**LLAMA-2 Model Documentation**

[1. Introduction: 3](#_Toc154674586)

[2. LLAMA-2 Model Structure 3](#_Toc154674587)

[3. LLAMA-2 Capabilities: 4](#_Toc154674588)

[3.1. Conversational Dialogue: 4](#_Toc154674589)

[3.2. Emotional Understanding: 5](#_Toc154674590)

[3.3. Domain Knowledge: 5](#_Toc154674591)

[4. LLAMA-2 Applications 5](#_Toc154674592)

[4.1. Chatbots: 5](#_Toc154674593)

[4.2. Language Translation: 5](#_Toc154674594)

[4.3. Content Generation: 5](#_Toc154674595)

[4.4. Sentiment Analysis: 5](#_Toc154674596)

[4.5. Medical and Healthcare: 5](#_Toc154674597)

[5. Training process of LLAMA-2: 6](#_Toc154674598)

[5.1. Pretraining Data: 6](#_Toc154674599)

[5.2. Training Details: 6](#_Toc154674600)

[6. LLAMA-2 Model Variants: 7](#_Toc154674601)

[6.1. LLama-2: 7](#_Toc154674602)

[6.2. Llama 2-Chat: 7](#_Toc154674603)

[7. LLAMA-2 Pretrained Model Evaluation: 8](#_Toc154674604)

[8. Conclusion 8](#_Toc154674605)

[9. Difference in Llama-1 and Llama-2 🦙 9](#_Toc154674606)

[9.1. Capabilities: 9](#_Toc154674607)

[9.2. Training Data: 9](#_Toc154674608)

[9.3. Response Style: 9](#_Toc154674609)

[9.4. Use Cases: 9](#_Toc154674610)

[9.5. Domain Knowledge: 9](#_Toc154674611)

[9.6. Emotional Understanding: 9](#_Toc154674612)

[10. Difference in Llama-2 and Llama2-chat 🦙 10](#_Toc154674613)

[10.1. Capabilities: 10](#_Toc154674614)

[10.2 Training Data: 10](#_Toc154674615)

[10.3. Response Style: 10](#_Toc154674616)

[11. Is LLAMA-2 training data does include healthcare-related texts: 11](#_Toc154674617)

# 1. Introduction:

Large Language Models (LLMs) have shown great potential as AI assistants that can excel in complex reasoning tasks requiring expert knowledge across a wide range of fields, including specialized domains such as programming and creative writing. They can interact with humans through intuitive chat interfaces, which has led to rapid and widespread adoption among the general public.

In simple the document talks about Large Language Models (LLMs), powerful AI systems that excel in various tasks. While their training seems straightforward, it involves complex processes and high computational requirements. Existing pretrained LLMs are compared, with closed "product" models like ChatGPT being highly fine-tuned for human preferences, making them more-usable and safe. The document introduces Llama 2, a new family of LLMs, sharing insights into their development, safety measures, and fine-tuning methodology.

Llama 2 is a recent advancement in artificial intelligence technology that has gained significant attention in the field of natural language processing. This paper aims to provide an overview of Llama 2, its capabilities, and its potential applications in various industries. Llama 2- performs well on benchmarks and is designed with safety considerations. The openness in sharing methodologies aims to contribute to responsible LLM development and safety improvement in the AI community. Unique observations during Llama 2's development.

# 2. LLAMA-2 Model Structure

LLama2 uses a modified version of the transformer architecture called the " Multi-Head Attention" mechanism. This mechanism allows the model to jointly attend to information from different representation subspaces at different positions. In other words, the model can attend to different aspects of the input text simultaneously, which improves its ability to capture complex patterns and relationships. LLama2 uses **12 layers in its encoder-decoder** architecture.

LLama2's structure includes input embeddings, encoder, decoder, positional encoding, multi-head attention, layer normalization, and output linear layer.

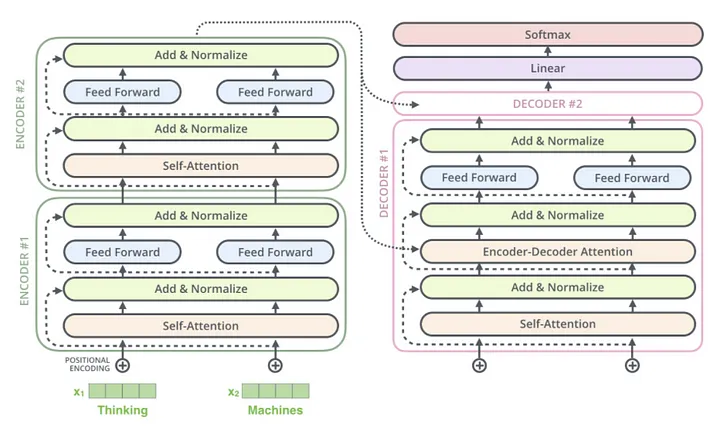


Figure 1. Encoder-Decoder architecture

* **Input embeddings:** numerical representation of input text.
* **Encoder:** series of identical layers with self-attention and feed-forward neural network (FFNN).
* **Decoder:** series of identical layers with self-attention and FFNN, generates output text.
* **Positional encoding:** adds unique fixed vector to each input sequence, preserving order.
* **Multi-head attention:** attends to different representation subspaces at different positions.
* **Layer normalization:** normalizes activations to zero mean and unit variance.
* **Output linear layer:** transforms decoder output to final output space.

# 3. LLAMA-2 Capabilities:

Llama 2 has several capabilities that make it an advanced tool for natural language processing. Some of its key capabilities include:

## 3.1. Conversational Dialogue:

Llama 2 is optimized for conversational dialogue, which means it can handle multi-turn conversations more effectively than other language models. It can understand and respond to user input in a more human-like way, making it suitable for applications such as chatbots, virtual assistants, and language translation.

## 3.2. Emotional Understanding:

Llama 2 has been trained to understand and respond to emotions, which means it can generate responses that are appropriate to the user's emotional state. This capability makes it suitable for applications such as customer service, tech support, and counseling.

## 3.3. Domain Knowledge:

Llama 2 has been trained on a larger and more diverse dataset than other language models, which means it has a broader knowledge base and can respond to a wider range of topics and questions.

# 4. LLAMA-2 Applications

Llama 2 has various potential applications across different industries, Llama 2 has various applications in natural language processing tasks, and its ability to generate coherent and contextually appropriate responses makes it a powerful tool for a wide range of applications.

## 4.1. Chatbots:

Llama 2 can be used to build chatbots that can understand and respond to user queries in a more human-like way. This can be useful for customer service, tech support, and other applications where it's important to provide quick and accurate responses to user inquiries.

## 4.2. Language Translation:

Llama 2 can be used to improve language translation systems, allowing them to better understand the nuances of language and provide more accurate translations.

## 4.3. Content Generation:

Llama 2 can be used to generate content, such as articles, blog posts, or social media posts, that is more engaging and natural-sounding than content generated by other AI models.

## 4.4. Sentiment Analysis:

Llama 2 can be used to analyze text to determine the sentiment behind it, such as determining whether a piece of text expresses a positive, negative, or neutral sentiment.

## 4.5. Medical and Healthcare:

Llama 2 can be used in the medical and healthcare industry to generate informative and accurate responses to patient inquiries, provide medical information, and assist in clinical decision-making.

# 5. Training process of LLAMA-2:

In creating the new Llama 2 models, the initial pretraining approach followed Touvron et al. (2023), utilizing an optimized auto-regressive transformer. However, several changes were implemented for enhanced performance. This included more robust data cleaning, updates to data mixes, training on 40% more total tokens, doubling the context length, and incorporating grouped-query attention (GQA) to improve the scalability of inference.

## 5.1. Pretraining Data:

The training corpus for our models comprises a fresh combination of data gathered from publicly accessible sources, excluding any data sourced from Meta's products or services. To prioritize privacy, we took measures to exclude data from websites known for containing extensive personal information. The training process involved 2 trillion tokens, striking a balance between performance and cost. Additionally, we up-sampled the most reliable sources to enhance knowledge representation and mitigate the occurrence of hallucinations.

The model is trained on a large and diverse dataset of text, including books, articles, and websites, to generate human-like language. Llama 2 is designed to understand and respond to user input in a more human-like way, making it a versatile tool for various applications.

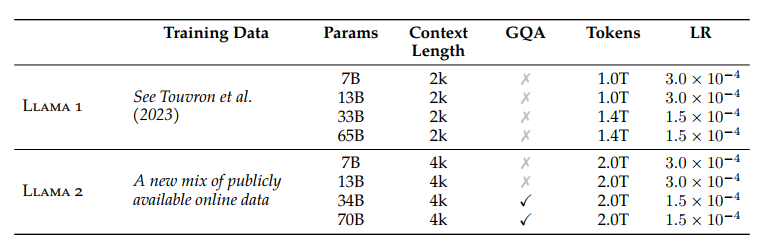


Figure 2. Training Data Comparison

## 5.2. Training Details:

In the development of Llama 2, we largely retain the pretraining settings and model architecture from Llama 1. The model is built on the standard transformer architecture (Vaswani et al., 2017), incorporating pre-normalization with RMSNorm (Zhang and Sennrich, 2019), the SwiGLU activation function (Shazeer, 2020), and rotary positional embeddings (RoPE, Su et al., 2022). Notable architectural differences include an increased context length and the adoption of grouped-query attention (GQA).

The tokenizer used is consistent with Llama 1, employing a byte pair encoding (BPE) algorithm (Sennrich et al., 2016) implemented with Sentence Piece (Kudo and Richardson, 2018). This tokenizer splits numbers into individual digits and uses bytes to handle unknown UTF-8 characters, resulting in a total vocabulary size of 32k tokens.

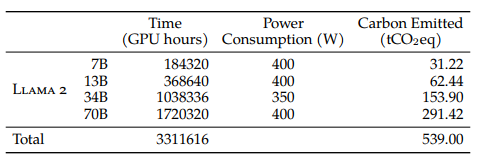


Figure 3. Training Hardware & Carbon Footprint

For pretraining, models are trained on Meta's Research Super Cluster (RSC) and internal production clusters, both equipped with NVIDIA A100s. Notable differences between the clusters include the type of interconnect (Quantum InfiniBand for RSC and RoCE for the production cluster) and variations in per-GPU power consumption caps (400W for RSC and 350W for the production cluster). This two-cluster setup allows a comparison of different interconnect types for large-scale training.

# 6. LLAMA-2 Model Variants:

## 6.1. LLama-2:

An updated version of Llama 1, trained on a new mix of publicly available data. We also increased the size of the pretraining corpus by 40%, doubled the context length of the model, and adopted grouped-query attention (Ainslie et al., 2023). We are releasing variants of Llama 2 with **7B, 13B, and 70B parameters**. We have also trained **34B variants**, which we report on in this paper but are **not releasing**.

## 6.2. Llama 2-Chat:

A fine-tuned version of Llama 2 that is optimized for dialogue use cases. We release variants of this model with **7B, 13B, and 70B** parameters as well

# 7. LLAMA-2 Pretrained Model Evaluation:

In the evaluation of Llama 2 pretrained models, along with Llama 1, MosaicML Pretrained Transformer (MPT), and Falcon models, standard academic benchmarks are used. The evaluations utilize an internal library, and results for MPT and Falcon models are compared between internal evaluations and publicly reported results. The overall performance across various benchmarks is summarized in Figure 4.

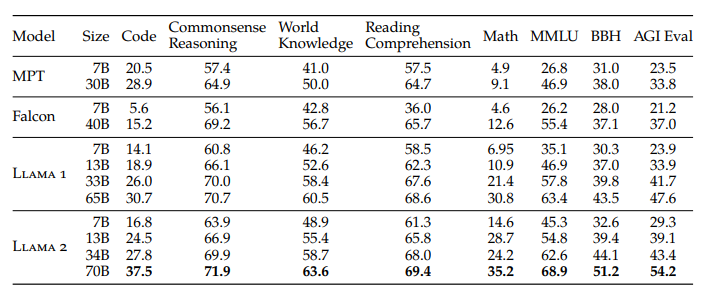


Figure 4. Overall performance on grouped academic benchmarks compared to open-source base models

# 8. Conclusion

In this study, we have introduced Llama 2, a new family of pretrained and fine-tuned models with scales of 7 billion to 70 billion parameters. These models have demonstrated their competitiveness with existing open-source chat models, as well as competency that is equivalent to some proprietary models on evaluation-sets we examined, although they **still lag behind other models like GPT-4**. We meticulously elaborated on the methods and techniques applied in achieving our models, with a heavy emphasis on their alignment with the principles of helpfulness and safety. To contribute more significantly to society and foster the pace of research, they have responsibly given open access to Llama 2 and Llama 2-Chat.

# 9. Difference in Llama-1 and Llama-2 🦙

Llamas 1 and Llamas 2 are both AI models developed by Meta AI, but they have some key differences in terms of their capabilities, training data, and use cases. Here are the main differences between the two models:

## 9.1. Capabilities:

Llamas 1 is a simpler AI model that is designed to generate human-like text based on a given prompt or input. It can be used for applications such as chatbots, language translation, and content generation. Llamas 2, on the other hand, is a more advanced model that can understand and respond to user input in a more human-like way. It can be used for applications such as chatbots, language translation, text summarization, and content generation.

## 9.2. Training Data:

Llamas 1 is trained on a dataset of around 10 billion parameters, while LLama 2 is trained on a larger and more diverse dataset of over 200 billion parameters.

## 9.3. Response Style:

Llamas 1 generates responses that are more formal and professional, while Llamas 2 generates responses that are more conversational and human-like.

## 9.4. Use Cases:

Llamas 1 can be used for applications such as customer service, language translation, and content generation. Llamas 2 can be used for a wider range of applications, including chatbots, virtual assistants, language translation, text summarization, and content generation.

## 9.5. Domain Knowledge:

Llamas 2 has been trained on a larger and more diverse dataset, which means it has a broader knowledge base and can respond to a wider range of topics and questions.

## 9.6. Emotional Understanding:

Llamas 2 has been trained to understand and respond to emotions, which means it can generate responses that are appropriate to the user's emotional state. Llamas 1 does not have this capability.

**In summary**, Llamas 2 is a more advanced and versatile AI model than Llamas 1, with a broader range of capabilities, a larger dataset, and a more conversational response style. Llamas 1 is a simpler and more focused AI model that is better suited for specific applications such as chatbots, language translation, and content generation.

# 10. Difference in Llama-2 and Llama2-chat 🦙

llama 2 and LLama2-Chat are both AI models developed by Meta AI, but they have different capabilities and use cases. Here are the main differences between the two models:

## 10.1. Capabilities:

llama 2 is a general-purpose AI model that can understand and respond to user input.

LLama2-Chat, on the other hand, is a variant of llama 2 that is specifically designed for chat applications. It is optimized for conversational dialogue and can handle multi-turn conversations more effectively than llama 2.

## 10.2 Training Data:

llama 2 is trained on a larger and more diverse dataset than LLama2-Chat. llama 2 is trained on a dataset of over 200 billion parameters.

while LLama2-Chat is trained on a dataset of around 10 billion parameters.

## 10.3. Response Style:

llama 2 generates responses that are more professional response.

while LLama2-Chat generates responses that are more conversational and casual. LLama2-Chat is designed to mimic the way humans chat with each other, using colloquial language and slang.

**In summary,** llama 2 is a more advanced and versatile AI model that can be used for a wide range of applications, while LLama2-Chat is a more specialized model that is specifically designed for chat applications and generates more conversational responses.

# 11. Is LLAMA-2 training data does include healthcare-related texts:

llama 2, the AI model developed by Meta AI, is trained on a large and diverse dataset that includes a wide range of topics and domains, **including medicine and healthcare**. However, it's important to note that the training data used for llama 2 is not exclusively medical or hospital-like data.

The training data for llama 2 includes a variety of **text sources, including books, articles, websites, and other publicly available texts**. This dataset is designed to be representative of the **broad range of topics and language styles** that the model will encounter in real-world applications.

While the training data does include medical and healthcare-related texts, it is not the only domain or topic that is represented. The dataset is diverse and includes texts from many **different fields, such as science, technology, politics, entertainment, and more**.

By training on a diverse dataset, llama 2 is able to learn patterns and relationships across a wide range of topics and language styles, which allows it to generate more accurate and informative responses to user queries.

That being said, if a hospital or medical organization were to use llama 2 for a specific application, they may need **to fine-tune** **the model** or adapt the training data to ensure that it is optimized for the specific domain and task at hand. This is because, while llama 2 is a highly advanced AI model, it is still a general-purpose model that may not be perfectly suited for a specific application without some **additional customization**.