

gemma3

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

qwq

QwQ is the reasoning model of the Qwen series.

```
tools 32b

1.3M Pulls \( \backsigma \) 8 Tags \( \backsigma \) Updated 4 weeks ago
```

deepseek-r1

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

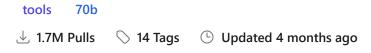
mistral-small3.1

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.

llama3.3

https://ollama.com/search 1/33

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



phi4

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



Ilama3.2

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

```
tools 1b 3b

13.2M Pulls \bigcirc 63 Tags \bigcirc 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

$\precedure$ 90M Pulls $\infty$ 93 Tags $\precedure$ Updated 4 months ago
```

nomic-embed-text

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

```
embedding

$\precedure$ 22.4M Pulls $\sigma$ 3 Tags $\text{D}$ Updated 13 months ago
```

https://ollama.com/search 2/33

mistral

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

llama3

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

```
8b 70b 	o 7.8M Pulls 	o 68 Tags 	o 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.

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 \bot 5M Pulls \bigcirc 196 Tags \bigcirc 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.

qwen

https://ollama.com/search 3/33

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

gemma

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

```
2b 7b 	begin{array}{lll} & 	ext{ } & 	ext{ }
```

qwen2

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

```
tools 0.5b 1.5b 7b 72b

4.2M Pulls \bigcirc 97 Tags \bigcirc Updated 7 months ago
```

gemma2

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

llama2

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

phi3

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

```
3.8b 14b
```

https://ollama.com/search 4/33

mxbai-embed-large

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

codellama

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

llama3.2-vision

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

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 \rightarrow 17 Tags \rightarrow Updated 8 months ago
```

tinyllama

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

https://ollama.com/search 5/33

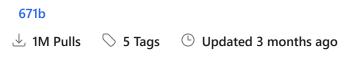
minicpm-v

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



deepseek-v3

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



mixtral

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

Ilama2-uncensored

Uncensored Llama 2 model by George Sung and Jarrad Hope.

starcoder2

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

bge-m3

https://ollama.com/search 6/33

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

deepseek-coder-v2

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

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.

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.

snowflake-arctic-embed

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

https://ollama.com/search 7/33

deepseek-coder

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

llava-llama3

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

smollm2

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

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.

https://ollama.com/search

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

$\precestrightarrow$ 538.6K Pulls $\sigma$ 70 Tags $\mathbb{C}$ 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

$\dsum 528.4K Pulls \quad \text{21 Tags} \quad \text{Updated 2 months ago}$
```

openthinker

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

phi

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

all-minilm

Embedding models on very large sentence level datasets.

wizardlm2

https://ollama.com/search 9/33

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

dolphin-mistral

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



orca-mini

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

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.

command-r

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

```
tools 35b

290.2K Pulls \bigcirc 32 Tags \bigcirc Updated 7 months ago
```

hermes3

https://ollama.com/search 10/33

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

```
tools 3b 8b 70b 405b

274.7K Pulls \bigcirc 65 Tags \bigcirc Updated 4 months ago
```

yi

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

phi3.5

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

codestral

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

zephyr

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

smollm

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

```
135m 360m 1.7b
```

https://ollama.com/search 11/33

granite-code

A family of open foundation models by IBM for Code Intelligence

wizard-vicuna-uncensored

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

starcoder

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

vicuna

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

mistral-openorca

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

https://ollama.com/search

moondream

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



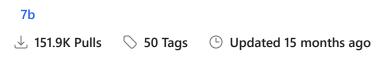
llama2-chinese

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



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.



codegeex4

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

deepseek-Ilm

An advanced language model crafted with 2 trillion bilingual tokens.

openhermes

https://ollama.com/search

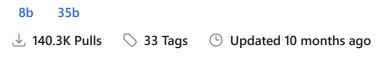
OpenHermes 2.5 is a 7B model fine-tuned by Teknium on Mistral with fully open datasets.

codeqwen

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

aya

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



deepseek-v2

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

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.

glm4

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

9b

https://ollama.com/search 14/33

ightharpoonup 128.2K Pulls ightharpoonup 32 Tags ightharpoonup 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.

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).

tinydolphin

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

nous-hermes2

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

command-r-plus

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

tools 104b

https://ollama.com/search 15/33

wizardcoder

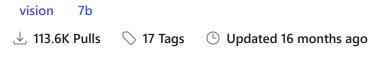
State-of-the-art code generation model

33b



bakllava

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



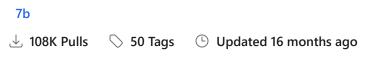
stablelm2

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.



neural-chat

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



https://ollama.com/search 16/33

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.

wizard-math

Model focused on math and logic problems

```
7b 13b 70b \ 102.9K Pulls \bigcirc 64 Tags \bigcirc 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 \bigcirc 5 Tags \bigcirc 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).

sqlcoder

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

https://ollama.com/search 17/33

llama3-gradient

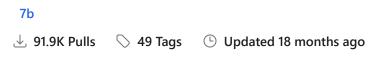
This model extends LLama-3 8B's context length from 8k to over 1m tokens.

bge-large

Embedding model from BAAI mapping texts to vectors.

samantha-mistral

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



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

$\delta$ 89.5K Pulls $\sigma$ 33 Tags $\text{\text{$\Gamma$}}$ Updated 2 months ago
```

dolphincoder

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

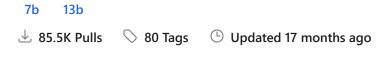
https://ollama.com/search

Ilava-phi3

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

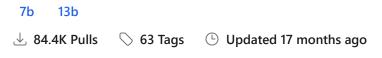
xwinlm

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



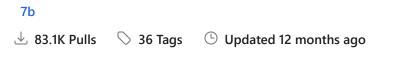
nous-hermes

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



starling-lm

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



phind-codellama

Code generation model based on Code Llama.

solar

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

https://ollama.com/search 19/33

10.7b



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.

yarn-llama2

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

athene-v2

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

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

1. 77.8K Pulls \quad 33 Tags \quad \text{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
```

https://ollama.com/search

 $begin{array}{c} & begin{array}{c} & begin{arra$

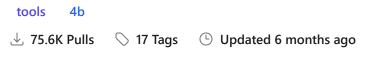
wizardlm

General use model based on Llama 2.



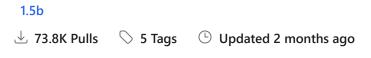
nemotron-mini

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



deepscaler

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



falcon

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

dolphin-phi

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

https://ollama.com/search 21/33

nemotron

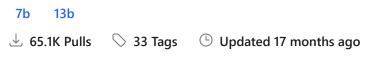
Llama-3.1-Nemotron-70B-Instruct is a large language model customized by NVIDIA to improve the helpfulness of LLM generated responses to user queries.

```
tools 70b

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.



granite3.2

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

```
tools 2b 8b

4 65.1K Pulls  $\sigma$ 9 Tags  $\mathbb{G}$ Updated 6 weeks ago
```

wizardlm-uncensored

Uncensored version of Wizard LM model

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

$\psi$ 60.8K Pulls $\sigma$ 33 Tags $\mathbb{O}$ Updated 8 months ago
```

https://ollama.com/search 22/33

stable-beluga

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

paraphrase-multilingual

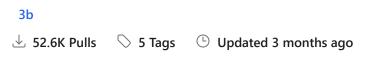
Sentence-transformers model that can be used for tasks like clustering or semantic search.

snowflake-arctic-embed2

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

smallthinker

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



deepseek-v2.5

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

https://ollama.com/search 23/33

aya-expanse

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

```
tools 8b 32b

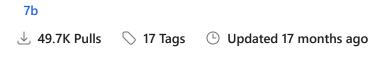
$\dslip \ 51.9K \ Pulls \quad \ 33 \ Tags \quad \ \ \ \ \ \ Updated 5 \ months \ ago
```

meditron

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

medllama2

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



falcon3

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

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 \bigcirc 33 Tags \bigcirc Updated 4 months ago
```

llama-pro

https://ollama.com/search 24/33

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

ightharpoonup 46.7K Pulls ightharpoonup 33 Tags ightharpoonup Updated 15 months ago

yarn-mistral

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

nexusraven

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

codeup

Great code generation model based on Llama2.

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 \bigcirc 33 Tags \bigcirc Updated 2 months ago

everythingIm

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

https://ollama.com/search 25/33

nous-hermes2-mixtral

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

```
8x7b

39.6K Pulls \quad 18 Tags \quad 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.

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.

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.

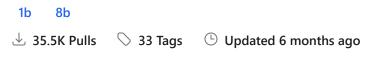
https://ollama.com/search 26/33

reader-Im

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

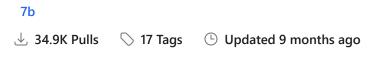
llama-guard3

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



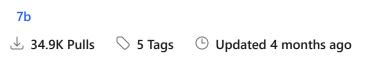
mathstral

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



marco-o1

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



solar-pro

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

falcon2

https://ollama.com/search 27/33

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



stablelm-zephyr

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



± 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

ightharpoonup 32.8K Pulls ightharpoonup 18 Tags ightharpoonup Updated 16 months ago

mistrallite

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

7b

codebooga

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

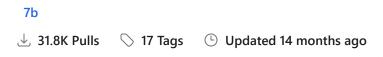
34b

ightharpoonup 32.1K Pulls ightharpoonup 16 Tags ightharpoonup Updated 17 months ago

duckdb-nsql

https://ollama.com/search 28/33

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

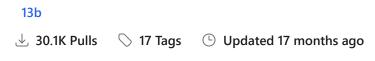


granite-embedding

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

wizard-vicuna

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



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.

opencoder

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

deepcoder

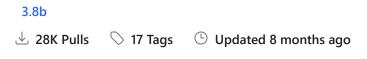
https://ollama.com/search 29/33

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



nuextract

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



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.

megadolphin

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

bespoke-minicheck

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

notux

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

https://ollama.com/search 30/33

8x7b



open-orca-platypus2

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

13b



notus

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

7b



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

$\precedure$ 24.2K Pulls $\sigma$ 20 Tags $\sigma$ Updated 7 days ago
```

tulu3

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

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
```

https://ollama.com/search 31/33

2	9	Tags

(L) Updated 7 weeks ago

goliath

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

业 23.4K Pulls



Updated 17 months ago

firefunction-v2

An open weights function calling model based on Llama 3, competitive with GPT-40 function calling capabilities.

tools 70b



(L) Updated 9 months ago

dbrx

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

132b



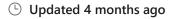


granite3-guardian

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

2b 8b





alfred

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

40b



sailor2

32/33 https://ollama.com/search

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

command-a

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

```
tools 111b

4 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.



Blog Download Docs

GitHub Discord X (Twitter) Meetups

© 2025 Ollama Inc.

https://ollama.com/search 33/33