Fireworks AI

Qwen / Qwen Qwq 32b Preview

accounts/fireworks/models/qwen-qwq-32b-preview

Serverless LLM Chat

Open in Playground

Details Playground

Qwen QwQ model focuses on advancing AI reasoning, and showcases the power of open models to match closed frontier model performance.QwQ-32B-Preview is an experimental release, comparable to o1 and surpassing GPT-4o and Claude 3.5 Sonnet on analytical and reasoning abilities across GPQA, AIME, MATH-500 and LiveCodeBench benchmarks. Note: This model is served experimentally as a serverless model. If you're deploying in production, be aware that Fireworks may undeploy the model with short notice.

Serverless API

Qwen Qwq 32b Preview 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 **qwen-qwq-32b-preview**.
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/qwen-qwq-32b-preview",
  "max_tokens": 4096,
  "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))
```

On-demand deployments

On-demand deployments allow you to use **Qwen Qwq 32b Preview** 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

Deploy this model

Model Details

Created by	bchen@fireworks.ai
Created	11/27/2024
Visibility	Public
Kind	Base model
Model size	32B parameters
Function Calling	Not supported
Provider	Qwen
Hugging Face	Visit link