Open Source LLMs

Exploring Different Access Methods

Venkata Reddy AI Classes

https://www.youtube.com/@VenkataReddyAIClasses/playlists

Contents

- OpenAI
- Cohere
- HuggingFace Hub
- Replicate
- Groq

OpenAI Model - Paid Version

- Get your OpenAI API key here https://platform.openai.com/usage

```python
import os
os.environ['OPENAI_API_KEY'] = "Your own OPENAI_API_KEY"

#Better way
from google.colab import userdata
os.environ['OPENAI_API_KEY'] = userdata.get("OPENAI_API_KEY")
```

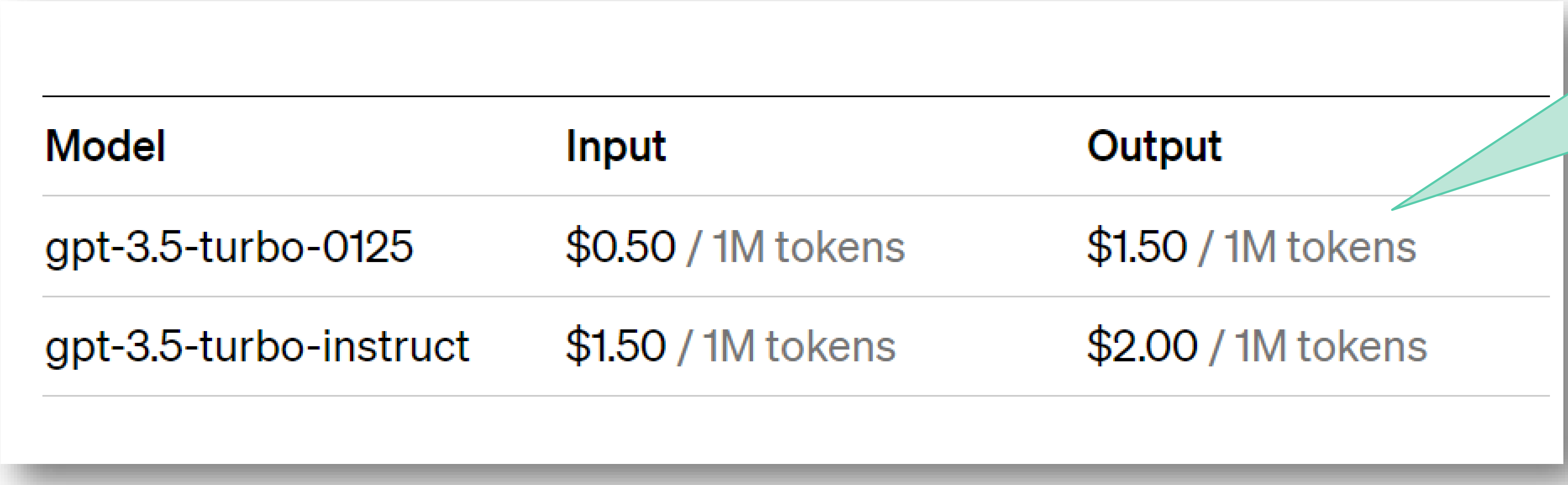
```
from langchain.llms import OpenAI

llm=OpenAI(temperature=0.9, max_tokens=256)
response = llm.invoke("Write a 4 line poem on AI")
print(response)
```

- temperature: Set to 0.9, which controls the randomness of the output.
  - A higher temperature results in more varied and unpredictable outputs,
  - while a lower temperature produces more deterministic and conservative outputs.
  - This is often used in generative tasks to balance between creativity and relevance.
- max_tokens: Set to 256, which specifies the maximum number of tokens (words or pieces of words) that the model can generate in a single response.

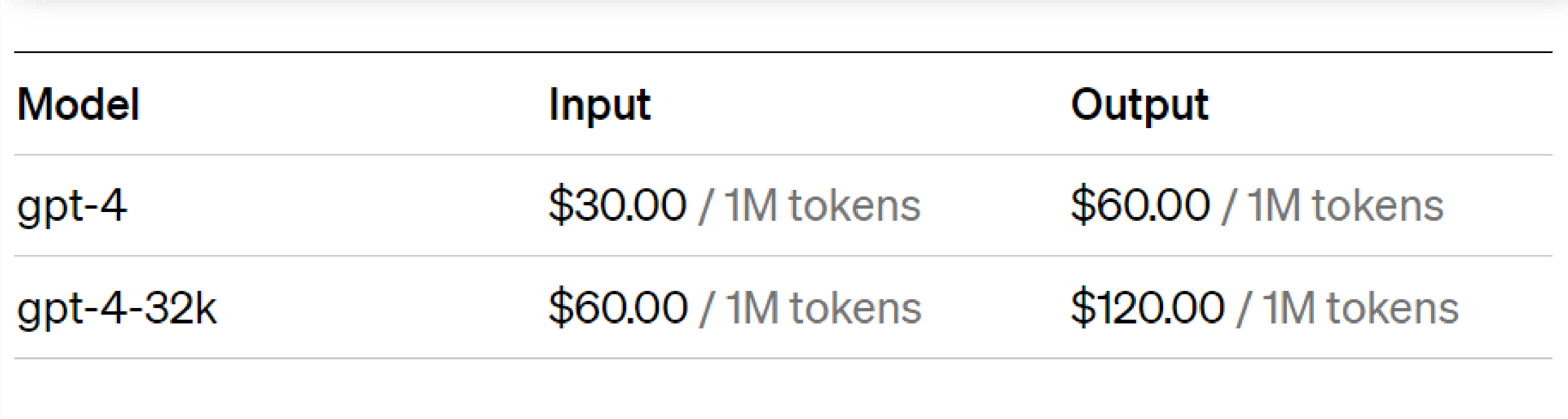
```
llm=OpenAI(temperature=0)
response = llm.invoke("What is overfitting in Machine Learning? Explain it
to a layman")
print(response)
```

OpenAI API is NOT Free



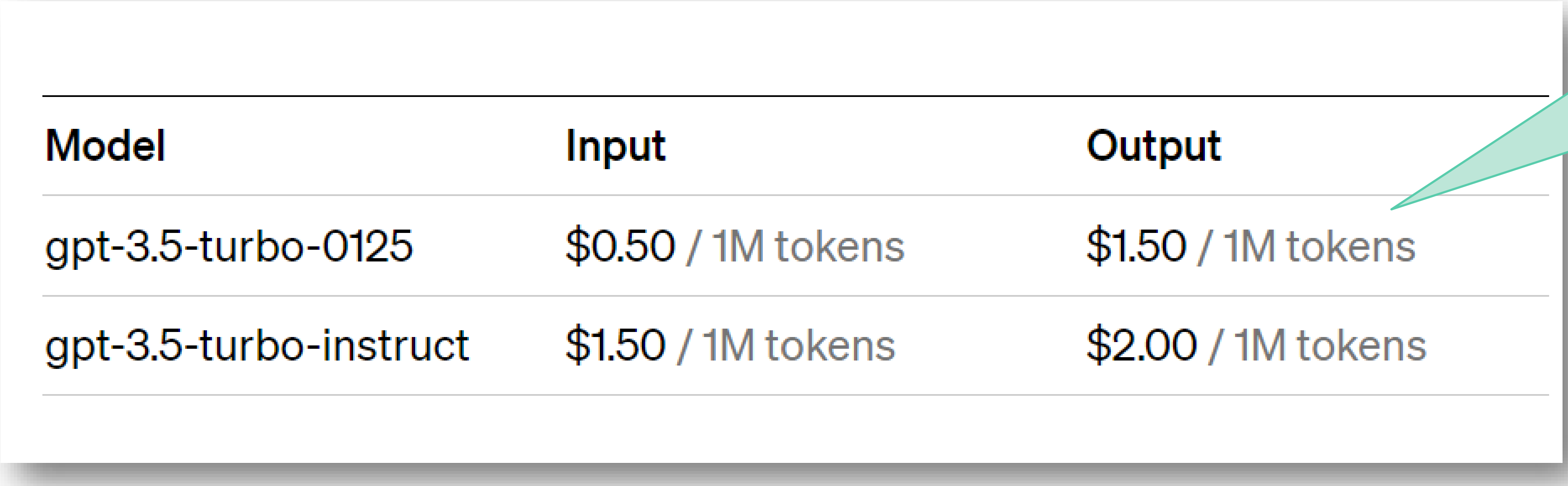
Model	Input	Output
gpt-3.5-turbo-0125	$0.50 / 1M tokens	$1.50 / 1M tokens
gpt-3.5-turbo-instruct	$1.50 / 1M tokens	$2.00 / 1M tokens

This is sufficient to explore all the concepts and applications around LLMs and GenAI



Model	Input	Output
gpt-4	$30.00 / 1M tokens	$60.00 / 1M tokens
gpt-4-32k	$60.00 / 1M tokens	$120.00 / 1M tokens

Is OpenAI costly?



Model	Input	Output
gpt-3.5-turbo-0125	$0.50 / 1M tokens	$1.50 / 1M tokens
gpt-3.5-turbo-instruct	$1.50 / 1M tokens	$2.00 / 1M tokens

This is sufficient to explore all the concepts and applications around LLMs and GenAI

- With 10 dollars we can play with nearly 15 million input and output tokens.
- Imagine if an average question(interaction) has 1000 tokens, then we can interact 15,000 times. –It is not very expensive

Cohere

- Get your Cohere Trail API key here https://dashboard.cohere.com/api-keys

```
os.environ['COHERE_API_KEY'] = "Your own COHERE_API_KEY"
#Better way
os.environ['COHERE_API_KEY'] = userdata.get("COHERE_API_KEY")

from langchain.llms import Cohere
llm = Cohere(temperature=0.9, max_tokens=256)
response = llm.invoke("Write a 4 line poem on AI")
print(response)

llm=Cohere(temperature=0)
response = llm.invoke("What is overfitting in Machine Learning? Explain it to a layman")
print(response)
```

Open source models

- Mistral Model (Mistral 7B, Mixtral8-7B)
- LLama (Llam2, Llama3)
- Bloom by Hugging Face
- Falcon 180B
- Opt 175B
- Xgen-7B
- Vicuna-13B

HuggingFace models

https://huggingface.co/mistralai

```
os.environ['HUGGINGFACEHUB_API_TOKEN'] = "Your own HUGGINGFACEHUB_API_TOKEN"
#Better way
os.environ['HUGGINGFACEHUB_API_TOKEN'] = userdata.get("HUGGINGFACEHUB_API_TOKEN")

from langchain.llms import HuggingFaceHub

repo_id="mistralai/Mistral-7B-Instruct-v0.2"

llm = HuggingFaceHub(
    repo_id=repo_id,
    model_kwargs={"temperature": 0.9, "max_length": 256},
)

response = llm.invoke("Write a 4 line poem on AI")
print(response)
```

Mistral AI models

```
repo_id="mistralai/Mistral-7B-Instruct-v0.2"

llm = HuggingFaceHub(
    repo_id=repo_id,
    model_kwargs={"temperature": 0.3, "max_length": 1000},
)

response = llm.invoke("How to pick a stock based on Revenue, Profit and
profit margin trends?")
print(response)
```

Llama from Hugging Facehub

- Llama from Hugging Facehub https://huggingface.co/meta-llama
- You need to fill the contact info and wait for the approval. https://huggingface.co/meta-llama/Meta-Llama-3.1-8B

```
repo_id="meta-llama/Meta-Llama-3.1-8B"

llm = HuggingFaceHub(
    repo_id=repo_id,
    model_kwargs={"temperature": 0.9},
)

response = llm.invoke("What are some ways to boost creativity?")
print(response)
```

#Throws an error
The model meta-llama/Meta-Llama-3.1-8B is too large to be loaded automatically (16GB > 10GB). Please use Spaces (https://huggingface.co/spaces) or Inference Endpoints (https://huggingface.co/inference-endpoints).

Replicate

- Run and fine-tune open-source models with Replicate's API. https://replicate.com/home
- Deploy custom models at scale using one line of code.
- Avoid managing infrastructure or learning machine learning details.
- Use open-source models or package your own.
- Choose to make models public or keep them private.
- Start with any open-source model with just one line of code.
- Replciate API Token
  - On top Left >>> Home>>Click on your id>> API Tokens https://replicate.com/account/api-tokens

```
!pip install replicate

os.environ["REPLICATE_API_TOKEN"] = userdata.get("REPLICATE_API_TOKEN")


from langchain.llms import Replicate

replicate_llm = Replicate(
    model="meta/meta-llama-3.1-405b-instruct",
    model_kwargs={"temperature": 0.6},
)

response = replicate_llm.invoke("What are some good strategies for
studying?")
print(response)
```

Groq

- https://groq.com/
- Developed the LPU(Language Processing Unit) chip to run LLMs faster and cheaper.
- LPU delivers fast, affordable, and energy-efficient AI solutions.
- Offers Groq Cloud to try open-source LLMs like Llama3 or Mixtral.
- Allows free use of Llama3 or Mixtral in apps via Groq API Key with rate limits.
- Models on Groq https://console.groq.com/docs/models
- Get your Groq API key https://console.groq.com/keys

```
!pip install langchain-groq
```

```
os.environ["GROQ_API_KEY"] = userdata.get("GROQ_API_KEY")
```

```
from langchain_groq import ChatGroq
llm=ChatGroq(
    model="llama3-70b-8192"
)
result=llm.invoke("what are the top 10 quotes about ignorance?")
print(result)
```

Many more ways

https://python.langchain.com/v0.1/docs/integrations/llms/

Thank you