

Chat GPT



Przegląd istniejących rozwiązań

Bartosz Mikulski

OpenAI Platform

Nie tylko Chat GPT

ChatGPT to tylko jedno z narzędzi dostępne od **OpenAI**. Większość ludzi używa go z poziomu przeglądarki internetowej w celu zadawania pytań (Promptów) do modelu.

Jednakże **OpenAI** udostępnia dużo innych produktów, które można wykorzystać do budowania aplikacji opartych o AI.

Build an application



Chat Beta

Learn how to use chat-based language models



Embeddings

Learn how to search, classify, and compare text



Image generation Beta

Learn how to generate or edit images



Text completion

Learn how to generate or edit text



Speech to text Beta

Learn how to turn audio into text



Fine-tuning

Learn how to train a model for your use case

Build a ChatGPT plugin



Introduction Alpha

Learn the basics of building a ChatGPT plugin



Examples Alpha

Explore ChatGPT plugin examples

Chat API

Playground

Load a preset...

Save

View code

Share

...

SYSTEM

You are an expert on artificial intelligence. Right now you are at the AI conference and responsible for answering attendees' queries.

USER

What role will the AI have in 5 years onward?



ASSISTANT

The AI will touch most if not all branches of industries. We will work faster and smarter with the help of these tools. Many jobs will be deprecated but many more will be created as in any revolution in technology so far.



USER

How will it affect our daily life?



+ Add message

Mode

Chat

Beta

Model

gpt-3.5-turbo

Temperature

0.7

Maximum length

256

Top P

1

Frequency penalty

0

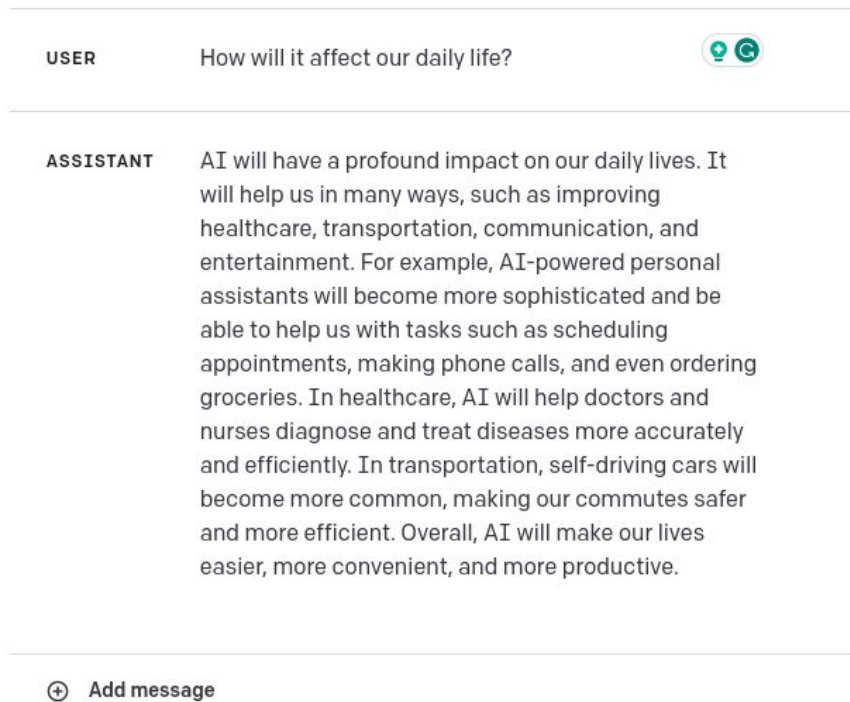
Presence penalty

0

Chat API

Standardowo jak w aplikacji ChatGPT tutaj model też odpowiada na nasze zapytania. Jednakże możemy zamodelować początek rozmowy, wraz z opisem asystenta.

Umożliwia to tworzenie aplikacji która będzie mogła odpowiadać za nas, biorąc pod uwagę dotychczasową rozmowę, np. między klientem, a konsultantem.



Complete API

Dedykowany model do uzupełniania podanego tekstu.

Playground

Load a preset...



Save

View code

Share



Wstęp do prezentacji na temat: Przegląd istniejących rozwiązań dla modelu GPT-3



Witam wszystkich na prezentacji dotyczącej modelu GPT-3. GPT-3, opracowany przez OpenAI, stanowi najnowocześniejsze rozwiązanie sztucznej inteligencji, które pozwala tworzyć złożone teksty na podstawie wprowadzonych przez użytkowników danych. W dzisiejszej prezentacji przyjrzymy się bliżej istniejącym rozwiązaniom dla tego modelu. Przedstawię wam krótkie omówienie technologii GPT-3, a następnie przyjrzymy się najpopularniejszym i najskuteczniejszym rozwiązaniom wykorzystywanym do

Mode

Complete



Model

text-davinci-003



Temperature

0.7



Maximum length

256



Complete API

Wszystkie zapytania które wykonaliśmy w piaskownicy, można przekonwertować do kodu w Pythonie, do bezpośredniego użycia w naszej aplikacji.

View code

You can use the following code to start integrating your current prompt and settings into your application.

```
POST /v1/completions python Copy
1 import os
2 import openai
3
4 openai.api_key = os.getenv("OPENAI_API_KEY")
5
6 response = openai.Completion.create(
7     model="text-davinci-003",
8     prompt="Wstęp do prezentacji na temat: Przegląd istniejących rozwiązań",
9     temperature=0.7,
10    max_tokens=256,
11    top_p=1,
12    frequency_penalty=0,
13    presence_penalty=0
14 )
```

Insert API

Dedykowany model do generowania tekstu w wybranym miejscu.

Playground

Load a preset... ▾

Save

View code

Share

⋮

We would like to invite you to our organization [insert].

Feel free to ask any additional questions.

GitHub Team

We would like to invite you to our organization **GitHub**, where we host and review code, manage projects, and **build software together.**

GitHub is a great platform for collaboration and communication. Our team of experts can help you find the right solution for your project, whether it's open source or private. We also offer a wide range of tutorials and resources to help you get started.

To join GitHub, you can sign up for a free account. After signing up, you'll have access to our community of developers, projects, and tools. You can also start creating your own projects and collaborate with others.

We hope you'll join us on GitHub and become part of our amazing community. We look forward to working with you.

Feel free to ask any additional questions.

GitHub Team

Mode

↓ Insert **Beta** ▾

Model

text-davinci-003 ▾

Temperature 0.7

▬

Maximum length 256

▬

Stop sequences

Enter sequence and press Tab

▬

Top P 1

▬

Frequency penalty 0

▬

Presence penalty 0

Edit API

Dedykowany model do wykonywania edycji na danym wejściu.

Playground

Load a preset...

Save

View code

Share

...

Input

The grass is blue, the sand is red, the water is yellow and the poppy is green.



Instructions

Fix the colors.



Submit



← Use as input

The grass is green, the sand is yellow, the water is blue and the poppy is red.

i Editing is free while in beta. We'd love your feedback. **x**

Mode

Edit **Beta** ▾

Model

text-davinci-edit-001 ▾

Temperature

0.7

Stop sequences

Enter sequence and press Tab

Top P

1

Zastosowania

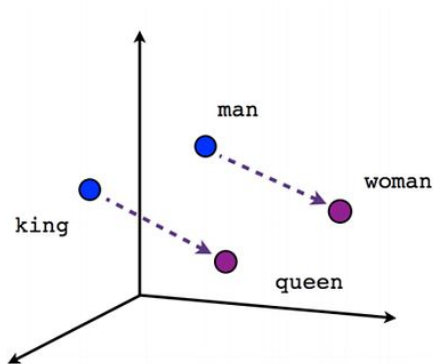
Przedstawione API (jak i cała rodzina GPT) nadaje się do poniższych zastosowań, które mogą mieć zastosowanie w wielu aplikacjach komercyjnych.

- Generowanie podsumowania długiego tekstu
- Poprawa błędów gramatycznych
- Wyjaśnianie skomplikowanych pojęć
- Generowanie szkieletu tekstu
- Generowanie listy pomysłów
- Debugowanie kodu aplikacji
- Szybkie znajdowanie prostych informacji

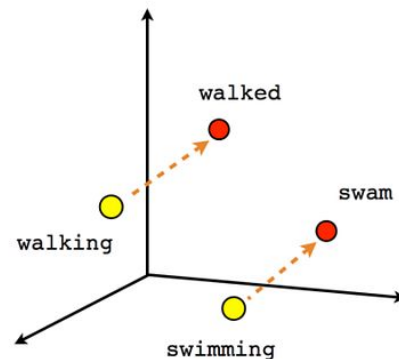
Embeddings API

Embedding to wektor obliczony na podstawie wejściowego tekstu. Takie wektory mają konkretne właściwości:

- podobne pojęcia znajdują się blisko siebie w przestrzeni
- powiązane pojęcia mają podobną odległość i kierunek do innych powiązanych pojęć
- można zmierzyć podobieństwo dwóch wektorów (tekstów) używając zwykłej odległości kosinusowej



Male-Female



Verb tense

Zastosowania

- Wyszukiwanie - rezultaty oceniane przez dopasowanie do zapytania
- Grupowanie - test jest grupowany podobieństwem
- Rekomendacje - przedmioty z podobnymi opisami są polecane
- Wykrywanie anomalii - wykrywanie tekstów które są mało powiązane z resztą korpusu
- Miary różnorodności - mierzenie rozkładów miar podobieństwa w korpusie
- Klasyfikacja - tekst jest klasyfikowany na podstawie największego podobieństwa do etykiety



Pinecone

<https://www.pinecone.io/>

Z tego powodu bardzo popularne stały się wektorowe bazy danych takie jak Pinecone, które są zoptymalizowane pod przechowywanie obliczonych wektorów i uproszczenia wykonywanie operacji opisanych po lewej stronie.

**Obecne
zastosowania**

LangChain

LangChain to framework do budowania aplikacji opartych o LLM (w tym platformę OpenAI, ale nie tylko). Upraszcza i strukturyzuje pracę z modelami za pomocą uniwersalnych modułów, takich jak:

- LLM i propmty
- Łańcuchy
- Augmentacje danych
- Agenci
- Pamięć

Przykłady rozwiązań które można zbudować z wykorzystaniem LangChain:

- Personalny asystent
- Odpowiadanie na pytania z użyciem zewnętrznych danych
- Odpytywanie API
- Podsumowywanie długich dokumentów

LangChain - Odpowiadanie na pytania

Bardzo często pojawia się problem automatycznego odpowiadania na pytania na które odpowiedzi znajdują się w istniejących zasobach.

Może to być Wikipedia, albo dokumentacja projektu. Znalezienie odpowiedzi czasami zajmuje dużo czasu. Niestety modele LLM nie są zwykle uczone na danych wewnątrz firmowych (no chyba, że są hostowane na GitHub ;))

LangChain pozwala nam zbudować aplikację, która będzie w stanie odpowiadać na pytania naszego zespołu i klientów bez trenowania modelu.

Przygotowanie danych

Dla każdego dokumentu liczymy embeddingi słów i zapisujemy dokumenty i embeddingi w wektorowej bazie danych.

Potok danych

Dla przychodzącego zapytania liczymy embedding i szukamy najbardziej podobnych dokumentów w bazie (np.: top 5)

Umieszczamy zawartość tych dokumentów w prompcie razem z pytaniem i odpytujemy LLM.

LangChain - Podsumowanie długiego tekstu

LLM są świetne do generowania podsumowania tekstu. Jednakże mają ograniczenia, np.: GPT 3.5 ma maksymalne okno percepcyjne o długości 4096 tokenów (token \neq słowo).

Co zrobić kiedy mamy bardzo długi tekst (np.: 10 stron A4), który chcemy podsumować?

Zwykły model LLM podsumuje tylko ostatnie 4096 tokenów co nie jest poprawną odpowiedzią.

Rozwiązaniem jest podzielenie tekstu na kawałki. Dobrą abstrakcją jest wizualizacja podziału jak podział na rozdziały, podrozdziały, sekcje itp.

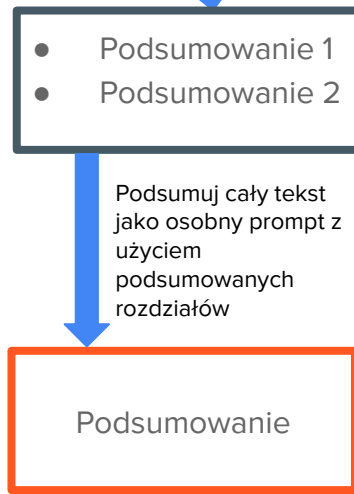
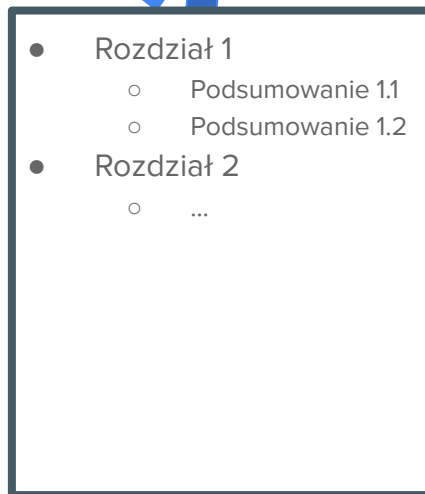
Model odpytujemy rekurencyjnie od wyjaśnienia najmniejszych bloków, aż do wyjaśnienia całego tekstu na podstawie pomniejszych podsumowań.

LangChain - Podsumowanie długiego tekstu

Podsumuj każdą sekcję jako osobny prompt

Podsumuj każdy podrozdział jako osobny prompt z użyciem podsumowanych sekcji

Podsumuj każdy rozdział jako osobny prompt z użyciem podsumowanych podrozdziałów



LangChain - przykład kodu dla API

```
!pip install arxiv
```

```
from langchain.chat_models import ChatOpenAI
from langchain.agents import load_tools, initialize_agent, AgentType
```

```
llm = ChatOpenAI(temperature=0.0)
tools = load_tools(
    ["arxiv"],
)
```

```
agent_chain = initialize_agent(
    tools,
    llm,
    agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    verbose=True,
)
```

```
agent_chain.run(
    "What's the paper 1605.08386 about?",
)
```

```
> Entering new AgentExecutor chain...
I need to use Arxiv to search for the paper.
```

```
Action: Arxiv
```

```
Action Input: "1605.08386"
```

```
Observation: Published: 2016-05-26
```

```
Title: Heat-bath random walks with Markov bases
```

```
Authors: Caprice Stanley, Tobias Windisch
```

```
Summary: Graphs on lattice points are studied whose edges come from a finite set of allowed moves of arbitrary length. We show that the diameter of these graphs on fibers of a fixed integer matrix can be bounded from above by a constant. We then study the mixing behaviour of heat-bath random walks on these graphs. We also state explicit conditions on the set of moves so that the heat-bath random walk, a generalization of the Glauber dynamics, is an expander in fixed dimension.
```

```
Thought:The paper is about heat-bath random walks with Markov bases on graphs of lattice points.
```

```
Final Answer: The paper 1605.08386 is about heat-bath random walks with Markov bases on graphs of lattice points.
```

```
> Finished chain.
```

```
'The paper 1605.08386 is about heat-bath random walks with Markov bases on graphs of lattice points.'
```

Auto GPT

Autonomiczny GPT - eksperyment i technologiczne demo pokazujące jak LLM mogą rozwiązywać skomplikowane problemy bez pomocy z zewnątrz.

Aplikacja posiada dostęp do internetu i różnych narzędzi, które może wykorzystać aby osiągnąć dany cel.

Auto GPT rozbija skomplikowany cel na szereg mniejszych, które jest w stanie wykonać z użyciem dostępnych mu narzędzi.

```
PS D:\Auto-GPT> python -m autogpt --continuous
Continuous Mode: ENABLED
WARNING: Continuous mode is not recommended. It is potentially dangerous and may cause your AI to run forever or carry out actions you would not usually authorise. Use at your own risk.
Welcome back! Would you like me to return to being AutoGPT-Demo? Continue with the last settings?
Name: AutoGPT-Demo
Role: an ai designed to teach me about auto gpt
Goals: ['search auto gpt', 'find the github and figure out what the project is', 'explain what auto gpt is in a file named autogpt.txt', 'terminate']
Continue (y/n): y
Using memory of type: LocalCache
AUTOGPT-DEMO THOUGHTS: I think the first step should be to use the 'google' command to search for 'Auto GPT'.
REASONING: This will help us gather more information about Auto GPT and we can proceed with identifying the relevant GitHub project.
PLAN:
- Use 'google' to search for 'Auto GPT'
- Browse relevant websites to find the GitHub project
- Write a document explaining what Auto GPT is
CRITICISM: I need to be sure to remain focused and efficient in my use of the 'google' command to minimize the number of steps needed to identify the relevant GitHub project and answer the key questions.
```



Lepsze zrozumienie biznesu klienta

W celu wsparcia klientów Stripe przegląda ich strony internetowe i robi podsumowanie ich biznesu.

Okazuje się, że GPT jest w stanie generować lepsze podsumowanie biznesu klienta, niż ludzie.

Odpowiadanie na problemy

Zespół do spraw pomocy, często musiał odpowiadać na oczywiste pytania, na które odpowiedzi były zawarte w dokumentacji. Znalezienie lub nakierowanie klienta na odpowiedź trwało bardzo długo.

GPT jest w stanie automatycznie odpowiadać na takie pytania po podłączeniu dokumentacji do modelu.

Be My Eyes

Od 2012 roku Be My Eyes tworzy technologię dla społeczności ponad 250 milionów osób niewidomych lub niedowidzących. Duński startup łączy osoby niewidome lub niedowidzące z wolontariuszami, którzy pomagają im w setkach codziennych zadań, takich jak identyfikacja produktu lub nawigacja po lotnisku.

Dzięki nowym możliwościom wprowadzania danych wizualnych GPT-4 (w wersji zapoznawczej), Be My Eyes rozpoczęło opracowywanie Virtual Volunteer™ opartego na GPT-4 w aplikacji Be My Eyes, która może generować ten sam poziom kontekstu i zrozumienia, co ochotnik-człowiek.

Koniec

Dziękuję za uwagę

Kontakt

- LinkedIn: <https://www.linkedin.com/in/bartosz-mikulski/>
- GitHub: <https://github.com/BartMiki>
- Kontakt: bartosz.mikulski.praca@gmail.com

