Models

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

The 7B model released by Mistral AI, updated to version 0.2.

751.9K Pulls 68 Tags Updated 7 weeks ago

gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

1.4M Pulls 102 Tags Updated 5 weeks ago

mixtral

A set of Mixture of Experts (MoE) model with open weights by Mistral AI in 8x7b and 8x22b parameter sizes.

232.1K Pulls 69 Tags Updated 11 days ago

llama2

Llama 2 is a collection of foundation language models ranging from 7B to 70B parameters.

1.5M Pulls 102 Tags Updated 3 months ago

codegemma

CodeGemma is a collection of powerful, lightweight models that can perform a variety of coding tasks like fill-in-the-middle code completion, code generation, natural language understanding, mathematical reasoning, and instruction following.

68.8K Pulls 85 Tags Updated 10 days ago

command-r

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

35K Pulls 17 Tags Updated 6 weeks ago

command-r-plus

Command R+ is a powerful, scalable large language model purpose-built to excel at real-world enterprise use cases.

28.3K Pulls 6 Tags Updated 3 weeks ago

---

llava

🌋 LLaVA is a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding. Updated to version 1.6.

210.2K Pulls 98 Tags Updated 3 months ago

---

dbrx

DBRX is an open, general-purpose LLM created by Databricks.

5,938 Pulls 7 Tags Updated 3 weeks ago

---

codellama

A large language model that can use text prompts to generate and discuss code.

435.5K Pulls 199 Tags Updated 3 days ago

---

qwen

Qwen 1.5 is a series of large language models by Alibaba Cloud spanning from 0.5B to 110B parameters

334.1K Pulls 379 Tags Updated 2 weeks ago

---

dolphin-mixtral

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

233.7K Pulls 87 Tags Updated 10 days ago

---

llama2-uncensored

Uncensored Llama 2 model by George Sung and Jarrad Hope.

184K Pulls 34 Tags Updated 6 months ago

---

deepseek-coder

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

134.1K Pulls 102 Tags 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.

123.2K Pulls 17 Tags Updated 7 months ago

---

nomic-embed-text

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

105.4K Pulls 3 Tags Updated 2 months ago

---

dolphin-mistral

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

98.3K Pulls 120 Tags Updated 6 weeks ago

phi

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

96K Pulls 18 Tags Updated 3 months ago

orca-mini

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

92K Pulls 119 Tags Updated 6 months ago

nous-hermes2

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

78.7K Pulls 33 Tags Updated 4 months ago

zephyr

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

68.3K Pulls 40 Tags Updated 4 weeks ago

llama2-chinese

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

61.2K Pulls 35 Tags Updated 6 months ago

wizard-vicuna-uncensored

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

58.2K Pulls 49 Tags Updated 6 months ago

---

vicuna

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

55.2K Pulls 111 Tags Updated 6 months ago

---

starcoder2

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

54.9K Pulls 67 Tags Updated 12 days ago

---

openhermes

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

49.1K Pulls 35 Tags Updated 4 months ago

---

tinyllama

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

48.6K Pulls 36 Tags 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.

44.8K Pulls 50 Tags Updated 4 months ago

---

starcoder

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

43.5K Pulls 100 Tags Updated 6 months ago

---

tinydolphin

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

42.5K Pulls 18 Tags Updated 3 months ago

---

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.

40.6K Pulls 54 Tags Updated 3 days ago

---

wizardcoder

State-of-the-art code generation model

39.6K Pulls 67 Tags Updated 4 months ago

---

yi

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

39.1K Pulls 174 Tags Updated yesterday

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.

37.6K Pulls 36 Tags Updated 7 weeks ago

mxbai-embed-large

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

35K Pulls 4 Tags Updated 7 days ago

neural-chat

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

32.7K Pulls 50 Tags Updated 6 weeks ago

phind-codellama

Code generation model based on Code Llama.

29.5K Pulls 49 Tags Updated 4 months ago

wizard-math

Model focused on math and logic problems

28.7K Pulls 64 Tags Updated 4 months ago

starling-lm

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

26.6K Pulls 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.

24.9K Pulls 38 Tags Updated 6 months ago

---

dolphincoder

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

24.3K Pulls 35 Tags Updated 5 weeks ago

---

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 33 Tags Updated 5 months ago

---

nous-hermes

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

23.3K Pulls 63 Tags Updated 6 months ago

---

dolphin-phi

2.7B uncensored Dolphin model by Eric Hartford, based on the Phi language model by Microsoft Research.

22.4K Pulls 15 Tags Updated 4 months ago

---

sqlcoder

SQLCoder is a code completion model fined-tuned on StarCoder for SQL generation tasks

21.7K Pulls 48 Tags Updated 3 months ago

---

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.

21.3K Pulls 84 Tags Updated 7 days ago

---

solar

A compact, yet powerful 10.7B large language model designed for single-turn conversation.

20.1K Pulls 32 Tags Updated 4 months ago

---

deepseek-llm

An advanced language model crafted with 2 trillion bilingual tokens.

18K Pulls 64 Tags Updated 5 months ago

---

yarn-llama2

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

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.

17.6K Pulls 17 Tags Updated 5 months ago

codeqwen

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

17.2K Pulls 21 Tags Updated 3 weeks ago

medllama2

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

17K Pulls 17 Tags Updated 6 months ago

samantha-mistral

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

16.7K Pulls 49 Tags Updated 7 months ago

all-minilm

Embedding models on very large sentence level datasets.

16.6K Pulls 10 Tags Updated 7 days ago

wizardlm-uncensored

Uncensored version of Wizard LM model

16.3K Pulls 18 Tags Updated 6 months ago

---

nous-hermes2-mixtral

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

16.2K Pulls 18 Tags Updated 3 months ago

---

llama3-gradient

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

15.5K Pulls 35 Tags Updated 9 days ago

---

stable-beluga

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

15.2K Pulls 49 Tags Updated 6 months ago

---

xwinlm

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

15.1K Pulls 80 Tags Updated 6 months ago

---

codeup

Great code generation model based on Llama2.

14.7K Pulls 19 Tags Updated 6 months ago

everythinglm

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

13.9K Pulls 18 Tags Updated 4 months ago

wizardlm

General use model based on Llama 2.

13.8K Pulls 73 Tags Updated 4 weeks ago

yarn-mistral

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

13.8K Pulls 33 Tags Updated 4 months ago

meditron

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

12.9K Pulls 22 Tags Updated 5 months ago

llama-pro

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

12.7K Pulls 33 Tags Updated 4 months ago

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.

10.7K Pulls 18 Tags Updated 5 months ago

---

stablelm-zephyr

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

10.6K Pulls 17 Tags Updated 4 months ago

---

nexusraven

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

10.3K Pulls 32 Tags Updated 3 months ago

---

codebooga

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

10.1K Pulls 16 Tags Updated 6 months ago

---

mistrallite

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

9,531 Pulls 17 Tags Updated 6 months ago

---

wizard-vicuna

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

8,944 Pulls 17 Tags Updated 6 months ago

---

goliath

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

7,251 Pulls 16 Tags Updated 5 months ago

---

open-orca-platypus2

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

7,012 Pulls 17 Tags Updated 6 months ago

---

snowflake-arctic-embed

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

6,925 Pulls 16 Tags Updated 3 weeks ago

---

notux

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

6,709 Pulls 18 Tags Updated 4 months ago

---

megadolphin

MegaDolphin-2.2-120b is a transformation of Dolphin-2.2-70b created by interleaving the model with itself.

6,619 Pulls 19 Tags Updated 4 months ago

---

duckdb-nsql

7B parameter text-to-SQL model made by MotherDuck and Numbers Station.

6,581 Pulls 17 Tags Updated 3 months ago

---

moondream

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

6,421 Pulls 19 Tags Updated yesterday

---

llama3-chatqa

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

5,807 Pulls 35 Tags Updated 3 days ago

---

notus

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

5,755 Pulls 18 Tags Updated 4 months ago

---

llava-llama3

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

5,599 Pulls 4 Tags Updated 6 days ago

alfred

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

4,959 Pulls 7 Tags Updated 5 months ago

llava-phi3

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

3,190 Pulls 4 Tags Updated 6 days ago

falcon2

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

1,154 Pulls 17 Tags Updated 17 hours ago