Fireworks AI

Qwen / QwQ-32B

accounts/fireworks/models/qwq-32b 📋

Serverless   LLM   Tunable   Chat   **Open in Playground**

Details   Playground

Medium-sized reasoning model from Qwen.

Serverless API

QwQ-32B 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 **qwq-32b.** 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/qwq-32b",
  "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

QwQ-32B 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 **QwQ-32B** 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	3/5/2025
Visibility	Public
Kind	Base model
Model size	32B parameters
Fine-tuning	Supported
Serverless LoRA Deployment	Not supported
Function Calling	Not supported
Provider	Qwen
Hugging Face	Visit link