Llama (Large Language Model Meta AI) is a family of autoregressive large language models (LLMs), released by Meta AI starting in February 2023.

Four model sizes were trained for the first version of LLaMA: 7, 13, 33, and 65 billion parameters. LLaMA's developers reported that the 13B parameter model's performance on most NLP benchmarks exceeded that of the much larger GPT-3 (with 175B parameters) and that the largest model was competitive with state of the art models such as PaLM and Chinchilla. In contrast, the most powerful LLMs have generally been accessible only through limited APIs (if at all), Meta released LLaMA's model weights to the research community under a noncommercial license. Within a week of LLaMA's release, its weights were leaked to the public on 4chan via BitTorrent.

In July 2023, Meta released several models such as Llama 2, using 7, 13, and 70 billion parameters.

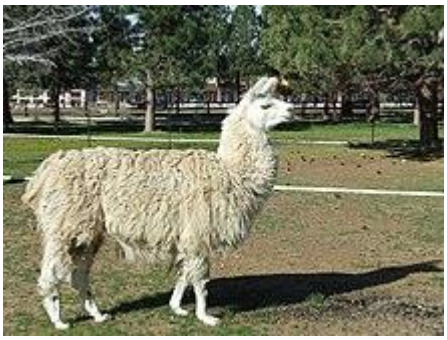
Models

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. The model architecture remains largely unchanged from that of LLaMA-1 models, but 40% more data was used to train the foundational models. The accompanying preprint also mentions a model with 34B parameters that might be released in the future upon satisfying safety targets.

Llama-2 includes foundational models and models fine-tuned for dialog, called Llama-2 Chat. In a further departure from LLaMA-1, 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).

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.