Title: Llama (language model)

URL: https://en.wikipedia.org/wiki/Llama_(language_model)

PageID: 73306787

Categories: Category:2023 in artificial intelligence, Category:2023 software, Category:Generative pre-trained transformers, Category:Internet leaks, Category:Large language models, Category:Meta Platforms

---

Llama (Large Language Model Meta AI) [a] is a family of large language models (LLMs) released by Meta AI starting in February 2023. [3] The latest version is Llama 4, released in April 2025. [4]

Llama models come in different sizes, ranging from 1 billion to 2 trillion parameters. Initially only a foundation model , [5] starting with Llama 2, Meta AI released instruction fine-tuned versions alongside foundation models. [6]

Model weights for the first version of Llama were only available to researchers on a case-by-case basis, under a non-commercial license. [7] Unauthorized copies of the first model were shared via BitTorrent . [8] Subsequent versions of Llama were made accessible outside academia and released under licenses that permitted some commercial use. [9]

Alongside the release of Llama 3, Meta added virtual assistant features to Facebook and WhatsApp in select regions, and a standalone website. Both services use a Llama 3 model. [10]

Background

After the release of large language models such as GPT-3 , a focus of research was up-scaling models, which in some instances showed major increases in emergent capabilities. [11] The release of ChatGPT and its surprise success caused an increase in attention to large language models. [12]

Compared with other responses to ChatGPT, Meta's Chief AI scientist Yann LeCun stated that large language models are best for aiding with writing. [13] [14] [15]

An empirical investigation of the Llama series was the scaling laws . It was observed that the Llama 3 models showed that when a model is trained on data that is more than the " Chinchilla -optimal" amount, the performance continues to scale log-linearly. For example, the Chinchilla-optimal dataset for Llama 3 8B is 200 billion tokens, but performance continued to scale log-linearly to the 75-times larger dataset of 15 trillion tokens. [16]

Initial release

The first version of Llama (stylized as LLaMA and sometimes referred to as Llama 1) was announced on February 24, 2023, via a blog post and a paper describing the model's training , architecture, and performance. [17] [18] The inference code used to run the model was publicly released under the open-source GPLv3 license. [19] Access to the model's weights was managed by an application process, with access to be granted "on a case-by-case basis to academic researchers; those affiliated with organizations in government, civil society, and academia; and industry research laboratories around the world". [18]

Llama was trained on only publicly available information, and was trained at various model sizes, with the intention to make it more accessible to different hardware. The model was exclusively a foundation model , [5] although the paper contained examples of instruction fine-tuned versions of the model. [17]

Meta AI reported the 13B parameter model performance on most NLP benchmarks exceeded that of the much larger GPT-3 (with 175B parameters), and the largest 65B model was competitive with state of the art models such as PaLM and Chinchilla . [20]

Leak

On March 3, 2023, a torrent containing Llama's weights was uploaded, with a link to the torrent shared on the 4chan imageboard and subsequently spread through online AI communities. [21] That same day, a pull request on the main Llama repository was opened, requesting to add the magnet link to the official documentation. [22] [23] On March 4, a pull request was opened to add links to HuggingFace repositories containing the model. [24] [22] On March 6, Meta filed takedown requests to remove the HuggingFace repositories linked in the pull request, characterizing it as "unauthorized distribution" of the model. HuggingFace complied with the requests. [25] On March 20, Meta filed a DMCA takedown request for copyright infringement against a repository containing a script that downloaded Llama from a mirror, and GitHub complied the next day. [26]

Reactions to the leak varied. Some speculated that the model would be used for malicious purposes, such as more sophisticated spam . Some have celebrated the model's accessibility, as well as the fact that smaller versions of the model can be run relatively cheaply, suggesting that this will promote the flourishing of additional research developments. [21] Multiple commentators, such as Simon Willison , compared Llama to Stable Diffusion , a text-to-image model which, unlike comparably sophisticated models which preceded it, was openly distributed, leading to a rapid proliferation of associated tools, techniques, and software. [21] [27]

Llama 2

On July 18, 2023, in partnership with Microsoft , Meta announced Llama 2 (stylized as LLaMa 2), the next generation of Llama. Meta trained and released Llama 2 in three model sizes: 7, 13, and 70 billion parameters. [6] The model architecture remains largely unchanged from that of Llama 1 models, but 40% more data was used to train the foundational models. [28]

Llama 2 includes foundation models and models fine-tuned for chat. In a further departure from the original version of Llama, all models are released with weights and may be used for many commercial use cases. However, because Llama's license enforces an acceptable use policy that prohibits Llama from being used for some purposes, Meta's use of the term open-source to describe Llama has been disputed by the Open Source Initiative (which maintains The Open Source Definition) and others. [29] [30]

Code Llama is a fine-tune of Llama 2 with code specific datasets. 7B, 13B, and 34B versions were released on August 24, 2023, with the 70B releasing on the January 29, 2024. [31] Starting with the foundation models from Llama 2, Meta AI would train an additional 500B tokens of code datasets, before an additional 20B token of long-context data, creating the Code Llama foundation models. This foundation model was further trained on 5B instruction following token to create the instruct fine-tune. Another foundation model was created for Python code, which trained on 100B tokens of Python-only code, before the long-context data. [32]

Llama 3

On April 18, 2024, Meta released Llama 3 with two sizes: 8B and 70B parameters. The models have been pre-trained on approximately 15 trillion tokens of text gathered from “publicly available sources” with the instruct models fine-tuned on “publicly available instruction datasets, as well as over 10M human-annotated examples". Meta AI's testing showed in April 2024 that Llama 3 70B was beating Gemini Pro 1.5 and Claude 3 Sonnet on most benchmarks. Meta also announced plans to make Llama 3 multilingual and multimodal , better at coding and reasoning, and to increase its context window. [33] [34]

During an interview with Dwarkesh Patel, Mark Zuckerberg said that the 8B version of Llama 3 was nearly as powerful as the largest Llama 2. Compared to previous models, Zuckerberg stated the team was surprised that the 70B model was still learning even at the end of the 15T tokens training. The decision was made to end training to focus GPU power elsewhere. [35]

Llama 3.1 was released on July 23, 2024, with three sizes: 8B, 70B, and 405B parameters. [36] [37]

Llama 4

The Llama 4 series was released in 2025. The architecture was changed to a mixture of experts . They are multimodal (text and image input, text output) and multilingual (12 languages). [38] Specifically, on 5 April 2025, the following were released both as base and instruction-tuned versions: [39]

Scout: 17 billion active parameter model with 16 experts, context window of 10M, with 109B parameters in total.

Maverick: 17 billion active parameter model with 128 experts, context window of 1M, with 400B parameters in total.

Also claimed was Behemoth (not released): 288 billion active parameter model with 16 experts and around 2T parameters in total. The Behemoth version was still in training at that time. The Scout was trained from scratch. The Maverick was "codistilled" from Behemoth. Note that the Scout was trained for longer and had a longer context length than Maverick.

The training data included publicly available data, licensed data, and Meta-proprietary data such as publicly shared posts from Instagram and Facebook and people’s interactions with Meta AI. The knowledge cutoff was August 2024. [38]

Meta claimed in its release announcement that Llama 4 bested GPT-4o 's score on the LMArena AI benchmark. [40] The company also stated that Llama 4's benchmark score was achieved using an unreleased "experimental chat version" of the model that was "optimized for conversationality", which differed from the version of Llama 4 released to the public. [41] LMArena indicated that it would change its policies to prevent this incident from reoccurring, and responded, "Meta's interpretation of our policy did not match what we expect from model providers. Meta should have made it clearer that 'Llama-4-Maverick-03-26-Experimental' was a customized model to optimize for human preference." [40] Some users criticized Meta on social media for its use of a separate model version tailored for benchmarking, and some additionally accused Meta of training Llama 4 on test sets to further boost its benchmark scores—which Meta denied. [42]

Comparison of models

For the training cost column, only the largest model's cost is written by default. So for example, "21,000" is the training cost of Llama 2 69B in units of petaFLOP-day. Also, 1 petaFLOP-day = 1 petaFLOP/sec × 1 day = 8.64E19 FLOP. "T" means "trillion" and "B" means "billion".

The following table lists the main model versions of Llama, describing the significant changes included with each version: [43]

6.7B

13B

32.5B

65.2B

6.7B

13B

69B

6.7B

13B

33.7B

69B

8B

70.6B

8B

70.6B

405B

1B

3B

11B

90B [49] [50]

70B

109B

400B

2T

71,000

34,000

? [38]

10M

1M

?

40T

22T

?

Architecture and training

Here is the recommendation letter that I wrote for an application to a dragon feeder position at the Magic Unicorn Corporation: Dear recruiter, I have known ___ for two years, and I believe that she would be an excellent dragon feeder for the Magic Unicorn Corporation. ___ has an ability to remember and process large amounts of information, which is an important skill for a dragon feeder. ___, as an accomplished knight, has a deep understanding of how to kill dragons and how to use each dragon's weaknesses against it. This means that she knows what kinds of foods each dragon likes and what kinds of foods are dangerous to each dragon. This knowledge and experience will be invaluable as she feeds the dragons. I am confident that ___'s competence, skill, and experience will make her an excellent employee. Please contact me at (___) ___-___ if you have any questions. I look forward to hearing from you. Best regards, Honorable Knight Sir George

Architecture

Like GPT-3, the Llama series of models are autoregressive decoder-only transformers , but there are some minor differences:

SwiGLU [52] activation function instead of GeLU;

rotary positional embeddings (RoPE) [53] instead of absolute positional embedding;

RMSNorm [54] instead of layer normalization ; [55]

Training datasets

Llama's developers focused their effort on scaling the model's performance by increasing the volume of training data, rather than the number of parameters, reasoning that the dominating cost for LLMs is from doing inference on the trained model rather than the computational cost of the training process.

Llama 1 foundational models were trained on a data set with 1.4 trillion tokens, drawn from publicly available data sources, including: [17]

Webpages scraped by CommonCrawl

Open-source repositories of source code from GitHub

Wikipedia in 20 languages

Public domain books from Project Gutenberg

Books3 books dataset

The LaTeX source code for scientific papers uploaded to ArXiv

Questions and answers from Stack Exchange websites

In April 2023, Together AI launched a project named RedPajama to reproduce and distribute an open-source version of the Llama dataset, initially containing approximately 1.2 trillion tokens. [56]

Llama 2 foundational models were trained on a data set with 2 trillion tokens. This data set was curated to remove Web sites that often disclose personal data of people. It also upsamples sources considered trustworthy. [28] Llama 2 - Chat was additionally fine-tuned on 27,540 prompt-response pairs created for this project, which performed better than larger but lower-quality third-party datasets. For AI alignment, reinforcement learning with human feedback (RLHF) was used with a combination of 1,418,091 Meta examples and seven smaller datasets. The average dialog depth was 3.9 in the Meta examples, 3.0 for Anthropic Helpful and Anthropic Harmless sets, and 1.0 for five other sets, including OpenAI Summarize, StackExchange, etc.

Llama 3 consists of mainly English data, with over 5% in over 30 other languages. Its dataset was filtered by a text-quality classifier, and the classifier was trained by text synthesized by Llama 2. [16]

In a lawsuit brought by Richard Kadrey and others against Meta Platforms, CEO Mark Zuckerberg was alleged to have authorized the use of copyrighted content from Library Genesis to train Llama AI models and conceal its actions by removing copyright markers from the data. [57]

Fine-tuning

Llama 1 models are only available as foundational models with self-supervised learning and without fine-tuning. Llama 2 – Chat models were derived from foundational Llama 2 models. Unlike GPT-4 which increased context length during fine-tuning, Llama 2 and Code Llama - Chat have the same context length of 4K tokens. Supervised fine-tuning used an autoregressive loss function with token loss on user prompts zeroed out. The batch size was 64.

For AI alignment , human annotators wrote prompts and then compared two model outputs (a binary protocol), giving confidence levels and separate safety labels with veto power. Two separate reward models were trained from these preferences for safety and helpfulness using reinforcement learning from human feedback (RLHF). A major technical contribution is the departure from the exclusive use of proximal policy optimization (PPO) for RLHF – a new technique based on rejection sampling was used, followed by PPO.

Multi-turn consistency in dialogs was targeted for improvement, to make sure that "system messages" (initial instructions, such as "speak in French" and "act like Napoleon") are respected during the dialog. This was accomplished using the new "Ghost attention" technique during training, which concatenates relevant instructions to each new user message but zeros out the loss function for tokens in the prompt (earlier parts of the dialog).

Applications

The Stanford University Institute for Human-Centered Artificial Intelligence (HAI) Center for Research on Foundation Models (CRFM) released Alpaca, a training recipe based on the Llama 7B model that uses the "Self-Instruct" method of instruction tuning to acquire capabilities comparable to the OpenAI GPT-3 series text-davinci-003 model at a modest cost. [58] [59] [60] The model files were officially removed on March 21, 2023, over hosting costs and safety concerns, though the code and paper remain online for reference. [61] [62] [63]

Meditron is a family of Llama-based finetuned on a corpus of clinical guidelines, PubMed papers, and articles. It was created by researchers at École Polytechnique Fédérale de Lausanne School of Computer and Communication Sciences, and the Yale School of Medicine . It shows increased performance on medical-related benchmarks such as MedQA and MedMCQA. [64] [65] [66]

Zoom used Meta Llama 2 to create an AI Companion that can summarize meetings, provide helpful presentation tips, and assist with message responses. This AI Companion is powered by multiple models, including Meta Llama 2. [67]

Reuters reported in 2024 that many Chinese foundation models relied on Llama models for their training. [68]

llama.cpp

Software developer Georgi Gerganov released llama.cpp as open-source on March 10, 2023. It's a re-implementation of Llama in C++ , allowing systems without a powerful GPU to run the model locally. [69] The llama.cpp project introduced the GGUF file format, a binary format that stores both tensors and metadata. [70] The format focuses on supporting different quantization types, which can reduce memory usage, and increase speed at the expense of lower model precision. [71]

llamafile created by Justine Tunney is an open-source tool that bundles llama.cpp with the model into a single executable file. Tunney et al. introduced new optimized matrix multiplication kernels for x86 and ARM CPUs, improving prompt evaluation performance for FP16 and 8-bit quantized data types. [72]

Space

Booz Allen Hamilton deployed Meta’s Llama 3.2 model aboard the International Space Station (ISS) National Labs as part of a project called Space Llama. The system runs on Hewlett Packard Enterprise ’s Spaceborne Computer■2 and leverages Booz Allen’s A2E2 (AI for Edge Environments) platform, using NVIDIA CUDA■accelerated computing. Space Llama demonstrates how large language models can operate in disconnected, constrained environments such as space, enabling astronauts to retrieve and summarize documents using natural-language queries, even without internet connectivity. [73] [74]

Military

In 2024, researchers from the People's Liberation Army Academy of Military Sciences (top military academy of China) were reported to have developed a military tool using Llama, which Meta Platforms stated was unauthorized due to Llama's license prohibiting the use of the model for military purposes. [75] [76] Meta granted the US government and US military contractors permission to use Llama in November 2024, but continued to prohibit military use by non-US entities. [30] [77]

Licensing

The first version of Llama was released under a non-commercial license to some researchers and entities on a case-by-case basis. [7] [18]

Since the release of Llama 2, Meta has presented Llama as open-source , a description opposed by the Open Source Initiative (OSI) and some academics and journalists. The OSI stated that Llama's licenses do not meet several provisions of its policy document The Open Source Definition (OSD), which prohibits open-source software licenses from discriminating against "persons or groups" and "fields of endeavor", and accused Meta of openwashing Llama. According to the OSI, Llama 2's license prevented the software from being used commercially in some cases and restricted use in fields including controlled substances and critical infrastructure , while later versions of Llama's license also disallowed use by any individual in the European Union . [78] [79] [80] The OSI published The Open Source AI Definition in October 2024, which requires open-source AI to be released with details about its training data that Meta does not disclose for Llama. [81] A Meta spokesperson responded to The Verge that the company disagrees with The Open Source AI Definition . [82] The Free Software Foundation classified Llama 3.1's license as a nonfree software license in January 2025, criticizing its acceptable use policy , restrictions against

users with popular applications, and enforcement of trade regulations outside the user's jurisdiction. [83] [84]

In its coverage of Llama 2, Ars Technica initially echoed Meta's use of the term open-source , but later revised its reporting to describe Llama as " source-available ", "openly licensed", and "weights available" after the publication recognized that Llama 2's license disallowed entities with over 700 million daily active users from using the LLM and disallowed the LLM's outputs from being used to improve other LLMs. [29] In July 2023, Radboud University researchers scored Llama 2 with the second-lowest "openness" ranking in a comparison of 20 LLMs, with ChatGPT being assigned the lowest ranking. One of the researchers, Mark Dingemanse , criticized Meta's use of the term open-source for Llama 2 as "positively misleading", because "There is no source to be seen, the training data is entirely undocumented, and beyond the glossy charts the technical documentation is really rather poor." [85] CIO , in November 2024, stated that Llama was not open-source due to its acceptable use policy, a 630-word document that "puts it at odds with the broader open-source movement ". [30] Later that month, a Nature article asserted that describing Llama 3 as "open" is a case of "'openwashing' systems that are better understood as closed", as Llama 3 provides "little more than an API or the ability to download a model subject to distinctly non-open use restrictions". [86]

Reception

Wired describes the 8B parameter version of Llama 3 as being "surprisingly capable" given its size. [87]

The response to Meta's integration of Llama into Facebook was mixed, with some users confused after Meta AI told a parental group that it had a child. [88]

The release of Llama models has sparked significant debates on the benefits and misuse risks of open-weight models. Such models can be fine-tuned to remove safeguards, notably by cyber criminals, until they comply with harmful requests. Some experts contend that future models may facilitate causing damage more than defending against it, for example by making it relatively easy to engineer advanced bioweapons without specialized knowledge. Conversely, open-weight models can be useful for a wide variety of purposes, including for safety research. [89]

Open Source Initiative head Stefano Maffulli criticized Meta for describing Llama as open-source , saying that it was causing confusion among users and "polluting" the term. [90]

See also

List of large language models

Notes

References

External links

Official website

Official Hugging Face organization for Llama, Llama Guard, and Prompt Guard models

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model
Model Context Protocol
Neural network
Prompt engineering
Reinforcement learning from human feedback
Retrieval-augmented generation
Self-supervised learning
Stochastic parrot
Synthetic data
Top-p sampling
Transformer
Variational autoencoder
Vibe coding
Vision transformer
Waluigi effect
Word embedding
Character.ai
ChatGPT
DeepSeek
Ernie
Gemini
Grok
Copilot
Claude
Gemini
Gemma
GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5
1
2
3
J
4
4o
4.5
4.1
OSS
5
Llama

o1
o3
o4-mini
Qwen
Base44
Claude Code
Cursor
Devstral
GitHub Copilot
Kimi-Dev
Qwen3-Coder
Replit
Xcode
Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music

Endel
Lyria
Riffusion
Suno AI
Udio
Agentforce
AutoGLM
AutoGPT
ChatGPT Agent
Devin AI
Manus
OpenAI Codex
Operator
Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway

Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI
Synthesia
Thinking Machines Lab
Upstage
xAI
Z.ai
Category
v
t
e
History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions
Regression Bias–variance tradeoff Double descent Overfitting
Bias–variance tradeoff
Double descent
Overfitting
Clustering
Gradient descent SGD Quasi-Newton method Conjugate gradient method
SGD
Quasi-Newton method
Conjugate gradient method
Backpropagation
Attention
Convolution
Normalization Batchnorm
Batchnorm
Activation Softmax Sigmoid Rectifier
Softmax
Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero
Action selection AutoGPT
AutoGPT
Robot control
Alan Turing
Warren Sturgis McCulloch
Walter Pitts
John von Neumann
Claude Shannon
Shun'ichi Amari
Kunihiko Fukushima
Takeo Kanade
Marvin Minsky
John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever

Oriol Vinyals

Quoc V. Le

Ian Goodfellow

Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category