



Hackathon Setup guide

Leverage 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

- Sign up at featherless.ai,
- go to the Pricing page
- Select Feather Premium and fill in coupon code DATATHON

featherless.ai

ModelsResources ▾PricingChat● Status

Search models...

Profile

← Upgrade your plan

Billing Information

Full name

Country or region

Belgium ▾

Address

Payment Information

Card

Google Pay

Bank

Cash App Pay

Secure, fast checkout with Link ▾

Card number

1234 1234 1234 1234

MASTERCARD

VISA

AMERICAN EXPRESS

Expiry date

MM / YY

Security code

CVC

123

Feather Premium

\$25.00 /month

Max. ∞B

Change

🔑 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

You will be charged the following:

1 × Feather Premium (at \$25.00 / month)	\$25.00
Amount due now	\$25.00

Got a coupon or referral code?

Enter promotion code here

Apply

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 user's profile is shown in the top right corner with the name 'Darin Verheijke'. 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 the following data:

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

A yellow button labeled 'New API Key' is located at the top right of the table. A red arrow points to this button. Another red arrow points to the 'API Keys' option in the user menu, which is also highlighted with a red box. The footer of the page includes copyright information for Featherless.ai and links to Discord, F.A.Q., About, Terms, Privacy, Cookies, and a settings icon.

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' page. 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 in the top right corner. The main content area is titled 'Models' and shows a grid of model cards. On the left, there is a 'MODEL FAMILY' sidebar with a search bar and a list of model families including Deepseek 3, Gemma 3, Gemma 2, Gemma, GLM 4.6, GLM 4, GPT OSS, GPT-SW3, GPT2-SW3, Ling 2, Llama 3.3, Llama 3.2, Llama 3.1, Llama 3, Llama 2, Mellum, Mistral 3.1, and Mistral 3. Each model card in the grid includes a 'Warm' status indicator, the model name, the provider (e.g., Qwen, Gensyn, meta-llama, TinyLlama), and download statistics (hearts and download count). The grid shows models like qwen25-7b-lc, qwen3-0b6, qwen25-0b5, llama31-8b, llama32-1b, and qwen25-3b.

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

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 hackathon project! We can't wait to see what you build!