

AI Pyramid OpenAI-Compatible API Chat Example

This example explains how to deploy a StackFlow chat model and the OpenAI-compatible API on AI Pyramid, then invoke the local model through cURL, the Python SDK, and Chatbox.

Preparation

1. Refer to [AI Pyramid Software Package Update](#) to complete the installation of the following model packages and software packages.

```
apt install lib-llm llm-sys llm-llm llm-openai-api
```

```
apt install llm-model-qwen2.5-1.5b-int4-ax650
```

Note

After installing a new model each time, you need to manually execute **systemctl restart llm-openai-api** to update the model list.

Curl Call

```
curl http://127.0.0.1:8000/v1/models \  
-H "Content-Type: application/json"
```

```
curl http://127.0.0.1:8000/v1/chat/completions \  
-H "Content-Type: application/json" \  
-H "Authorization: Bearer sk-xxxxxxx" \  
-d '{  
  "model": "qwen2.5-1.5B-Int4-ax650",  
  "messages": [  
    {"role": "developer", "content": "You are a helpful home assistant."},  
    {"role": "user", "content": "Write a one-sentence bedtime story about a unicorn."}  
  ]  
'
```

```
root@m5stack-AI-Pyramid: ~
File Edit View Search Terminal Help
root@m5stack-AI-Pyramid:~# curl http://127.0.0.1:8000/v1/chat/completions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer sk-xxxxxxx" \
-d '{
  "model": "qwen2.5-1.5B-Int4-ax650",
  "messages": [
    {"role": "developer", "content": "You are a helpful home assistant."},
    {"role": "user", "content": "Write a one-sentence bedtime story about a un
icorn."}
  ]
}'
{"id": "chatcmpl-c351698a-0681-43f5-b2e2-68d0e92a7028", "object": "chat.completion"
, "created": 1767929302, "model": "qwen2.5-1.5B-Int4-ax650", "choices": [{"message": {"
role": "assistant", "content": "A gentle unicorn trotted through the forest after a
long day of exploring the world, feeling content in the knowledge that the star
s above promised another wonderful night in the great, vast sky."}]}]root@m5stac
k-AI-Pyramid:~#
```

Python Call

```
from openai import OpenAI
client = OpenAI(
    api_key="sk-",
    base_url="http://127.0.0.1:8000/v1"
)

client.models.list()
print(client.models.list())
```

```

from openai import OpenAI
client = OpenAI(
    api_key="sk-",
    base_url="http://127.0.0.1:8000/v1"
)

completion = client.chat.completions.create(
    model="qwen2.5-1.5B-Int4-ax650",
    messages=[
        {"role": "developer", "content": "You are a helpful home assistant."},
        {"role": "user", "content": "Turn on the light!"}
    ]
)

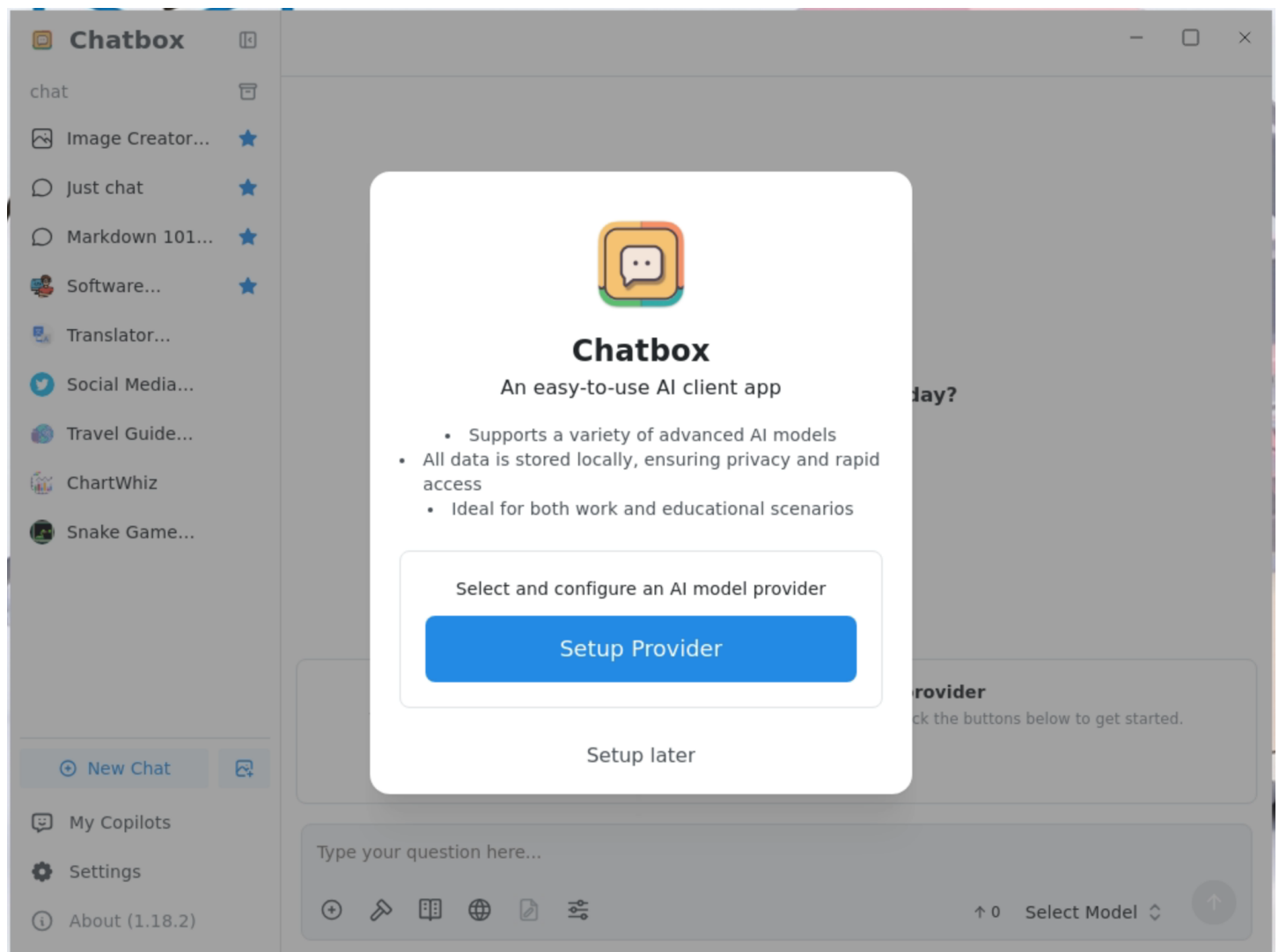
print(completion.choices[0].message)

```

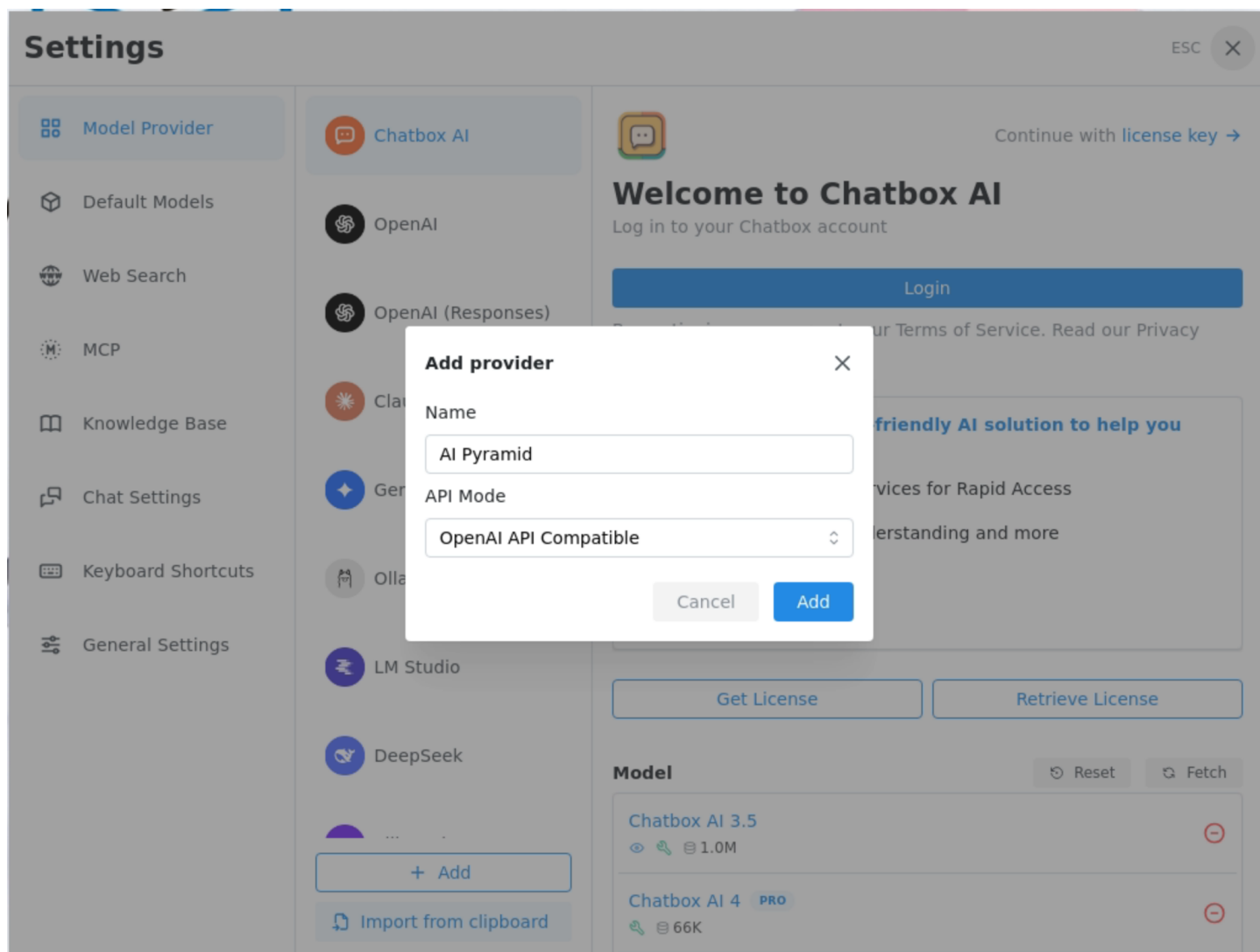
ChatBox Call

Get [ChatBox](#)

Click **Setup Provider** to add a model provider.



In Add provider, set Name to **AI Pyramid**, and select **OpenAI API Compatible** for API Mode.



In API Host, enter the IP address and API path of AI Pyramid, then retrieve and add the installed models.

Settings

ESC

Model Provider

Default Models

Web Search

MCP

Knowledge Base

Chat Settings

Keyboard Shortcuts

General Settings

Chatbox AI

OpenAI

OpenAI (Responses)

Claude

Gemini

Ollama

LM Studio

DeepSeek

...

+ Add

Import from clipboard

AI Pyramid

Name

AI Pyramid

API Mode

OpenAI API Compatible

API Key

Check

API Host

http://127.0.0.1:8000/v1

API Path

/chat/completions

http://127.0.0.1:8000/v1/chat/completions

Improve Network Compatibility

Model

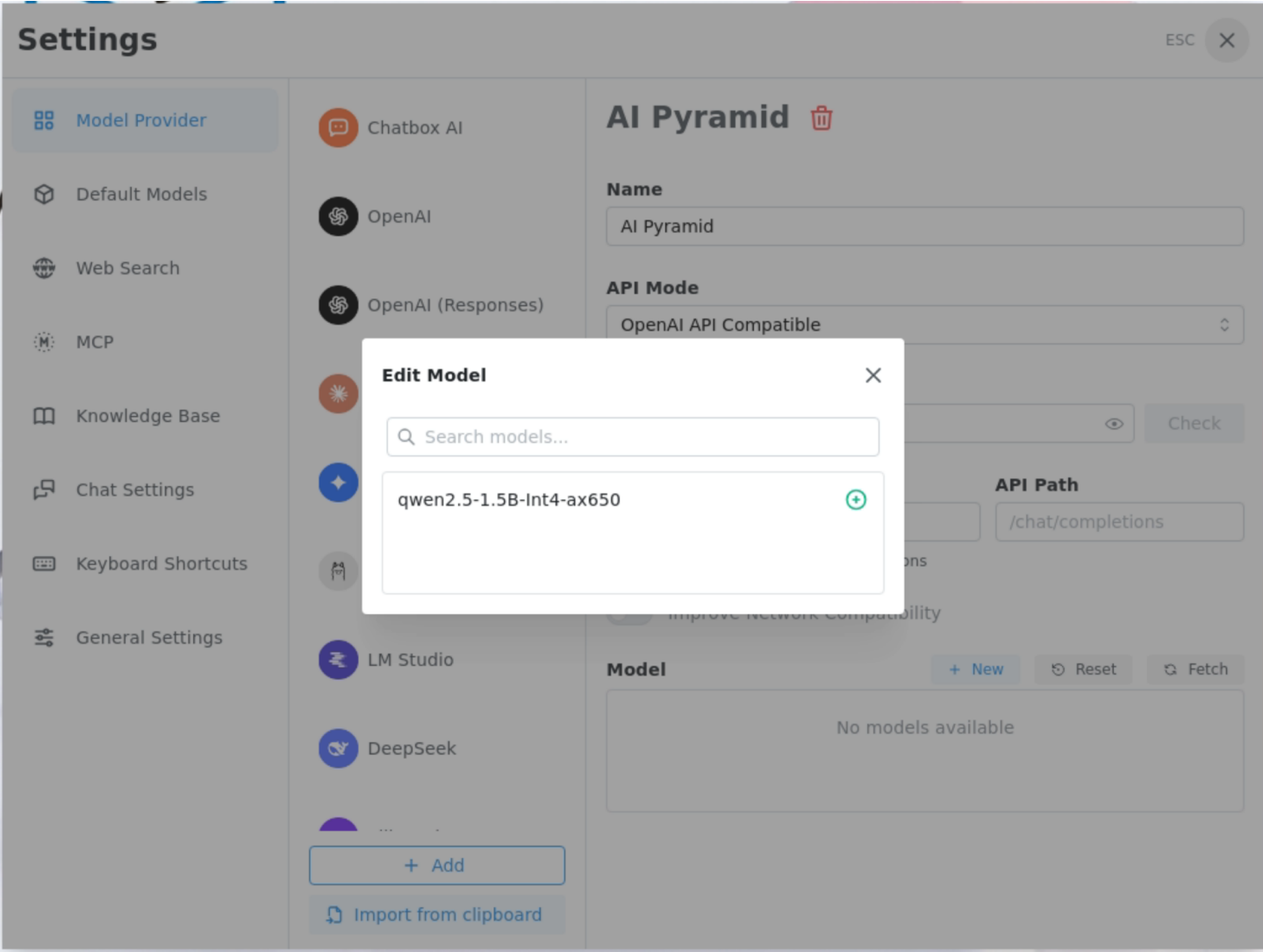
+ New

Reset

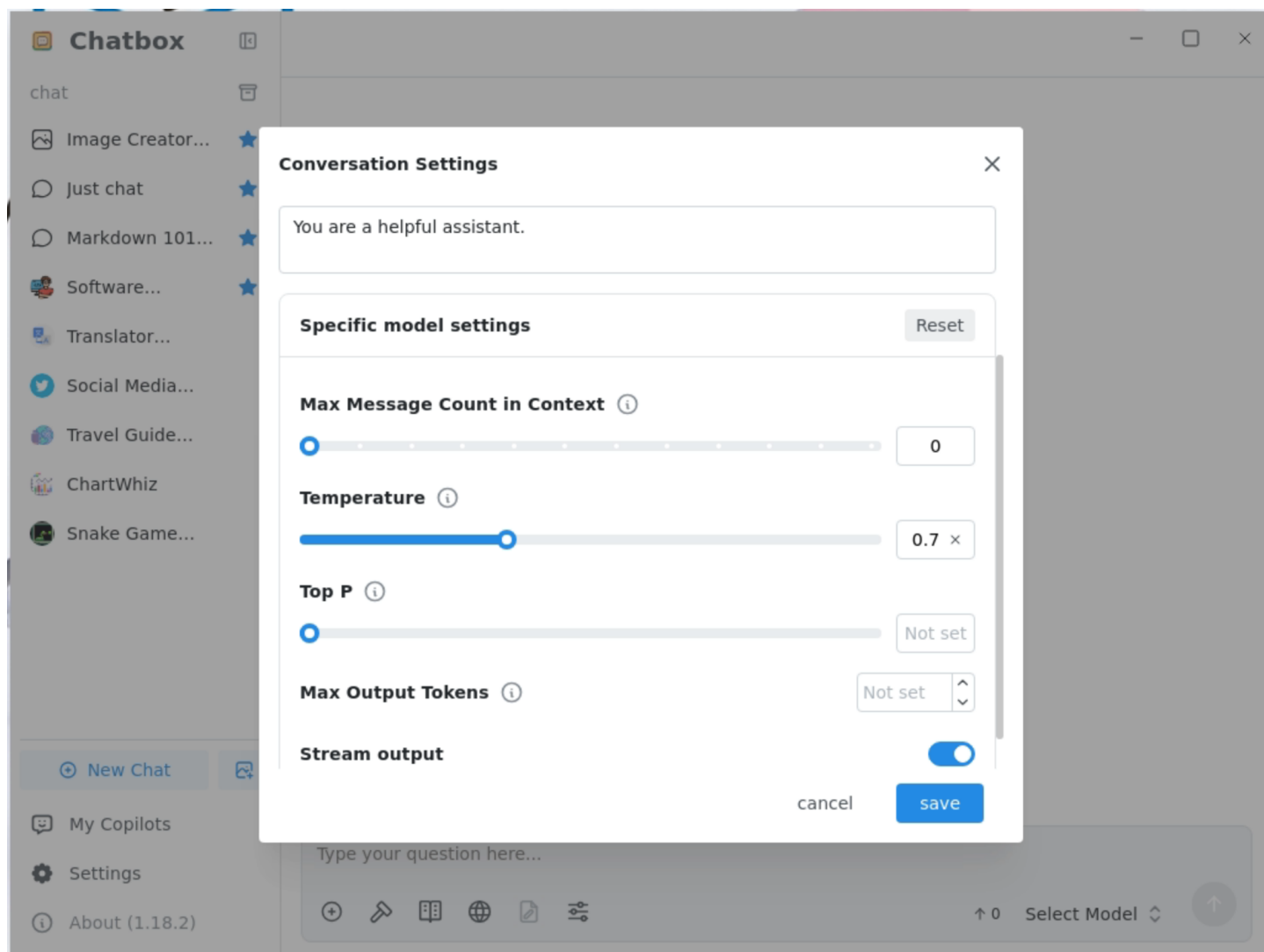
Fetch

No models available

Add the **qwen2.5-1.5B-Int4-ax650** model provided by **LLM8850**.



Set the maximum context message length to 0.



Streaming output is supported.

Chatbox

chat

Image Creator...

Just chat

Markdown 101...

Software...

[ERROR: Extra...

Translator...

Social Media...

Travel Guide...

ChartWhiz

Snake Game...

New Chat

My Copilots

Settings

About (1.18.2)

[ERROR: Extra data: line 2 column 1...

You are a helpful assistant.

Write a one-sentence bedtime story about a unicorn.

< 5 / 5 >

"Once upon a time, a unicorn lived in a magical meadow with friends. She had a shiny mane and beautiful blue eyes, and she loved to dance and sing to the happy tunes that the trees hummed by her feet."

tokens used: 76, model: AI Pyramid (qwen2.5-1.5B-Int4-ax650)

Type your question here...

↑ 0 qwen2.5-1.5B-Int4-ax650

....