

## Meta Llama / Llama 3 8B Instruct

accounts/fireworks/models/llama-v3-8b-instruct 

Serverless

LLM

Tunable

Chat

[Open in Playground](#)

[Details](#) [Playground](#)

Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes. The Llama 3 instruction tuned models are optimized for dialogue use cases and outperform many of the available open source chat models on common industry benchmarks.

### Serverless API

**Llama 3 8B Instruct** is available via Fireworks' serverless API, where you pay per token. There are several ways to call the Fireworks API, including Fireworks' Python client, the REST API, or OpenAI's Python client.

See below for easy generation of calls and a description of the raw REST API for making API requests. See the [Querying text models](#) docs for details.

Try it

### API Examples

Generate a model response using the chat endpoint of **llama-v3-8b-instruct**. [API reference](#)

[Python](#) [Typescript](#) [Java](#) [Go](#) [Shell](#) [Chat](#) [Completion](#)



```
import requests
import json

url = "https://api.fireworks.ai/inference/v1/chat/completions"
payload = {
    "model": "accounts/fireworks/models/llama-v3-8b-instruct",
    "max_tokens": 1024,
    "top_p": 1,
    "top_k": 40,
    "presence_penalty": 0,
    "frequency_penalty": 0,
    "temperature": 0.6,
    "messages": [
        {
            "role": "user",
            "content": "Hello, how are you?"
        }
    ]
}
headers = {
    "Accept": "application/json",
    "Content-Type": "application/json",
    "Authorization": "Bearer <API_KEY>"
}
requests.request("POST", url, headers=headers, data=json.dumps(payload))
```

## Fine-tuning

**Llama 3 8B Instruct** can be fine-tuned on your data to create a model with better response quality. Fireworks uses low-rank adaptation (LoRA) to train a model that can be served efficiently at inference time.

See the [Fine-tuning](#) guide for details.

Fine-tune this model


# On-demand deployments

On-demand deployments allow you to use **Llama 3 8B Instruct** on dedicated GPUs with Fireworks' high-performance serving stack with high reliability and no rate limits.

See the [On-demand deployments](#)  guide for details.

Deploy this Base model

 [Open in model playground](#)

 [Finetune this model](#)

 [Deploy this model](#)

## Model Details

Created by	yingliu@fireworks.ai
Created	4/18/2024
Visibility	Public
Kind	Base model
Model size	8B parameters
Fine-tuning	Supported
Serverless LoRA Deployment	Not supported
Function Calling	Not supported
Provider	Meta Llama
Hugging Face	<a href="#">Visit link</a>