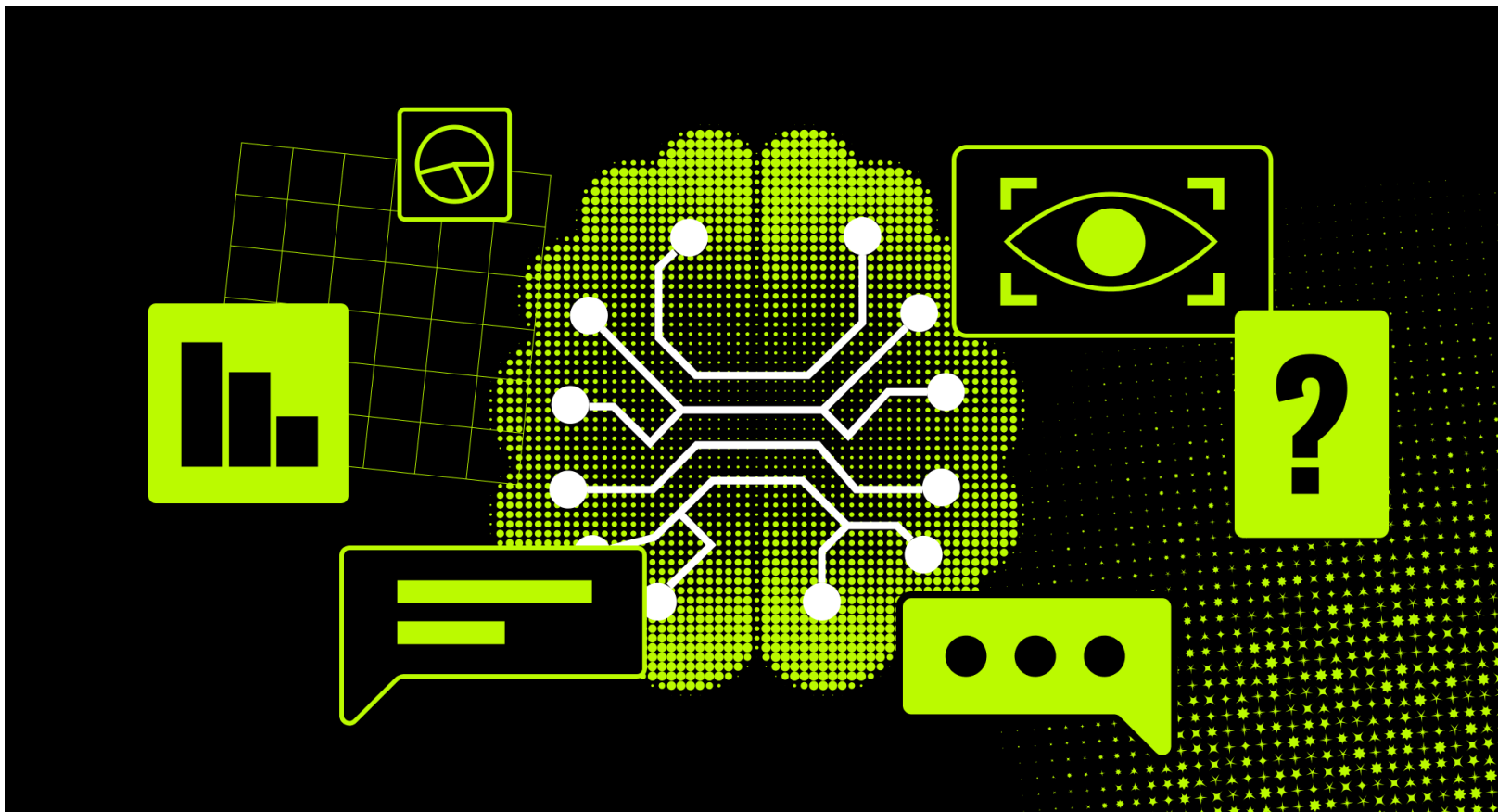
































LLM API



LLM API Providers Leaderboard

		FEATURES ↗	MODEL INTELLIGENCE ↗	PRICE ↗	OUTPUT TOKENS/S ↗	LATENCY ↗		
API PROVIDER ↕	MODEL ↕	CONTEXT WINDOW ↕	ARTIFICIAL ANALYSIS INTELLIGENCE INDEX ↕	BLENDED USD/1M Tokens ↕	MEDIAN Tokens/s ↕	MEDIAN First Chunk (s) ↕	FURTHER ANALYSIS	
 Microsoft Azure	 o3-mini (high)	200k	66	\$1.93	11.4	92.05	🔗 Model	🔗 Providers
 OpenAI	 o3-mini	200k	63	\$1.93	194.0	12.99	🔗 Model	🔗 Providers
 Microsoft Azure	 o3-mini	200k	63	\$1.93	30.7	34.12	🔗 Model	🔗 Providers
 OpenAI	 o1	200k	62	\$26.25	39.7	26.09	🔗 Model	🔗 Providers
 Microsoft Azure	 o1	200k	62	\$26.25	36.5	28.83	🔗 Model	🔗 Providers
 deepseek	 DeepSeek R1	64k	60	\$0.96	25.3	11.46	🔗 Model	🔗 Providers
 aws	 DeepSeek R1	128k	60	\$2.36	84.5	0.43	🔗 Model	🔗 Providers
NEBIUS	 DeepSeek R1 Base	128k	60	\$1.20	9.6	0.94	🔗 Model	🔗 Providers
NEBIUS	 DeepSeek R1 Fast	128k	60	\$3.00	62.3	0.66	🔗 Model	🔗 Providers
 CentML	 DeepSeek R1	128k	60	\$3.99	69.2	0.55	🔗 Model	🔗 Providers
 Microsoft Azure	 DeepSeek R1	128k	60	\$0.00	17.2	1.02	🔗 Model	🔗 Providers
 Fireworks AI	 DeepSeek R1	128k	60	\$4.25	88.1	0.73	🔗 Model	🔗 Providers
 deepinfra	 DeepSeek R1 (Turbo, FP4)	33k	60	\$3.00	43.3	0.25	🔗 Model	🔗 Providers
 deepinfra	 DeepSeek R1	64k	60	\$1.16	8.2	0.77	🔗 Model	🔗 Providers
 FriendliAI	 DeepSeek R1	128k	60	\$4.00	43.6	0.43	🔗 Model	🔗 Providers
 Novita	 DeepSeek R1 Turbo	64k	60	\$1.15	31.9	0.79	🔗 Model	🔗 Providers

<https://artificialanalysis.ai/leaderboards/providers>

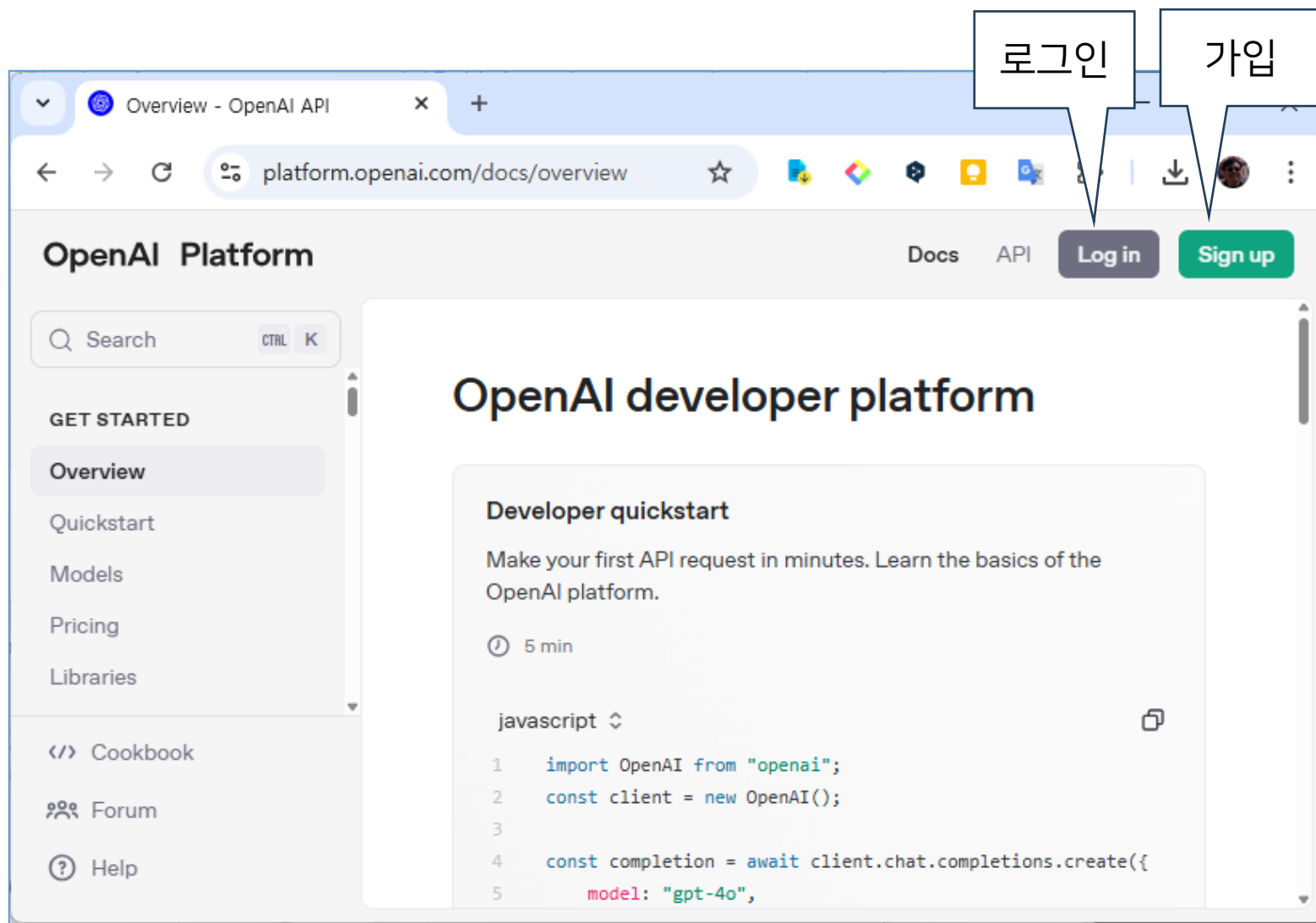
OpenAI API



OpenAI

<https://platform.openai.com/>

사이트 접속 및 회원가입



OpenAI API (유료)

<https://platform.openai.com/settings/organization/billing/overview>

The screenshot shows the OpenAI API billing overview page. The left sidebar contains a 'SETTINGS' section with 'Billing' highlighted. The main content area shows the 'Billing' overview with a credit balance of \$10 and an 'Add to credit balance' button highlighted with a red box. Below this, the 'Payment methods' section is also highlighted with a red box. Two modal windows are open: 'Add payment method' and 'Add to credit balance'.

Billing Overview

Pay as you go

Credit balance ⓘ
\$10

Add to credit balance Cancel plan

ⓘ Auto recharge is off
When your credit balance reaches \$0, your API requests will automatically keep your credit balance topped up.
Enable auto recharge

Payment methods
Add or change payment method

Preferences
Manage billing information

Pricing
View pricing and FAQs

Add payment method

Add your credit card details below. This card will be saved to your account and can be removed at any time.

Card information

Card number MM / YY CVC

Name on card

Billing address

Country Address line 1 Address line 2 City Postal code State, county, province, or region

☐ Set as default payment method

Cancel **Add payment method**

Add to credit balance

Amount to add

\$ 5

Enter an amount between \$5 and \$900 Model pricing

Payment method

+ Add payment method

Cancel **Continue**

API 사용 한도

<https://platform.openai.com/docs/guides/rate-limits>

<https://platform.openai.com/docs/models/gpt-4o-mini>



GPT-4o mini

Default



Fast, affordable small model for focused tasks

TIER	RPM	RPD	TPM	BATCH QUEUE LIMIT
Free	3	200	40,000	-
Tier 1	500	10,000	200,000	2,000,000
Tier 2	5,000	-	2,000,000	20,000,000
Tier 3	5,000	-	4,000,000	40,000,000
Tier 4	10,000	-	10,000,000	1,000,000,000
Tier 5	30,000	-	150,000,000	15,000,000,000

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

- 1 token \sim 4 chars in English
- 1 token \sim $\frac{3}{4}$ words
- 100 tokens \sim 75 words

참고 :

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

OpenAI API Key 생성

<https://platform.openai.com/settings/organization/api-keys>

API keys - OpenAI API

platform.openai.com/settings/organization/api-keys

Personal / Default project

SETTINGS

- Your profile
- ORGANIZATION
 - General
 - API keys
 - Admin keys
 - Members
 - Projects

API keys

As an owner of this organization, you can view and manage all API keys in this organization.

Do not share your API key with others or expose it in the browser or other client-side code. To protect your account's security, disable any API key that has leaked publicly.

View usage per API key on the [Usage page](#).

NAME	SECRET KEY	LAST USED ⓘ	PROJECT ACCESS	CREATED BY
HamKey	sk-...a88A	Never	Default project	Danny Park

+ Create new secret key

Create new secret key

Owned by

☒ You ☐ Service account

This API key is tied to your user and can make requests against the selected project. If you are removed from the organization or project, this key will be disabled.

Name Optional

Project

✓ Default project

☒ All ☐ Restricted ☐ Read only

OpenAI 모델

<https://platform.openai.com/docs/models>

Featured models



GPT-4.5 Preview

Largest and most capable GPT model



o3-mini

Fast, flexible, intelligent reasoning model



GPT-4o

Fast, intelligent, flexible GPT model

Reasoning models o-series models that excel at complex, multi-step tasks.



o3-mini

Fast, flexible, intelligent reasoning model



o1

High-intelligence reasoning model



o1-mini

A faster, more affordable reasoning model than o1

Cost-optimized models Smaller, faster models that cost less to run.



GPT-4o mini

Fast, affordable small model for focused tasks



GPT-4o mini Audio

Smaller model capable of audio inputs and outputs

DALL·E Models that can generate and edit images, given a natural language prompt.



DALL·E 3

Our latest image generation model



DALL·E 2

Our first image generation model

Text-to-speech Models that can convert text into natural sounding spoken audio.



TTS-1

Text-to-speech model optimized for speed



TTS-1 HD

Text-to-speech model optimized for quality

Whisper Model that can transcribe and translate audio into text.



Whisper

General-purpose speech recognition model

OpenAI 요금제

<https://platform.openai.com/docs/pricing>

Text tokens

Price per 1M tokens · Batch API price ☐

Model	Input	Cached input	Output
gpt-4.5-preview ↳ gpt-4.5-preview-2025-02-27	\$75.00	\$37.50	\$150.00
gpt-4o ↳ gpt-4o-2024-08-06	\$2.50	\$1.25	\$10.00
gpt-4o-audio-preview ↳ gpt-4o-audio-preview-2024-12-17	\$2.50	-	\$10.00
gpt-4o-realtime-preview ↳ gpt-4o-realtime-preview-2024-12-17	\$5.00	\$2.50	\$20.00
gpt-4o-mini ↳ gpt-4o-mini-2024-07-18	\$0.15	\$0.075	\$0.60
gpt-4o-mini-audio-preview ↳ gpt-4o-mini-audio-preview-2024-12-17	\$0.15	-	\$0.60
gpt-4o-mini-realtime-preview ↳ gpt-4o-mini-realtime-preview-2024-12-17	\$0.60	\$0.30	\$2.40
o1 ↳ o1-2024-12-17	\$15.00	\$7.50	\$60.00
o3-mini ↳ o3-mini-2025-01-31	\$1.10	\$0.55	\$4.40

Embeddings

Price per 1M tokens · Batch API price ☐

Model	Cost
text-embedding-3-small	\$0.02
text-embedding-3-large	\$0.13
text-embedding-ada-002	\$0.10

Image generation

Price per image

Model	Quality	1024x1024	1024x1792
DALL·E 3	Standard	\$0.04	\$0.08
	HD	\$0.08	\$0.12
Model	256x256	512x512	1024x1024
DALL·E 2	\$0.016	\$0.018	\$0.02

Other models

Price per 1M tokens · Batch API price ☐

Model	Input	Output
chatgpt-4o-latest	\$5.00	\$15.00
gpt-4-turbo ↳ gpt-4-turbo-2024-04-09	\$10.00	\$30.00
gpt-4 ↳ gpt-4-0613	\$30.00	\$60.00
gpt-4-32k	\$60.00	\$120.00
gpt-3.5-turbo ↳ gpt-3.5-turbo-0125	\$0.50	\$1.50

토큰(Token)

<https://platform.openai.com/tokenizer>

The screenshot shows the OpenAI Platform Tokenizer web interface. At the top, there's a navigation bar with 'OpenAI Platform', 'Docs', 'API reference', 'Log in', and 'Sign up' buttons. Below this, there are tabs for different models: 'GPT-4o & GPT-4o mini' (selected), 'GPT-3.5 & GPT-4', and 'GPT-3 (Legacy)'. A text box contains a paragraph about how OpenAI's models process text using tokens. Below the text box are 'Clear' and 'Show example' buttons. A summary table shows 'Tokens' as 52 and 'Characters' as 280. Below this, the input text is displayed with individual tokens highlighted in different colors. At the bottom, there are 'Text' and 'Token IDs' buttons. A helpful rule of thumb is provided at the very bottom.

OpenAI Platform Docs API reference Log in Sign up

GPT-4o & GPT-4o mini GPT-3.5 & GPT-4 GPT-3 (Legacy)

OpenAI's large language models process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens. Learn more.

Clear Show example

Tokens	Characters
52	280

OpenAI's large language models process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens. Learn more.

Text Token IDs

A helpful rule of thumb is that one token generally corresponds to roughly ¾ of a word (so 1000 tokens is roughly 750 words).

Tokens

52

Characters

280

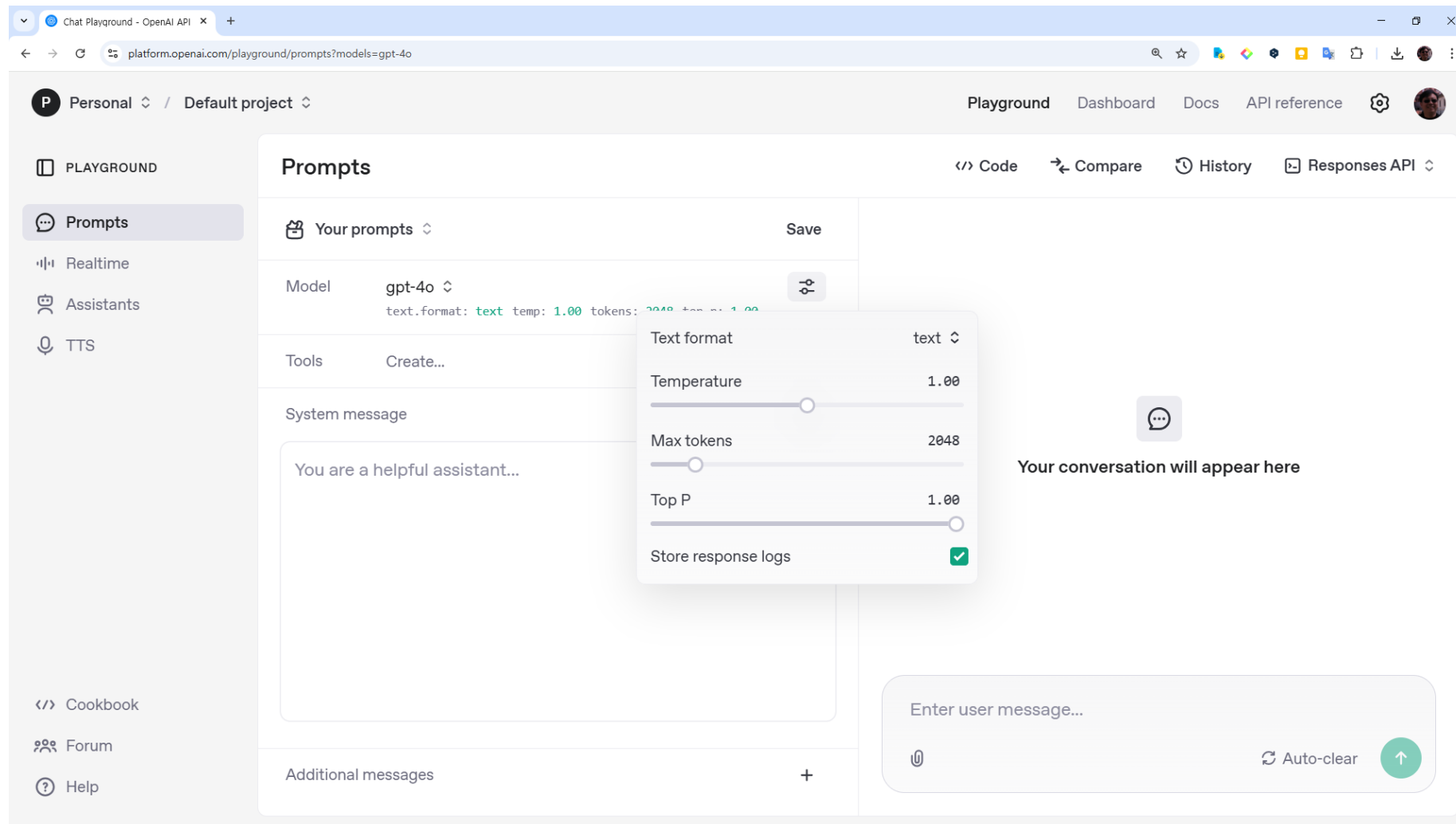
[6447, 17527, 885, 4410, 6439, 7015, 2273, 2201, 2360, 20290, 11, 1118, 553, 5355, 45665, 328, 9862, 2491, 306, 261, 920, 328, 2201, 13, 623, 7015, 4484, 316, 4218, 290, 39535, 14321, 2870, 1879, 20290, 11, 326, 19383, 540, 24168, 290, 2613, 6602, 306, 261, 16281, 328, 20290, 13, 15983, 945, 13]

Text

Token IDs

플레이그라운드

<https://platform.openai.com/playground>



- Temperature : 값이 낮을수록 가장 높은 확률의 다음 토큰을 선택하고, 값이 높아지면 무작위성이 높아짐
- Max Tokens
모델이 생성하는 토큰 최대 길이
- Top P : 값이 높으면 모델이 가능성이 낮은 단어를 포함하여 더 다양한 출력을 얻을 수 있음

API 사용 방법

Step 1: Setup Python

✓ Install Python

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

✓ Setup a virtual environment (optional)

`python -m venv venv`

Windows : `venv\Scripts\activate`

Unix or Mac : `source venv/bin/activate`

✓ Install the OpenAI Python library

`pip install 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

```
1 import OpenAI from "openai";
2 const client = new OpenAI();
3
4 const completion = await client.chat.completions.create({
5   model: "gpt-4o",
6   messages: [
7     {
8       role: "user",
9       content: "Write a one-sentence bedtime story about a unicorn.",
10     },
11   ],
12 });
13
14 console.log(completion.choices[0].message.content);
```

OpenAI API 실습자료



openai_api.ipynb

information_retrieval.ipynb

ReAct.ipynb

pe-lecture.ipynb

colab

OpenAI Cookbook examples

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

The screenshot shows the GitHub interface for the 'openai-cookbook' repository, specifically the 'examples' directory. The browser address bar shows 'github.com/openai/openai-cookbook/tree/main/examples'. The repository name 'openai / openai-cookbook' is visible at the top. Below the repository name, there are tabs for 'Code', 'Issues' (35), 'Pull requests' (39), 'Actions', 'Security', and 'Insights'. The 'main' branch is selected. A search bar 'Go to file' is present. A commit by 'erikakettleon-openai' is highlighted, titled 'One way translation - update images (#1735)'. Below this, a table lists the contents of the 'examples' directory.

Name	Last commit message	Last commit date
..		
agents_sdk	Add Cookbook: Using the OpenAI Agents SDK to Automate Stripe Dispute ...	last week
azure	[typo] replace words (#1399)	5 months ago
book_translation	update latex_book to use tiktoken, gpt4o, modified chunk sizes and ad...	last month
chatgpt	Add support scope disclaimer (#1713)	2 weeks ago
dalle	Fix syntax error in DALL-E notebook (#1036)	last year
data	File Search with Responses (#1708)	2 weeks ago
evaluation	Clean up the organization of the SQL generation notebook (#1655)	2 months ago
fine-tuned_qa	fix: possessive error in Markdown cell (#1234)	7 months ago
gpt4o	Small spelling fix (#1594)	2 months ago
multimodal	Model swap to GPT-4o (#1601)	2 months ago
o1	fix: small typo in Update Using reasoning for data validation in nuph...	5 months ago

Prompt Engineering with Llama 2&3



<https://learn.deeplearning.ai/courses/prompt-engineering-with-llama-2/lesson/bg26k/introduction>

허깅페이스(Hugging Face) 오픈소스 모델 사용



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

Streamlit으로 AI앱 만들기

<https://streamlit.io/>

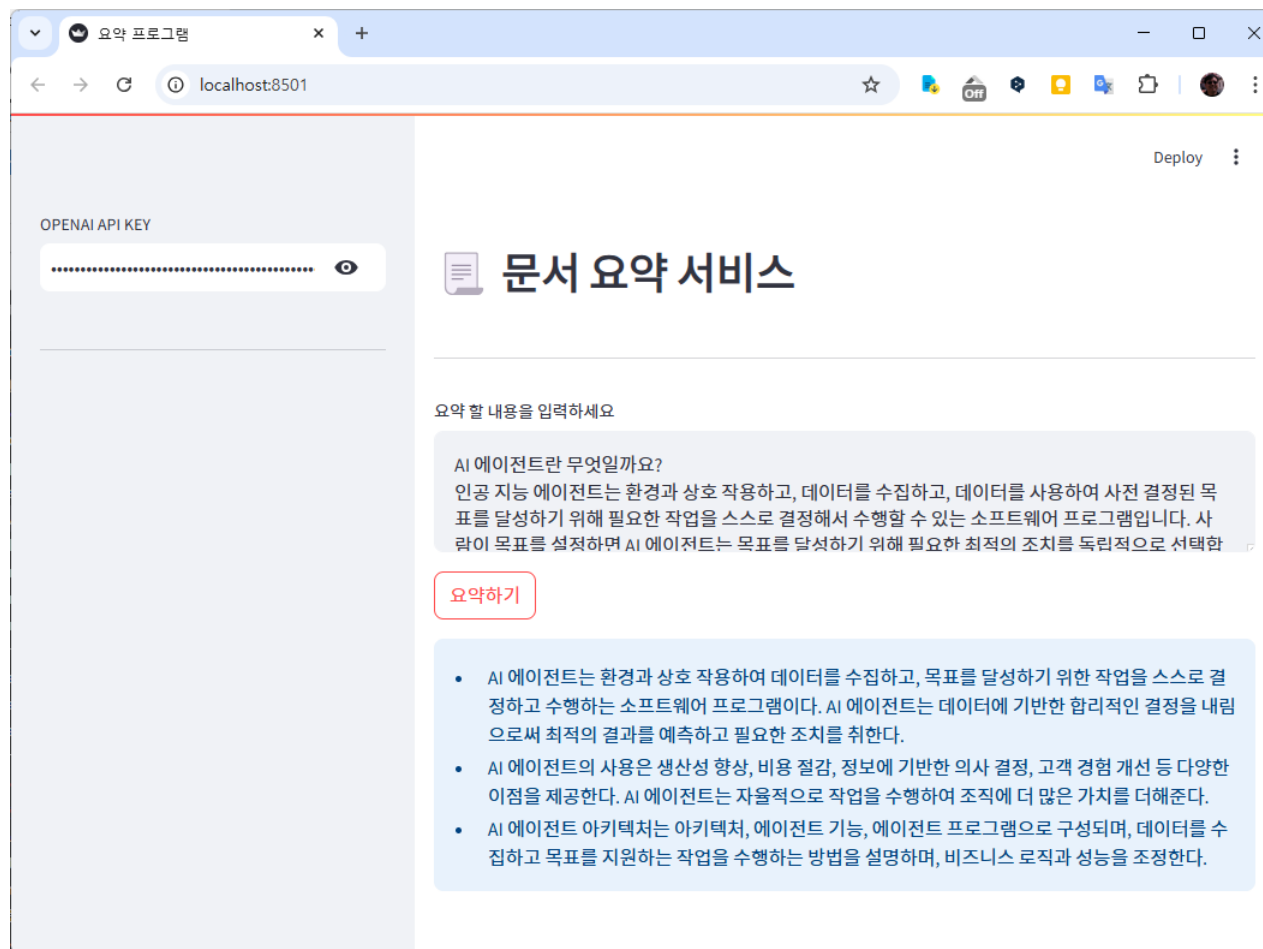
Streamlit은 데이터 과학, 머신러닝, 분석 프로젝트를 위한 웹 애플리케이션을 만드는 과정을 간소화하고, 신속하게 웹 애플리케이션을 만들 수 있게 설계된 오픈소스입니다.

■ 설치

```
pip install streamlit
```

■ 실행

```
streamlit run st_summerize_app.py
```



Thank you 😊