LLaMA (Large Language Model Meta AI) is a family of open-source language models developed by Meta, designed for tasks like text generation, translation, and summarization. It comes in various sizes, with 3B, 7B, 11B, and 90B parameters, offering high efficiency and competitive performance. LLaMA models are available for free, but running them incurs cloud or hardware costs. These models are open-source, making them accessible for research and development. They are designed to perform well with fewer parameters, making them more efficient than similar large-scale models like GPT.

**LLaMA 3.2 11B**

LLaMA 3.2 11B is a powerful language model with 11 billion parameters, designed for text generation, summarization, and conversational tasks. It's part of Meta AI's LLaMA series and can be accessed via Hugging Face.

**Key Model Specifications**

* **Parameters**: 11 billion
* **Layers**: 32-40 transformer layers
* **Embedding Size**: 4096-8192 dimensions
* **Attention Heads**: 32-64 heads
* **Architecture**: Transformer-based

**Features**

* **Scalable**: Useful for text generation, question answering, and conversation.
* **Efficient**: Optimized for performance and memory usage.
* **Versatile**: Can handle multiple NLP tasks like code generation and translation.

**###**code

import requests

API\_URL = "https://api-inference.huggingface.co/models/meta-llama/Llama-3.2-11B"

headers = {"Authorization": "Bearer your\_huggingface\_api\_key"}

def query(payload):

response = requests.post(API\_URL, headers=headers, json=payload)

return response.json()

# Example prompt

data = query({"inputs": "Tell me a fun fact about space."})

print(data[0]["generated\_text"])

**Applications**

* Chatbots and conversational agents
* Text summarization and translation
* Creative content generation
* Code completion and generation

And we can use all of these models depending on our needs

