**ʌ⅄ Fireworks AI**

# Meta Llama / Llama 3.1 8B Instruct

accounts/fireworks/models/llama-v3p1-8b-instruct 📋

[ Serverless ]  [ LLM ]  [ Tunable ]  [ Chat ]                    **Open in Playground**

Details    Playground

The Meta Llama 3.1 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction tuned generative models in 8B, 70B and 405B sizes. The Llama 3.1 instruction tuned text only models (8B, 70B, 405B) are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

## Serverless API

**Llama 3.1 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-v3p1-8b-instruct**.
API reference ⬈

Python   Typescript   Java   Go   Shell        Chat   Completion                    🗗

```python
import requests
import json

url = "https://api.fireworks.ai/inference/v1/chat/completions"
payload = {
  "model": "accounts/fireworks/models/llama-v3p1-8b-instruct",
  "max_tokens": 16384,
  "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 3.1 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 3.1 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 | 7/22/2024 |
| Visibility | Public |
| Kind | Base model |
| Model size | 8B parameters |
| Fine-tuning | Supported |
| Serverless LoRA Deployment | Supported |
| Function Calling | Not supported |
| Provider | Meta Llama |
| Hugging Face | Visit link |

Fireworks AI

𝕏 📷 ▶ 💼 👾

| Pages | Company | Legal |
|---|---|---|
| Home | Blog | Trust Center ↗ |
| Pricing | Careers ↗ | Terms |