Llama (language model)

Llama (acronym for Large Language Model Meta AI, and formerly stylized as LLaMA) is a family of autoregressive large language models (LLMs) released by Meta AI starting in February 2023.[2][3] The latest version is Llama 3, released in April 2024.[4]
Model weights for the first version of Llama were made available to the research community under a non-commercial license, and access was granted on a case-by-case basis.[5][3] Unauthorized copies of the model were shared via BitTorrent. In response, Meta AI issued DMCA takedown requests against repositories sharing the link on GitHub.[6][7] Subsequent versions of Llama were made accessible outside academia and released under licenses that permitted some commercial use.[8][9] Llama models are trained at different parameter sizes, typically ranging between 7B and 70B.[4] Originally, Llama was only available as a foundation model.[10] Starting with Llama 2, Meta AI started releasing instruction fine-tuned versions alongside foundation models.[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.[11]

Developer(s)	Meta AI
Initial release	February 24, 2023; 16 months ago
Stable release	Llama 3 / April 18, 2024; 2 months ago
Repository	github.com/meta-llama/llama 3
Written in	Python
Type	Large language model GPT Foundation model
License	Meta Llama 3 Community License[1]
Website	llama.meta.com

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.[12] The release of ChatGPT and its surprise success caused an increase in attention to large language models.[13]
Compared with other responses to ChatGPT, Meta's Chief AI scientist Yann LeCun stated that large language models are best for aiding with writing.[14][15][16]

Initial release

LLaMA was announced on February 24, 2023, via a blog post and a paper describing the model's training, architecture, and performance.[2][3] The inference code used to run the model was publicly released under the open-source GPLv3 license.[17] 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".[3]
Llama was trained on only publicly available information, and was trained at various different model sizes, with the intention to make it more accessible to different hardware.
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.[2]

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.[6]
That same day, a pull request on the main LLaMA repository was opened, requesting to add the magnet link to the official documentation.[18][19] On March 4, a pull request was opened to add links to HuggingFace repositories containing the model.[20][18] 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.[21] 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.[7]
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.[6] 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.[6][22]

Llama 2

On July 18, 2023, in partnership with Microsoft, Meta announced Llama 2, the next generation of Llama. Meta trained and released Llama 2 in three model sizes: 7, 13, and 70 billion parameters.[9] The model architecture remains largely unchanged from that of LLaMA-1 models, but 40% more data was used to train the foundational models.[23] The accompanying preprint[23] also mentions a model with 34B parameters that might be released in the future upon satisfying safety targets.

Llama 2 includes foundation models and models fine-tuned for chat. In a further departure from LLaMA, all models are released with weights and are free for many commercial use cases. However, due to some remaining restrictions, Meta's description of LLaMA as open source has been disputed by the Open Source Initiative (known for maintaining the Open Source Definition).[24]

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.[25] 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.[26]

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 plans on releasing multimodal models, models capable of conversing in multiple languages, and models with larger context windows. A version with 400B+ parameters is currently being trained.[4]

Meta AI's testing shows that Llama 3 70B beats Gemini, and Claude in most benchmarks.[27][28] During an interview with Dwarkesh Patel, Mark Zuckerberg said the 8B version of Llama 3 was nearly as powerful as the largest Llama 2. That Llama 3 had increased priority in coding abilities based on what was learned from CodeLlama. Compared to previous models, Zuckerberg stated the team was surprised that the 70B model was still learning even at the end on training in the 15T tokens. The decision was made to end training to focus GPU power elsewhere, Zuckerberg stated more research needs to be done on AI data scaling.

When asked if Meta would continue to open-source models, Zuckerberg stated only if it aligned with Meta's strategy. Zuckerberg floated the possibility of smaller versions of the Llama models for application-specific usecases[29]

Comparison of models

Name	Release date	Parameters	Training cost (petaFLOP-day)	Context length	Corpus size	Commercial viability?
LLaMA	February 24, 2023	• 6.7B • 13B • 32.5B • 65.2B	6,300[30]	2048	1–1.4T	No
Llama 2	July 18, 2023	• 6.7B • 13B • 69B	21,000[31]	4096	2T	Yes
Code Llama	August 24, 2023	• 6.7B • 13B • 33.7B • 69B				
Llama 3	April 18, 2024	• 8B • 70.6B • 400B+ (unreleased)	100,000[32][33]	8192	15T	

Architecture and training

Architecture

LLaMA uses the transformer architecture, the standard architecture for language modeling since 2018.
There are minor architectural differences. Compared to GPT-3, LLaMA

- uses SwiGLU[34] activation function instead of GeLU;
- uses rotary positional embeddings[35] instead of absolute positional embedding;
- uses root-mean-squared layer-normalization[36] instead of standard layer-normalization.[37]
- increases context length to 8k in Llama 3 (compared to 4k in Llama 2 and 2k in Llama 1 and GPT-3)

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:[2]

- Webpages scraped by CommonCrawl
- Open source repositories of source code from GitHub
- Wikipedia in 20 different 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

On April 17, 2023, TogetherAI launched a project named RedPajama to reproduce and distribute an open source version of the LLaMA dataset.[38] The dataset has approximately 1.2 trillion tokens and is publicly available for download.[39]

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.[23] 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.[4]

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.[40][41][42] The model files were officially removed on March 21st 2023 over hosting costs and safety concerns, though the code and paper remain online for reference.[43][44][45]

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.[46][47][48]

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.[49]

llama.cpp

Main article: 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.[50] The llama.cpp project introduced the GGUF file format, a binary format that stores both tensors and metadata.[51] The format focuses on supporting different quantization types, which can reduce memory usage, and increase speed at the expense of lower model precision.[52]

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.[53]

Reception

Wired describes the 8B parameter version of Llama 3 as being "surprisingly capable" given its size.[54]

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.[55]

According to the Q4 2023 Earnings transcript, Meta adopted the strategy of open weights to improve on model safety, iteration speed, increase adoption among developers and researchers, and to become the industry standard. Llama 5, 6, and 7 are planned for the future.[56]