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.

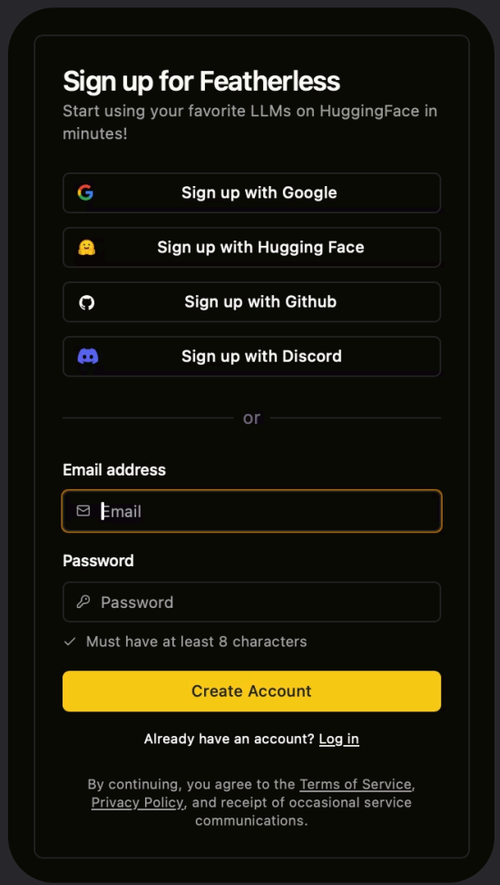
Get Started in 3 Steps

Sign up for Feather Premium (no card required)

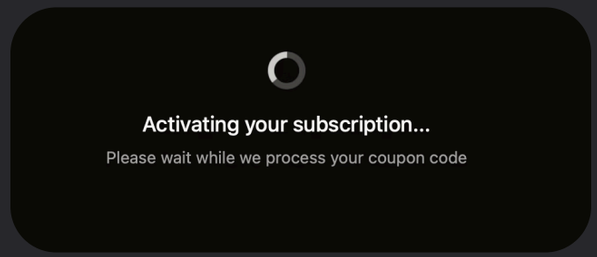
→ Sign up at **featherless.ai**

→ Click the **coupon redemption link here** - (**CTRLHACKDEL** will automatically be entered)

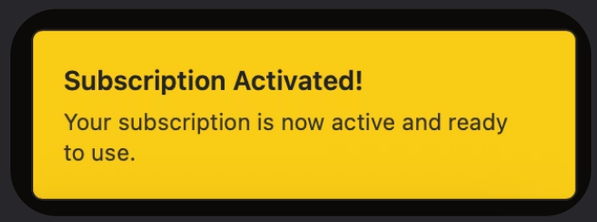
→ Enjoy one month for free - you may need to refresh the page to see the changes



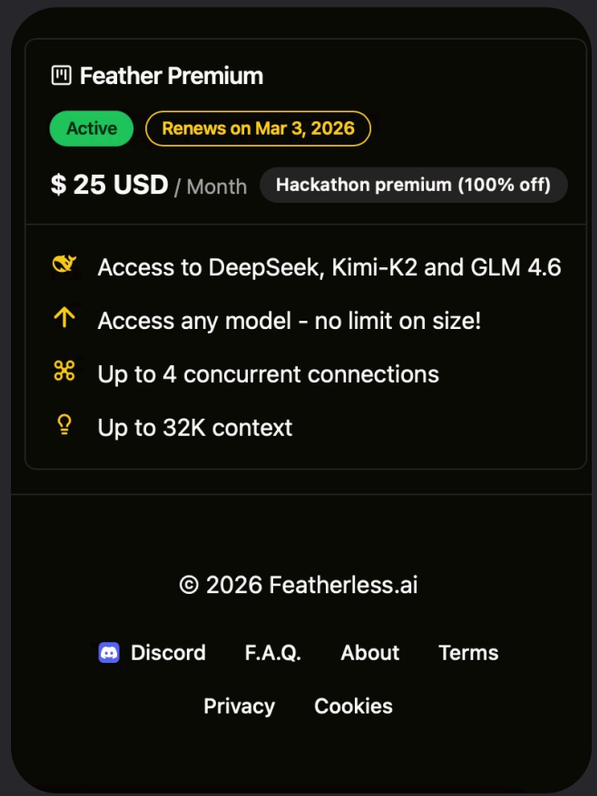
You'll see the following message once you click the coupon link



Then you're good to go!



Have fun!

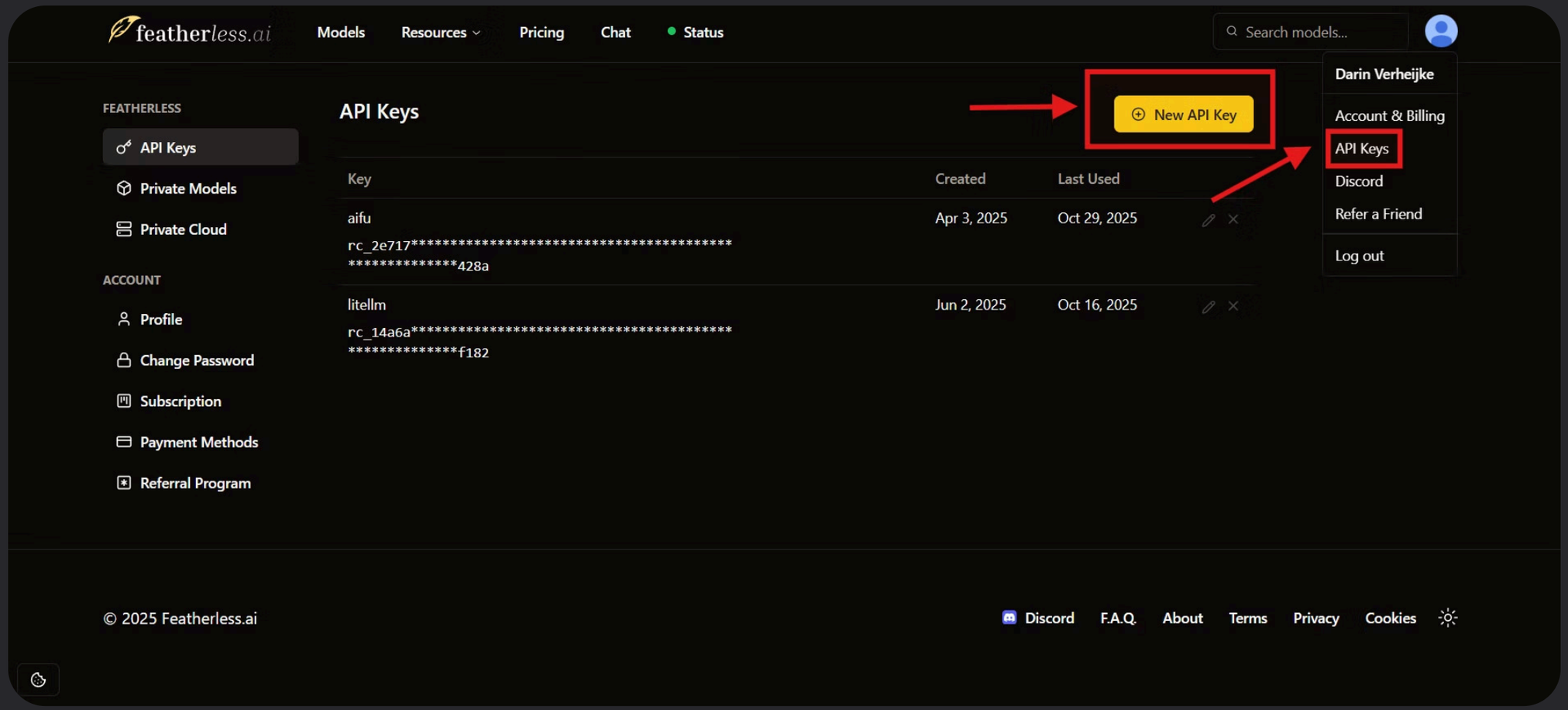


Get Started in 3 Steps

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



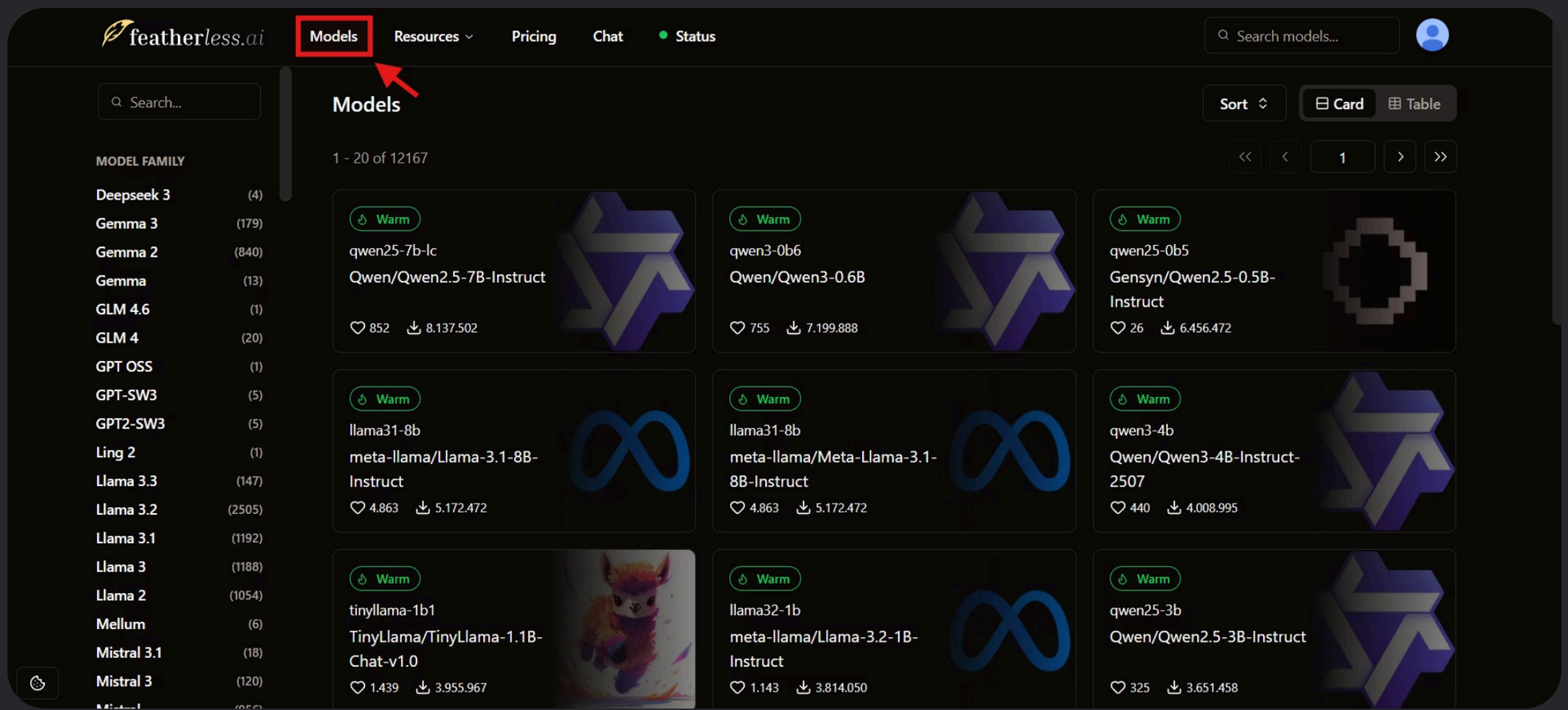
Get Started in 3 Steps

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!



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

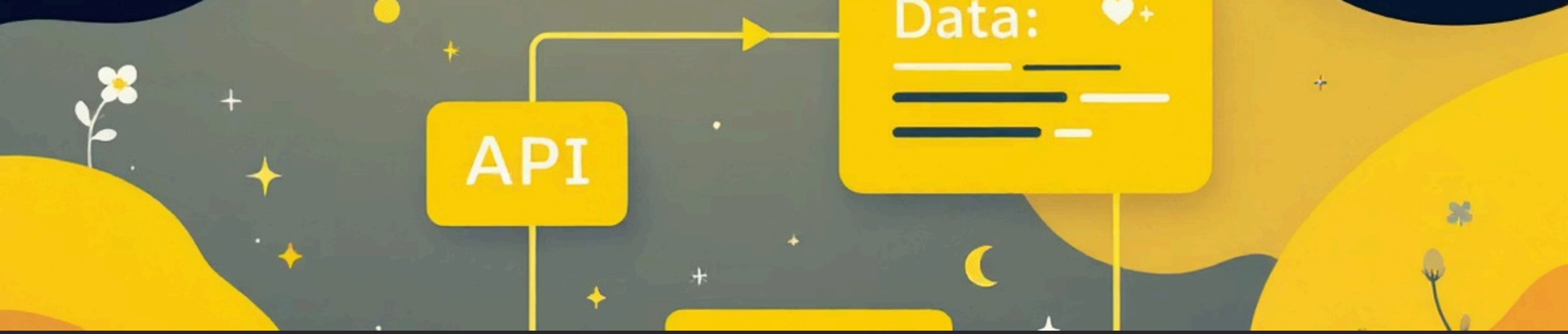
Using Python Requests

```
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)
```



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