



Embedding

Vision

Tools

Popular



gemma3

The current, most capable model that runs on a single GPU.

[vision](#) [1b](#) [4b](#) [12b](#) [27b](#)[↓ 3.5M Pulls](#) [🏷 17 Tags](#) [🕒 Updated 3 weeks ago](#)

qwq

QwQ is the reasoning model of the Qwen series.

[tools](#) [32b](#)[↓ 1.3M Pulls](#) [🏷 8 Tags](#) [🕒 Updated 4 weeks ago](#)

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

[1.5b](#) [7b](#) [8b](#) [14b](#) [32b](#) [70b](#) [671b](#)[↓ 36.8M Pulls](#) [🏷 29 Tags](#) [🕒 Updated 2 months ago](#)

mistral-small3.1

Building upon Mistral Small 3, Mistral Small 3.1 (2503) adds state-of-the-art vision understanding and enhances long context capabilities up to 128k tokens without compromising text performance.

[vision](#) [tools](#) [24b](#)[↓ 28.2K Pulls](#) [🏷 5 Tags](#) [🕒 Updated 8 days ago](#)

llama3.3

New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.

[tools](#) [70b](#)

↓ 1.7M Pulls 🏷 14 Tags ⌚ Updated 4 months ago

phi4

Phi-4 is a 14B parameter, state-of-the-art open model from Microsoft.

[14b](#)

↓ 1.6M Pulls 🏷 5 Tags ⌚ Updated 3 months ago

llama3.2

Meta's Llama 3.2 goes small with 1B and 3B models.

[tools](#) [1b](#) [3b](#)

↓ 13.2M Pulls 🏷 63 Tags ⌚ Updated 6 months ago

llama3.1

Llama 3.1 is a new state-of-the-art model from Meta available in 8B, 70B and 405B parameter sizes.

[tools](#) [8b](#) [70b](#) [405b](#)

↓ 90M Pulls 🏷 93 Tags ⌚ Updated 4 months ago

nomic-embed-text

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

[embedding](#)

↓ 22.4M Pulls 🏷 3 Tags ⌚ Updated 13 months ago

mistral

The 7B model released by Mistral AI, updated to version 0.3.

[tools](#) [7b](#)

↓ 11.7M Pulls 🏷 84 Tags ⌚ Updated 8 months ago

llama3

Meta Llama 3: The most capable openly available LLM to date

[8b](#) [70b](#)

↓ 7.8M Pulls 🏷 68 Tags ⌚ Updated 10 months ago

qwen2.5

Qwen2.5 models are pretrained on Alibaba's latest large-scale dataset, encompassing up to 18 trillion tokens. The model supports up to 128K tokens and has multilingual support.

[tools](#) [0.5b](#) [1.5b](#) [3b](#) [7b](#) [14b](#) [32b](#) [72b](#)

↓ 6.7M Pulls 🏷 133 Tags ⌚ Updated 6 months ago

qwen2.5-coder

The latest series of Code-Specific Qwen models, with significant improvements in code generation, code reasoning, and code fixing.

[tools](#) [0.5b](#) [1.5b](#) [3b](#) [7b](#) [14b](#) [32b](#)

↓ 5M Pulls 🏷 196 Tags ⌚ Updated 5 months 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.

[vision](#) [7b](#) [13b](#) [34b](#)

↓ 4.9M Pulls 🏷 98 Tags ⌚ Updated 14 months ago

qwen

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

[0.5b](#) [1.8b](#) [4b](#) [7b](#) [14b](#) [32b](#) [72b](#) [110b](#)

↓ 4.6M Pulls 🏷 379 Tags ⌚ Updated 11 months ago

gemma

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

[2b](#) [7b](#)

↓ 4.5M Pulls 🏷 102 Tags ⌚ Updated 12 months ago

qwen2

Qwen2 is a new series of large language models from Alibaba group

[tools](#) [0.5b](#) [1.5b](#) [7b](#) [72b](#)

↓ 4.2M Pulls 🏷 97 Tags ⌚ Updated 7 months ago

gemma2

Google Gemma 2 is a high-performing and efficient model available in three sizes: 2B, 9B, and 27B.

[2b](#) [9b](#) [27b](#)

↓ 4M Pulls 🏷 94 Tags ⌚ Updated 8 months ago

llama2

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

[7b](#) [13b](#) [70b](#)

↓ 3.2M Pulls 🏷 102 Tags ⌚ Updated 15 months ago

phi3

Phi-3 is a family of lightweight 3B (Mini) and 14B (Medium) state-of-the-art open models by Microsoft.

[3.8b](#) [14b](#)

↓ 3M Pulls ↗ 72 Tags ⌚ Updated 8 months ago

mxlbai-embed-large

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

[embedding](#) [335m](#)

↓ 2.4M Pulls ↗ 4 Tags ⌚ Updated 11 months ago

codellama

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

[7b](#) [13b](#) [34b](#) [70b](#)

↓ 1.9M Pulls ↗ 199 Tags ⌚ Updated 9 months ago

llama3.2-vision

Llama 3.2 Vision is a collection of instruction-tuned image reasoning generative models in 11B and 90B sizes.

[vision](#) [11b](#) [90b](#)

↓ 1.8M Pulls ↗ 9 Tags ⌚ Updated 5 months ago

mistral-nemo

A state-of-the-art 12B model with 128k context length, built by Mistral AI in collaboration with NVIDIA.

[tools](#) [12b](#)

↓ 1.5M Pulls ↗ 17 Tags ⌚ Updated 8 months ago

tinylama

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


[1.1b](#)

↓ 1.4M Pulls ↗ 36 Tags ⌚ Updated 15 months ago

minicpm-v

A series of multimodal LLMs (MLLMs) designed for vision-language understanding.

[vision](#) [8b](#)

↓ 1M Pulls  17 Tags  Updated 4 months ago

deepseek-v3

A strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token.

[671b](#)

↓ 1M Pulls  5 Tags  Updated 3 months ago

mixtral

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

[tools](#) [8x7b](#) [8x22b](#)

↓ 951.4K Pulls  70 Tags  Updated 3 months ago

llama2-uncensored

Uncensored Llama 2 model by George Sung and Jarrad Hope.

[7b](#) [70b](#)

↓ 935.9K Pulls  34 Tags  Updated 17 months ago

starcoder2

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

[3b](#) [7b](#) [15b](#)

↓ 916.4K Pulls  67 Tags  Updated 7 months ago

bge-m3

BGE-M3 is a new model from BAAI distinguished for its versatility in Multi-Functionality, Multi-Linguality, and Multi-Granularity.

[embedding](#) [567m](#)

↓ 863K Pulls 🏷 3 Tags ⌚ Updated 8 months ago

deepseek-coder-v2

An open-source Mixture-of-Experts code language model that achieves performance comparable to GPT4-Turbo in code-specific tasks.

[16b](#) [236b](#)

↓ 782.2K Pulls 🏷 64 Tags ⌚ Updated 7 months ago

dolphin3

Dolphin 3.0 Llama 3.1 8B 🐬 is the next generation of the Dolphin series of instruct-tuned models designed to be the ultimate general purpose local model, enabling coding, math, agentic, function calling, and general use cases.

[8b](#)

↓ 773.3K Pulls 🏷 5 Tags ⌚ Updated 3 months ago

olmo2

OLMo 2 is a new family of 7B and 13B models trained on up to 5T tokens. These models are on par with or better than equivalently sized fully open models, and competitive with open-weight models such as Llama 3.1 on English academic benchmarks.

[7b](#) [13b](#)

↓ 729.1K Pulls 🏷 9 Tags ⌚ Updated 3 months ago

snowflake-arctic-embed

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

[embedding](#) [22m](#) [33m](#) [110m](#) [137m](#) [335m](#)

↓ 708.1K Pulls 🏷 16 Tags ⌚ Updated 12 months ago

deepseek-coder

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

[1.3b](#) [6.7b](#) [33b](#)

↓ 658.6K Pulls 🏷 102 Tags ⌚ Updated 15 months ago

llava-llama3

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

[vision](#) [8b](#)

↓ 650K Pulls 🏷 4 Tags ⌚ Updated 11 months ago

smollm2

SmolLM2 is a family of compact language models available in three size: 135M, 360M, and 1.7B parameters.

[tools](#) [135m](#) [360m](#) [1.7b](#)

↓ 574.6K Pulls 🏷 49 Tags ⌚ Updated 5 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.

[2b](#) [7b](#)

↓ 564.5K Pulls 🏷 85 Tags ⌚ Updated 9 months 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.

[8x7b](#) [8x22b](#)

↓ 538.6K Pulls  70 Tags  Updated 3 months ago

mistral-small

Mistral Small 3 sets a new benchmark in the “small” Large Language Models category below 70B.

[tools](#) [22b](#) [24b](#)

↓ 528.4K Pulls  21 Tags  Updated 2 months ago

openthinker

A fully open-source family of reasoning models built using a dataset derived by distilling DeepSeek-R1.

[7b](#) [32b](#)

↓ 519.4K Pulls  15 Tags  Updated 11 days ago

phi

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

[2.7b](#)

↓ 513.4K Pulls  18 Tags  Updated 15 months ago

all-minilm

Embedding models on very large sentence level datasets.

[embedding](#) [22m](#) [33m](#)

↓ 382.4K Pulls  10 Tags  Updated 11 months ago

wizardlm2

State of the art large language model from Microsoft AI with improved performance on complex chat, multilingual, reasoning and agent use cases.

[7b](#) [8x22b](#)

↓ 361.5K Pulls 🏷 22 Tags ⌚ Updated 12 months ago

dolphin-mistral

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

[7b](#)

↓ 335.7K Pulls 🏷 120 Tags ⌚ Updated 12 months ago

orca-mini

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

[3b](#) [7b](#) [13b](#) [70b](#)

↓ 328.6K Pulls 🏷 119 Tags ⌚ Updated 17 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.

[8b](#) [70b](#)

↓ 310.8K Pulls 🏷 53 Tags ⌚ Updated 11 months ago

command-r

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

[tools](#) [35b](#)

↓ 290.2K Pulls 🏷 32 Tags ⌚ Updated 7 months ago

hermes3

Hermes 3 is the latest version of the flagship Hermes series of LLMs by Nous Research

[tools](#) [3b](#) [8b](#) [70b](#) [405b](#)

↓ 274.7K Pulls 🏷 65 Tags ⌚ Updated 4 months ago

yi

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

[6b](#) [9b](#) [34b](#)

↓ 273.2K Pulls 🏷 174 Tags ⌚ Updated 11 months ago

phi3.5

A lightweight AI model with 3.8 billion parameters with performance overtaking similarly and larger sized models.

[3.8b](#)

↓ 263K Pulls 🏷 17 Tags ⌚ Updated 7 months ago

codestral

Codestral is Mistral AI's first-ever code model designed for code generation tasks.

[22b](#)

↓ 252.6K Pulls 🏷 17 Tags ⌚ Updated 7 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.

[7b](#) [141b](#)

↓ 242.9K Pulls 🏷 40 Tags ⌚ Updated 12 months ago

smollm

🍌 A family of small models with 135M, 360M, and 1.7B parameters, trained on a new high-quality dataset.

[135m](#) [360m](#) [1.7b](#)

↓ 217.3K Pulls 🏷 94 Tags ⌚ Updated 7 months ago

granite-code

A family of open foundation models by IBM for Code Intelligence

[3b](#) [8b](#) [20b](#) [34b](#)

↓ 200.1K Pulls 🏷 162 Tags ⌚ Updated 7 months ago

wizard-vicuna-uncensored

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

[7b](#) [13b](#) [30b](#)

↓ 194.1K Pulls 🏷 49 Tags ⌚ Updated 17 months ago

starcoder

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

[1b](#) [3b](#) [7b](#) [15b](#)

↓ 192.3K Pulls 🏷 100 Tags ⌚ Updated 17 months ago

vicuna

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

[7b](#) [13b](#) [33b](#)

↓ 179.4K Pulls 🏷 111 Tags ⌚ Updated 17 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.

[7b](#)

↓ 168.9K Pulls 🏷 17 Tags ⌚ Updated 18 months ago

moondream

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

[vision](#) [1.8b](#)

↓ 165.4K Pulls ↗ 18 Tags ⌚ Updated 11 months ago

llama2-chinese

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

[7b](#) [13b](#)

↓ 151.9K Pulls ↗ 35 Tags ⌚ Updated 17 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.

[7b](#)

↓ 151.9K Pulls ↗ 50 Tags ⌚ Updated 15 months ago

codegeex4

A versatile model for AI software development scenarios, including code completion.

[9b](#)

↓ 142.2K Pulls ↗ 17 Tags ⌚ Updated 9 months ago

deepseek-llm

An advanced language model crafted with 2 trillion bilingual tokens.

[7b](#) [67b](#)

↓ 140.8K Pulls ↗ 64 Tags ⌚ Updated 16 months ago

openhermes

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

↓ 140.4K Pulls ↗ 35 Tags ⌚ Updated 15 months ago

codeqwen

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

7b

↓ 140.3K Pulls ↗ 30 Tags ⌚ Updated 9 months ago

aya

Aya 23, released by Cohere, is a new family of state-of-the-art, multilingual models that support 23 languages.

8b 35b

↓ 140.3K Pulls ↗ 33 Tags ⌚ Updated 10 months ago

deepseek-v2

A strong, economical, and efficient Mixture-of-Experts language model.

16b 236b

↓ 138.5K Pulls ↗ 34 Tags ⌚ Updated 9 months ago

mistral-large

Mistral Large 2 is Mistral's new flagship model that is significantly more capable in code generation, mathematics, and reasoning with 128k context window and support for dozens of languages.

tools 123b

↓ 131.1K Pulls ↗ 32 Tags ⌚ Updated 4 months ago

glm4

A strong multi-lingual general language model with competitive performance to Llama 3.

9b

↓ 128.2K Pulls 🏷 32 Tags ⌚ Updated 9 months ago

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.

[3b](#)

↓ 126.3K Pulls 🏷 36 Tags ⌚ Updated 12 months ago

qwen2-math

Qwen2 Math is a series of specialized math language models built upon the Qwen2 LLMs, which significantly outperforms the mathematical capabilities of open-source models and even closed-source models (e.g., GPT4o).

[1.5b](#) [7b](#) [72b](#)

↓ 124.6K Pulls 🏷 52 Tags ⌚ Updated 7 months ago

tinydolphin

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

[1.1b](#)

↓ 124.5K Pulls 🏷 18 Tags ⌚ Updated 14 months ago

nous-hermes2

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

[10.7b](#) [34b](#)

↓ 124.3K Pulls 🏷 33 Tags ⌚ Updated 15 months ago

command-r-plus

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

[tools](#) [104b](#)

↓ 122.9K Pulls 🏷 21 Tags ⌚ Updated 7 months ago

wizardcoder

State-of-the-art code generation model

33b

↓ 120.2K Pulls 🏷 67 Tags ⌚ Updated 15 months ago

bakllava

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

vision 7b

↓ 113.6K Pulls 🏷 17 Tags ⌚ Updated 16 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.

1.6b 12b

↓ 110.6K Pulls 🏷 84 Tags ⌚ Updated 11 months ago

neural-chat

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

7b

↓ 108K Pulls 🏷 50 Tags ⌚ Updated 16 months ago

reflection

A high-performing model trained with a new technique called Reflection-tuning that teaches a LLM to detect mistakes in its reasoning and correct course.

70b

↓ 104.5K Pulls 🏷 17 Tags ⌚ Updated 7 months ago

wizard-math

Model focused on math and logic problems

7b 13b 70b

↓ 102.9K Pulls 🏷 64 Tags ⌚ Updated 16 months ago

phi4-mini

Phi-4-mini brings significant enhancements in multilingual support, reasoning, and mathematics, and now, the long-awaited function calling feature is finally supported.

tools 3.8b

↓ 102.8K Pulls 🏷 5 Tags ⌚ Updated 6 weeks ago

llama3-chatqa

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

8b 70b

↓ 101.4K Pulls 🏷 35 Tags ⌚ Updated 11 months ago

sqlcoder

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

7b 15b

↓ 100.4K Pulls 🏷 48 Tags ⌚ Updated 14 months ago

llama3-gradient

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


[8b](#) [70b](#)

↓ 100K Pulls  35 Tags  Updated 11 months ago

bge-large

Embedding model from BAAI mapping texts to vectors.

[embedding](#) [335m](#)

↓ 96.9K Pulls  3 Tags  Updated 8 months ago

samantha-mistral

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

[7b](#)

↓ 91.9K Pulls  49 Tags  Updated 18 months ago

granite3.1-dense

The IBM Granite 2B and 8B models are text-only dense LLMs trained on over 12 trillion tokens of data, demonstrated significant improvements over their predecessors in performance and speed in IBM's initial testing.

[tools](#) [2b](#) [8b](#)

↓ 89.5K Pulls  33 Tags  Updated 2 months ago

dolphincoder

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

[7b](#) [15b](#)

↓ 86.8K Pulls  35 Tags  Updated 12 months ago

llava-phi3

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

[vision](#) [3.8b](#)

↓ 86.3K Pulls 🏷 4 Tags ⌚ Updated 11 months ago

xwinlm

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

[7b](#) [13b](#)

↓ 85.5K Pulls 🏷 80 Tags ⌚ Updated 17 months ago

nous-hermes

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

[7b](#) [13b](#)

↓ 84.4K Pulls 🏷 63 Tags ⌚ Updated 17 months ago

starling-lm

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

[7b](#)

↓ 83.1K Pulls 🏷 36 Tags ⌚ Updated 12 months ago

phind-codellama

Code generation model based on Code Llama.

[34b](#)

↓ 83.1K Pulls 🏷 49 Tags ⌚ Updated 15 months ago

solar

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

[10.7b](#) 80.6K Pulls  32 Tags  Updated 16 months ago

yi-coder

Yi-Coder is a series of open-source code language models that delivers state-of-the-art coding performance with fewer than 10 billion parameters.

[1.5b](#) [9b](#) 80.3K Pulls  67 Tags  Updated 7 months ago

yarn-llama2

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

[7b](#) [13b](#) 80K Pulls  67 Tags  Updated 17 months ago

athene-v2

Athene-V2 is a 72B parameter model which excels at code completion, mathematics, and log extraction tasks.

[tools](#) [72b](#) 79K Pulls  17 Tags  Updated 5 months ago

granite3-dense

The IBM Granite 2B and 8B models are designed to support tool-based use cases and support for retrieval augmented generation (RAG), streamlining code generation, translation and bug fixing.

[tools](#) [2b](#) [8b](#) 77.8K Pulls  33 Tags  Updated 4 months ago

internlm2

InternLM2.5 is a 7B parameter model tailored for practical scenarios with outstanding reasoning capability.

[1m](#) [1.8b](#) [7b](#) [20b](#)

↓ 77.5K Pulls 🏷 65 Tags ⌚ Updated 8 months ago

wizardlm

General use model based on Llama 2.

↓ 76.5K Pulls 🏷 73 Tags ⌚ Updated 17 months ago

nemotron-mini

A commercial-friendly small language model by NVIDIA optimized for roleplay, RAG QA, and function calling.

[tools](#) [4b](#)

↓ 75.6K Pulls 🏷 17 Tags ⌚ Updated 6 months ago

deepscaler

A fine-tuned version of Deepseek-R1-Distilled-Qwen-1.5B that surpasses the performance of OpenAI's o1-preview with just 1.5B parameters on popular math evaluations.

[1.5b](#)

↓ 73.8K Pulls 🏷 5 Tags ⌚ Updated 2 months ago

falcon

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

[7b](#) [40b](#) [180b](#)

↓ 72.3K Pulls 🏷 38 Tags ⌚ Updated 17 months ago

dolphin-phi

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

[2.7b](#)

↓ 70.4K Pulls 🏷 15 Tags ⌚ Updated 15 months ago

nemotron

Llama-3.1-Nemotron-70B-Instruct is a large language model customized by NVIDIA to improve the helpfulness of LLM generated responses to user queries.

[tools](#) [70b](#)

↓ 70K Pulls 🏷 17 Tags ⌚ Updated 6 months 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.

[7b](#) [13b](#)

↓ 65.1K Pulls 🏷 33 Tags ⌚ Updated 17 months ago

granite3.2

Granite-3.2 is a family of long-context AI models from IBM Granite fine-tuned for thinking capabilities.

[tools](#) [2b](#) [8b](#)

↓ 65.1K Pulls 🏷 9 Tags ⌚ Updated 6 weeks ago

wizardlm-uncensored

Uncensored version of Wizard LM model

[13b](#)

↓ 63.3K Pulls 🏷 18 Tags ⌚ Updated 17 months ago

llama3-groq-tool-use

A series of models from Groq that represent a significant advancement in open-source AI capabilities for tool use/function calling.

[tools](#) [8b](#) [70b](#)

↓ 60.8K Pulls 🏷 33 Tags ⌚ Updated 8 months ago

stable-beluga

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

[7b](#) [13b](#) [70b](#)

↓ 59.9K Pulls  49 Tags ⌚ Updated 17 months ago

paraphrase-multilingual

Sentence-transformers model that can be used for tasks like clustering or semantic search.

[embedding](#) [278m](#)

↓ 57.6K Pulls  3 Tags ⌚ Updated 8 months ago

snowflake-arctic-embed2

Snowflake's frontier embedding model. Arctic Embed 2.0 adds multilingual support without sacrificing English performance or scalability.

[embedding](#) [568m](#)

↓ 57.4K Pulls  3 Tags ⌚ Updated 4 months ago

smallthinker

A new small reasoning model fine-tuned from the Qwen 2.5 3B Instruct model.

[3b](#)

↓ 52.6K Pulls  5 Tags ⌚ Updated 3 months ago

deepseek-v2.5

An upgraded version of DeepSeek-V2 that integrates the general and coding abilities of both DeepSeek-V2-Chat and DeepSeek-Coder-V2-Instruct.



[236b](#)

↓ 52.2K Pulls  7 Tags ⌚ Updated 7 months ago

aya-expanse

Cohere For AI's language models trained to perform well across 23 different languages.


[tools](#) [8b](#) [32b](#)

↓ 51.9K Pulls  33 Tags  Updated 5 months ago

meditron

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

[7b](#) [70b](#)

↓ 50.7K Pulls  22 Tags  Updated 16 months ago

medllama2

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


[7b](#)

↓ 49.7K Pulls  17 Tags  Updated 17 months ago

falcon3

A family of efficient AI models under 10B parameters performant in science, math, and coding through innovative training techniques.

[1b](#) [3b](#) [7b](#) [10b](#)

↓ 48.1K Pulls  17 Tags  Updated 3 months ago

granite3-moe

The IBM Granite 1B and 3B models are the first mixture of experts (MoE) Granite models from IBM designed for low latency usage.

[tools](#) [1b](#) [3b](#)

↓ 47.2K Pulls  33 Tags  Updated 4 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.

↓ 46.7K Pulls 🏷 33 Tags ⌚ Updated 15 months ago

yarn-mistral

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

7b

↓ 45.9K Pulls 🏷 33 Tags ⌚ Updated 17 months ago

nexusraven

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

13b

↓ 42.5K Pulls 🏷 32 Tags ⌚ Updated 15 months ago

codeup

Great code generation model based on Llama2.

13b

↓ 40.5K Pulls 🏷 19 Tags ⌚ Updated 17 months ago

granite3.1-moe

The IBM Granite 1B and 3B models are long-context mixture of experts (MoE) Granite models from IBM designed for low latency usage.

tools 1b 3b

↓ 39.8K Pulls 🏷 33 Tags ⌚ Updated 2 months ago

everythinglm

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

13b

↓ 39.7K Pulls 🏷 18 Tags ⌚ Updated 15 months ago

nous-hermes2-mixtral

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

8x7b

↓ 39.6K Pulls 🏷 18 Tags ⌚ Updated 3 months ago

granite3.2-vision

A compact and efficient vision-language model, specifically designed for visual document understanding, enabling automated content extraction from tables, charts, infographics, plots, diagrams, and more.

vision tools 2b

↓ 39.3K Pulls 🏷 5 Tags ⌚ Updated 6 weeks ago

shieldgemma

ShieldGemma is set of instruction tuned models for evaluating the safety of text prompt input and text output responses against a set of defined safety policies.

2b 9b 27b

↓ 39K Pulls 🏷 49 Tags ⌚ Updated 6 months ago

exaone3.5

EXAONE 3.5 is a collection of instruction-tuned bilingual (English and Korean) generative models ranging from 2.4B to 32B parameters, developed and released by LG AI Research.

2.4b 7.8b 32b

↓ 36.4K Pulls 🏷 13 Tags ⌚ Updated 4 months ago

reader-lm

A series of models that convert HTML content to Markdown content, which is useful for content conversion tasks.

[0.5b](#) [1.5b](#)

↓ 36K Pulls 🏷 33 Tags ⌚ Updated 7 months ago

llama-guard3

Llama Guard 3 is a series of models fine-tuned for content safety classification of LLM inputs and responses.

[1b](#) [8b](#)

↓ 35.5K Pulls 🏷 33 Tags ⌚ Updated 6 months ago

mathstral

MathStral: a 7B model designed for math reasoning and scientific discovery by Mistral AI.

[7b](#)

↓ 34.9K Pulls 🏷 17 Tags ⌚ Updated 9 months ago

marco-o1

An open large reasoning model for real-world solutions by the Alibaba International Digital Commerce Group (AIDC-AI).

[7b](#)

↓ 34.9K Pulls 🏷 5 Tags ⌚ Updated 4 months ago

solar-pro

Solar Pro Preview: an advanced large language model (LLM) with 22 billion parameters designed to fit into a single GPU

[22b](#)

↓ 34.5K Pulls 🏷 18 Tags ⌚ Updated 6 months ago

falcon2

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

11b

↓ 33.5K Pulls 🏷 17 Tags ⌚ Updated 11 months ago

stablelm-zephyr

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

3b

↓ 33.2K Pulls 🏷 17 Tags ⌚ Updated 15 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.

7b

↓ 32.8K Pulls 🏷 18 Tags ⌚ Updated 16 months ago

mistrallite

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

7b

↓ 32.1K Pulls 🏷 17 Tags ⌚ Updated 17 months ago

codebooga

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

34b

↓ 32.1K Pulls 🏷 16 Tags ⌚ Updated 17 months ago

duckdb-nsql

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

7b

↓ 31.8K Pulls 🏷 17 Tags ⌚ Updated 14 months ago

granite-embedding

The IBM Granite Embedding 30M and 278M models are text-only dense biencoder embedding models, with 30M available in English only and 278M serving multilingual use cases.

embedding 30m 278m

↓ 31K Pulls 🏷 6 Tags ⌚ Updated 3 months ago

wizard-vicuna

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

13b

↓ 30.1K Pulls 🏷 17 Tags ⌚ Updated 17 months ago

command-r7b

The smallest model in Cohere's R series delivers top-tier speed, efficiency, and quality to build powerful AI applications on commodity GPUs and edge devices.

tools 7b

↓ 30K Pulls 🏷 5 Tags ⌚ Updated 2 months ago

opencoder

OpenCoder is an open and reproducible code LLM family which includes 1.5B and 8B models, supporting chat in English and Chinese languages.

1.5b 8b

↓ 29.6K Pulls 🏷 9 Tags ⌚ Updated 4 months ago

deepcoder

DeepCoder is a fully open-Source 14B coder model at O3-mini level, with a 1.5B version also available.

[1.5b](#) [14b](#)

↓ 28.3K Pulls  9 Tags  Updated 7 days ago

nuextract

A 3.8B model fine-tuned on a private high-quality synthetic dataset for information extraction, based on Phi-3.

[3.8b](#)

↓ 28K Pulls  17 Tags  Updated 8 months ago

exaone-deep

EXAONE Deep exhibits superior capabilities in various reasoning tasks including math and coding benchmarks, ranging from 2.4B to 32B parameters developed and released by LG AI Research.

[2.4b](#) [7.8b](#) [32b](#)

↓ 27.9K Pulls  13 Tags  Updated 3 weeks ago

megadolphin

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

[120b](#)

↓ 25.7K Pulls  19 Tags  Updated 15 months ago

bespoke-minicheck


A state-of-the-art fact-checking model developed by Bespoke Labs.

[7b](#)

↓ 25.7K Pulls  17 Tags  Updated 6 months ago



notux

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

[8x7b](#) 24.9K Pulls  18 Tags  Updated 15 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.

[13b](#) 24.4K Pulls  17 Tags  Updated 17 months ago

notus

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

[7b](#) 24.2K Pulls  18 Tags  Updated 15 months ago

cogito

Cogito v1 Preview is a family of hybrid reasoning models by Deep Cogito that outperform the best available open models of the same size, including counterparts from LLaMA, DeepSeek, and Qwen across most standard benchmarks.

[tools](#) [3b](#) [8b](#) [14b](#) [32b](#) [70b](#) 24.2K Pulls  20 Tags  Updated 7 days ago

tulu3

Tulu 3 is a leading instruction following model family, offering fully open-source data, code, and recipes by the The Allen Institute for AI.

[8b](#) [70b](#) 24K Pulls  9 Tags  Updated 3 months ago

r1-1776

A version of the DeepSeek-R1 model that has been post trained to provide unbiased, accurate, and factual information by Perplexity.

[70b](#) [671b](#)

↓ 23.4K Pulls 🏷 9 Tags ⌚ Updated 7 weeks ago

goliath

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

↓ 23.4K Pulls 🏷 16 Tags ⌚ Updated 17 months ago

firefunction-v2

An open weights function calling model based on Llama 3, competitive with GPT-4o function calling capabilities.

[tools](#) [70b](#)

↓ 20.2K Pulls 🏷 17 Tags ⌚ Updated 9 months ago

dbrx

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

[132b](#)

↓ 18.9K Pulls 🏷 7 Tags ⌚ Updated 12 months ago

granite3-guardian

The IBM Granite Guardian 3.0 2B and 8B models are designed to detect risks in prompts and/or responses.

[2b](#) [8b](#)

↓ 17.8K Pulls 🏷 10 Tags ⌚ Updated 4 months ago

alfred

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

[40b](#)

↓ 16.4K Pulls 🏷 7 Tags ⌚ Updated 17 months ago

sailor2

Sailor2 are multilingual language models made for South-East Asia.
Available in 1B, 8B, and 20B parameter sizes.

[1b](#) [8b](#) [20b](#)

↓ 10.8K Pulls 🏷 13 Tags ⌚ Updated 4 months ago

command-a

111 billion parameter model optimized for demanding enterprises that require fast, secure, and high-quality AI

[tools](#) [111b](#)

↓ 6,989 Pulls 🏷 5 Tags ⌚ Updated 4 weeks ago

command-r7b-arabic

A new state-of-the-art version of the lightweight Command R7B model that excels in advanced Arabic language capabilities for enterprises in the Middle East and Northern Africa.

[tools](#) [7b](#)

↓ 4,848 Pulls 🏷 5 Tags ⌚ Updated 6 weeks ago

[Blog](#) [Download](#) [Docs](#)

[GitHub](#) [Discord](#) [X \(Twitter\)](#) [Meetups](#)

© 2025 Ollama Inc.