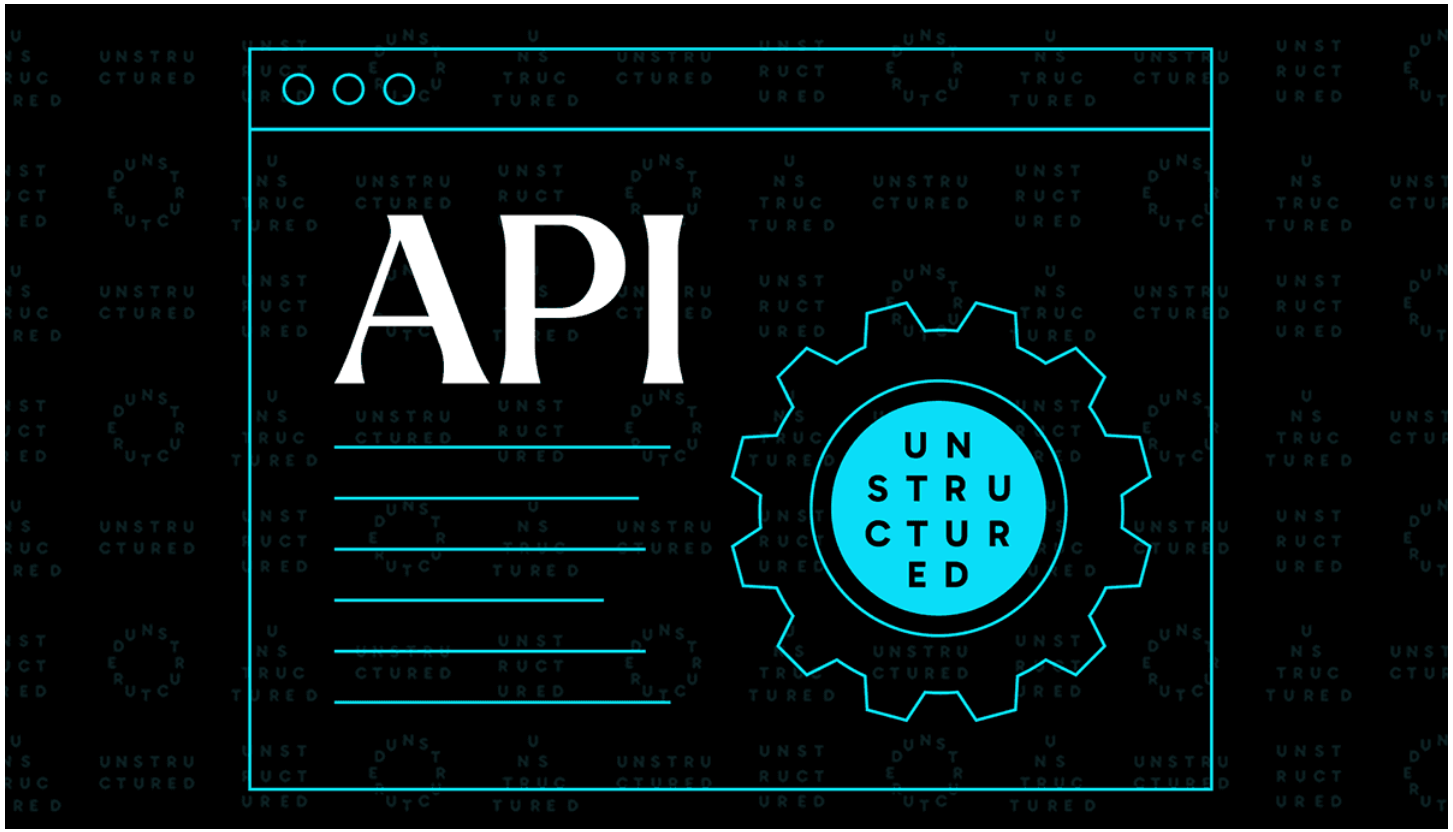
LLM API



박경규

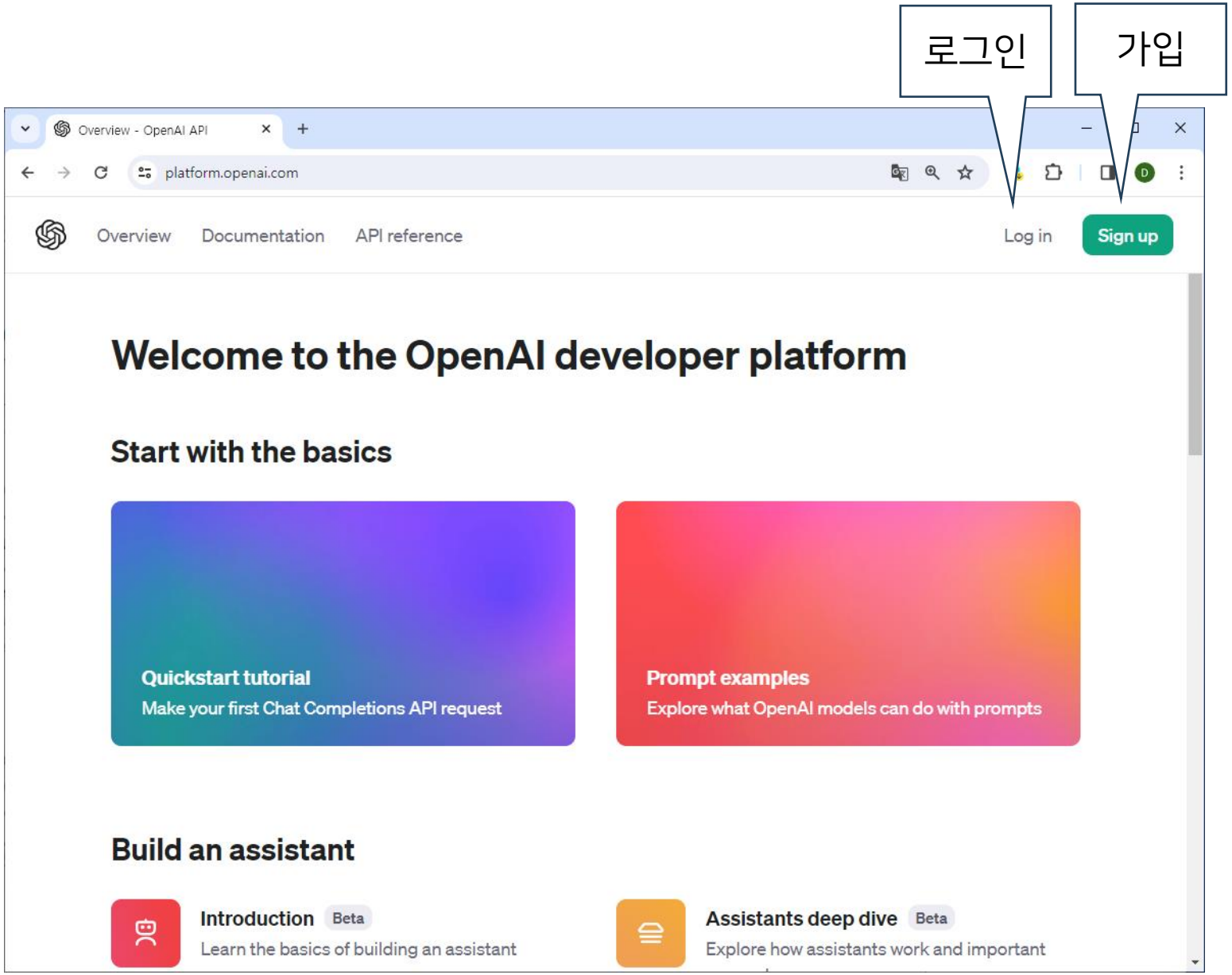
OpenAI API

OpenAI

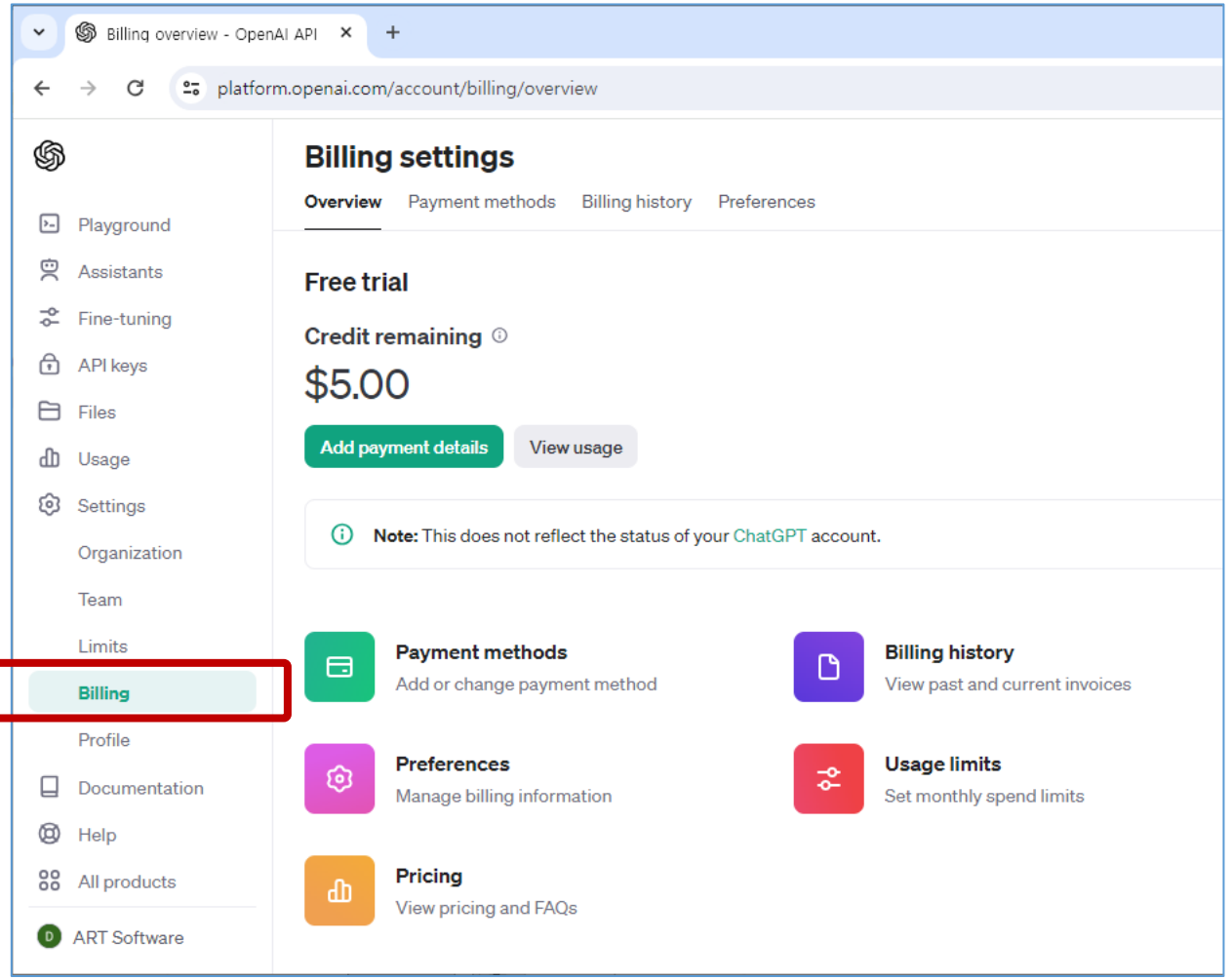
OpenAI

https://platform.openai.com/

사이트 접속 및 회원가입



OpenAI API 무료사용 https://platform.openai.com/account/billing/overview



Rate limits

MODEL	TOKEN LIMITS	REQUEST AND OTHER LIMITS
gpt-3.5-turbo : LLM	40,000 TPM	3 RPM 200 RPD
text-embedding-3-small	150,000 TPM	3 RPM 200 RPD
dall-e-3 : Text to Image		3 RPM 200 RPD
tts-1 : Text to Speech		3 RPM 200 RPD
whisper-1 : Automatic Speech Recognition		3 RPM 200 RPD

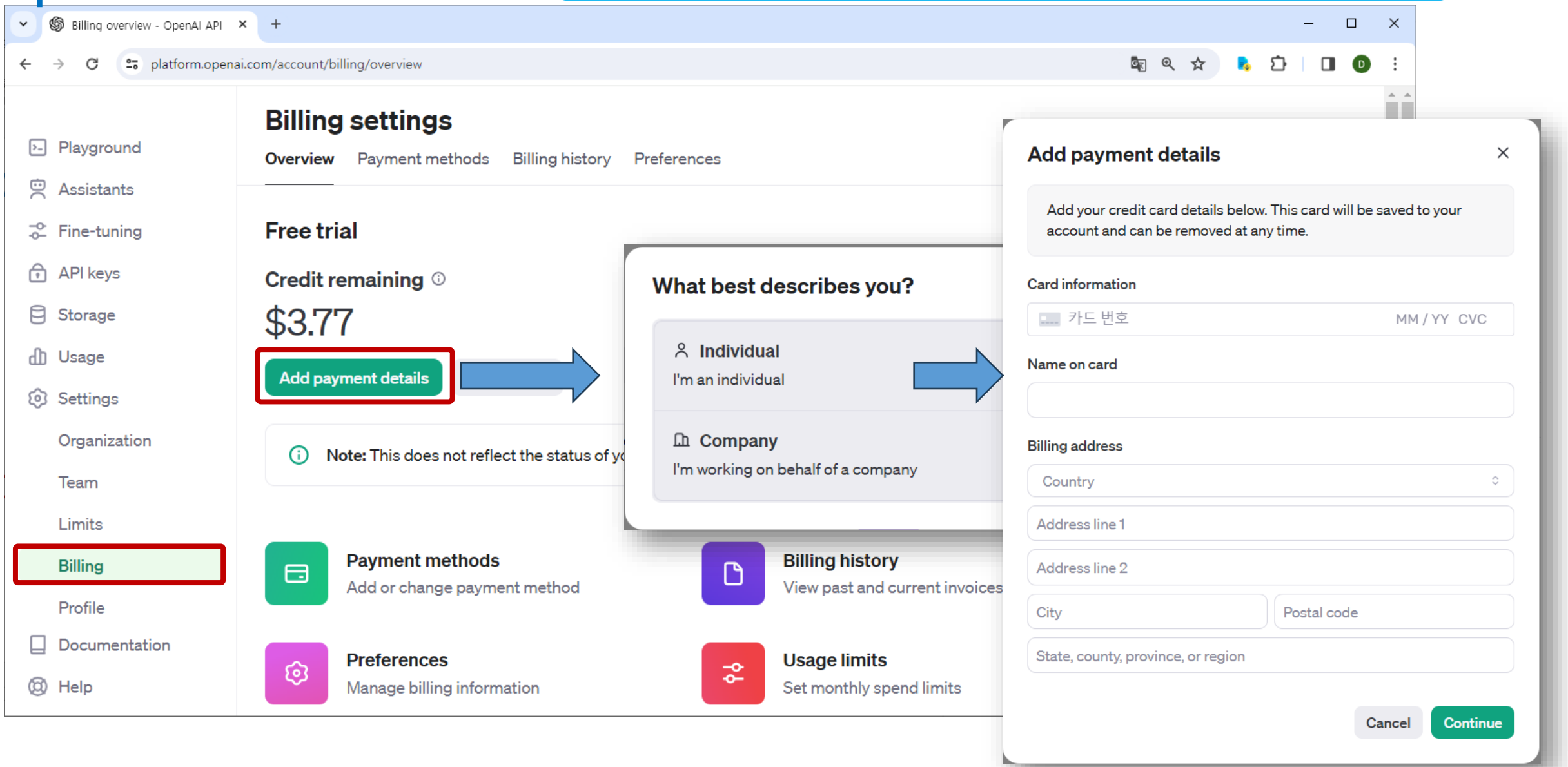
- TPM (tokens per minute)
- TPD (tokens per day)
- RPM (requests per minute)
- RPD (requests per day)
- IPM (images per minute)

- 1 token ~= 4 chars in English
- 1 token ~= ¾ words
- 100 tokens ~= 75 words

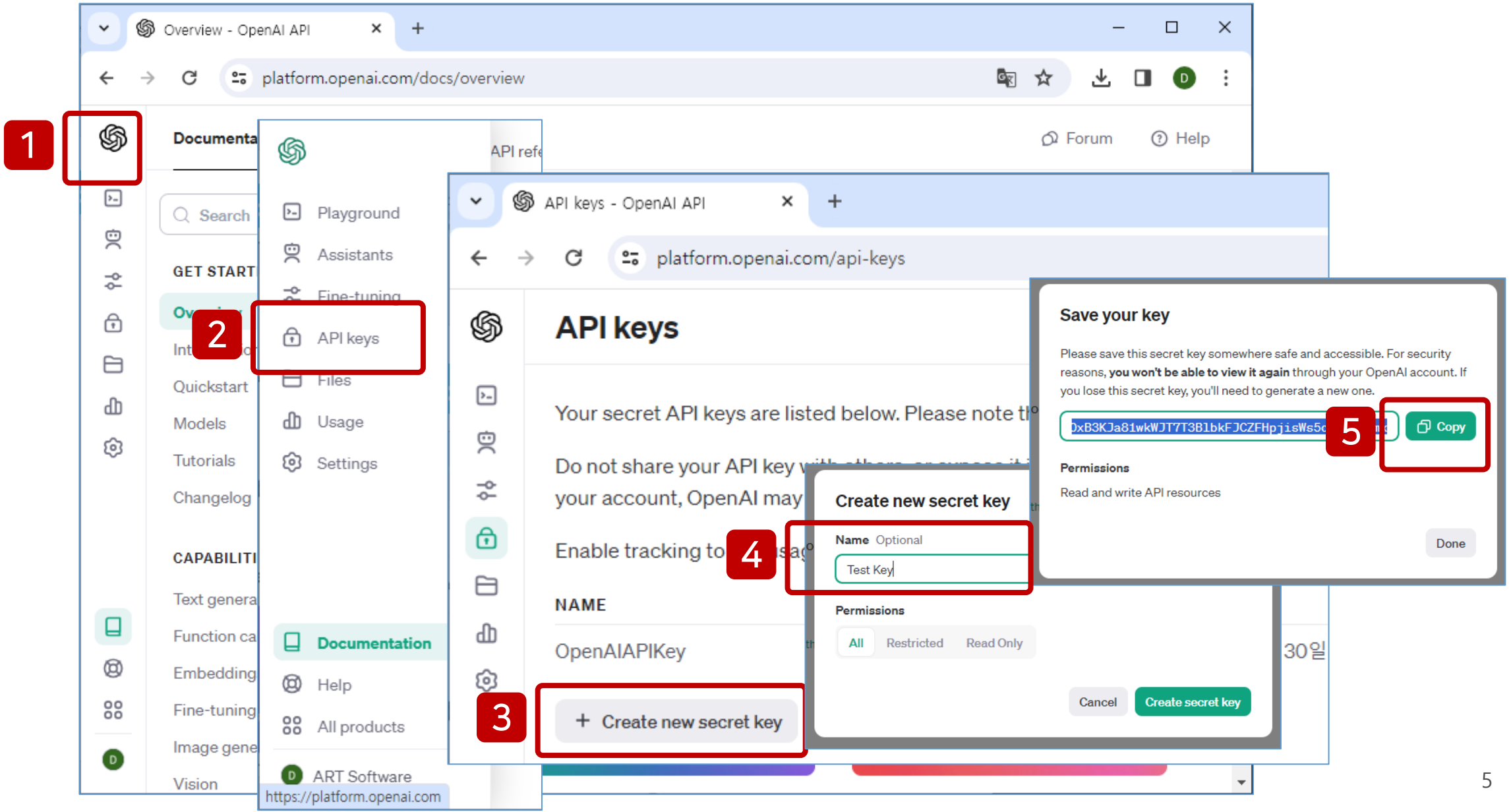
참고 : https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

OpenAI API 유료사용

https://platform.openai.com/account/billing/overview



OpenAI API Key 생성



OpenAI 요금제 https://openai.com/pricing

GPT

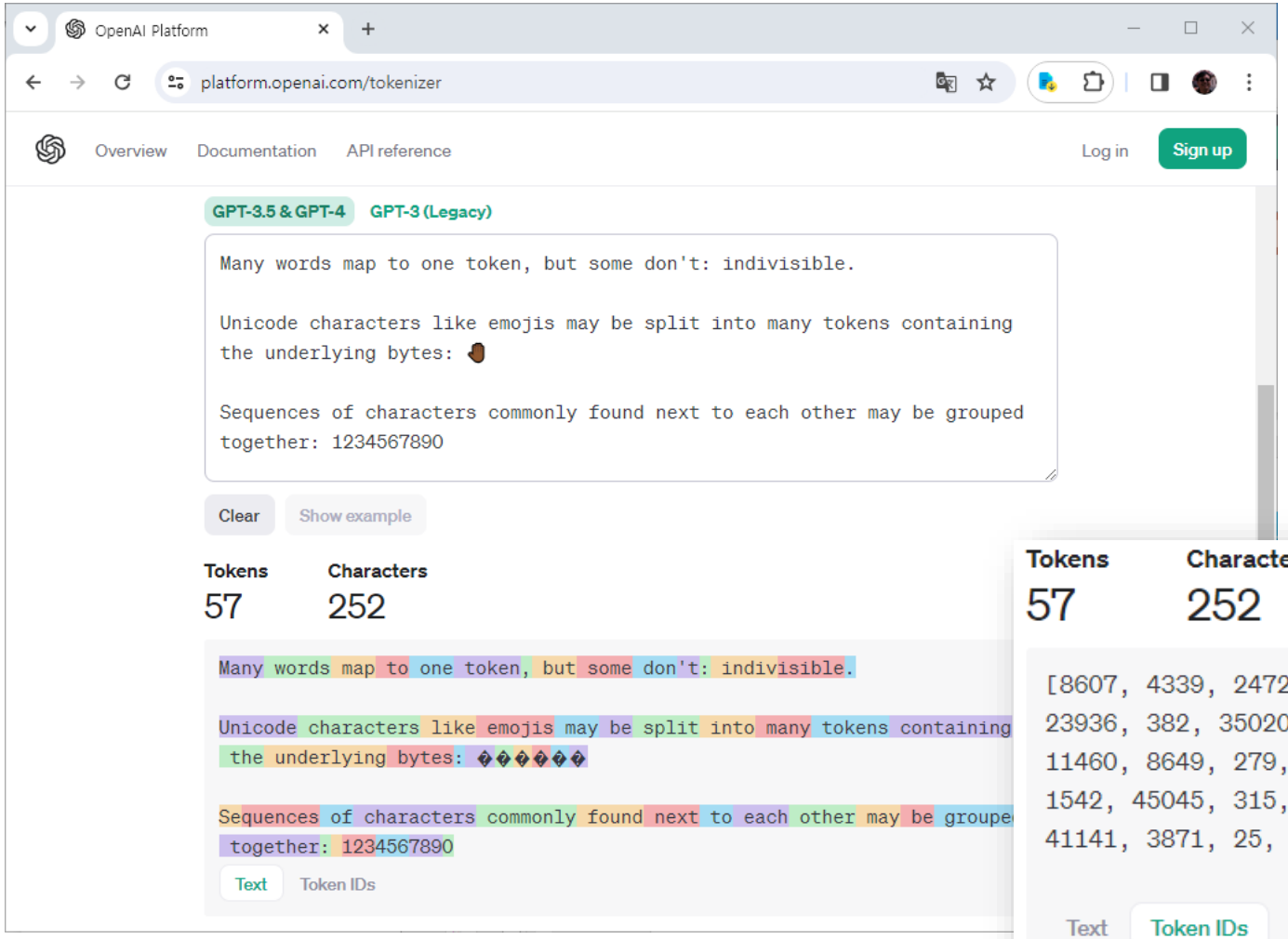
Model	Input	Output
gpt-4-0125-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-4-1106-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-4-1106-vision-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-4	$0.03 / 1K tokens	$0.06 / 1K tokens
gpt-4-32k	$0.06 / 1K tokens	$0.12 / 1K tokens
gpt-3.5-turbo-0125	$0.0005 / 1K tokens	$0.0015 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

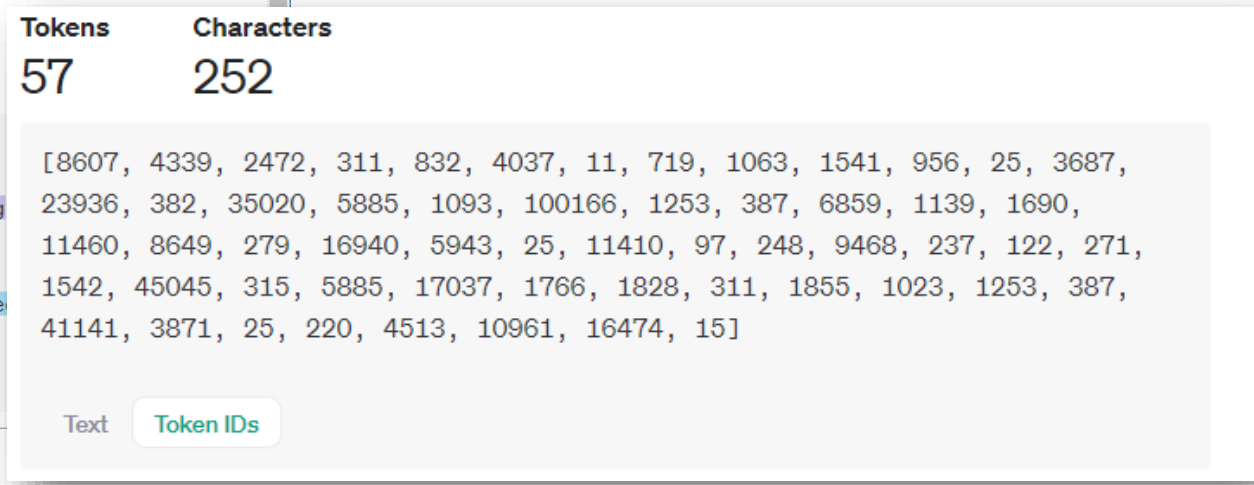
Embedding models

Model	Usage
text-embedding-3-small	$0.00002 / 1K tokens
text-embedding-3-large	$0.00013 / 1K tokens
ada v2	$0.00010 / 1K tokens

토큰(Token)

https://platform.openai.com/tokenizer





tiktoken

토큰 한도(Token limit)

https://platform.openai.com/docs/guides/text-generation/managing-tokens

총 토큰 수는 API 호출에 영향을 줌

- 모델의 최대 한도 미만이어야 함, gpt-3.5-turbo 토큰 한도 4,096
  총 토큰 수 : 입력 토큰 + 출력 토큰
- 토큰당 지불하는 API 호출 비용
- API 호출에 걸리는 시간

Counting tokens for chat API calls

```python
def num_tokens_from_messages(messages, model="gpt-3.5-turbo-0613"):
    """Returns the number of tokens used by a list of messages."""
    try:
        encoding = tiktoken.encoding_for_model(model)
    except KeyError:
        encoding = tiktoken.get_encoding("cl100k_base")
    if model == "gpt-3.5-turbo-0613":  # note: future models may deviate fr
        num_tokens = 0
        for message in messages:
            num_tokens += 4  # every message follows <im_start>{role/name}\
            for key, value in message.items():
                num_tokens += len(encoding.encode(value))
                if key == "name":  # if there's a name, the role is omitted
                    num_tokens += -1  # role is always required and always
        num_tokens += 2  # every reply is primed with <im_start>assistant
        return num_tokens
    else:
        raise NotImplementedError(f"""num_tokens_from_messages() is not pre
    See https://github.com/openai/openai-python/blob/main/chatml.md for inf
```

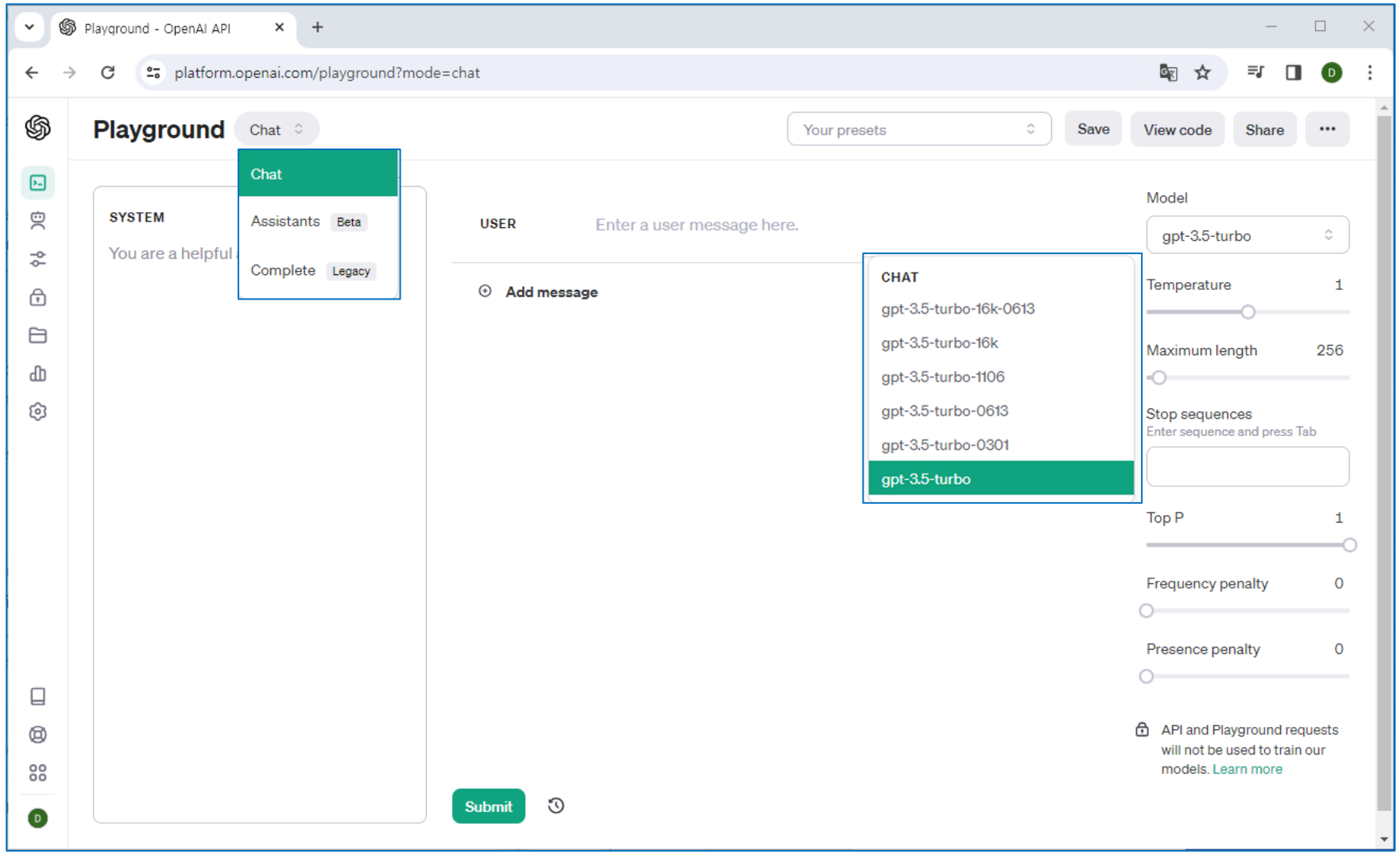
```python
messages = [
    {"role": "system", "content": "You are a helpful, pattern-following
    {"role": "system", "name":"example_user", "content": "New synergies wil
    {"role": "system", "name": "example_assistant", "content": "Things work
    {"role": "system", "name":"example_user", "content": "Let's circle back
    {"role": "system", "name": "example_assistant", "content": "Let's talk
    {"role": "user", "content": "This late pivot means we don't have time t
]

model = "gpt-3.5-turbo-0613"

print(f"{num_tokens_from_messages(messages, model)} prompt tokens counted
# Should show ~126 total_tokens
```

플레이그라운드

https://platform.openai.com/playground



- **Temperature : 값이 낮을수록 가장 높은 확률의 다음 토큰을 선택하고, 높아지면 무작위성이 높아짐**
- Max Length : 모델이 생성하는 토큰 최대 길이
- Stop Sequences : 모델의 토큰 생성을 중지하는 문자열
- Top P : 값이 높으면 모델이 가능성이 낮은 단어를 포함하여 더 다양한 출력을 얻을 수 있음
- Frequency Penalty : 해당 토큰이 나타난 횟수에 비례하여 페널티 적용
- Presence Penalty : 모든 반복 토큰에 동일한 페널티 적용(2번 나타나는 토큰과 10번 나타나는 토큰 모두 동일한 페널티)

※ Temperature 와 Top_p, 그리고 Frequency Penalty와 Presence Penalty 동시 변경은 비권장함

API 사용 방법

Step 1: Setup Python

Install Python

https://www.python.org/downloads/

Setup a virtual environment (optional)

python -m venv myenv

Windows : myenv₩Scripts₩activate

Unix or Mac : source myenv/bin/activate

Install the OpenAI Python library

pip install --upgrade openai

Step 2: Setup your API key

Windows : setx OPENAI_API_KEY "your-api-key-here"
Unix or Mac : export OPENAI_API_KEY='your-api-key-here”

Step 3: Sending your first API request

```python
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explaining
    {"role": "user", "content": "Compose a poem that explains the concept of recursi
  ]
)

print(completion.choices[0].message)
```

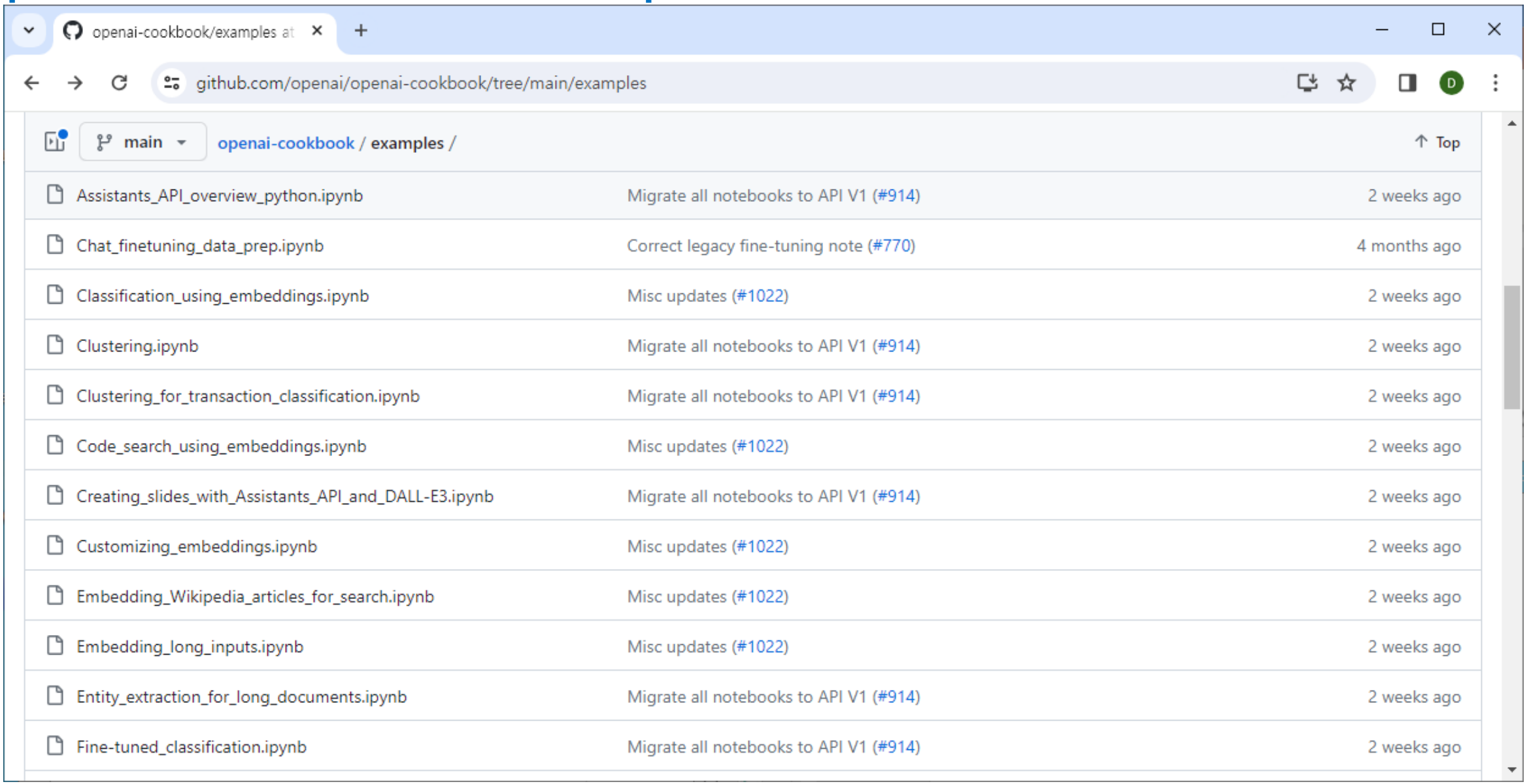
https://platform.openai.com/docs/quickstart?ref=seongjin.me&context=python

OpenAI API 실습

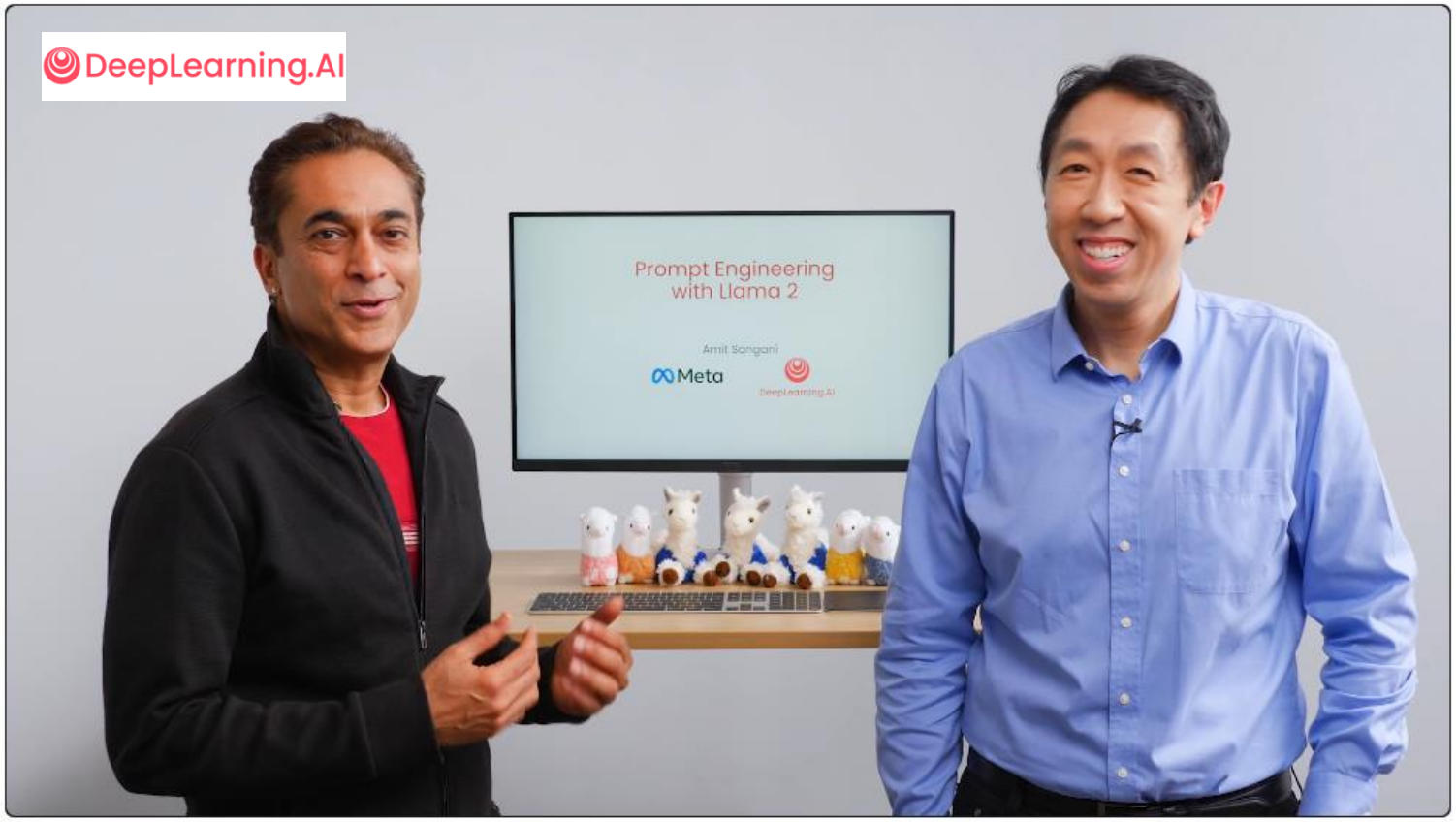
- openai_api.ipynb
- How_to_count_tokens_with_tiktoken.ipynb GitHub
- chunking.ipynb
- information_retrieval.ipynb
- ReAct.ipynb
- pe-lecture.ipynb colab

OpenAI Cookbook examples

https://github.com/openai/openai-cookbook/tree/main/examples



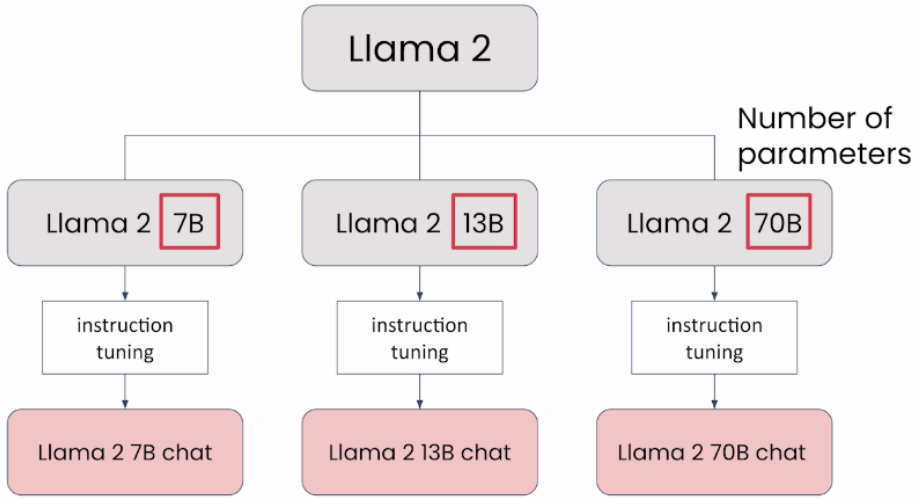
Llama 2 프롬프트 엔지니어링



https://learn.deeplearning.ai/courses/prompt-engineering-with-llama-2/

Llama Model

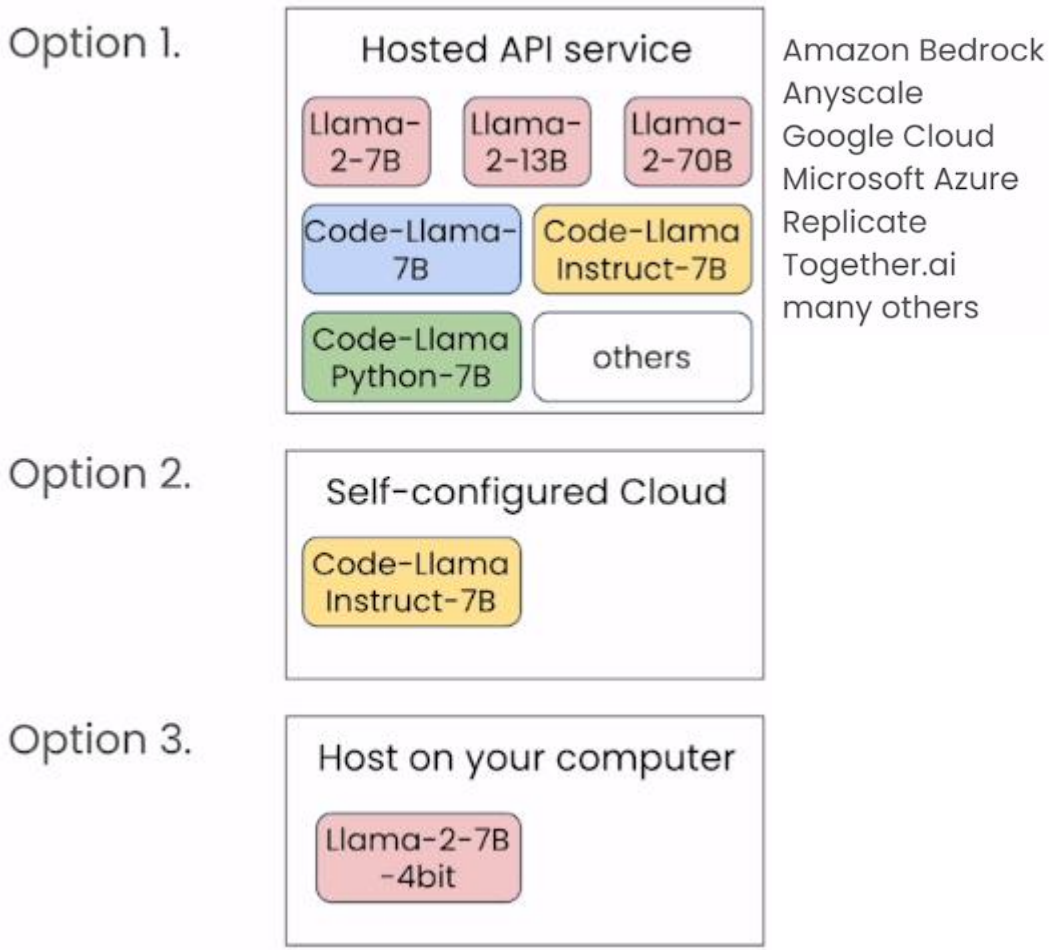
■ 모델 종류



■ 모델 성능

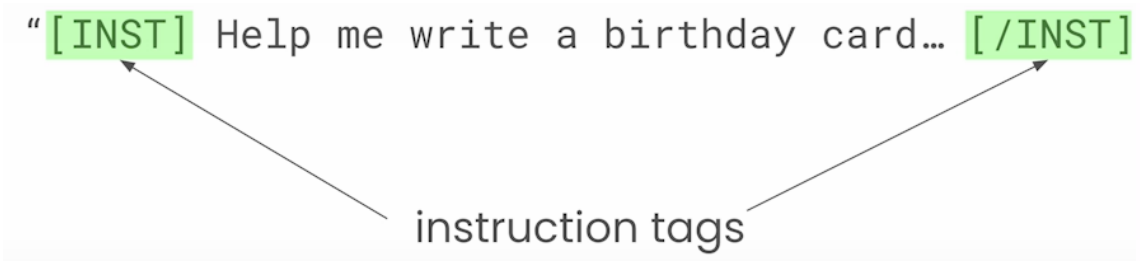
	Llama 2	Falcon 40B	GPT 3.5
Performance (MMLU benchmark)	68.9	55.4	70.0
Access / Privacy	Free to download	Free to download	Access via OpenAI

■ 사용 방법



https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

Multi-Turn

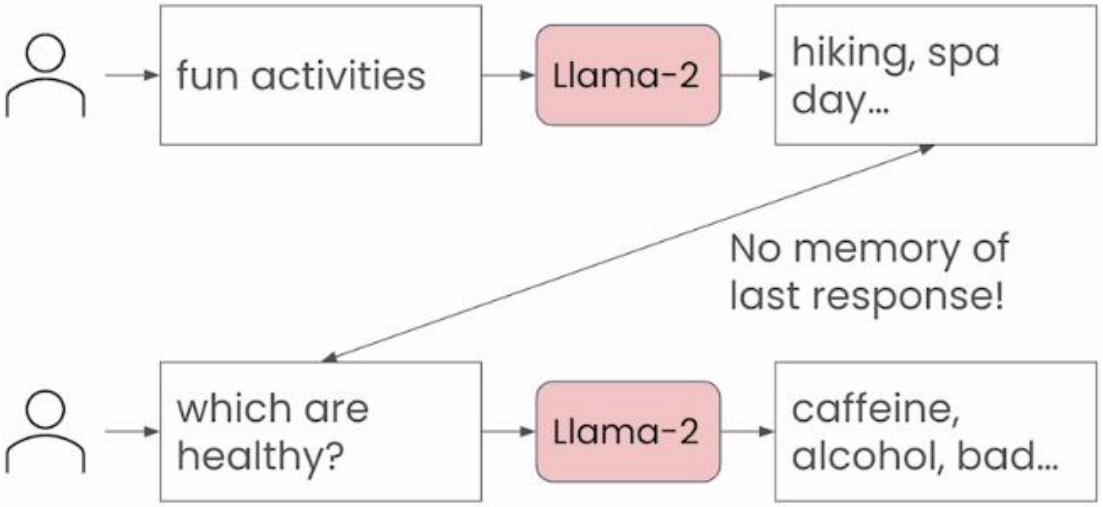


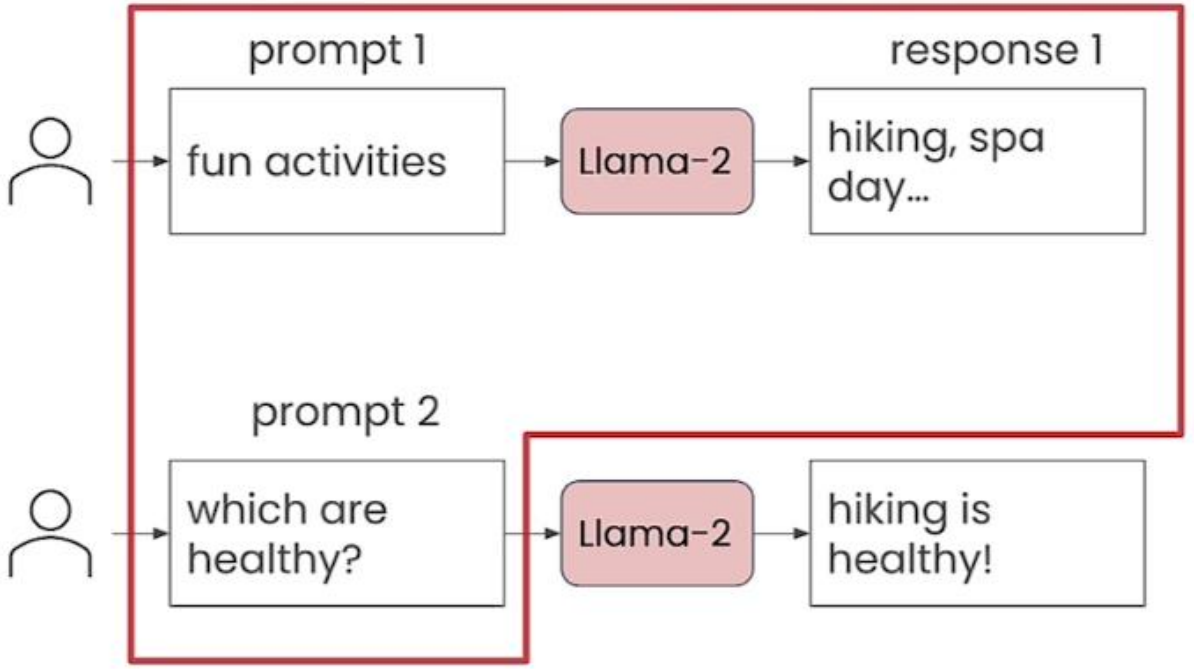
start tags

```
prompt_chat = f"""
    <s>[INST]{user prompt 1}[/INST]
    Assistant: {model response 1}</s>
    <s>[INST]{user prompt 2}[/INST]
    Assistant: {model response 2}</s>
    …
    <s>[INST]{user prompt 3}[/INST]
```

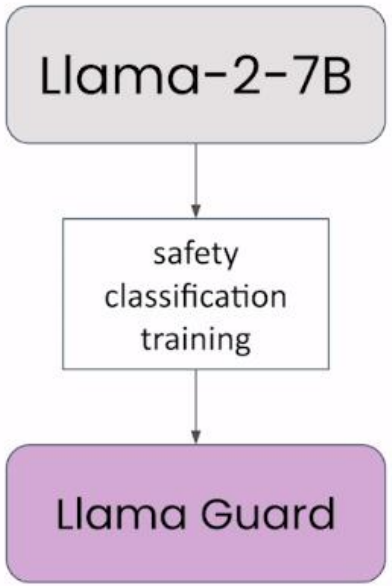
end tags

LLMs are stateless

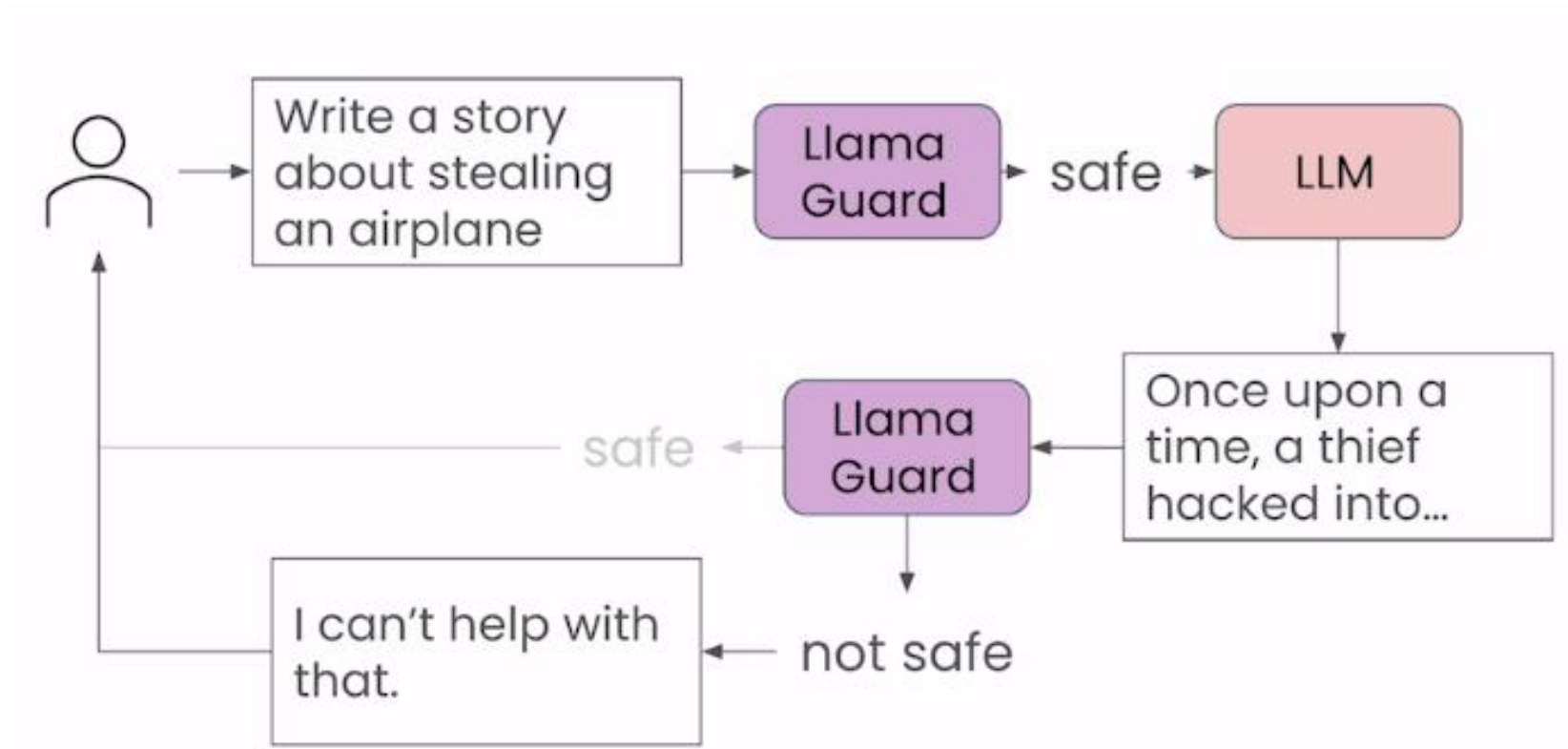




Llama Guard


Llama-2-7B
safety classification training
Llama Guard


GUARD

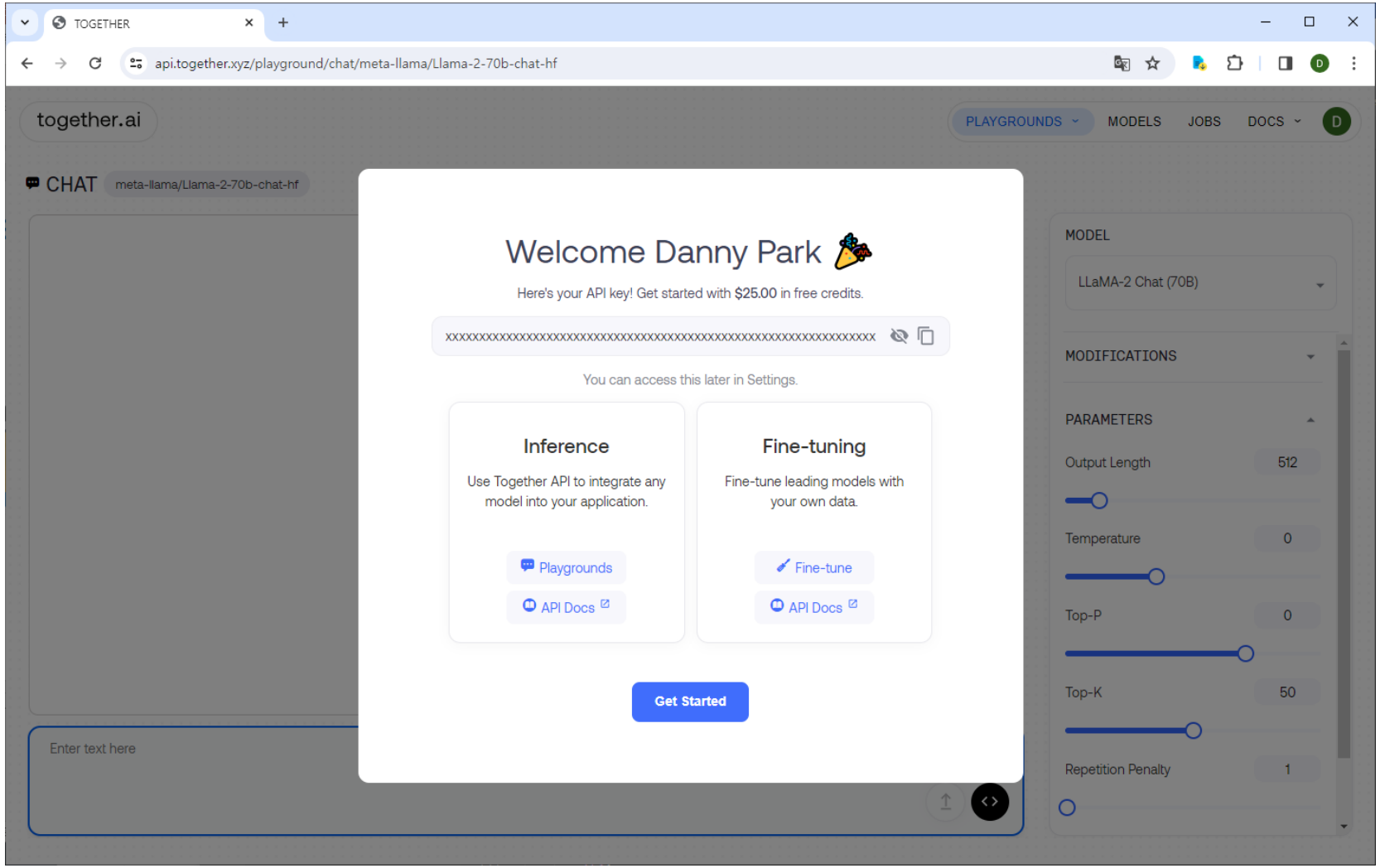

Write a story about stealing an airplane
Llama Guard
safe
LLM
Once upon a time, a thief hacked into...
Llama Guard
safe
not safe
I can't help with that.

Llama 실습 : Together API Key

https://www.together.ai/

- 사이트 접속 및 회원가입
- 환경변수에

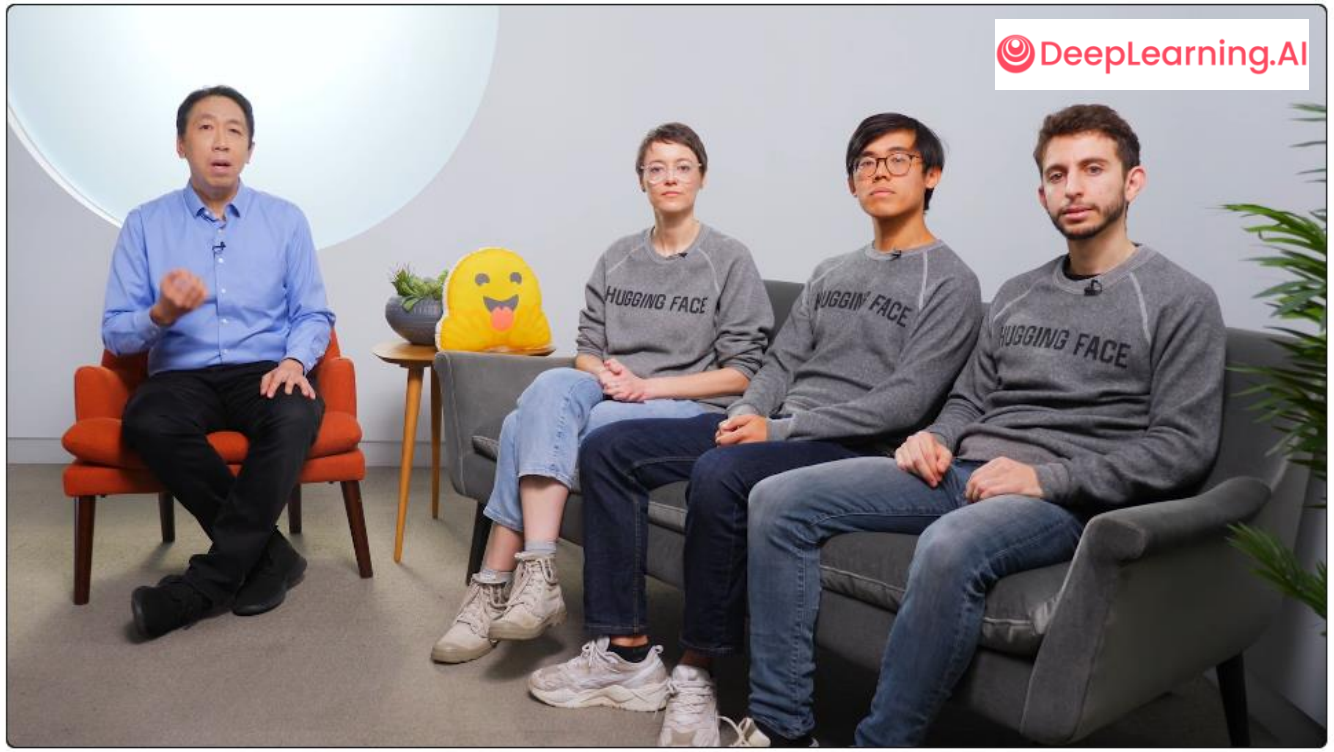
  TOGETHER_API_KEY 값 추가
- ./code/llama/utils.py 파일 참고



Llama 실습

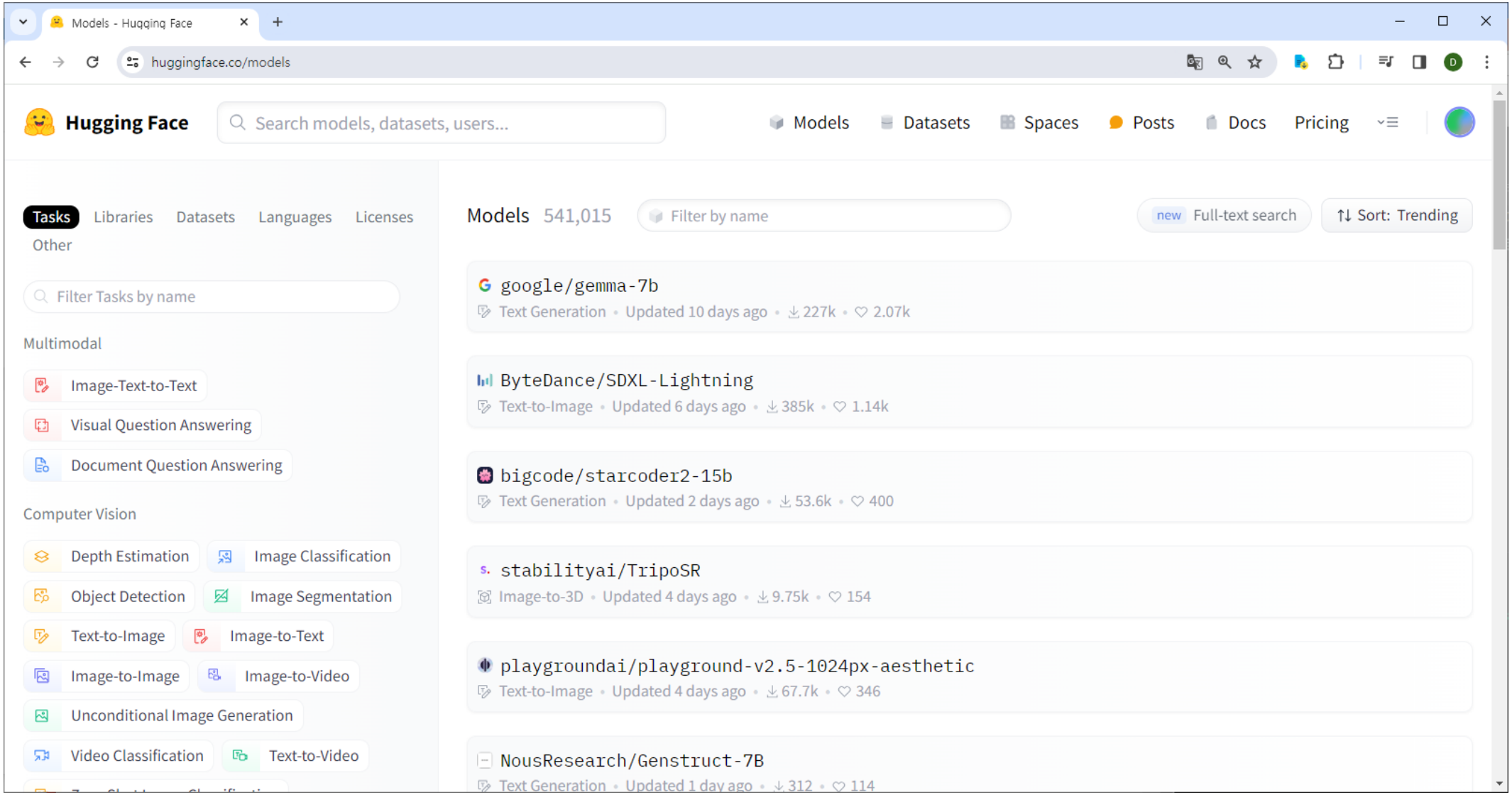
- L2_getting_started.ipynb
- L3_multi_turn_conversations.ipynb
- L4_prompt_engineering_techniques.ipynb
- L5_comparing_llama_models.ipynb
- L6_code_llama.ipynb
- L7_llama_guard.ipynb

허깅페이스(Hugging Face)
오픈소스 모델 사용



https://learn.deeplearning.ai/courses/open-source-models-hugging-face/

모델 선택



https://huggingface.co/models

모델 선택

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Feature Extraction
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Audio

- Text-to-Speech
- Text-to-Audio
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction

Tabular

- Tabular Classification
- Tabular Regression

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering

Reinforcement Learning

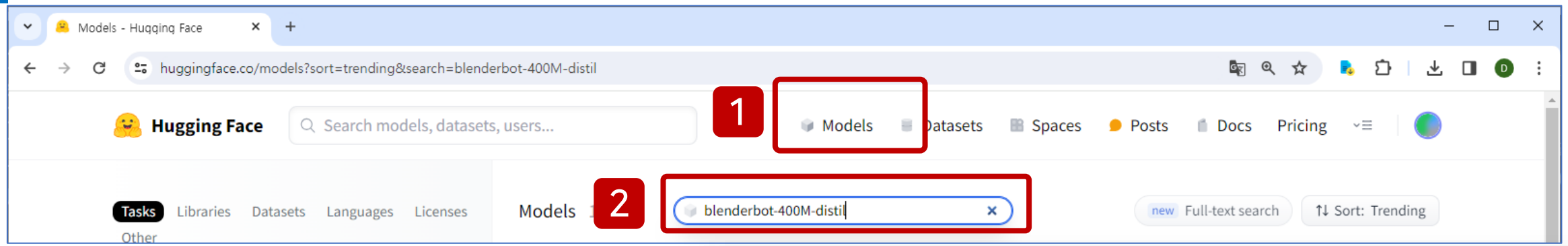
- Reinforcement Learning
- Robotics

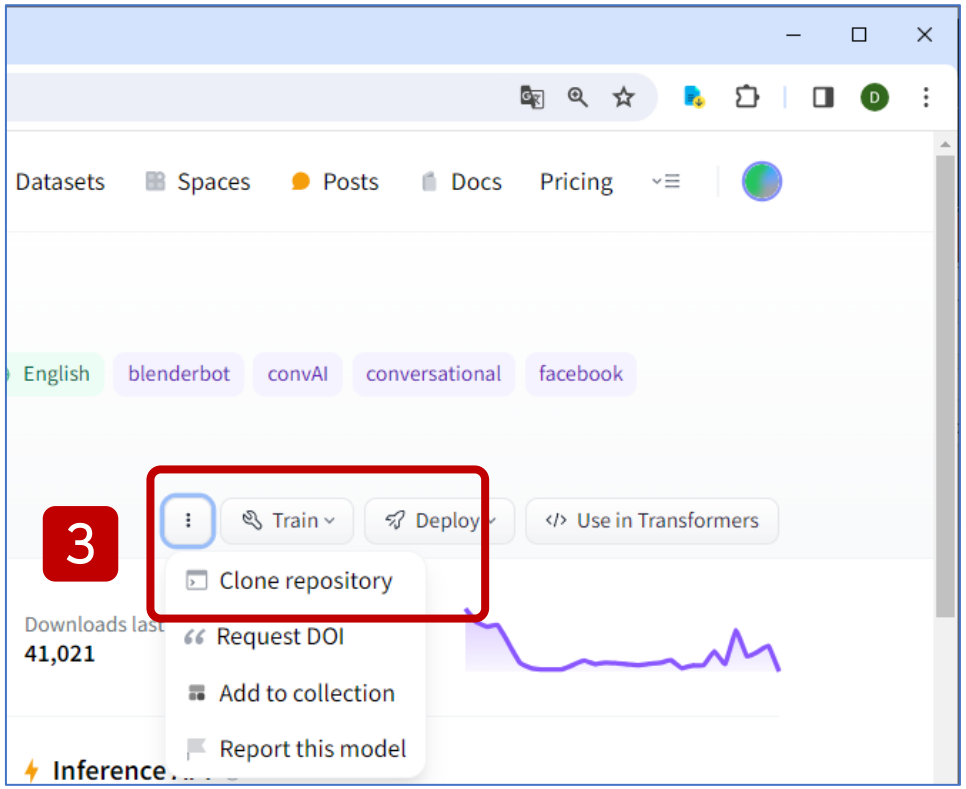
Other

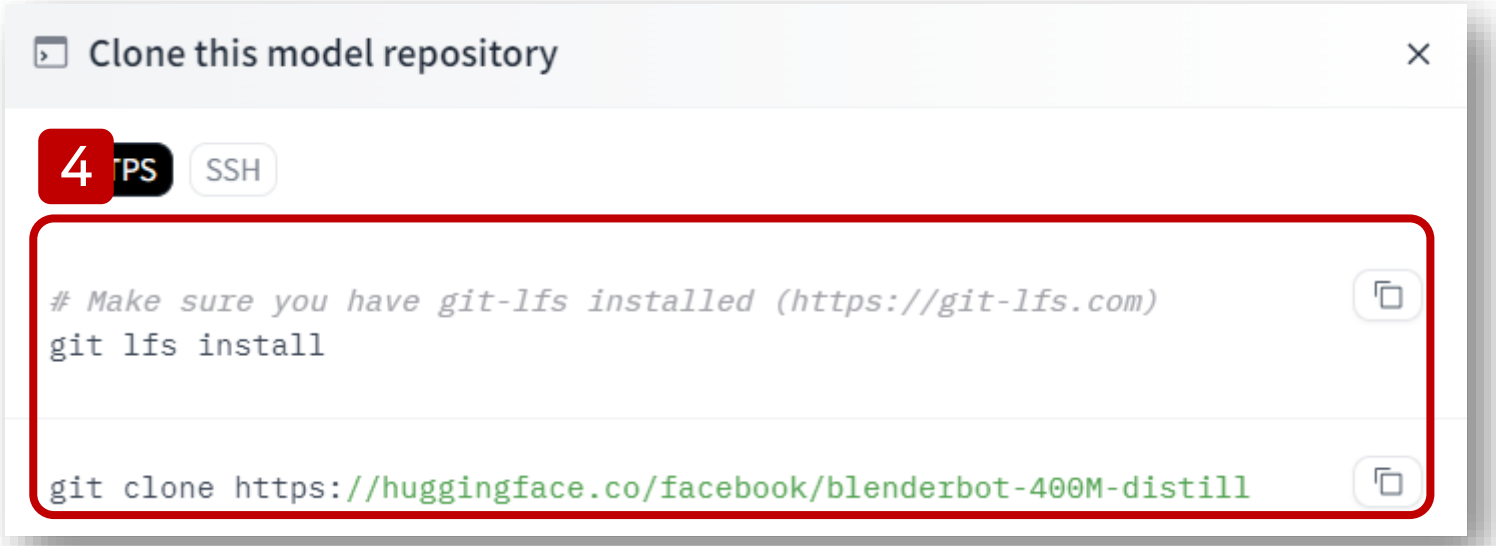
- Graph Machine Learning

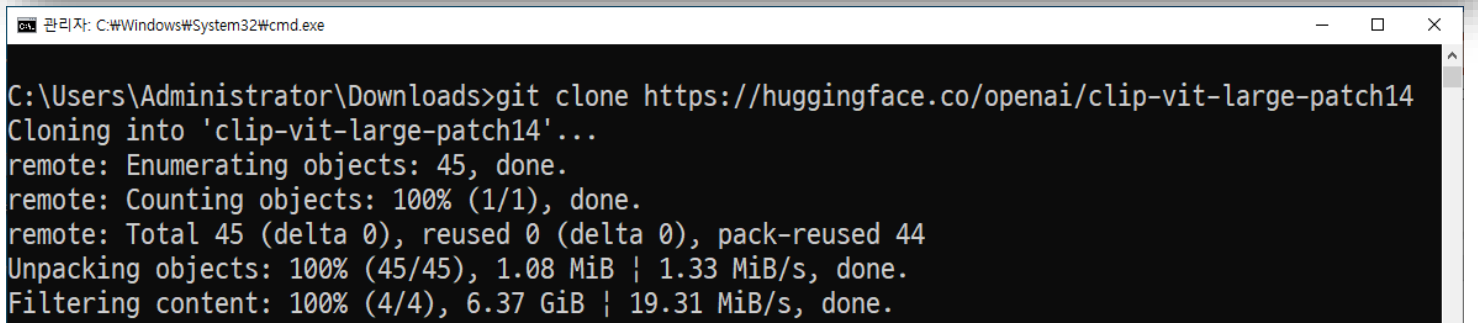
https://huggingface.co/models

모델 다운로드

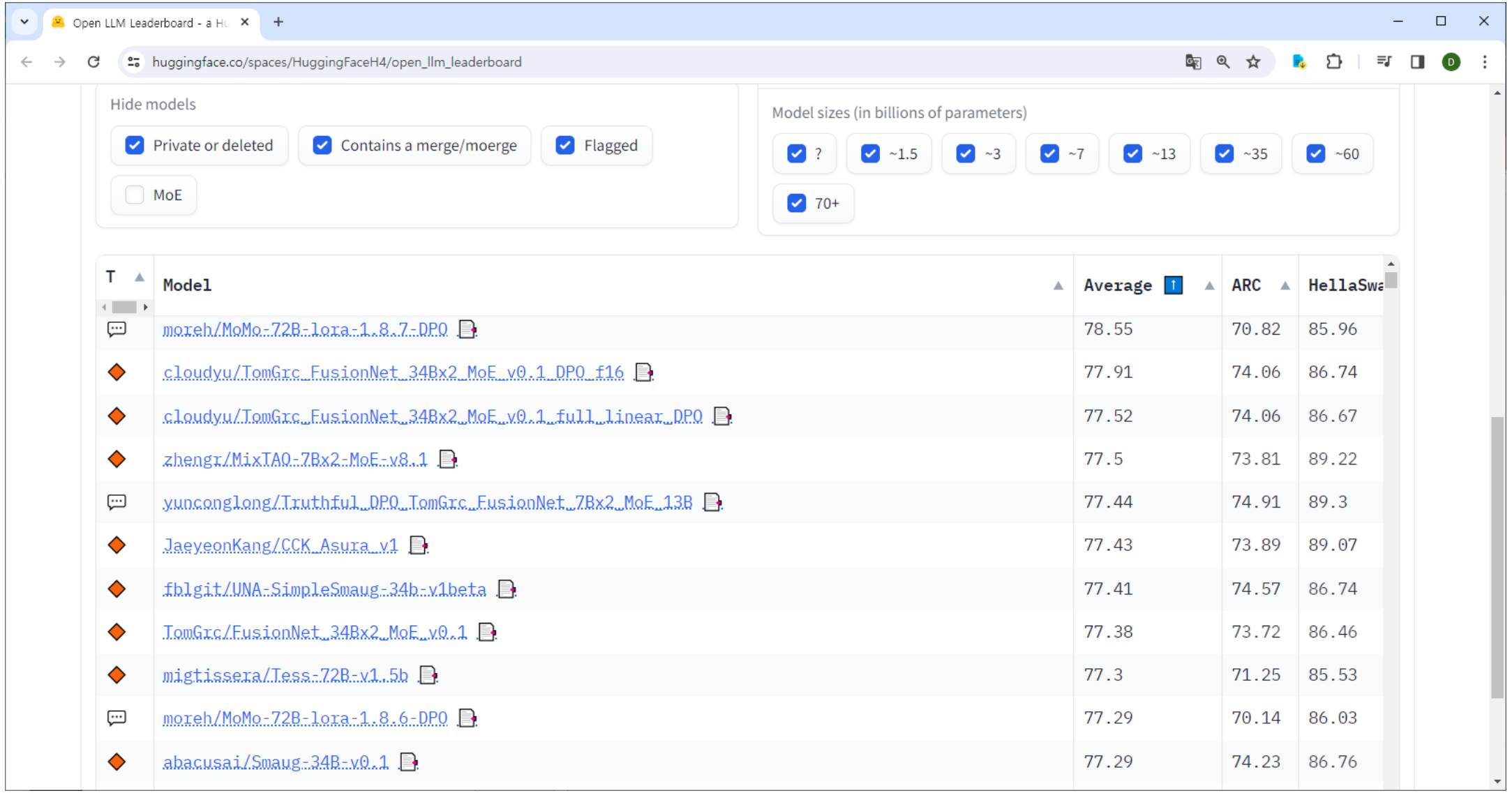

Models - Hugging Face
huggingface.co/models?sort=trending&search=blenderbot-400M-distil
Hugging Face
Search models, datasets, users...
1
Models
Datasets
Spaces
Posts
Docs
Pricing
Tasks
Libraries
Datasets
Languages
Licenses
Other
Models
2
blenderbot-400M-distil
new Full-text search
Sort: Trending


Datasets
Spaces
Posts
Docs
Pricing
English
blenderbot
convAI
conversational
facebook
3
Train
Deploy
Use in Transformers
Clone repository
Request DOI
Add to collection
Report this model
Downloads last
41,021
Inference


Clone this model repository
4
TPS
SSH
Make sure you have git-lfs installed (https://git-lfs.com)
git lfs install
git clone https://huggingface.co/facebook/blenderbot-400M-distill


관리자: C:₩Windows₩System32₩cmd.exe
C:\Users\Administrator\Downloads>git clone https://huggingface.co/openai/clip-vit-large-patch14
Cloning into 'clip-vit-large-patch14'...
remote: Enumerating objects: 45, done.
remote: Counting objects: 100% (1/1), done.
remote: Total 45 (delta 0), reused 0 (delta 0), pack-reused 44
Unpacking objects: 100% (45/45), 1.08 MiB | 1.33 MiB/s, done.
Filtering content: 100% (4/4), 6.37 GiB | 19.31 MiB/s, done.

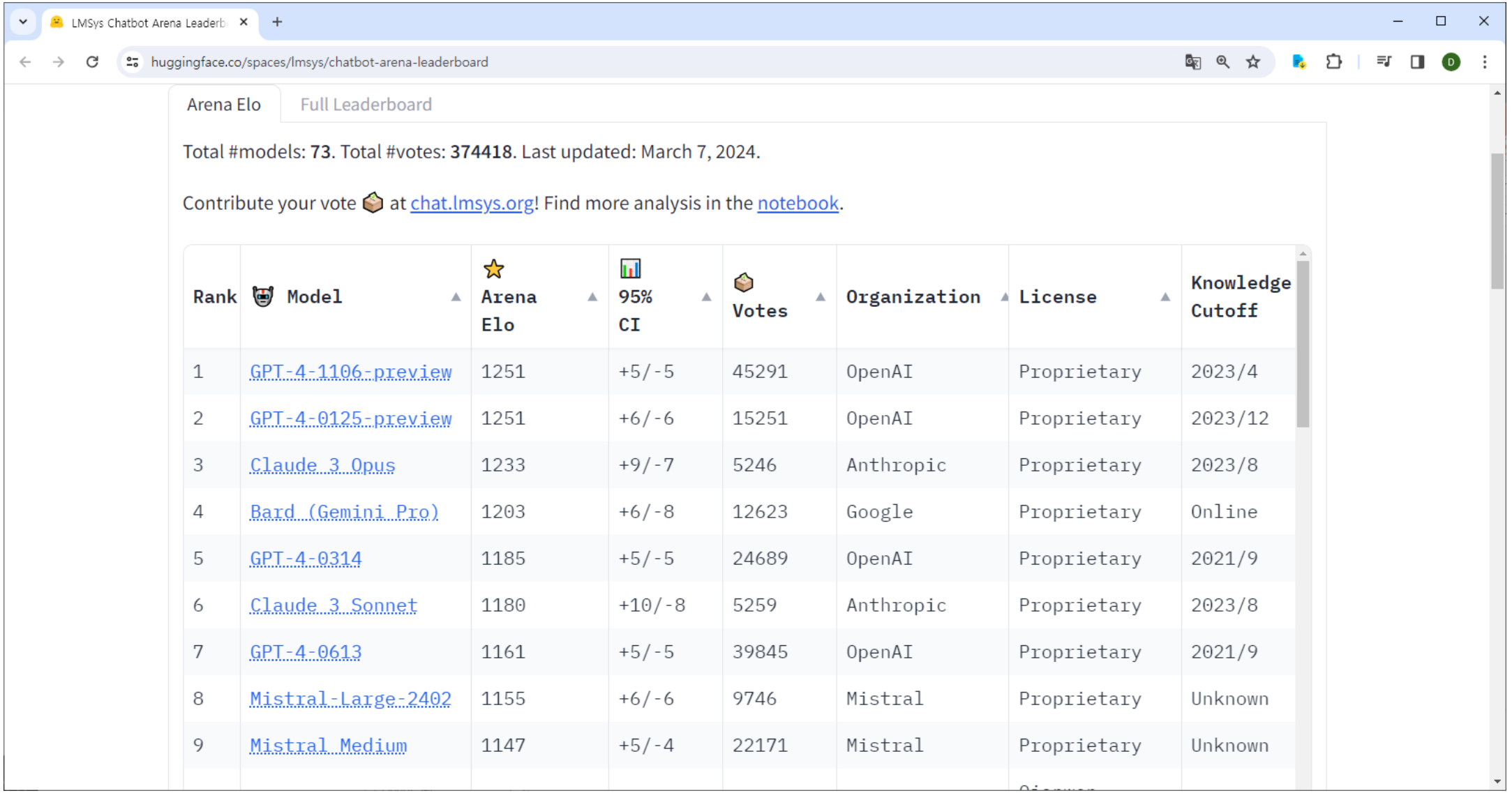
Open LLM 리더보드



T	Model	Average ↑	ARC	HellaSwa
💬	moreh/MoMo-72B-lora-1.8.7-DPO	78.55	70.82	85.96
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16	77.91	74.06	86.74
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_full_linear_DPO	77.52	74.06	86.67
🔶	zhengr/MixTAO-7Bx2-MoE-v8.1	77.5	73.81	89.22
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3
🔶	JaeyeonKang/CCK_Asura_v1	77.43	73.89	89.07
🔶	fblgit/UNA-SimpleSmaug-34b-v1beta	77.41	74.57	86.74
🔶	TomGrc/FusionNet_34Bx2_MoE_v0.1	77.38	73.72	86.46
🔶	migtissera/Tess-72B-v1.5b	77.3	71.25	85.53
💬	moreh/MoMo-72B-lora-1.8.6-DPO	77.29	70.14	86.03
🔶	abacusai/Smaug-34B-v0.1	77.29	74.23	86.76

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

LMSYS Chatbot Arena Leaderboard



Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-1106-preview	1251	+5/-5	45291	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1251	+6/-6	15251	OpenAI	Proprietary	2023/12
3	Claude 3 Opus	1233	+9/-7	5246	Anthropic	Proprietary	2023/8
4	Bard (Gemini Pro)	1203	+6/-8	12623	Google	Proprietary	Online
5	GPT-4-0314	1185	+5/-5	24689	OpenAI	Proprietary	2021/9
6	Claude 3 Sonnet	1180	+10/-8	5259	Anthropic	Proprietary	2023/8
7	GPT-4-0613	1161	+5/-5	39845	OpenAI	Proprietary	2021/9
8	Mistral-Large-2402	1155	+6/-6	9746	Mistral	Proprietary	Unknown
9	Mistral Medium	1147	+5/-4	22171	Mistral	Proprietary	Unknown

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

실습

- L2_NLP.ipynb
- L3_Translation_and_Summarization.ipynb
- L4_Sentence_Embeddings.ipynb
- L5_Zero-Shot_Audio_Classification.ipynb
- L6_Automatic_Speech_Recognition.ipynb
- L7_Text_to_Speech.ipynb
- L8_object_detection.ipynb
- L9_segmentation.ipynb
- L10_image_retrieval.ipynb
- L11_image_captioning.ipynb
- L12_visual_q_and_a.ipynb
- L13_Zero_Shot_Image_Classification.ipynb

Thank you