

Filter by name	Featured	~
llama3		
Meta Llama 3: The most capable openly available LLM to date		
↓ 1.2M Pulls ○ 67 Tags ○ Updated 6 days ago		
phi3		
Phi-3 Mini is a 3.8B parameters, lightweight, state-of-the-art open model by Microsoft.		
⊥ 167.2K Pulls 🦴 6 Tags 🕒 Updated 2 weeks ago		
wizardlm2		
State of the art large language model from Microsoft AI with improved performance on complex chat, multilingual, reasoning and agent use cases.		
↓ 48.7K Pulls ○ 22 Tags ○ Updated 3 weeks ago		
mistral		

gemma

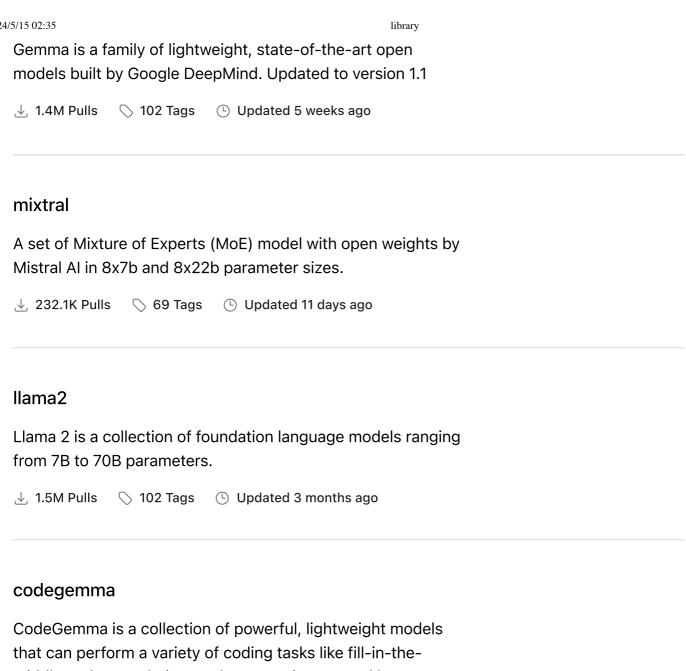
→ 751.9K Pulls

📏 68 Tags

https://ollama.com/library 1/17

Updated 7 weeks ago

2024/5/15 02:35



middle code completion, code generation, natural language understanding, mathematical reasoning, and instruction following.

command-r

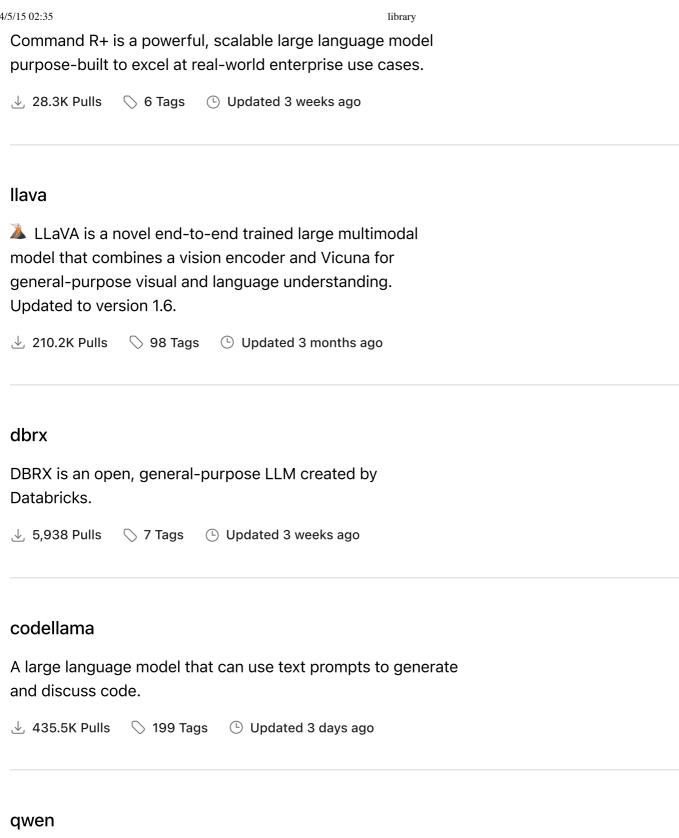
Command R is a Large Language Model optimized for conversational interaction and long context tasks.

Updated 6 weeks ago

command-r-plus

2/17 https://ollama.com/library

2024/5/15 02:35



Cloud spanning from 0.5B to 110B parameters

Qwen 1.5 is a series of large language models by Alibaba

dolphin-mixtral

3/17 https://ollama.com/library

Uncensored, 8x7b and 8x22b fine-tuned models based on the Mixtral mixture of experts models that excels at coding tasks. Created by Eric Hartford.

ightharpoonup 233.7K Pulls ightharpoonup 87 Tags ightharpoonup Updated 10 days ago

llama2-uncensored

Uncensored Llama 2 model by George Sung and Jarrad Hope.

ightharpoonup 184K Pulls ightharpoonup 34 Tags ightharpoonup Updated 6 months ago

deepseek-coder

DeepSeek Coder is a capable coding model trained on two trillion code and natural language tokens.

 \downarrow 134.1K Pulls \bigcirc 102 Tags \bigcirc Updated 4 months ago

mistral-openorca

Mistral OpenOrca is a 7 billion parameter model, fine-tuned on top of the Mistral 7B model using the OpenOrca dataset.

nomic-embed-text

A high-performing open embedding model with a large token context window.

dolphin-mistral

The uncensored Dolphin model based on Mistral that excels at coding tasks. Updated to version 2.8.

https://ollama.com/library 4/17

library **→** 98.3K Pulls Updated 6 weeks ago

phi

Phi-2: a 2.7B language model by Microsoft Research that demonstrates outstanding reasoning and language understanding capabilities.

√ 18 Tags (L) Updated 3 months ago

orca-mini

A general-purpose model ranging from 3 billion parameters to 70 billion, suitable for entry-level hardware.

Updated 6 months ago

nous-hermes2

The powerful family of models by Nous Research that excels at scientific discussion and coding tasks.

zephyr

Zephyr is a series of fine-tuned versions of the Mistral and Mixtral models that are trained to act as helpful assistants.

(L) Updated 4 weeks ago ± 68.3K Pulls 40 Tags

llama2-chinese

Llama 2 based model fine tuned to improve Chinese dialogue ability.

 ↓ 61.2K Pulls (L) Updated 6 months ago √ 35 Tags

https://ollama.com/library 5/17

wizard-vicuna-uncensored

Wizard Vicuna Uncensored is a 7B, 13B, and 30B parameter model based on Llama 2 uncensored by Eric Hartford.

vicuna

General use chat model based on Llama and Llama 2 with 2K to 16K context sizes.

starcoder2

StarCoder2 is the next generation of transparently trained open code LLMs that comes in three sizes: 3B, 7B and 15B parameters.

openhermes

OpenHermes 2.5 is a 7B model fine-tuned by Teknium on Mistral with fully open datasets.

tinyllama

The TinyLlama project is an open endeavor to train a compact 1.1B Llama model on 3 trillion tokens.

 \bot 48.6K Pulls \bigcirc 36 Tags \bigcirc Updated 4 months ago

openchat

A family of open-source models trained on a wide variety of data, surpassing ChatGPT on various benchmarks. Updated to version 3.5-0106.

starcoder

StarCoder is a code generation model trained on 80+ programming languages.

tinydolphin

An experimental 1.1B parameter model trained on the new Dolphin 2.8 dataset by Eric Hartford and based on TinyLlama.

dolphin-llama3

Dolphin 2.9 is a new model with 8B and 70B sizes by Eric Hartford based on Llama 3 that has a variety of instruction, conversational, and coding skills.

wizardcoder

State-of-the-art code generation model

yi

Yi 1.5 is a high-performing, bilingual language model.

https://ollama.com/library 7/17

stable-code

Stable Code 3B is a coding model with instruct and code completion variants on par with models such as Code Llama 7B that are 2.5x larger.

ightharpoonup 37.6K Pulls ightharpoonup 36 Tags ightharpoonup Updated 7 weeks ago

mxbai-embed-large

State-of-the-art large embedding model from mixedbread.ai

neural-chat

A fine-tuned model based on Mistral with good coverage of domain and language.

 \bot 32.7K Pulls \bigcirc 50 Tags \bigcirc Updated 6 weeks ago

phind-codellama

Code generation model based on Code Llama.

wizard-math

Model focused on math and logic problems

ightharpoonup 28.7K Pulls ightharpoonup 64 Tags ightharpoonup Updated 4 months ago

starling-lm

https://ollama.com/library 8/17

Starling is a large language model trained by reinforcement learning from AI feedback focused on improving chatbot helpfulness.

	s 🚫 36 Tags	Updated	5 weeks ago
--	-------------	---------	-------------

falcon

A large language model built by the Technology Innovation Institute (TII) for use in summarization, text generation, and chat bots.

dolphincoder

A 7B and 15B uncensored variant of the Dolphin model family that excels at coding, based on StarCoder2.



orca2

Orca 2 is built by Microsoft research, and are a fine-tuned version of Meta's Llama 2 models. The model is designed to excel particularly in reasoning.

 $_$ 24.1K Pulls \bigcirc 33 Tags \bigcirc Updated 5 months ago

nous-hermes

General use models based on Llama and Llama 2 from Nous Research.

\downarrow	23.3K Pulls	Updated 6 months ago

dolphin-phi

https://ollama.com/library 9/17

2.7B uncensored Dolphin model by Eric Hartford, based on the Phi language model by Microsoft Research.				
sqlcoder				
SQLCoder is a code completion model fined-tuned on StarCoder for SQL generation tasks				
stablelm2				
Stable LM 2 is a state-of-the-art 1.6B and 12B parameter language model trained on multilingual data in English, Spanish, German, Italian, French, Portuguese, and Dutch.				
solar				
A compact, yet powerful 10.7B large language model designed for single-turn conversation.				
deepseek-Ilm				
An advanced language model crafted with 2 trillion bilingual tokens.				

yarn-llama2

An extension of Llama 2 that supports a context of up to 128k tokens.

https://ollama.com/library 10/17

± 17.9K Pulls ♦ 67 Tags • Updated 6 months ago

bakllava

BakLLaVA is a multimodal model consisting of the Mistral 7B base model augmented with the LLaVA architecture.

 \bot 17.6K Pulls \bigcirc 17 Tags \bigcirc Updated 5 months ago

codeqwen

CodeQwen1.5 is a large language model pretrained on a large amount of code data.

 \bot 17.2K Pulls \bigcirc 21 Tags \bigcirc Updated 3 weeks ago

medllama2

Fine-tuned Llama 2 model to answer medical questions based on an open source medical dataset.

samantha-mistral

A companion assistant trained in philosophy, psychology, and personal relationships. Based on Mistral.

ightharpoonup 16.7K Pulls ightharpoonup 49 Tags ightharpoonup Updated 7 months ago

all-minilm

Embedding models on very large sentence level datasets.

ightharpoonup 16.6K Pulls ightharpoonup 10 Tags ightharpoonup Updated 7 days ago

https://ollama.com/library 11/17

wizardlm-uncensored

Uncensored version of Wizard LM model

 \perp 16.3K Pulls \bigcirc 18 Tags \bigcirc Updated 6 months ago

nous-hermes2-mixtral

The Nous Hermes 2 model from Nous Research, now trained over Mixtral.

 \bot 16.2K Pulls \bigcirc 18 Tags \bigcirc Updated 3 months ago

llama3-gradient

This model extends LLama-3 8B's context length from 8k to over 1m tokens.

 $oldsymbol{\perp}$ 15.5K Pulls igtriangle 35 Tags igtriangle Updated 9 days ago

stable-beluga

Llama 2 based model fine tuned on an Orca-style dataset. Originally called Free Willy.

 \bot 15.2K Pulls \bigcirc 49 Tags \bigcirc Updated 6 months ago

xwinlm

Conversational model based on Llama 2 that performs competitively on various benchmarks.

codeup

Great code generation model based on Llama2.

https://ollama.com/library 12/17

(L) Updated 6 months ago

everythinglm

Uncensored Llama2 based model with support for a 16K context window.

Updated 4 months ago

wizardlm

General use model based on Llama 2.

yarn-mistral

An extension of Mistral to support context windows of 64K or 128K.

⊥ 13.8K Pulls

meditron

Open-source medical large language model adapted from Llama 2 to the medical domain.

llama-pro

An expansion of Llama 2 that specializes in integrating both general language understanding and domain-specific knowledge, particularly in programming and mathematics.

https://ollama.com/library 13/17

magicoder

Magicoder is a family of 7B parameter models trained on
75K synthetic instruction data using OSS-Instruct, a novel
approach to enlightening LLMs with open-source code
snippets.

ightharpoonup 10.7K Pulls ightharpoonup 18 Tags ightharpoonup Updated 5 months ago

stablelm-zephyr

A lightweight chat model allowing accurate, and responsive output without requiring high-end hardware.

 \perp 10.6K Pulls \bigcirc 17 Tags \bigcirc Updated 4 months ago

nexusraven

Nexus Raven is a 13B instruction tuned model for function calling tasks.

codebooga

A high-performing code instruct model created by merging two existing code models.

 \bot 10.1K Pulls \bigcirc 16 Tags \bigcirc Updated 6 months ago

mistrallite

MistralLite is a fine-tuned model based on Mistral with enhanced capabilities of processing long contexts.

 \downarrow 9,531 Pulls \bigcirc 17 Tags \bigcirc Updated 6 months ago

https://ollama.com/library 14/17

wizard-vicuna

Wizard Vicuna is a 13B parameter model based on Llama 2 trained by MelodysDreamj.

goliath

A language model created by combining two fine-tuned Llama 2 70B models into one.

 \pm 7,251 Pulls \bigcirc 16 Tags \bigcirc Updated 5 months ago

open-orca-platypus2

Merge of the Open Orca OpenChat model and the GaragebAlnd Platypus 2 model. Designed for chat and code generation.

snowflake-arctic-embed

A suite of text embedding models by Snowflake, optimized for performance.

notux

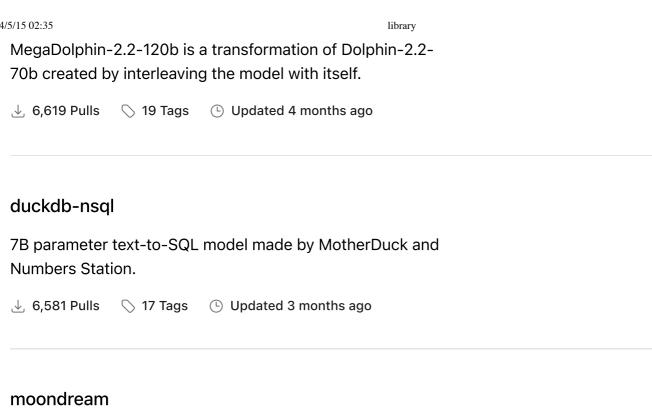
A top-performing mixture of experts model, fine-tuned with high-quality data.

 $\,\,$ $\,\,$ $\,\,$ 6,709 Pulls $\,\,$ $\,\,$ $\,$ 18 Tags $\,\,$ $\,$ $\,$ $\,$ Updated 4 months ago

megadolphin

https://ollama.com/library 15/17

2024/5/15 02:35



moondream2 is a small vision language model designed to run efficiently on edge devices.

Updated yesterday

Ilama3-chatqa

A model from NVIDIA based on Llama 3 that excels at conversational question answering (QA) and retrievalaugmented generation (RAG).

notus

A 7B chat model fine-tuned with high-quality data and based on Zephyr.

Updated 4 months ago

llava-llama3

A LLaVA model fine-tuned from Llama 3 Instruct with better scores in several benchmarks.

https://ollama.com/library 16/17

alfred

A robust conversational model designed to be used for both chat and instruct use cases.

Ilava-phi3

A new small LLaVA model fine-tuned from Phi 3 Mini.

falcon2

Falcon2 is an 11B parameters causal decoder-only model built by TII and trained over 5T tokens.

Blog Docs GitHub

Discord X (Twitter) Meetups

© 2024 Ollama

https://ollama.com/library 17/17