🧠 What is LLaMA?

LLaMA stands for **Large Language Model Meta AI**. It's a family of transformer-based language models developed by **Meta AI**.

- Released as **open-source** alternatives to GPT models
- Available in multiple sizes: 7B, 13B, and 65B parameters
- **Decoder-only architecture** (like GPT), optimized for text generation
- Trained on **public data only**, no private internet sources
- Powerful but **requires strong GPUs** for full-size models

Many newer GenAI systems are built on top of LLaMA-style architectures or fine-tuned variants of LLaMA.

🚀 Why Use LLaMA?

- **Free & Open Source**: No need to pay for API tokens
- **Offline Deployment**: Run locally without internet once downloaded
- **Community Supported**: Many versions on Hugging Face
- **Modular**: Easy to fine-tune and integrate with RAG or Agents

LLaMA models power projects like:

- TinyLLaMA
- Mistral
- Vicuna
- Alpaca
- and more

⚙️ Setup: Load a Lightweight LLaMA-Based Model

Full LLaMA models are heavy, so we will start with:

- TinyLlama/TinyLlama-1.1B-Chat-v1.0 --> a lightweight LLaMA variant
- Or optionally mistralai/Mistral-7B-Instruct-v0.2 if you have GPU

We'll use 🤗 Hugging Face's transformers library.

```
!pip install transformers -q
```

Load Model and Tokenizer

```
from transformers import AutoTokenizer, AutoModelForCausalLM, pipeline

# Model: Lightweight LLaMA-based model
model_name = "TinyLlama/TinyLlama-1.1B-Chat-v1.0"  # You can change this to Mistral for GPU

# Load tokenizer and model
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

# Create a text-generation pipeline
llm_pipeline = pipeline("text-generation", model=model, tokenizer=tokenizer)
```

Show hidden output

Generate Text

```
prompt = " Generative AI in simple words."
response = llm_pipeline(prompt, max_new_tokens=100, do_sample=True, temperature=0.7)[0]['generated_text']
print(response)
```

```
   Generative AI in simple words. It is artificial intelligence that can create content, which can help in the process of creative thinki

2. AI-Powered Chatbots   Chatbots are software applications that provide automated customer service and engagement. AI
```

⚡ Note for GPU Users (Colab)

If you're using Google Colab:

- Go to Runtime → Change runtime type → Select **GPU**
- Try loading larger models like:
  - mistralai/Mistral-7B-Instruct-v0.2
  - meta-llama/Llama-2-7b-chat-hf *(requires HF auth)*

Also install accelerate and bitsandbytes for better performance.

Summary

- LLaMA is Meta’s open-source LLM family
- It’s efficient, scalable, and customizable
- We used a lightweight LLaMA model (TinyLLaMA) to generate text
- Larger models like Mistral or LLaMA-2 can be tried with GPU

```
Start coding or generate with AI.
```