Yi-1.5-6B-Chat, Yi-1.5-9B, Yi-1.5-9B-Chat, Yi-1.5-34B, Yi-1.5-34B-Chat	Yi-1.5	6, 9, 34	4096	Apache 2.0	
DeepSeek-V2-Lite	2024/05	DeepSeek-V2-Lite, DeepSeek-V2-Lite-Chat	DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model	16	32k	Custom Free with usage restriction and models trained on DeepSeek outputs become DeepSeek	

| Phi-3 small/medium | 2024/05 | Phi-3-mini-4k-instruct, Phi-3-mini-128k-instruct, Phi-3-medium-4k-instruct, Phi-3-medium-128k-instruct | New models added to the Phi-3 family, available on Microsoft Azure, Technical Report | 7, 14 | 4096, 128k | MIT | |
| | | | | | | derivatives, subject to this license. | |

Open LLMs for code

Language Model	Release Date	Checkpoints	Paper/Blog	Params (B)	Context Length	Licence	Try it
SantaCoder	2023/01	santacoder	SantaCoder: don't reach for the stars!	1.1	2048	OpenRAIL-M v1	SantaCoder

CodeGen2	2023/04	codegen2 1B-16B	CodeGen2: Lessons for Training LLMs on Programming and Natural Languages	1 - 16	2048	Apache 2.0	
StarCoder	2023/05	starcoder	StarCoder: A State-of-the-Art LLM for Code, StarCoder: May the source be with you!	1.1-15	8192	OpenRAIL-M v1	
StarChat Alpha	2023/05	starchat-alpha	Creating a Coding Assistant with StarCoder	16	8192	OpenRAIL-M v1	
Replit Code	2023/05	replit-code-v1-3b	Training a SOTA Code LLM in 1 week and Quantifying the Vibes — with Reza Shabani of Replit	2.7	infinity? (ALiBi)	CC BY-SA-4.0	Replit-Code-v1-3B

CodeT5+	2023/05	CodeT5+	CodeT5+: Open Code Large Language Models for Code Understanding and Generation	0.22 - 16	512	BSD-3-Clause	Codet5+-6B
XGen-7B	2023/06	XGen-7B-8K-Base	Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length	7	8192	Apache 2.0	
CodeGen2.5	2023/07	CodeGen2.5-7B-multi	CodeGen2.5: Small, but mighty	7	2048	Apache 2.0	
DeciCoder-1B	2023/08	DeciCoder-1B	Introducing DeciCoder: The New Gold Standard in Efficient and Accurate Code Generation	1.1	2048	Apache 2.0	DeciCoder Demo

| Code Llama | 2023/08 | Inference Code for CodeLlama models | [Code Llama: Open Foundation Models for Code](#) | 7 - 34 | [4096](#) | [Custom](#) Free if you have under 700M users and you cannot use LLaMA outputs to train other LLMs besides LLaMA and its derivatives | [HuggingChat](#) |

Open LLM datasets for pre-training

Name	Release Date	Paper/Blog	Dataset	Tokens (T)	License
RedPajama	2023/04	[RedPajama, a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens](#)	[RedPajama-Data](#)	1.2	Apache 2.0
starcoderdata	2023/05	[StarCoder: A State-of-the-Art LLM for Code](#)	[starcoderdata](#)	0.25	Apache 2.0

Open LLM datasets for instruction-tuning

Name	Release Date	Paper/Blog	Dataset	Samples (K)	License
OIG (Open Instruction Generalist)	2023/03	THE OIG DATASET	OIG	44,000	Apache 2.0
databricks-dolly-15k	2023/04	Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM	databricks-dolly-15k	15	CC BY-SA-3.0
MPT-7B-Instruct	2023/05	Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs	dolly_hhrlhf	59	CC BY-SA-3.0

Open LLM datasets for alignment-tuning

Name	Release Date	Paper/Blog	Dataset	Samples (K)	License
OpenAssistant Conversations Dataset	2023/04	[OpenAssistant Conversations - Democratizing Large Language Model Alignment](#)	[oasst1](#)	161	Apache 2.0

Evals on open LLMs

- [Leaderboard by lmsys.org](#)
- [Evals by MosaicML](#)
- [Holistic Evaluation of Language Models (HELM)](#)
- [LLM-Leaderboard](#)
- [TextSynth Server Benchmarks](#)
- [Open LLM Leaderboard by Hugging Face](#)

What do the licences mean?

- [Apache 2.0](): Allows users to use the software for any purpose, to distribute it, to modify it, and to distribute modified versions of the software under the terms of the license, without concern for royalties.
- [MIT](): Similar to Apache 2.0 but shorter and simpler. Also, in contrast to Apache 2.0, does not require stating any significant changes to the original code.
- [CC BY-SA-4.0](): Allows (i) copying and redistributing the material and (ii) remixing, transforming, and building upon the material for any purpose, even commercially. But if you do the latter, you must distribute your contributions under the same license as the original. (Thus, may not be viable for internal teams.)
- [OpenRAIL-M v1](): Allows royalty-free access and flexible downstream use and sharing of the model and modifications of it, and comes with a set of use restrictions (see [Attachment A]())
- [BSD-3-Clause](): This version allows unlimited redistribution for any purpose as long as its copyright notices and the license's disclaimers of warranty are maintained.

GenerativeAI Roadmap Here:

https://god-level-python.notion.site/Neurons-to-GenerativeAI-Bootcamp-Live-Self-paced-a59ec2f641084c488179271fc077f0c4?pvs=4

Socials

Be part of 50,000+ like-minded AI professionals across the platform

→ LinkedIn → YouTube→ Twitter → Instagram → Medium

→ Telegram → Discord Server → GitHub

About Me

Himanshu Ramchandani

The last AI Consultant you'll ever work with.
Data & AI Solutions Consulting