anyscale

Anyscale Private Endpoints is a fast, scalable managed endpoint in your account.

- **Supports Llama2 7b-13b-70B LLMs**
- **Serve and Finetune APIs**
- **Pay per instance utilization across cloud providers**
- **Managed platform**

Every Organization's Challenge

OpenAI and Cohere use Ray to train their largest LLMs. Spotify built their next generation platform on Ray and LiveEO accelerates and optimizes their geospatial workloads by up to 65% on Anyscale.

cohere

OpenAI

LiveEO

Spotify

Anyscale Private Endpoints

Open LLMs such as Llama2 have made it easier for many organizations to build their own ChatGPT. Enterprises can start and innovate using Anyscale Private Endpoints securely without needing a team to manage complex and expensive infrastructure. Users can also customize and finetune on top of the Anyscale platform easily.

Challenges of deploying LLM applications

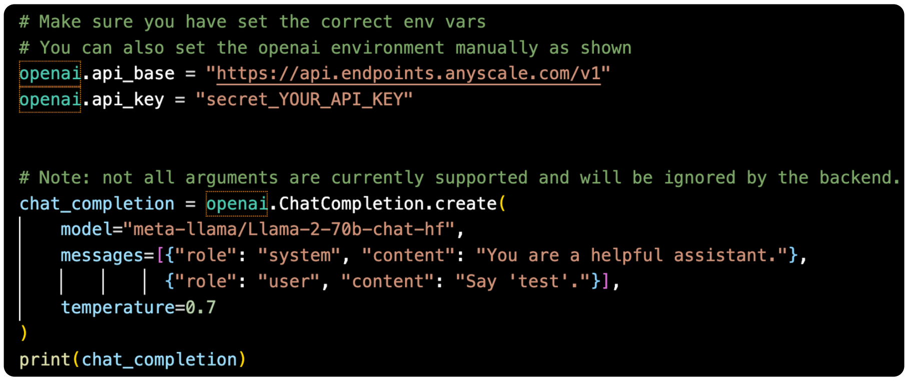
Over the past months, organizations are looking to innovate on the recent progress of foundational models but face several challenges:

- Overreliance on **closed source models** such as ChatGPT.
- **Data privacy** and enterprise **governance** requirements
- **Cost**
- **Lack of flexibility and customization**
- **Production application deployment requirements** - No control over latency, rate limit, or token per minute limitations.

Anyscale Private Endpoints Features

Anyscale Provide Endpoints provides:

- **Fast and Scalable APIs:** Anyscale Endpoints accelerates development cycles by providing a fast and scalable LLM API, enabling developers to iterate and innovate at unprecedented speeds.
- **State-of-the-Art Open LLMs:** Anyscale Endpoints empowers developers with access to SOTA open LLMs, including the acclaimed Llama-2 model, fueling the creation of innovative LLM applications.
- **Managed Platform:** Anyscale Private Endpoints' managed platform architecture eliminates infrastructure management burdens, allowing developers to focus solely on building core business LLM applications securely in their account.
- **OpenAI SDK compatible:** Change the environment variables of your existing applications and start using Anyscale Endpoint.



```python
# Make sure you have set the correct env vars
# You can also set the openai environment manually as shown
openai.api_base = "https://api.endpoints.anyscale.com/v1"
openai.api_key = "secret_YOUR_API_KEY"


# Note: not all arguments are currently supported and will be ignored by the backend.
chat_completion = openai.ChatCompletion.create(
    model="meta-llama/Llama-2-70b-chat-hf",
    messages=[{"role": "system", "content": "You are a helpful assistant."},
              {"role": "user", "content": "Say 'test'."}],
    temperature=0.7
)
print(chat_completion)
```

Fig. Migrate your OpenAI application with minimal code change. (Set the environment variables for Anyscale Private Endpoints)