



# Hackathon Setup guide

Leverage [featherless.ai](https://featherless.ai) - a serverless AI inference platform with unlimited access to thousands of open-source AI models. Focus on building amazing applications while we handle the infrastructure.

# Get Started in 3 Steps





Sign up for Feather Premium (no card required)


- Sign up at [featherless.ai](https://featherless.ai)
- Click the **coupon redemption link** [here](#) - (**ENVIRO** will automatically be entered)
- Enjoy one month for free - you may need to refresh the page to see the changes


## Sign up for Featherless

Start using your favorite LLMs on HuggingFace in minutes!

 Sign up with Google

 Sign up with Hugging Face

 Sign up with Github


 Sign up with Discord

or

Email address

 Email

Password

 Password

✓ Must have at least 8 characters

Create Account

Already have an account? [Log in](#)

By continuing, you agree to the [Terms of Service](#), [Privacy Policy](#), and receipt of occasional service communications.

You'll see the following message  
once you click the coupon link



Activating your subscription...

Please wait while we process your coupon code

Then you're good to go!

## Subscription Activated!

Your subscription is now active and ready to use.





Have fun!

## Feather Premium


Active

Renews on Mar 3, 2026

\$ 25 USD / Month Hackathon premium (100% off)

-  Access to DeepSeek, Kimi-K2 and GLM 4.6
-  Access any model - no limit on size!
-  Up to 4 concurrent connections
-  Up to 32K context

© 2026 Featherless.ai

 [Discord](#) [F.A.Q.](#) [About](#) [Terms](#)  
[Privacy](#) [Cookies](#)

# Get Started in 3 Steps



## Get your API key

→ After subscribing, go to the top right and click on API Keys

→ Here you can create an API key, keep this API key handy, you'll need it to access the service

The screenshot displays the Featherless.ai dashboard. The top navigation bar includes links for Models, Resources, Pricing, Chat, and Status. A search bar for models is located on the right. The left sidebar contains sections for FEATHERLESS (API Keys, Private Models, Private Cloud) and ACCOUNT (Profile, Change Password, Subscription, Payment Methods, Referral Program). The main content area is titled 'API Keys' and contains a table with columns for Key, Created, and Last Used. Two API keys are listed: one for 'aifu' created on Apr 3, 2025, and one for 'litellm' created on Jun 2, 2025. A yellow 'New API Key' button is visible in the top right of the main content area. A user menu is open on the far right, showing options like Account & Billing, API Keys, Discord, Refer a Friend, and Log out. Red annotations highlight the 'New API Key' button and the 'API Keys' option in the user menu.

Key	Created	Last Used
aifu rc_2e717***** *****428a	Apr 3, 2025	Oct 29, 2025
litellm rc_14a6a***** *****f182	Jun 2, 2025	Oct 16, 2025

# Get Started in 3 Steps



## Choose a model

- Go to our [model catalog](#)
- Pick a model for your use case and copy the model ID
- Test with a simple chat completion call!

The screenshot displays the featherless.ai website's 'Models' catalog. The navigation bar at the top includes 'Models' (highlighted with a red box and an arrow), 'Resources', 'Pricing', 'Chat', and 'Status'. A search bar is located on the right. The main content area shows a grid of model cards, each with a 'Warm' status indicator, a model name, a description, and download statistics. A sidebar on the left lists various model families with their respective counts.

**MODEL FAMILY**

- Deepseek 3 (4)
- Gemma 3 (179)
- Gemma 2 (840)
- Gemma (13)
- GLM 4.6 (1)
- GLM 4 (20)
- GPT OSS (1)
- GPT-SW3 (5)
- GPT2-SW3 (5)
- Ling 2 (1)
- Llama 3.3 (147)
- Llama 3.2 (2505)
- Llama 3.1 (1192)
- Llama 3 (1188)
- Llama 2 (1054)
- Mellum (6)
- Mistral 3.1 (18)
- Mistral 3 (120)

**Models** 1 - 20 of 12167

Model Name	Description	Hearts	Downloads
qwen25-7b-lc	Qwen/Qwen2.5-7B-Instruct	852	8,137,502
qwen3-0b6	Qwen/Qwen3-0.6B	755	7,199,888
qwen25-0b5	Gensyn/Qwen2.5-0.5B-Instruct	26	6,456,472
llama31-8b	meta-llama/Llama-3.1-8B-Instruct	4,863	5,172,472
llama31-8b	meta-llama/Meta-Llama-3.1-8B-Instruct	4,863	5,172,472
qwen3-4b	Qwen/Qwen3-4B-Instruct-2507	440	4,008,995
tinylama-1b1	TinyLlama/TinyLlama-1.1B-Chat-v1.0	1,439	3,955,967
llama32-1b	meta-llama/Llama-3.2-1B-Instruct	1,143	3,814,050
qwen25-3b	Qwen/Qwen2.5-3B-Instruct	325	3,651,458

# Popular Model Options



DeepSeek-V3-0324

Advanced reasoning and  
coding capabilities



Llama-3.3-70B-Instruct

Powerful general-purpose  
model



Kimi-K2-Instruct

Long context capabilities



Mistral-Nemo-Instruct

Fast and efficient processing



GLM-4.6

Bilingual capabilities

# Your First API Call

## Using Python Requests

```
import requests
response = requests.post(

url="https://api.featherless.ai/v1/chat/completi
ons",
  headers={
    "Authorization": "Bearer YOUR_API_KEY"
  },
  json={
    "model": "deepseek-ai/DeepSeek-V3-0324",
    "messages": [
      {"role": "user", "content": "Hello!"}
    ]
  }
)
print(response.json())
```

## Using OpenAI SDK

```
from openai import OpenAI
client = OpenAI(
  base_url="https://api.featherless.ai/v1",
  api_key="YOUR_API_KEY"
)
response = client.chat.completions.create(
  model="deepseek-ai/DeepSeek-V3-0324",
  messages=[
    {"role": "user", "content": "Hello!"}
  ]
)
print(response.choices[0].message.content)
```



API

Data:

# Core API Endpoints

`/v1/chat/completions`

**Best for:** Chatbots, virtual assistants, conversational applications

- Structured messages with roles (system, user, assistant)
- Maintains conversation context automatically
- Ideal for interactive user-assistant interactions

`/v1/completions`

**Best for:** Content generation, text transformation, data extraction

- Takes a single prompt string
- Direct control over prompt format
- Maximum flexibility for custom use cases

# Troubleshooting Common Errors



## 401 - Unauthenticated

API key not recognized. Verify you've copied it correctly or generate a new one from account settings.



## 403 - Unauthorized

Model is gated. Visit the model's page, click "Unlock Model," and agree to license terms.



## 500 - Internal Server Error

Request could not be processed. Check for unsupported parameters in API documentation.



## 503 - Service Unavailable

Insufficient capacity or cold model. Retry the request. If it persists after three attempts, report on Discord.

# Concurrency Limits

Featherless operates on a unique model of **capacity reservation** rather than token-based billing. Your subscription tier determines the maximum size and number of concurrent AI model inference calls you can make, ensuring consistent performance without unpredictable token costs.

## Model Concurrency Costs

7B to 15B	1	Qwen 2.5 7B, Llama2 13B
24B to 34B	2	Qwen 32B Coder, Mistral 3 24B
70B and 72B	4	Llama 3.3 70B, Qwen 2.5 72B
Deepseek v3, R1 & Kimi-K2	4	(Feather Premium only)

For a more advanced explanation and detailed examples, visit our [Concurrency Limits documentation](#).

# Application Guides

Explore our detailed application guides designed to help you seamlessly integrate Featherless.ai into your existing projects and workflows.

Find these guides and more on our documentation page:

[featherless.ai/docs/application-guides](https://featherless.ai/docs/application-guides)

## Coding

### Aider

Powering your in-terminal coding assistance with inference from featherless

### Cursor

Setting Up Cursor with Featherless.ai

### Roo Code

Roo Code (prev. Roo Cline) gives you a whole dev team of AI agents in your code editor.

### Cline

A fully collaborative AI partner that's open source, fully extensible, and designed to amplify developer impact.

## Workflow Automation

### n8n

Supercharge your AI workflows by connecting Featherless with n8n's powerful automation platform. Build sophisticated AI-powered automations that connect with 1000+ apps and services.

### Dify

Unleash 10,000+ open-source models in Dify with our dedicated plugin. Build powerful AI applications using serverless inference and low-code orchestration without the infrastructure headaches.

# Ready to Build!

You now have everything you need to start building amazing AI-powered applications. Here's your action plan:

## 1 Test Your Setup

Get your API key and test with a simple example

## 2 Explore Models

Find the right model for your use case in the catalog

## 3 Review Examples

Check the GitHub cookbook for code samples

## 4 Start Building

Begin your hackathon project with confidence

---

Quick Links: [Sign Up](#) · [Model Catalog](#) · [Documentation](#) · [GitHub](#)

**Good luck with your project!** We can't wait to see what you build!