

LLM Collection

This section consists of a collection and summary of notable and foundational LLMs.

Models

Model	Release Date	Size (B)	Checkpoints	Description
Falcon LLM	Sep 2023	7, 40, 180	Falcon-7B , Falcon-40B , Falcon-180B	Falcon LLM is a foundational large language model (LLM) with 180 billion parameters trained on 3500 Billion tokens. TII has now released Falcon LLM – a 180B model.
Mistral-7B-v0.1	Sep 2023	7	Mistral-7B-v0.1	Mistral-7B-v0.1 is a pretrained generative text model with 7 billion parameters. The model is based on a transformer architecture with features like Grouped-Query Attention, Byte-fallback BPE tokenizer and Sliding-Window Attention.
CodeLlama	Aug 2023	7, 13, 34	CodeLlama-7B , CodeLlama-13B , CodeLlama-34B	The Code Llama family is designed for general code synthesis and understanding. It is specifically tuned for instruction following and safer

Model	Release Date	Size (B)	Checkpoints	Description
				deployment. The models are auto-regressive and use an optimized transformer architecture. They are intended for commercial and research use in English and relevant programming languages.
Llama-2	Jul 2023	7, 13, 70	Llama-2-7B , Llama-2-13B , Llama-2-70B	LLaMA-2, developed by Meta AI, was released in July 2023 with models of 7, 13, and 70 billion parameters. It maintains a similar architecture to LLaMA-1 but uses 40% more training data. LLaMA-2 includes foundational models and dialog-fine-tuned models, known as LLaMA-2 Chat, and is available for many commercial uses, with some restrictions.
XGen-7B-8K	Jul 2023	7	XGen-7B-8K	The XGen-7B-8K, developed by Salesforce AI Research, is a 7B parameter language model.
Claude-2	Jul 2023	130	-	Claude 2 is a foundational LLM built by Anthropic, designed to be safer

Model	Release Date	Size (B)	Checkpoints	Description
				and more "steerable" than its previous version. It is conversational and can be used for a variety of tasks like customer support, Q&A, and more. It can process large amounts of text and is well-suited for applications that require handling extensive data, such as documents, emails, FAQs, and chat transcripts.
Tulu	Jun 2023	7, 13, 30, 65	Tulu-7B , Tulu-13B Tulu-30B , Tulu-65B	Tulu is a family of models developed by Allen Institute for AI. The models are LLaMa models that have been fine-tuned on a mixture of instruction datasets, including FLAN V2, CoT, Dolly, Open Assistant 1, GPT4-Alpaca, Code-Alpaca, and ShareGPT. They are designed to follow complex instructions across various NLP tasks
ChatGLM2-6B	Jun 2023	6	ChatGLM2-6B	ChatGLM2-6B is the second-generation version of the open-source bilingual (Chinese-English) chat model ChatGLM-6B. It has improved performance, longer

Model	Release Date	Size (B)	Checkpoints	Description
				context capabilities, more efficient inference, and an open license for academic and commercial use. The model uses a hybrid objective function and has been trained with 1.4T bilingual tokens. It shows substantial improvements in performance on various datasets compared to its first-generation counterpart.
Nous-Hermes-13B	Jun 2023	13	Nous-Hermes-13B	Nous-Hermes-13B is a language model fine-tuned by Nous Research on over 300,000 instructions.
Baize-v2	May 2023	7, 13	Baize-v2-13B	Baize-v2 is an open-source chat model developed by UCSD and Sun Yat-Sen University, fine-tuned with LoRA, and trained with supervised fine-tuning (SFT) and self-distillation with feedback (SDF).
RWKV-4-Raven	May 2023	1.5, 3, 7, 14	RWKV-4-Raven	RWKV-4-Raven is a series of models. These models are fine-tuned on various datasets like Alpaca, CodeAlpaca, Guanaco, GPT4All,

Model	Release Date	Size (B)	Checkpoints	Description
				and ShareGPT. They follow a 100% RNN architecture for the language model.
Guanaco	May 2023	7, 13, 33, 65	Guanaco-7B , Guanaco-13B , Guanaco-33B Guanaco-65B	Guanaco models are open-source chatbots fine-tuned through 4-bit QLoRA tuning of LLaMA base models on the OASST1 dataset. They are intended for research purposes. The models allow for cheap and local experimentation with high-quality chatbot systems.
PaLM_2	May 2023	-	-	A Language Model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM.
Gorilla	May 2023	7	Gorilla	Gorilla: Large Language Model Connected with Massive APIs
RedPajama-INCITE	May 2023	3, 7	RedPajama-INCITE	A family of models including base, instruction-tuned & chat models.
LIMA	May 2023	65	-	A 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully

Model	Release Date	Size (B)	Checkpoints	Description
				curated prompts and responses, without any reinforcement learning or human preference modeling.
Replit Code	May 2023	3	Replit Code	replit-code-v1-3b model is a 2.7B LLM trained on 20 languages from the Stack Dedup v1.2 dataset.
h2oGPT	May 2023	7, 12, 20, 40	h2oGPT	h2oGPT is a LLM fine-tuning framework and chatbot UI with document(s) question-answer capabilities.
CodeGen2	May 2023	1, 3, 7, 16	CodeGen2	Code models for program synthesis.
CodeT5 and CodeT5+	May 2023	16	CodeT5	CodeT5 and CodeT5+ models for Code Understanding and Generation from Salesforce Research.
StarCoder	May 2023	15	StarCoder	StarCoder: A State-of-the-Art LLM for Code
MPT	May 2023	7, 30	MPT-7B , MPT-30B	MosaicML's MPT models are open-source, commercially licensed Large Language Models, offering customizable AI solutions optimized for various NLP tasks.
DLite	May 2023	0.124 - 1.5	DLite-v2-1.5B	Lightweight instruction following models which exhibit

Model	Release Date	Size (B)	Checkpoints	Description
				ChatGPT-like interactivity.
WizardLM	Apr 2023	70, 30, 13	WizardLM-13B , WizardLM-30B , WizardLM-70B	WizardLM is a family of large language models designed to follow complex instructions. The models performs well in coding, mathematical reasoning, and open-domain conversations. The models are license-friendly and adopt a prompt format from Vicuna for multi-turn conversations. The models are developed by the WizardLM Team, designed for various NLP tasks.
FastChat-T5-3B	Apr 2023	3	FastChat-T5-3B	FastChat-T5 is an open-source chatbot trained by fine-tuning Flan-t5-xl (3B parameters) on user-shared conversations collected from ShareGPT. It's based on an encoder-decoder transformer architecture and can autoregressively generate responses to users' inputs.
GPT4All-13B-Snoozy	Apr 2023	13	GPT4All-13B-Snoozy	GPT4All-13B-Snoozy is a GPL licensed chatbot trained over a massive curated

Model	Release Date	Size (B)	Checkpoints	Description
				corpus of assistant interactions including word problems, multi-turn dialogue, code, poems, songs, and stories. It has been finetuned from Llama 13B and is developed by Nomic AI. The model is designed for assistant-style interaction data and is primarily in English.
Koala-13B	Apr 2023	13	Koala-13B	Koala-13B is a chatbot created by Berkeley AI Research (BAIR). It is fine-tuned on Meta's LLaMA and focuses on dialogue data scraped from the web. The model aims to balance performance and cost, providing a lighter, open-source alternative to models like ChatGPT. It has been trained on interaction data that includes conversations with highly capable closed-source models such as ChatGPT.
OpenAssistant (Llama family)	Apr 2023	30, 70	Llama2-30b-oasst , Llama2-70b-oasst	OpenAssistant-LLaMA models are language models from OpenAssistant's work on the Llama models.

Model	Release Date	Size (B)	Checkpoints	Description
				It supports CPU + GPU inference using GGML format and aims to provide an open-source alternative for instruction following tasks
Dolly	Apr 2023	3, 7, 12	Dolly-v2-3B , Dolly-v2-7B , Dolly-v2-12B	An instruction-following LLM, fine-tuned on a human-generated instruction dataset licensed for research and commercial use.
StableLM	Apr 2023	3, 7	StableLM-Alpha-3B , StableLM-Alpha-7B	Stability AI's StableLM series of language models
Pythia	Apr 2023	0.070 - 12	Pythia	A suite of 16 LLMs all trained on public data seen in the exact same order and ranging in size from 70M to 12B parameters.
Open Assistant (Pythia Family)	Mar 2023	12	Open Assistant	OpenAssistant is a chat-based assistant that understands tasks, can interact with third-party systems, and retrieve information dynamically to do so.
Med-PaLM 2	Mar 2023	-	-	Towards Expert-Level Medical Question Answering with Large Language Models

Model	Release Date	Size (B)	Checkpoints	Description
ChatGLM-6B	Mar 2023	6	ChatGLM-6B	ChatGLM-6B, is an open-source, Chinese-English bilingual dialogue model based on the General Language Model (GLM) architecture with 6.2 billion parameters. Despite its small size causing some factual or mathematical logic issues, it's adept for Chinese question-answering, summarization, and conversational tasks due to its training on over 1 trillion English and Chinese tokens
GPT-3.5-turbo	Mar 2023	175	-	GPT-3.5-Turbo is OpenAI's advanced language model optimized for chat but also works well for traditional completion tasks. It offers better performance across all aspects compared to GPT-3 and is 10 times cheaper per token.
Vicuna	Mar 2023	7, 13, 33	Vicuna-7B , Vicuna-13B	Vicuna is a family of auto-regressive language models based on the transformer architecture. It's fine-tuned from LLaMA and primarily intended for research

Model	Release Date	Size (B)	Checkpoints	Description
				on large language models and chatbots. It's developed by LMSYS and has a non-commercial license.
Alpaca-13B	Mar 2023	13	-	Alpaca is an instruction-following language model fine-tuned from Meta's LLaMA 7B. It's designed for academic research to address issues like misinformation and toxicity. Alpaca is trained on 52K instruction-following demonstrations and aims to be a more accessible option for academic study. It's not intended for commercial use due to licensing and safety concerns.
Claude-1	Mar 2023	137	-	Claude is foundational a large language model (LLM) built by Anthropic. It is designed to be a helpful, honest, and harmless AI assistant. It can perform a wide variety of conversational and text processing tasks and is accessible through a chat interface and API.

Model	Release Date	Size (B)	Checkpoints	Description
Cerebras-GPT	Mar 2023	0.111 - 13	Cerebras-GPT	Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster
BloombergGPT	Mar 2023	50	-	BloombergGPT: A Large Language Model for Finance
PanGu-Σ	Mar 2023	1085	-	PanGu-Σ: Towards Trillion Parameter Language Model with Sparse Heterogeneous Computing
GPT-4	Mar 2023	-	-	GPT-4 Technical Report
LLaMA	Feb 2023	7, 13, 33, 65	LLaMA	LLaMA: Open and Efficient Foundation Language Models
ChatGPT	Nov 2022	-	-	A model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
Galactica	Nov 2022	0.125 - 120	Galactica	Galactica: A Large Language Model for Science
mT0	Nov 2022	13	mT0-xxl	Crosslingual Generalization

Model	Release Date	Size (B)	Checkpoints	Description
				through Multitask Finetuning
BLOOM	Nov 2022	176	BLOOM	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
U-PaLM	Oct 2022	540	-	Transcending Scaling Laws with 0.1% Extra Compute
UL2	Oct 2022	20	UL2, Flan-UL2	UL2: Unifying Language Learning Paradigms
Sparrow	Sep 2022	70	-	Improving alignment of dialogue agents via targeted human judgements
Flan-T5	Oct 2022	11	Flan-T5-xxl	Scaling Instruction-Finetuned Language Models
AlexaTM	Aug 2022	20	-	AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2Seq Model
GLM-130B	Oct 2022	130	GLM-130B	GLM-130B: An Open Bilingual Pre-trained Model
OPT-IML	Dec 2022	30, 175	OPT-IML	OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization
OPT	May 2022	175	OPT-13B, OPT-66B	OPT: Open Pre-trained Transformer Language Models


Model	Release Date	Size (B)	Checkpoints	Description
PaLM	Apr 2022	540	-	PaLM: Scaling Language Modeling with Pathways
Tk-Instruct	Apr 2022	11	Tk-Instruct-11B	Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks
GPT-NeoX-20B	Apr 2022	20	GPT-NeoX-20B	GPT-NeoX-20B: An Open-Source Autoregressive Language Model
Chinchilla	Mar 2022	70	-	Shows that for a compute budget, the best performances are not achieved by the largest models but by smaller models trained on more data.
InstructGPT	Mar 2022	175	-	Training language models to follow instructions with human feedback
CodeGen	Mar 2022	0.350 - 16	CodeGen	CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis
AlphaCode	Feb 2022	41	-	Competition-Level Code Generation with AlphaCode
MT-NLG	Jan 2022	530	-	Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-

Model	Release Date	Size (B)	Checkpoints	Description
				Scale Generative Language Model
LaMDA	Jan 2022	137	-	LaMDA: Language Models for Dialog Applications
GLaM	Dec 2021	1200	-	GLaM: Efficient Scaling of Language Models with Mixture-of-Experts
Gopher	Dec 2021	280	-	Scaling Language Models: Methods, Analysis & Insights from Training Gopher
WebGPT	Dec 2021	175	-	WebGPT: Browser-assisted question-answering with human feedback
Yuan 1.0	Oct 2021	245	-	Yuan 1.0: Large-Scale Pre-trained Language Model in Zero-Shot and Few-Shot Learning
I0	Oct 2021	11	I0	Multitask Prompted Training Enables Zero-Shot Task Generalization
ELAN	Sep 2021	137	-	Finetuned Language Models Are Zero-Shot Learners
HyperCLOVA	Sep 2021	82	-	What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative

Model	Release Date	Size (B)	Checkpoints	Description
				Pretrained Transformers
ERNIE 3.0 Titan	Jul 2021	10	-	ERNIE 3.0 Titan: Exploring Larger-scale Knowledge Enhanced Pre-training for Language Understanding and Generation
Jurassic-1	Aug 2021	178	-	Jurassic-1: Technical Details and Evaluation
ERNIE 3.0	Jul 2021	10	-	ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation
Codex	Jul 2021	12	-	Evaluating Large Language Models Trained on Code
GPT-J-6B	Jun 2021	6	GPT-J-6B	A 6 billion parameter, autoregressive text generation model trained on The Pile.
CPM-2	Jun 2021	198	CPM	CPM-2: Large-scale Cost-effective Pre-trained Language Models
PanGu-α	Apr 2021	13	PanGu-α	PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation
mT5	Oct 2020	13	mT5	mT5: A massively multilingual pre-

Model	Release Date	Size (B)	Checkpoints	Description
				trained text-to-text transformer
BART	Jul 2020	-	BART	Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
GShard	Jun 2020	600	-	GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding
GPT-3	May 2020	175	-	Language Models are Few-Shot Learners
CTRL	Sep 2019	1.63	CTRL	CTRL: A Conditional Transformer Language Model for Controllable Generation
ALBERT	Sep 2019	0.235	ALBERT	A Lite BERT for Self-supervised Learning of Language Representations
XLNet	Jun 2019	-	XLNet	Generalized Autoregressive Pretraining for Language Understanding and Generation
I5	Oct 2019	0.06 - 11	Flan-I5	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
GPT-2	Nov 2019	1.5	GPT-2	Language Models are Unsupervised Multitask Learners

Model	Release Date	Size (B)	Checkpoints	Description
RoBERTa	Jul 2019	0.125 - 0.355	RoBERTa	A Robustly Optimized BERT Pretraining Approach
BERT	Oct 2018	-	BERT	Bidirectional Encoder Representations from Transformers
GPT	Jun 2018	-	GPT	Improving Language Understanding by Generative Pre-Training

 This section is under development.

Data adopted from [Papers with Code](#) and the recent work by [Zhao et al. \(2023\)](#).

Last updated on September 19, 2024