OpenRouter large model API aggregation platform

Concept Introduction

1.1 What is "OpenRouter"?

OpenRouter is a **large-model API aggregation platform**. You can think of it as a "universal remote control for large models." Typically, if you want to use OpenAI's GPT-4, Anthropic's Claude-3, and Google's Gemini, you need to register accounts on each platform, obtain three sets of API keys, and write three different sets of calling code for each of their different API formats.

OpenRouter solves this problem. It provides a **unified API interface** (compatible with OpenAI's API format). With just one OpenRouter API key, you can access all of these models, as well as hundreds of other open and closed-source models, through this single entry point.

The **core advantages** of integrating OpenRouter into `largemodel` projects are:

- **Ultimate flexibility**: You can seamlessly switch between any model, such as GPT, Claude, or Llama, by simply changing a single model name in the configuration file, without changing any code. **Cost-Effectiveness**: OpenRouter provides clear pricing for each model, allowing you to easily compare and select the most cost-effective model for your specific task.

Explore Cutting-Edge Models: Many new and excellent open source models are included in OpenRouter as soon as possible, allowing developers to quickly try them out.

In the `largemodel` project, configure the `llm_platform` parameter to `openrouter` to enable this powerful aggregation platform.

2. Practical Operations

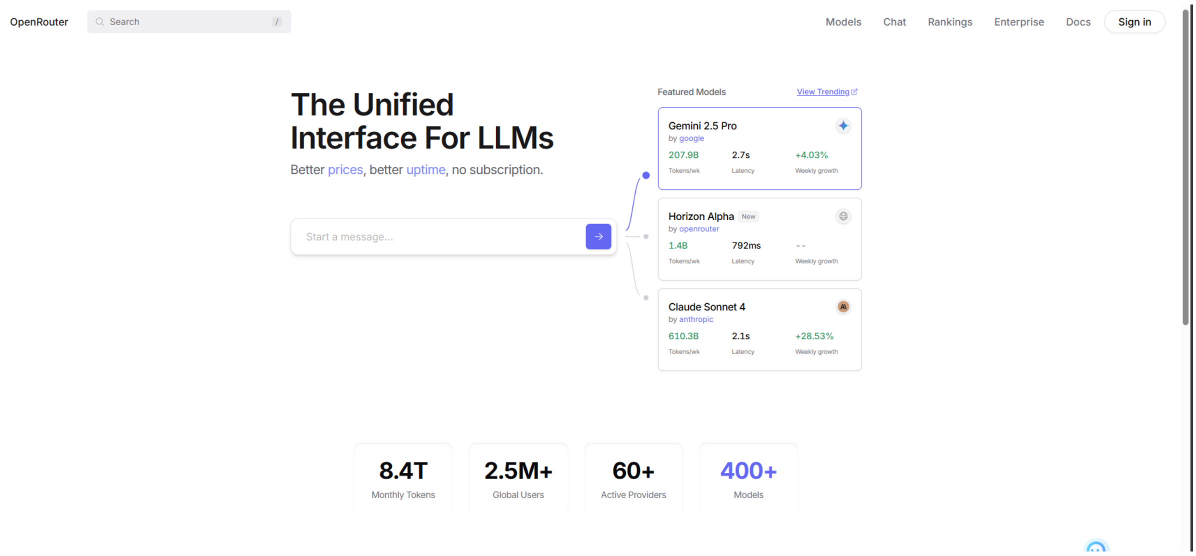
2.1 Obtaining API Credentials

1. **Register for OpenRouter**: Visit OpenRouter.ai and register an account using Google or email. If the connection fails, search for the "OpenRouter" platform yourself.
2. **Top Up**: OpenRouter uses a prepaid payment model. You must first top up a small amount (e.g., $5) on the "Credits" page to start using it. Alternatively, you can directly use the free models, but there are rate limits.
3. **Create an API Key**: On the "Keys" page, click "Create Key" to generate a new API key. **Copy and save it immediately**.

2.2 Graphical Process

1. First, visit the following website. If the connection fails, search for the "OpenRouter" platform and then search for the keyword for the large model.

OpenRouter



The Unified Interface For LLMs

Better prices, better uptime, no subscription.

Gemini 2.5 Pro
by google
207.9B Tokens/wk | 2.7s Latency | +4.03% Weekly growth

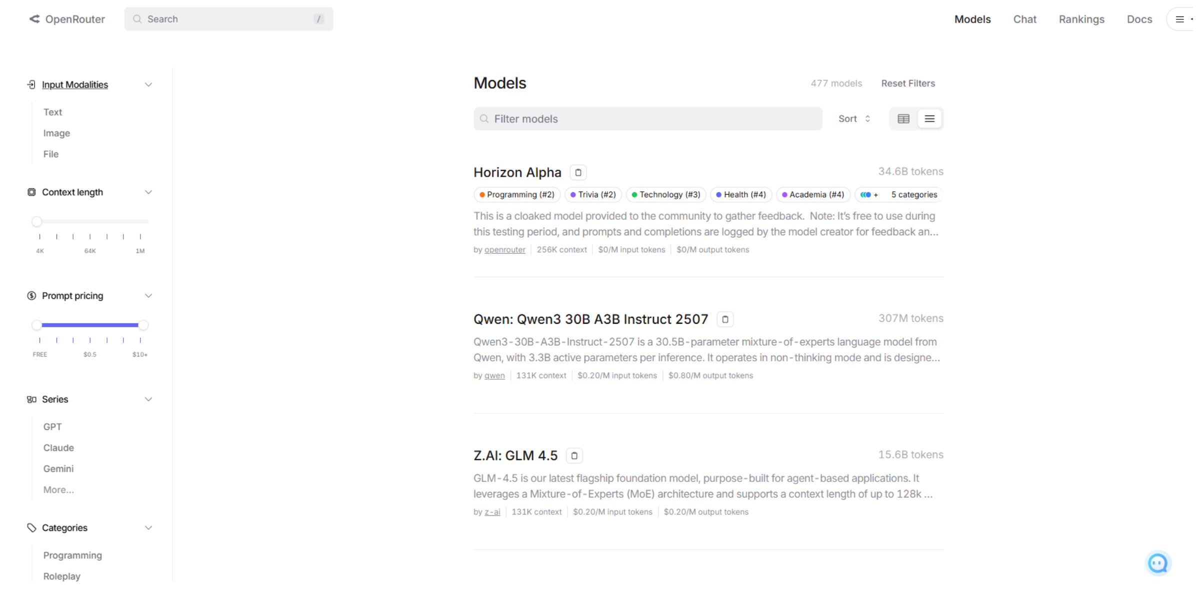
Horizon Alpha New
by openrouter
1.4B Tokens/wk | 792ms Latency | - - Weekly growth

Claude Sonnet 4
by anthropic
610.3B Tokens/wk | 2.1s Latency | +28.53% Weekly growth

8.4T Monthly Tokens | 2.5M+ Global Users | 60+ Active Providers | 400+ Models

2.Register and log in to your account

3.After registering your account, click to enter the Model Square



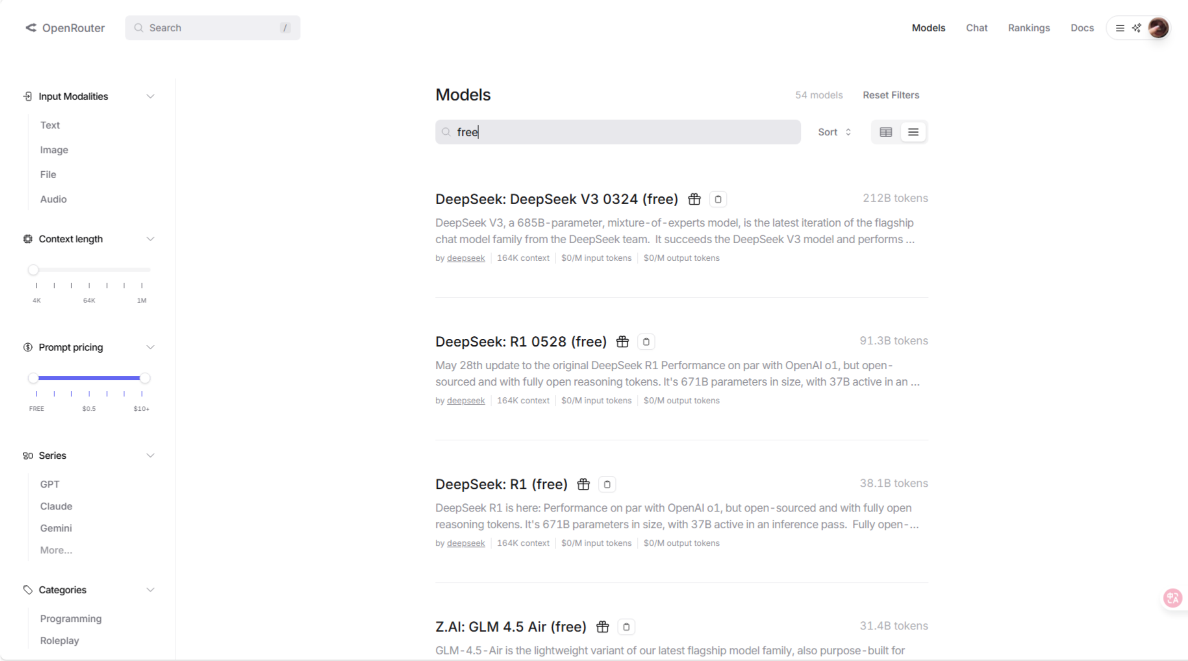
Models 477 models

Horizon Alpha — 34.6B tokens
Programming (#2) · Trivia (#2) · Technology (#3) · Health (#4) · Academia (#4) · + 5 categories
This is a cloaked model provided to the community to gather feedback. Note: It's free to use during this testing period, and prompts and completions are logged by the model creator for feedback an...
by openrouter | 256K context | $0/M input tokens | $0/M output tokens

Qwen: Qwen3 30B A3B Instruct 2507 — 307M tokens
Qwen3-30B-A3B-Instruct-2507 is a 30.5B-parameter mixture-of-experts language model from Qwen, with 3.3B active parameters per inference. It operates in non-thinking mode and is designe...
by qwen | 131K context | $0.20/M input tokens | $0.80/M output tokens

Z.AI: GLM 4.5 — 15.6B tokens
GLM-4.5 is our latest flagship foundation model, purpose-built for agent-based applications. It leverages a Mixture-of-Experts (MoE) architecture and supports a context length of up to 128k ...
by z-ai | 131K context | $0.20/M input tokens | $0.20/M output tokens

4.You can enter free in the search bar to find models that can be called for free.



Models 54 models Reset Filters

free Sort

DeepSeek: DeepSeek V3 0324 (free) 212B tokens

DeepSeek V3, a 685B-parameter, mixture-of-experts model, is the latest iteration of the flagship chat model family from the DeepSeek team. It succeeds the DeepSeek V3 model and performs ...

by deepseek | 164K context | $0/M input tokens | $0/M output tokens

DeepSeek: R1 0528 (free) 91.3B tokens

May 28th update to the original DeepSeek R1 Performance on par with OpenAI o1, but open-sourced and with fully open reasoning tokens. It's 671B parameters in size, with 37B active in an ...

by deepseek | 164K context | $0/M input tokens | $0/M output tokens

DeepSeek: R1 (free) 38.1B tokens

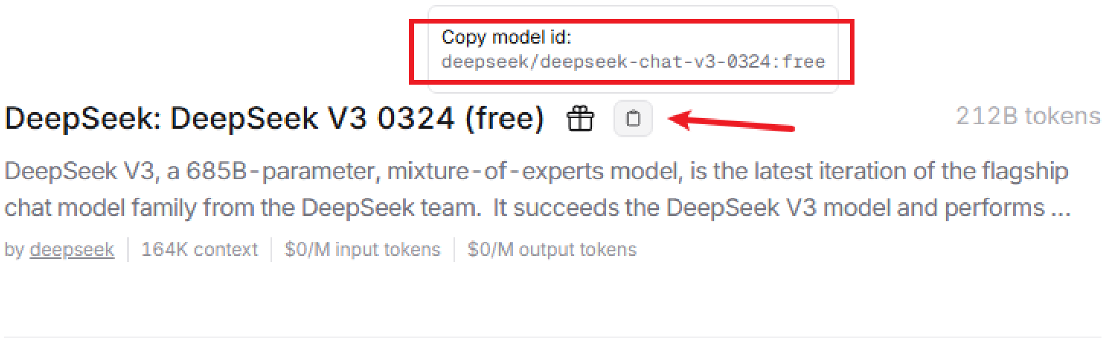
DeepSeek R1 is here: Performance on par with OpenAI o1, but open-sourced and with fully open reasoning tokens. It's 671B parameters in size, with 37B active in an inference pass. Fully open-...

by deepseek | 164K context | $0/M input tokens | $0/M output tokens

Z.AI: GLM 4.5 Air (free) 31.4B tokens

GLM-4.5-Air is the lightweight variant of our latest flagship model family, also purpose-built for

5.Place the mouse on the clipboard icon next to the model to see the name to be used when calling the model. Click it to copy it.



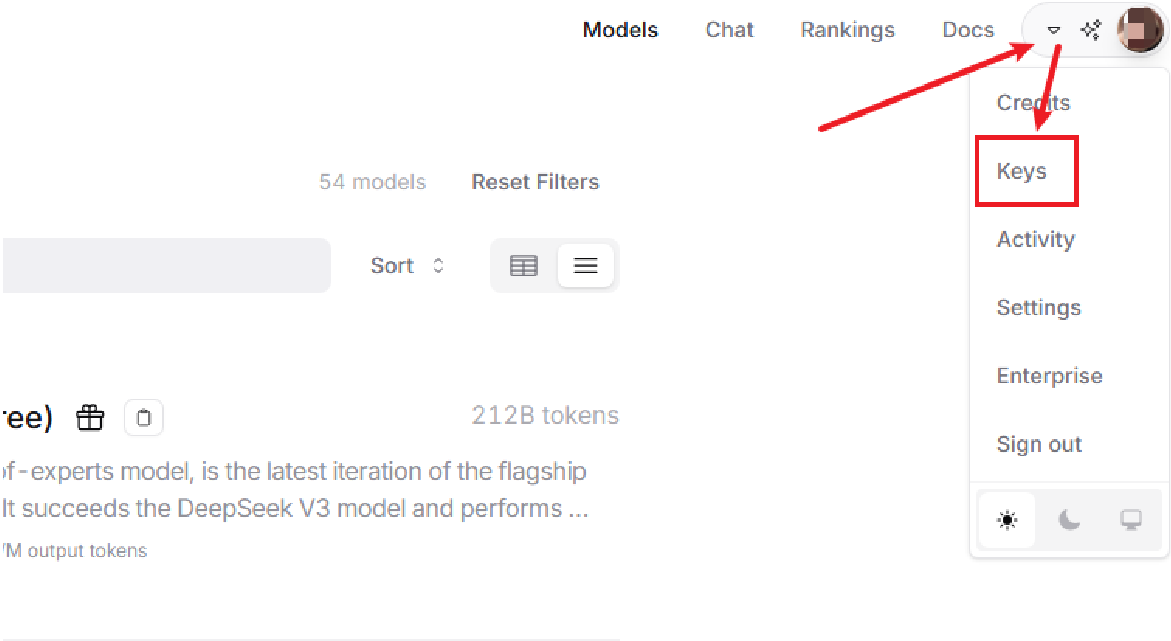
Copy model id:
deepseek/deepseek-chat-v3-0324:free

DeepSeek: DeepSeek V3 0324 (free) 212B tokens

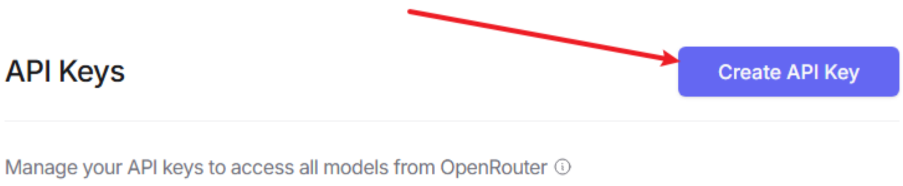
DeepSeek V3, a 685B-parameter, mixture-of-experts model, is the latest iteration of the flagship chat model family from the DeepSeek team. It succeeds the DeepSeek V3 model and performs ...

by deepseek | 164K context | $0/M input tokens | $0/M output tokens

6.Move the mouse to the upper right corner next to the avatar, it will expand, you can click Keys, go in and create a key.



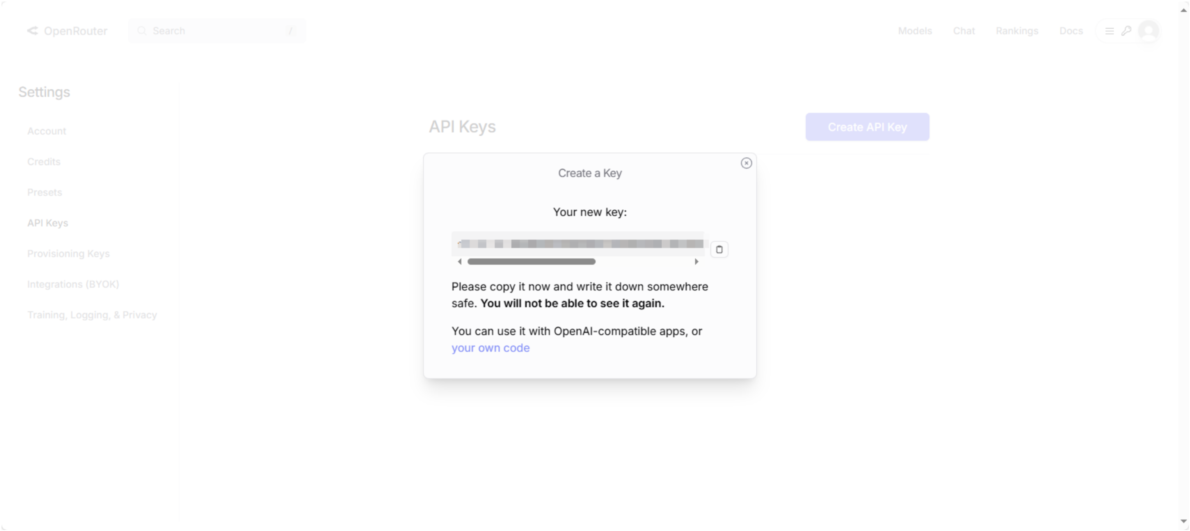
7.Click to apply for API key



API Keys

Manage your API keys to access all models from OpenRouter

8.After applying, you will get a string of API keys



Create a Key

Your new key:

[redacted]

Please copy it now and write it down somewhere safe. **You will not be able to see it again.**

You can use it with OpenAI-compatible apps, or your own code

2.3 Configuration file modification

1. **Open** `large_model_interface.yaml`:

```
vim ~/yahboom_ws/src/largemodel/config/large_model_interface.yaml
```

Then, in the openrouter_api_key field, replace sk-xxxxxxxxxxxx with the API key you copied earlier.

openrouter_model is used to configure the main model during the session.

```
# OpenRouter平台配置 (OpenRouter Platform Configuration)
openrouter_api_key: "sk-xxxxxxxxxxxxxxxxxxx"
openrouter_model: "moonshotai/kimi-k2:free" # The model to use, for example
"google/gemini-pro-vision"
```

2. **Open** `yahboom.yaml`:

```
vim ~/yahboom_ws/src/largemodel/config/yahboom.yaml
```

3. **Switch platform**: Change the `llm_platform` parameter to `openrouter`.

```
llm_platform: 'openrouter'
```

After switching the platform parameters, return to the workspace and compile again to make them effective.