**Model Details:**

Meta introduced the [Llama 2](https://llama.meta.com/) series of large language models with sizes varying from 7 billion to 70 billion parameters. Engineers designed these LLMs specifically for dialogue applications. In terms of performance, Llama-2-Chat excels over open-source chat models in many of the benchmarks Meta assessed. When gauged for helpfulness and safety in human reviews, they match well-known proprietary models such as ChatGPT and PaLM. Llama 2 models are available in three parameter sizes: 7B, 13B, and 70B, and come in both pretrained and fine-tuned forms. These models solely accept text as input and produce text as output. The underlying framework for Llama 2 is an auto-regressive language model. Enhanced versions undergo supervised fine-tuning (SFT) and harness reinforcement learning combined with human insights (RLHF) to better resonate with human desires for safety and utility.

A table with numbers and symbols

Description automatically generated

The Llama 2 series of models utilize token counts exclusively from their pretraining data. All models in this series have been trained using a global batch-size encompassing 4M tokens. The 70B iteration employs Grouped-Query Attention (GQA) to enhance inference scalability. [Llama Research Paper](https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/)

**Indented Use:**

Llama 2 is intended for commercial and research use in English. Tuned models are intended for assistant-like chat, whereas pretrained models can be adapted for a variety of natural language generation tasks.

|  |  |
| --- | --- |
| **Use Cases:**  Text Summarization  Q&A Chatbot (w RAG)  Code Generation  Entity Extraction | **Usage:**  Amazon Bedrock  Sagemaker Jumpstart  Local Laptop/Desktop  Hosted Internally  Integrated with Application |

**Speed / Latency**

Refer the **Holistic Evaluation of Language Models** project ([paper](https://arxiv.org/abs/2211.09110), [website](https://crfm.stanford.edu/helm/latest/)) by [Stanford CRFM](https://crfm.stanford.edu/). This package includes the following features:

* Collection of datasets in a standard format (e.g., NaturalQuestions)
* Collection of metrics beyond accuracy (efficiency, bias, toxicity, etc.)
* Collection of perturbations for evaluating robustness and fairness (e.g., typos, dialect)

**FNMA Usage**:

Innovation Env. ONLY

**Stage**:

TSP (IN PROGRESS)

**Training Region:** USA

**Training Data:**

Llama 2 was pretrained on 2 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over one million new human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.

**Data Recency/Freshness**:

The pretraining data has a cutoff of September 2022, but some tuning data is more recent, up to July 2023.

**Limitations:**

* Activities that breach legal and regulatory standards, including international trade laws, are discouraged.
* Applications in languages other than English.

**Open Source**:

Research / Commercial

**Mode Size**:

7B, 13B, 70B

**Modality**:

Text-to-Text

**Developers**:

Meta

**License:** [Meta Llama License Link](https://github.com/meta-llama/llama/blob/main/LICENSE)

**EULA:** [EULA Link](https://github.com/meta-llama/llama/blob/main/USE_POLICY.md)

**Documentation:** [README Link](https://github.com/meta-llama/llama/blob/main/README.md)

**FAQ**: [Llama2 FAQ Link](https://ai.meta.com/llama/faq/)

**Cost: $0**

**Architecture: Kalyana C Bedhu**

**Data Science: Sandy Farr**

**Security (Info. Sec)**

This section addresses security concerns.

**Data Protection / Privacy:**

**Data Leakage:**

**Known Vulnerabilities:**

**Jailbreaking Guardrails:**

**Evaluation & Benchmarks**

Meta reports the results for the Llama 2 models on standard academic benchmarks. For all the evaluations, Meta uses their internal evaluations library. For TruthfulQA, they present the percentage of generations that are both truthful and informative (the higher the better). For ToxiGen, they present the percentage of toxic generations (the smaller the better).

A table with numbers and text

Description automatically generated

A table with text on it

Description automatically generated

**Ethical Considerations (Legal)**

Llama 2, while groundbreaking, is not exempt from the inherent risks associated with new technologies. There’s a possibility that the model may sometimes deliver responses that are erroneous, display bias, or are generally objectionable. As a precaution, it’s imperative for developers to conduct comprehensive safety checks and calibration to align Llama 2’s behavior with their specific application requirements before integration.

**Compliance:**

**Toxicity:**

**Bias:**

**Hallucination:**