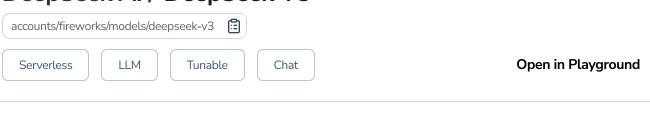


Details







A a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token from Deepseek.

Serverless API

Playground

DeepSeek V3 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 \Box docs for details.

Try it

API Examples

Generate a model response using the chat endpoint of deepseek-v3. API reference





```
import requests
import json
url = "https://api.fireworks.ai/inference/v1/chat/completions"
payload = {
  "model": "accounts/fireworks/models/deepseek-v3",
  "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?"
   }
  ٦
headers = {
  "Accept": "application/json",
  "Content-Type": "application/json",
  "Authorization": "Bearer <API_KEY>"
requests.request("POST", url, headers=headers, data=json.dumps(payload))
```

Fine-tuning

DeepSeek V3 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 **DeepSeek V3** 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

	<u>Open</u>	in	model	<u>playgr</u>	ound
--	-------------	----	-------	---------------	------

Finetune this model

Deploy this model

Model Details

Created by		yingliu@fireworks.ai		
Created			12/30/2024	
Visibility			Public	
Kind		Base model		
Model size		685B parameters		
Fine-tuning			Supported	
Serverless LoRA Deployment		Not supported		
Function Calling			Supported	
Provider			DeepSeek Al	
✓ Fireworks AI Hugging Face	Pages	Company	Legal <u>Visit link</u>	
	Home	Blog	Trust Center [☑]	
SOC 2 HIPAA	Pricing	Careers [☑]	Terms	
Powered by Vanta Powered by Vanta	Models		Privacy	
	Docs [☑]		Licenses	