



에 게시됨 더 나은 프로그래밍



티모시 무가이 따르다

3월 18일 · 11분 읽기 · ✨ · ▶ 듣다



구하다



맞춤형 지식 기반으로 나만의 맞춤형 ChatGPT를 구축하는 방법

ChatGPT 봇에 사용자 지정 데이터 소스를 제공하는 방법에 대한 단계별 가이드



[Unsplash](#) 의 [Christian Wiediger](#) 사진

ChatGPT는 대부분의 사람들이 다양한 작업을 자동화하기 위해 매일 사용하는 필수 도구가 되었습니다. 일정 기간 동안 ChatGPT를 사용했다면 잘못된 답변을 제공할 수 있고 일부 틈새 주제에 대한 컨텍스트가 0으로 제한된다는 것을 깨달았을 것입니다. 이것은 우리가 chatGPT를 활용하여 격차를 해소하고 ChatGPT가 더 많은 사용자 정의 데이터를 가질 수 있도록 하는 방법에 대한 질문을 제기합니다.

풍부한 지식은 우리가 매일 상호 작용하는 다양한 플랫폼, 즉 직장에서의 Confluence 위키 페이지, 슬랙 그룹, 회사 지식 기반, Reddit, Stack Overflow, 책, 뉴스레터 및 동료 가 공유하는 Google 문서를 통해 배포됩니다. 이러한 모든 정보 소스를 유지하는 것은 그 자체로 정규직입니다.

데이터 소스를 선택적으로 선택하고 해당 정보를 데이터와 함께 ChatGPT 대화에 쉽게 공급할 수 있다면 좋지 않을까요?

1. 신속한 엔지니어링을 통한 데이터 공급

ChatGPT를 확장하는 방법에 대해 알아보기 전에 ChatGPT를 수동으로 확장하는 방법과 문제가 무엇인지 살펴보겠습니다. ChatGPT를 확장하는 기존의 접근 방식은 신속한 엔지니어링을 통한 것입니다.

ChatGPT는 매우 간단합니다. 컨텍스트 인식. 먼저 실제 질문 앞에 원본 문서 콘텐츠를 추가하여 ChatGPT와 상호 작용해야 합니다.

다음 콘텐츠를 기반으로 질문을 드리겠습니다.

- 콘텐츠 시작-

ChatGPT 컨텍스트를 제공하는 매우 긴 텍스트

- 콘텐츠 끝-

이 접근 방식의 문제는 모델에 제한된 컨텍스트가 있다는 것입니다. GPT-3에 대해 약 4,097개의 토큰만 허용할 수 있습니다. 이 접근 방식은 항상 콘텐츠를 붙여넣어야 하는 상당히 수동적이고 지루한 프로세스이기 때문에 곧 벽에 부딪히게 될 것입니다.

ChatGPT에 삽입하고 싶은 수백 개의 PDF 문서가 있다고 상상해 보세요. 곧 페이지 문제 직면하게 될 것입니다. GPT-4가 GPT-3의 후속 제품이라고 생각할 수 있습니다. 2023년 3월 14일에 막 출시되었으며 GPT-3 프로세스 이미지의 약 8배인 25,000개의 단어를 처리할 수 있으며 GPT-3.5보다 훨씬 더 미묘한 명령을 처리할 수 있습니다. 이것은 여전히 데이터 입력 제한이라는 동일한 근본적인 문제를 가지고 있습니다. 이러한

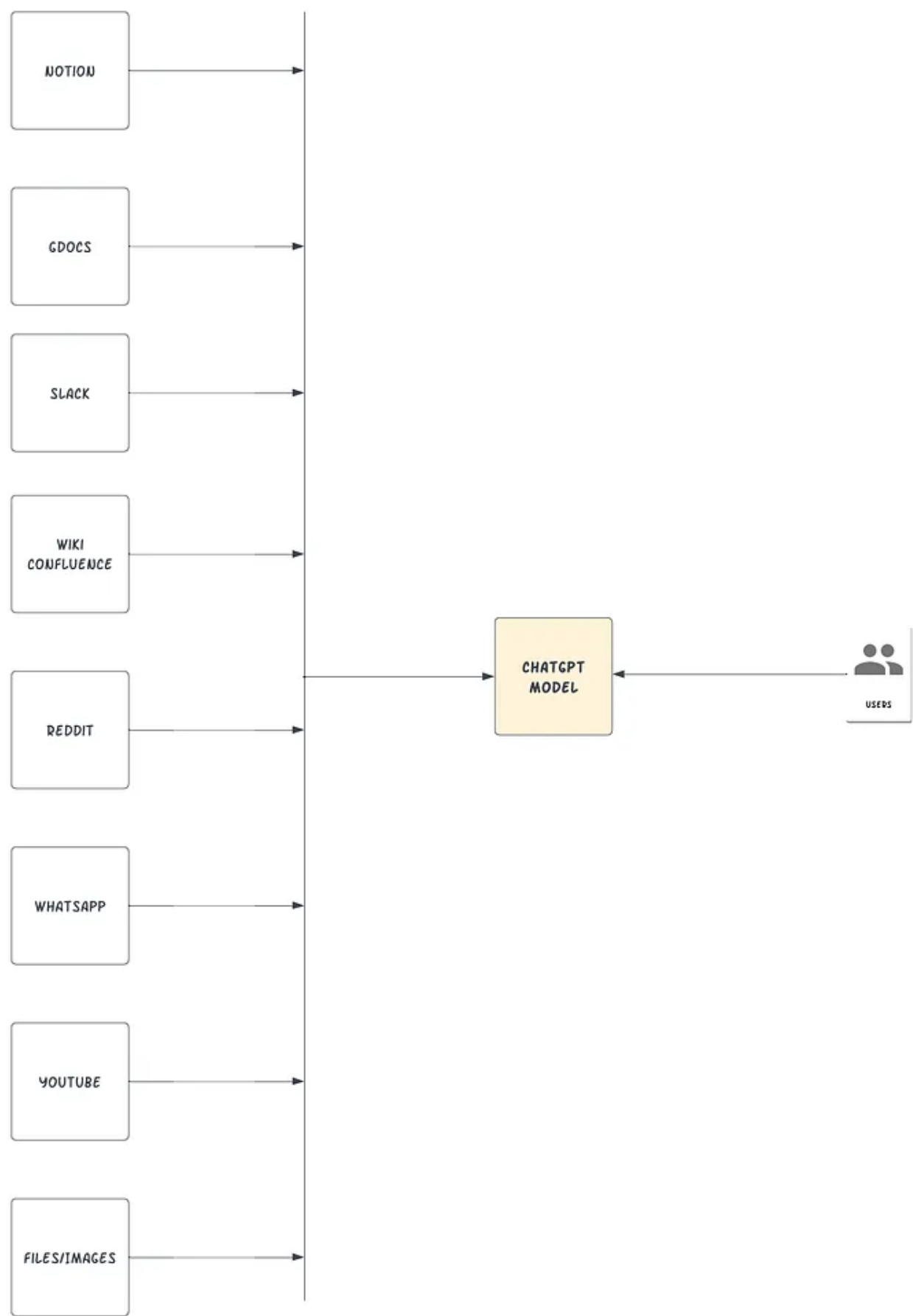
제한 사항 중 일부를 우회하는 방법은 무엇입니까? LlamaIndex라는 Python 라이브러리를 활용할 수 있습니다.

2. LlamaIndex(GPT 인덱스)로 ChatGPT 확장하기

GPT 인덱스라고도 하는 LlamaIndex는 LLM을 외부 데이터와 연결하기 위한 중앙 인터페이스를 제공하는 프로젝트입니다. 그래, 당신은 그것을 올바르게 읽었습니다.

LlamaIndex를 사용하면 아래 그림과 같은 것을 만들 수 있습니다.

DATA SOURCES



ChatGPT에 입력되는 사용자 정의 데이터 소스

LlamaIndex는 기존 데이터 소스 및 유형을 사용 가능한 데이터 커넥터(예: API, PDF, 문서, SQL 등)와 연결합니다. LlamaIndex는 구조화된 데이터와 구조화되지 않은 데이터에 대한 인덱스를 제공하여 LLM을 사용할 수 있도록 합니다. 이러한 인덱스는 일반적인 상용구와 문제점을 제거하여 컨텍스트 내 학습을 용이하게 합니다. 즉, 빠른 삽입을 위해 액세스 가능한 방식으로 컨텍스트를 보존합니다.

GPT-3 Davinci의 경우 4,096 토큰 한도, GPT-4의 경우 8,000 토큰 한도인 프롬프트 제한 처리는 컨텍스트가 너무 클 때 액세스가 훨씬 쉬워지고 사용자에게 상호 작용할 수 있는 방법을 제공하여 텍스트 분할 문제를 해결합니다. 지수. LlamaIndex는 또한 문서에서 관련 부분을 추출하여 프롬프트에 제공하는 프로세스를 추상화합니다.

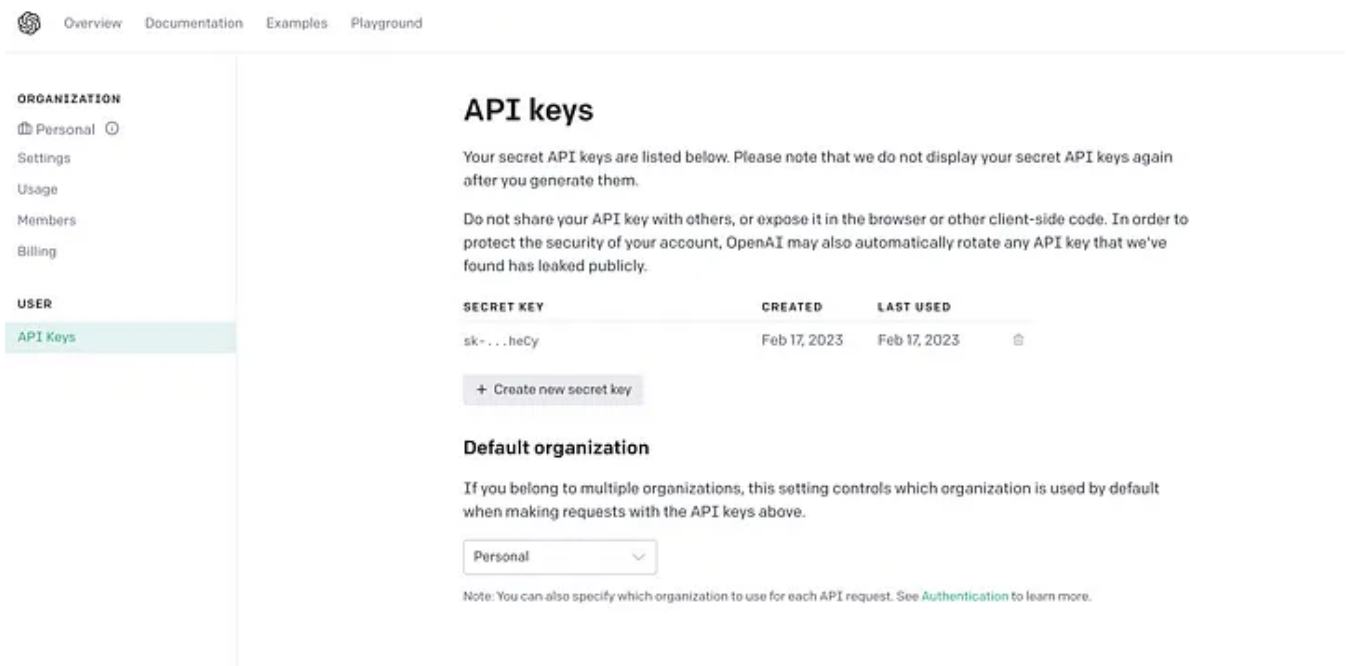
사용자 정의 데이터 소스를 추가하는 방법

이 섹션에서는 GPT "text-davinci-003" 및 LlamaIndex를 사용하여 기존 문서를 기반으로 Q&A 챗봇을 만듭니다.

전제 조건

시작하기 전에 다음에 대한 액세스 권한이 있는지 확인하십시오.

- 시스템에 설치된 Python ≥ 3.7
- OpenAI 웹사이트에서 찾을 수 있는 OpenAI API 키 . 싱글 사인온에 Gmail 계정을 사용할 수 있습니다.



The screenshot displays the OpenAI API Keys management interface. On the left, there is a sidebar with navigation links: Overview, Documentation, Examples, and Playground. Below these, there are sections for ORGANIZATION (Personal, Settings, Usage, Members, Billing) and USER (API Keys). The main content area is titled 'API keys' and contains the following information:

- A warning: "Your secret API keys are listed below. Please note that we do not display your secret API keys again after you generate them."
- A security notice: "Do not share your API key with others, or expose it in the browser or other client-side code. In order to protect the security of your account, OpenAI may also automatically rotate any API key that we've found has leaked publicly."
- A table of API keys:

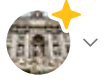
SECRET KEY	CREATED	LAST USED
sk-...heCy	Feb 17, 2023	Feb 17, 2023
- A button: "+ Create new secret key"
- A section for 'Default organization' with a dropdown menu set to 'Personal'.
- A note: "Note: You can also specify which organization to use for each API request. See Authentication to learn more."

- Google Docs에 업로드된 Word 문서 몇 개. LlamaIndex는 다양한 데이터 소스를 지원합니다. 이 자습서에서는 Google 문서도구를 시연합니다.

작동 방식

1. LlamaIndex로 문서 데이터 인덱스를 생성합니다.

앱에서 열기 ↗



이 인덱스를 활용하여 쿼리와 데이터가 얼마나 일치하는지에 따라 가장 적절한 섹션을 찾습니다. 그러면 정보가 프롬프트에 로드되고 GPT가 귀하의 질문에 응답하는 데 필요한 배경 지식을 갖도록 GPT로 전송됩니다.

4. 그런 다음 상황에 맞는 피드에 따라 ChatGPT에 요청할 수 있습니다.

가급적이면 가상 환경이나 conda 환경을 사용하여 mychatbot을 호출할 수 있는 Python 프로젝트용 새 폴더를 만듭니다.

먼저 종속성 라이브러리를 설치해야 합니다. 방법은 다음과 같습니다.

```
pip 설치 openai
pip 설치 라마-인덱스
pip 설치 google-auth-oauthlib
```



다음으로 Python에서 라이브러리를 가져오고 새 main.py 파일에서 OpenAI API 키를 설정합니다.

```
# 필요한 패키지 가져오기
import os
import pickle

from google.auth.transport.requests import Request

from google_auth_oauthlib.flow import InstalledAppFlow
from llama_index import GPTSimpleVectorIndex, download_loader
```

```
os.environ['OPENAI_API_KEY'] = 'SET-YOUR-OPEN-AI-API-KEY'
```

위 스니펫에서는 LlamaIndex  1.4K |  26 | ... 1AI에 액세스해야 하므로 명확성을 위해 환경 변수를 명시적으로 설정하고 있습니다. 일반적인 프로덕션 환경에서는 환경 변수, 저장소 또는 인프라에서 액세스할 수 있는 모든 비밀 관리 서비스에 키를 넣을 수 있습니다.

Google 문서를 검색하기 위해 Google 계정에 대해 인증하는 데 도움이 되는 함수를 구성해 보겠습니다.

```
def authorize_gdocs():
    google_oauth2_scopes = [
        "https://www.googleapis.com/auth/documents.readonly"
    ]
    cred = None
    if os.path.exists("token.pickle"):
        with open("token.pickle", "rb") as token:
            cred = pickle.load(token)
    if not cred or not cred.valid:
        if cred and cred.expired and cred.refresh_token:
            cred.refresh(Request())
        else:
            flow = InstalledAppFlow.from_client_secrets_file("credentials.json",
                                                            google_oauth2_scopes)
            cred = flow.run_local_server(port=0)
    with open("token.pickle", "wb") as token:
        pickle.dump(cred, token)
```

Google 문서 API를 활성화하고 Google 콘솔에서 자격 증명을 가져오려면 다음 단계를 따르세요.

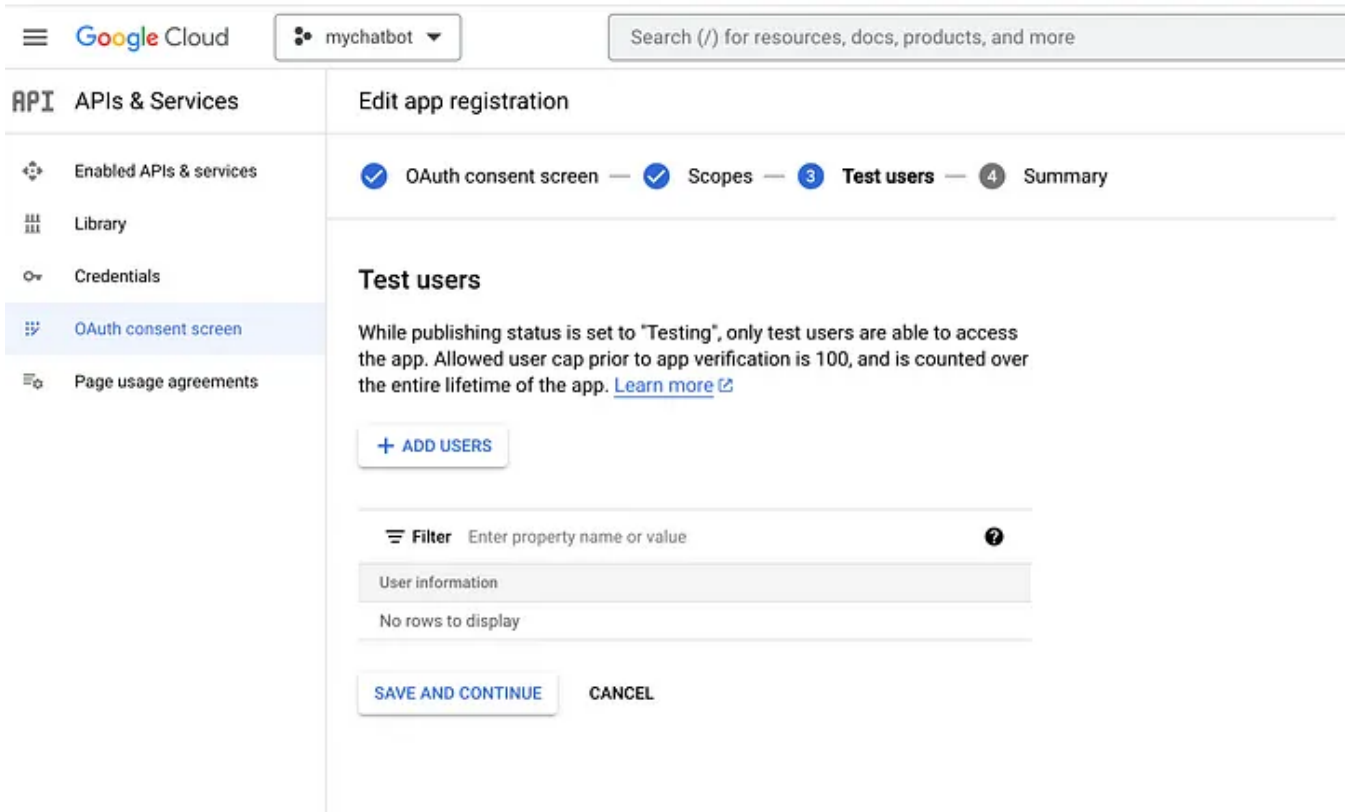
1. Google Cloud Console 웹사이트(console.cloud.google.com)로 이동합니다.
2. Create a new project if you haven't already. You can do this by clicking on the "Select a project" dropdown menu in the top navigation bar and selecting "New Project." Follow the prompts to give your project a name and select the organization you want to associate it with.
3. Once your project is created, please select it from the dropdown menu in the top navigation bar.
4. Go to the "APIs & Services" section from the left-hand menu and click on the "+ ENABLE APIS AND SERVICES" button at the top of the page.

5. Search for "Google Docs API" in the search bar and select it from the results list.
6. Click the "Enable" button to enable the API for your project.
7. Click on the OAuth consent screen menu and create and give your app a name, e.g., "mychatbot," then enter the support email, save, and add scopes.

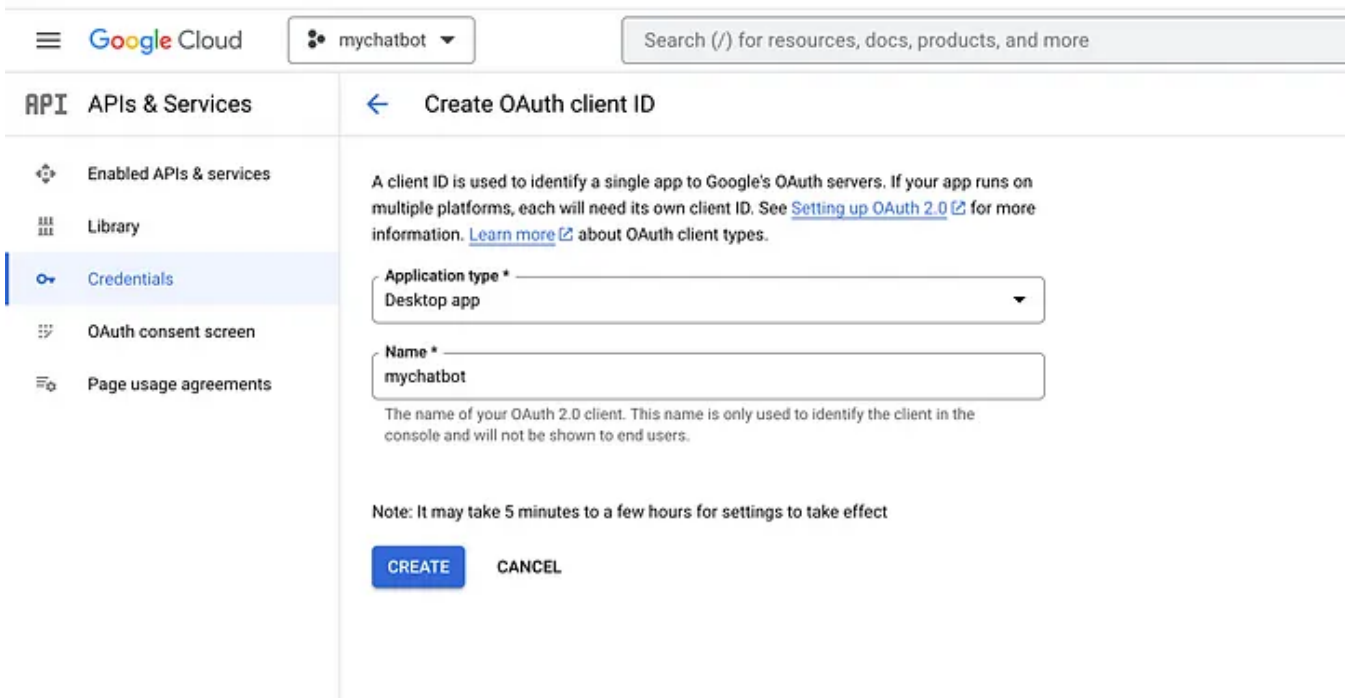
The screenshot shows the Google Cloud console interface. On the left, the 'APIs & Services' menu is visible, with 'OAuth consent screen' selected. The main area displays the 'Edit app registration' page, which includes tabs for 'OAuth consent screen', 'Scopes', and 'Test users'. The 'Scopes' tab is active, showing a list of selected scopes. A modal dialog titled 'Update selected scopes' is open, displaying a table of available scopes for the Google Docs API. The scope '.../auth/documents.readonly' is selected and highlighted with a red box. Below the table, there is a section for 'Manually add scopes' with a text input field and an 'ADD TO TABLE' button. At the bottom of the dialog is an 'UPDATE' button.

API	Scope	User-facing description	
<input type="checkbox"/>	Google Docs API	.../auth/documents	See, edit, create, and delete all your Google Docs documents
<input checked="" type="checkbox"/>	Google Docs API	.../auth/documents.readonly	See all your Google Docs documents
<input type="checkbox"/>	Google Docs API	.../auth/drive	See, edit, create, and delete all of your Google Drive files
<input type="checkbox"/>	Google Docs API	.../auth/drive.readonly	See and download all your Google Drive files
<input type="checkbox"/>	Google Docs API	.../auth/drive.file	See, edit, create, and delete only the specific Google Drive files you use with this app

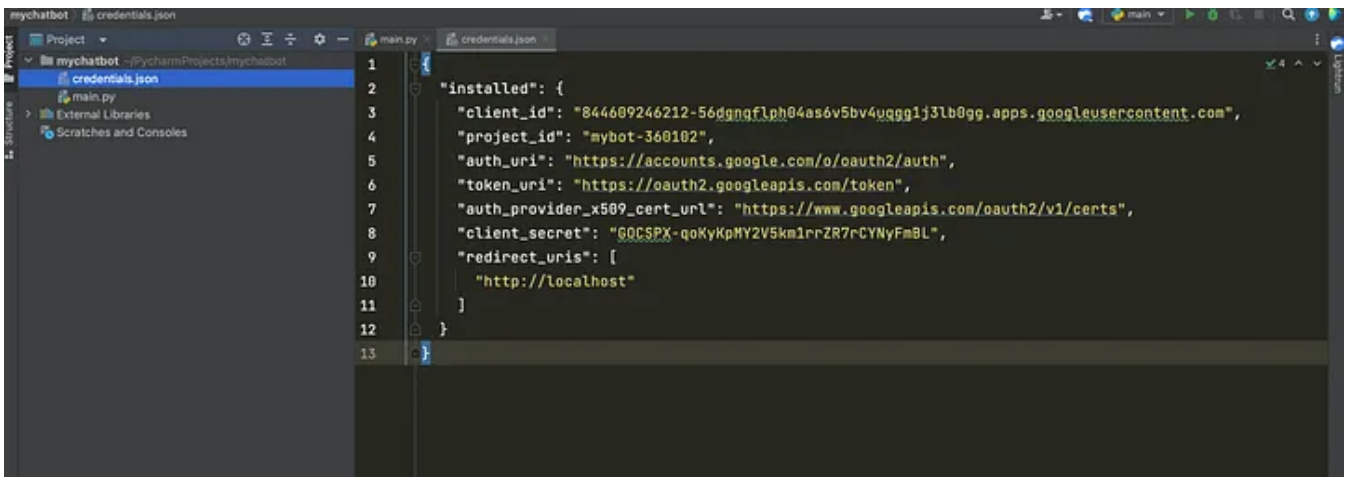
You must also add test users since this Google app will not be approved yet. This can be your own email.



You will then need to set up credentials for your project to use the API. To do this, go to the "Credentials" section from the left-hand menu and click "Create Credentials." Select "OAuth client ID" and follow the prompts to set up your credentials.



Once your credentials are set up, you can download the JSON file and store it in the root of your application, as illustrated below:



Example folder structure with google credentials in root

Once you have set up your credentials, you can access the Google Docs API from your Python project.

Go to your Google Docs, open up a few of them, and get the unique id that can be seen in your browser URL bar, as illustrated below:



Gdoc ID

Copy out the gdoc IDs and paste them into your code below. You can have N number of gdocs that you can index so ChatGPT has context access to your custom knowledge base. We will use the GoogleDocsReader plugin from the LlamaIndex library to load your documents.

```

# function to authorize or download latest credentials
authorize_gdocs()

# initialize LlamaIndex google doc reader
GoogleDocsReader = download_loader('GoogleDocsReader')

# list of google docs we want to index
gdoc_ids = ['1ofZ96nWEZYCJsterFqik_xNQTGFHtnc-7cYrf0dMPKQ']

loader = GoogleDocsReader()

```

```
# load gdocs and index them
documents = loader.load_data(document_ids=gdoc_ids)
index = GPTSimpleVectorIndex(documents)
```

LlamaIndex has a variety of data connectors covering services such as Notion, Obsidian, Reddit, Slack, etc. You can find the compressive list of available data connectors [here](#).

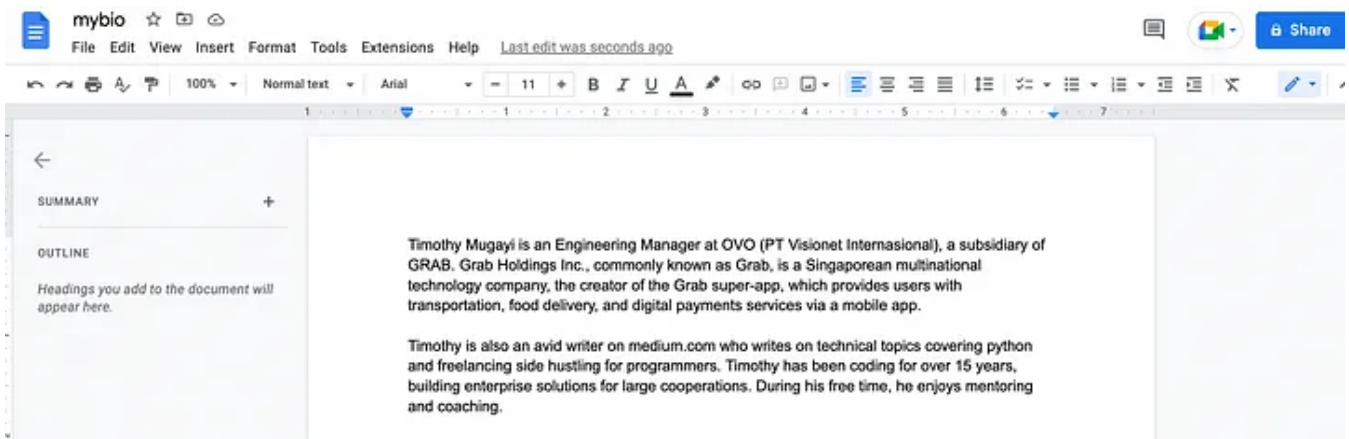
If you wish to save and load the index on the fly, you can use the following function calls. This will speed up the process of fetching from pre-saved indexes instead of making API calls to external sources.

```
# Save your index to a index.json file
index.save_to_disk('index.json')
# Load the index from your saved index.json file
index = GPTSimpleVectorIndex.load_from_disk('index.json')
```

Querying the index and getting a response can be achieved by running the following code below. Code can easily be extended into a rest API that connects to a UI where you can interact with your custom data sources via the GPT interface.

```
# Querying the index
while True:
    prompt = input("Type prompt...")
    response = index.query(prompt)
    print(response)
```

Given we have a Google Doc with details about me, information that's readily available if you publicly search on google.



We will interact directly with vanilla ChatGPT first to see what output it generates without injecting a custom data source.

M who is timothy mugayi hint he is a writer on medium

I'm sorry, but I cannot find any information about a person named "Timothy Mugayi" online. It's possible that he is not a well-known writer or public figure, or that he uses a pseudonym or a different name for his writing on Medium. Without additional information, I am unable to provide more context or details about this person.

That was a little disappointing! Let's try again.

```
INFO:google_auth_oauthlib.flow:"GET /?state=oz9XY8CE3LaLLsTxIz4sDgrHha4fEJ&code=40304112"
INFO:googleapiclient.discovery_cache:file_cache is only supported with oauth2client
INFO:root:> [build_index_from_documents] Total LLM token usage: 0 tokens
INFO:root:> [build_index_from_documents] Total embedding token usage: 175 tokens
Type prompt...who is timothy mugayi hint he is a writer on medium
```

```
INFO:root:> [query] Total LLM token usage: 300 tokens
INFO:root:> [query] Total embedding token usage: 14 tokens
Timothy Mugayi is an Engineering Manager at OVO (PT Visionet Internasional), a
last_token_usage=300
Type prompt...
```

```
Type prompt...Given you know who timothy mugayi is write an interesting introduction
```

```
Timothy Mugayi is an experienced and accomplished professional with a wealth of
last_token_usage=330
```

It can now infer answers using a new custom data source, accurately producing the following output.

We can take things further.

```
Type prompt...Write a cover letter for timothy mugayi for an upwork python project
INFO:root:> [query] Total LLM token usage: 436 tokens
INFO:root:> [query] Total embedding token usage: 30 tokens

Dear [Hiring Manager],

I am writing to apply for the Python project to build a custom ChatGPT bot with

I am currently an Engineering Manager at OVO (PT Visionet Internasional), a sub

I am confident that I can deliver a high-quality product that meets the require

Thank you for your time and consideration.

Sincerely,
Timothy Mugayi
last_token_usage=436
Type prompt...
```

LlamaIndex will internally accept your prompt, search the index for pertinent chunks, and then pass both your prompt and the pertinent chunks to the ChatGPT model. The procedures above demonstrate a fundamental first use of LlamaIndex and GPT for answering questions. Yet, there is much more you can do. You are only limited by your creativity when configuring LlamaIndex to utilize a different large language model (LLM), using a different type of index for various activities, or updating old indices with a new index programmatically.

Here is an example of changing the LLM model explicitly. This time we tap into another Python package that comes bundled with LlamaIndex called langchain.

```

llm_predictor = LLMPredictor(llm=OpenAI(temperature=0, model_name="text-davinci-001")

# define prompt configuraiton
# set maximum input size
max_input_size = 4096
# set number of output tokens
num_output = 256
# set maximum chunk overlap
max_chunk_overlap = 20
prompt_helper = PromptHelper(max_input_size, num_output, max_chunk_overlap)

index = GPTSimpleVectorIndex(
    documents, llm_predictor=llm_predictor, prompt_helper=prompt_helper
)

```

If you want to keep tabs on your OpenAI free or paid credits, you can navigate to the OpenAI [dashboard](#) and check how much credit is left.

Creating an index, inserting into an index, and querying an index will use tokens. Hence, it's always important to ensure you output token usage for tracking purposes when building your custom bots.

```

last_token_usage = index.llm_predictor.last_token_usage

print(f"last_token_usage={last_token_usage}")

```

Final Thoughts

LlamaIndex와 결합된 ChatGPT는 자체 문서 소스를 기반으로 지식을 추론할 수 있는 맞춤형 ChatGPT 챗봇을 구축하는 데 도움이 될 수 있습니다. ChatGPT 및 기타 LLM은 매우 강력하지만 LLM 모델을 확장하면 훨씬 더 세련된 경험을 제공하고 고객 지원 지원 또는 심지어 스팸 분류기와 같은 실제 비즈니스 사용 사례를 구축하는 데 사용할 수 있는 대화형 챗봇을 구축할 가능성이 열립니다. 실시간 데이터를 제공할 수 있다는 점을 감안할 때 특정 기간까지 훈련되는 ChatGPT 모델의 일부 제한 사항을 평가할 수 있습니다.

전체 소스 코드는 이 [GitHub 저장소](#)를 참조할 수 있습니다.

도메인을 이해하는 사용자 지정 *ChatGPT* 봇을 구축하려는 경우 댓글 섹션에 메시지를 드롭하고 연결해 보겠습니다.

프로그램 작성

채팅

소프트웨어 개발

일체 포함

파이썬

Enjoy the read? Reward the writer.^{Beta}

Your tip will go to Timothy Mugayi through a third-party platform of their choice, letting them know you appreciate their story.


Give a tip

Sign up for Coffee Bytes

By Better Programming

A newsletter covering the best programming articles published across Medium [Take a look.](#)

Emails will be sent to kobong9431@gmail.com. [Not you?](#)

 Get this newsletter