Getting Started

The Vector Inference Platform is available to all Vector Institute community members. For an up-to-date list of available models and their specifications, visit inference.vectorinstitute.ai.

Getting an API Key

Access is managed via invite links. To get started:

1. Contact the AI Engineering team to request access. They will create your account and send you an invite email.
2. Click the link in the invite email. It is valid for **48 hours** and takes you to the Vector Proxy dashboard.
3. Your API key is generated and displayed **once** — copy and save it immediately. The key will not be shown again.

API keys have the format `vp_xxxxxxxx.yyyyyy...`.

Prerequisites

Install the OpenAI Python client:

```
pip install openai
```

Usage

The platform exposes an OpenAI-compatible API at **https://proxy.vectorinstitute.ai/v1**. Use it as a drop-in replacement for any OpenAI client by changing the `base_url` and `model` parameters.

```
from openai import OpenAI

client = OpenAI(
    base_url="https://proxy.vectorinstitute.ai/v1",
    api_key="vp_xxxxxxxx.yyyyyy..."
)

stream = client.chat.completions.create(
    model="<model-id>",  # see inference.vectorinstitute.ai for available models
    messages=[{"role": "user", "content": "Explain attention mechanisms in transformers."}],
    stream=True,
)
```

```
for chunk in stream:
    if chunk.choices:
        print(chunk.choices[0].delta.content or "", end="", flush=True)
```

You can also use curl:

```
curl https://proxy.vectorinstitute.ai/v1/chat/completions \
  -H "Authorization: Bearer vp_xxxxxxxx.yyyyyy..." \
  -H "Content-Type: application/json" \
  -d '{
    "model": "<model-id>",
    "messages": [{"role": "user", "content": "Hello!"}]
  }'
```

Listing Available Models

Retrieve the current list of enabled models programmatically:

```
curl https://proxy.vectorinstitute.ai/v1/models \
  -H "Authorization: Bearer vp_xxxxxxxx.yyyyyy..."
```

Or visit inference.vectorinstitute.ai for a visual overview.

Revision #11
Created 2026-01-08 20:54:51 UTC
Updated 2026-05-26 22:09:18 UTC by Amrit Krishnan