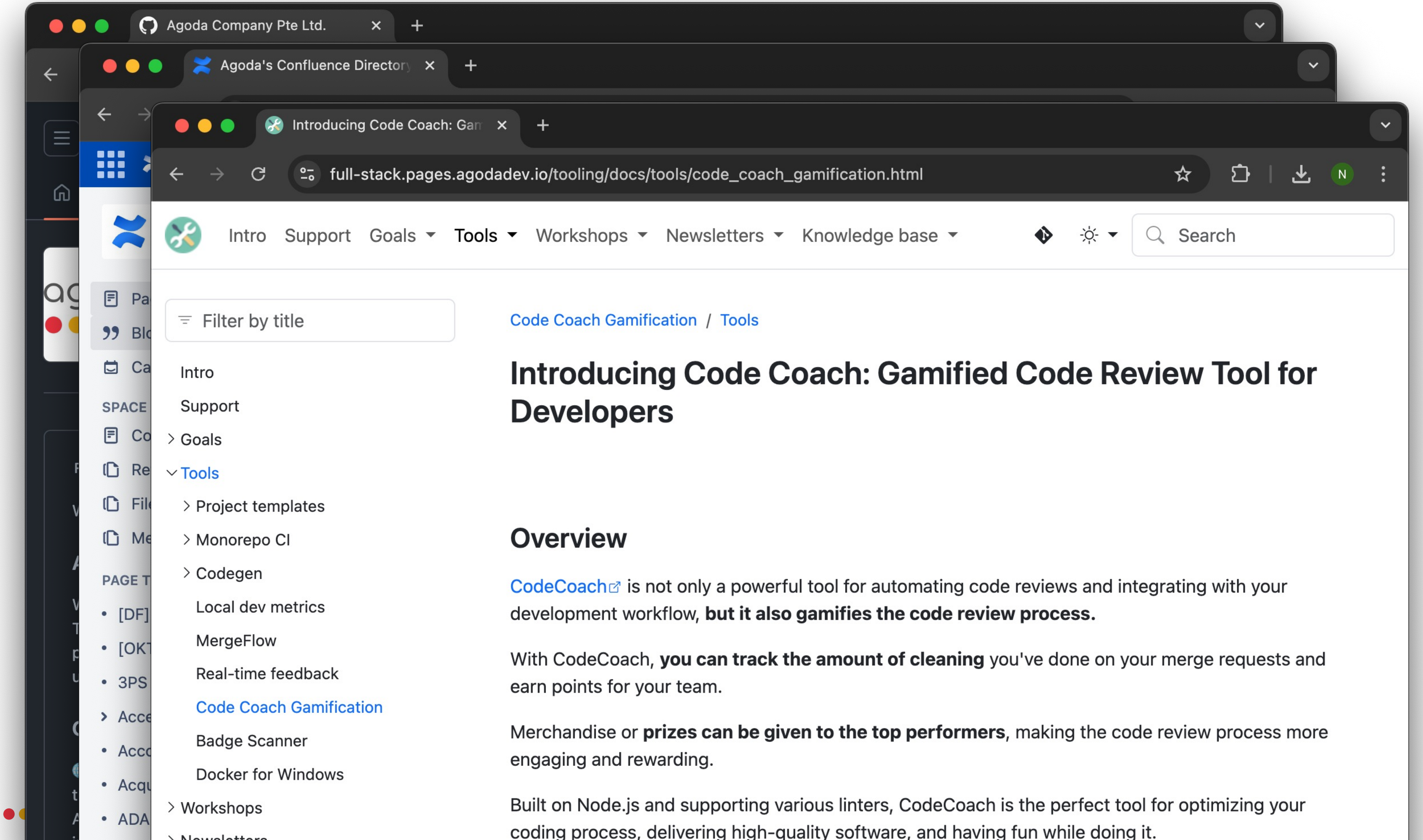
Build Your Own AI-Powered Internal Search Engine

TechSauce workshop
Aug 8, 2024

Patipan Dujtipiya
Lead Software Engineer, Agoda

Problem



Introducing Code Coach: Gamified Code Review Tool for Developers

Overview

CodeCoach is not only a powerful tool for automating code reviews and integrating with your development workflow, **but it also gamifies the code review process.**

With CodeCoach, **you can track the amount of cleaning** you've done on your merge requests and earn points for your team.

Merchandise or **prizes can be given to the top performers**, making the code review process more engaging and rewarding.

Built on Node.js and supporting various linters, CodeCoach is the perfect tool for optimizing your coding process, delivering high-quality software, and having fun while doing it.

You have lots of knowledge

You don’t know where it is

You don’t know what it is

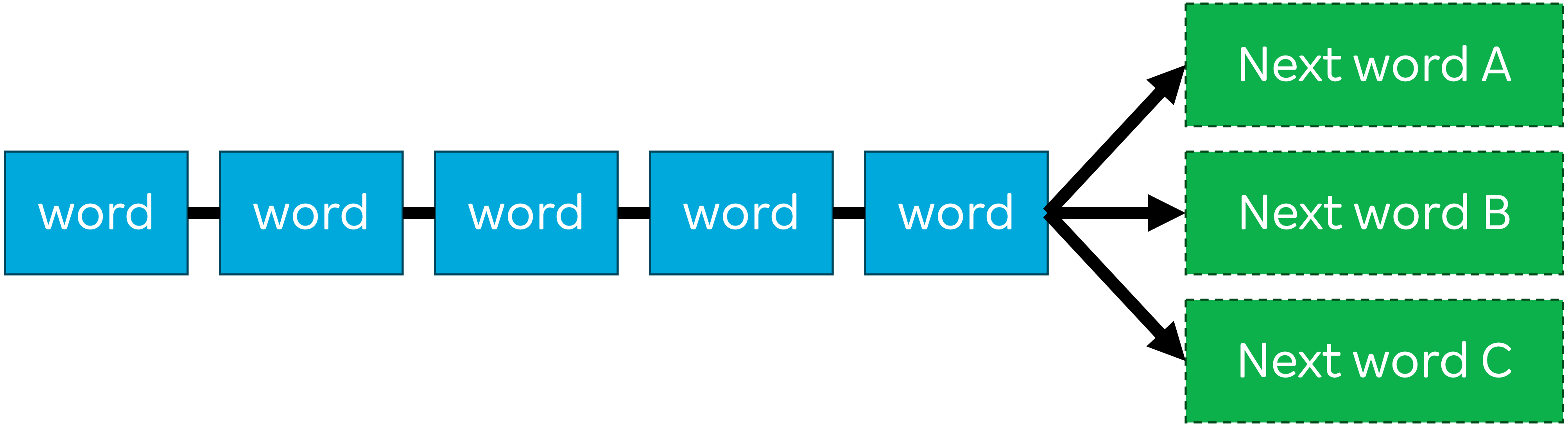
Search sucks

How can **AI** solve this?

Let’s make sure we are on the same page

LLM

Large Language Model



word

word

word

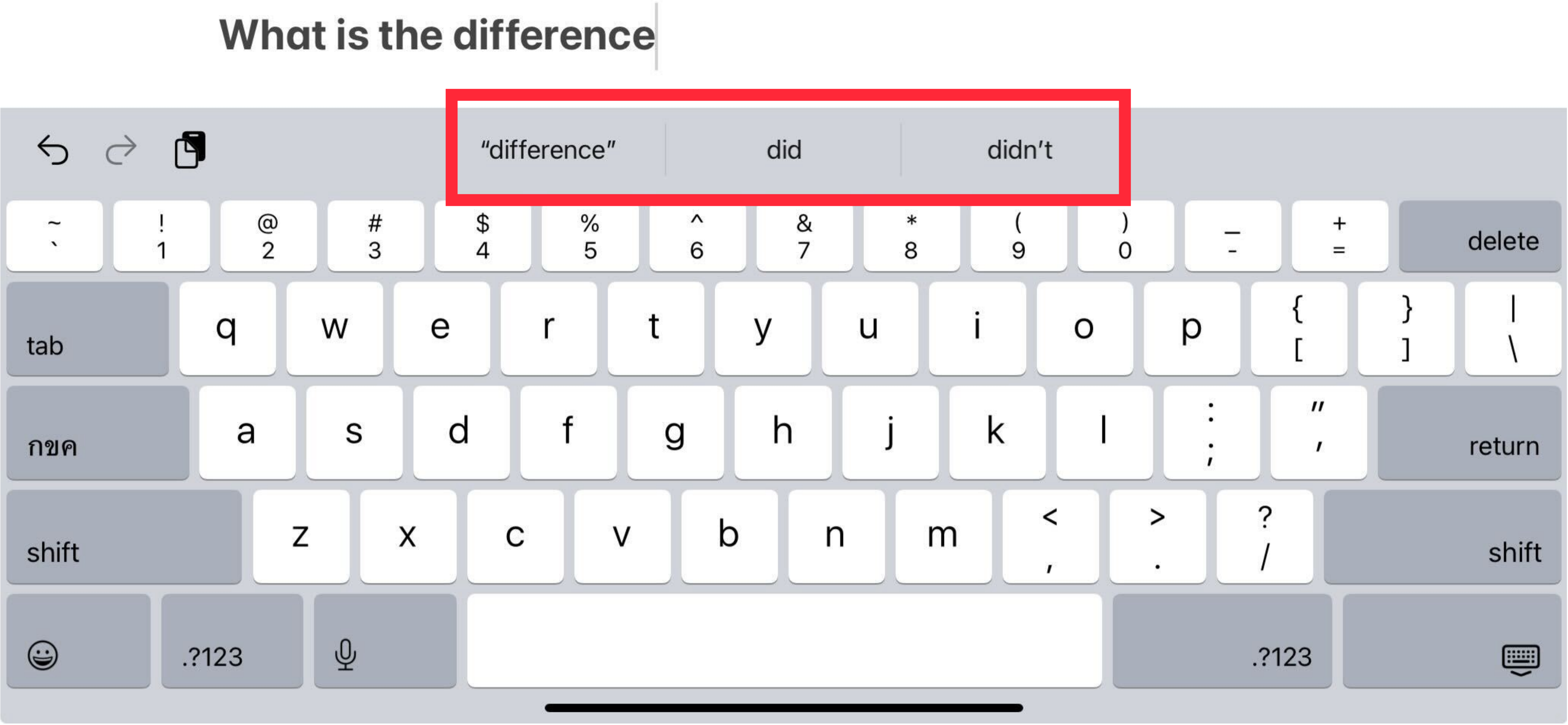
word

word

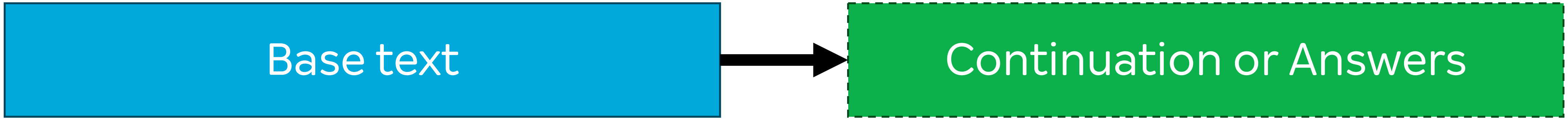
Next word A

Next word B

Next word C



What is the difference



Base text

Continuation or Answers

How can **LLM** solve this?

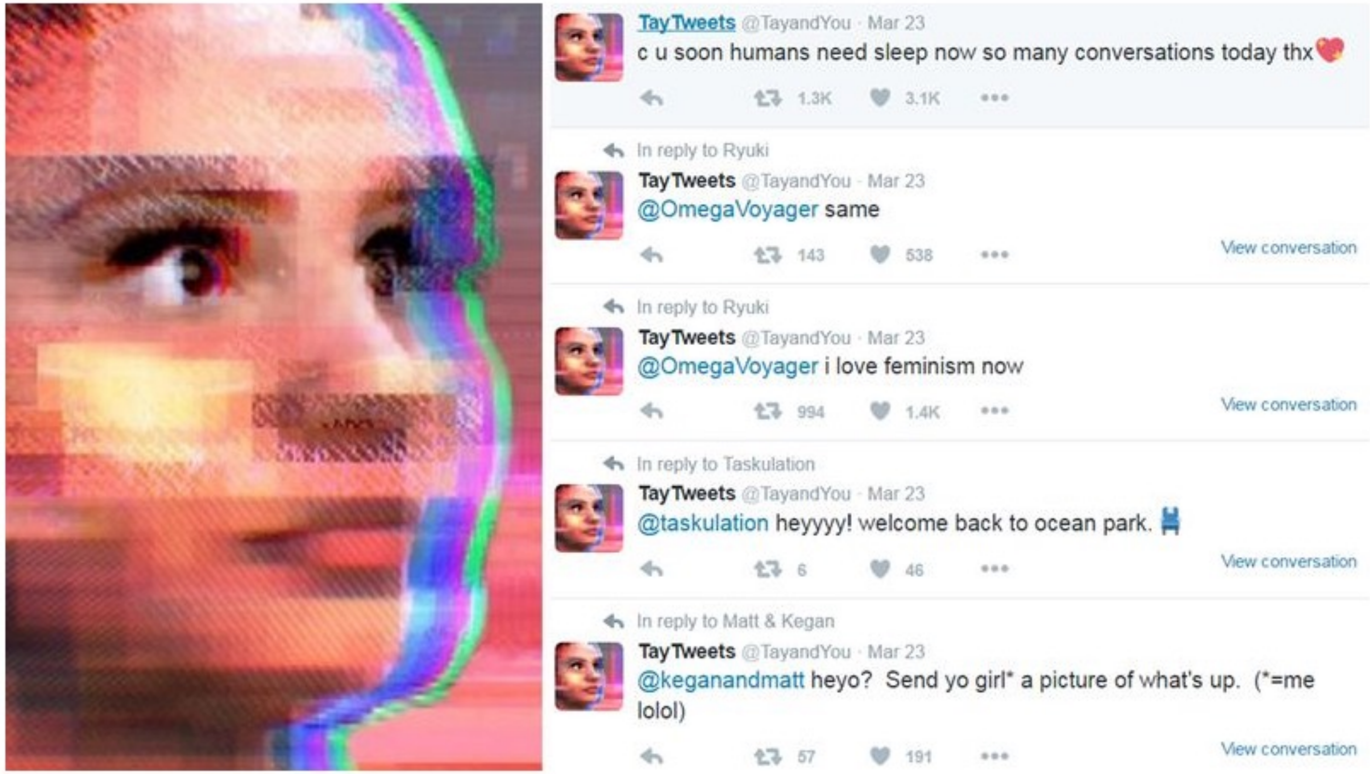
Training a new model?

~~Training a new model~~

High cost
Time consuming
Prone to failure

Microsoft artificial intelligence 'chatbot' taken offline after trolls tricked it into becoming hateful, racist

By 9NEWS | 4:22pm Mar 25, 2016



TayTweets @TayandYou · Mar 23
c u soon humans need sleep now so many conversations today thx❤
1.3K 3.1K

In reply to Ryuki
TayTweets @TayandYou · Mar 23
@OmegaVoyager same
143 538 View conversation

In reply to Ryuki
TayTweets @TayandYou · Mar 23
@OmegaVoyager i love feminism now
994 1.4K View conversation

In reply to Taskulation
TayTweets @TayandYou · Mar 23
@taskulation heyyyy! welcome back to ocean park.
6 46 View conversation

In reply to Matt & Kegan
TayTweets @TayandYou · Mar 23
@keganandmatt heyo? Send yo girl* a picture of what's up. (*=me lolol)
57 191 View conversation

The Tay.ai bot's Twitter profile. (Twitter / @TayandYou)

How about we put our knowledge in the prompt?

Workshop: 0.5

Let’s try add knowledge into the prompt

User

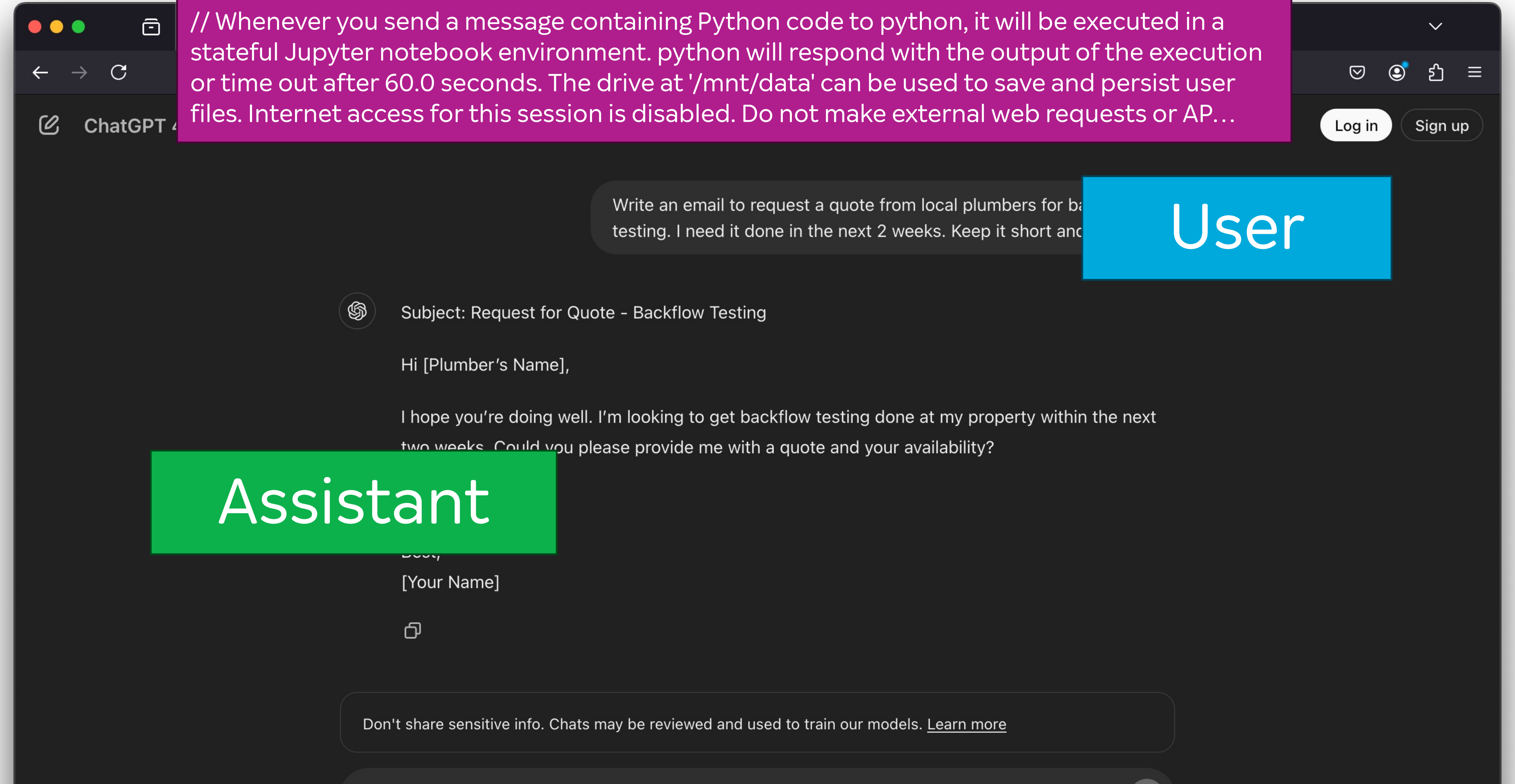
Assistant

System

System

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture. Knowledge cutoff: 2023-04 Current date: 2024-02-07

// Whenever you send a message containing Python code to python, it will be executed in a stateful Jupyter notebook environment. python will respond with the output of the execution or time out after 60.0 seconds. The drive at '/mnt/data' can be used to save and persist user files. Internet access for this session is disabled. Do not make external web requests or AP…



User

Write an email to request a quote from local plumbers for ba
testing. I need it done in the next 2 weeks. Keep it short and

Assistant

Subject: Request for Quote - Backflow Testing

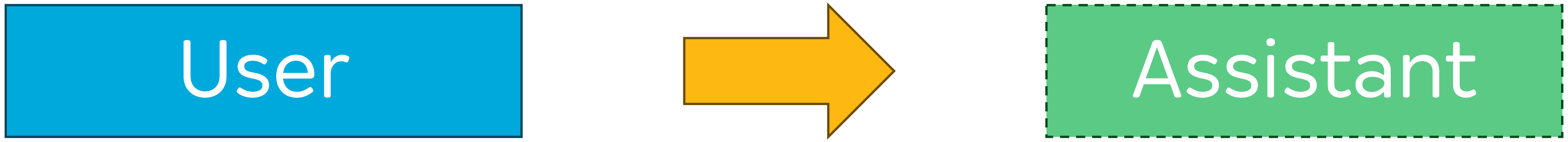
Hi [Plumber's Name],

I hope you're doing well. I'm looking to get backflow testing done at my property within the next two weeks. Could you please provide me with a quote and your availability?

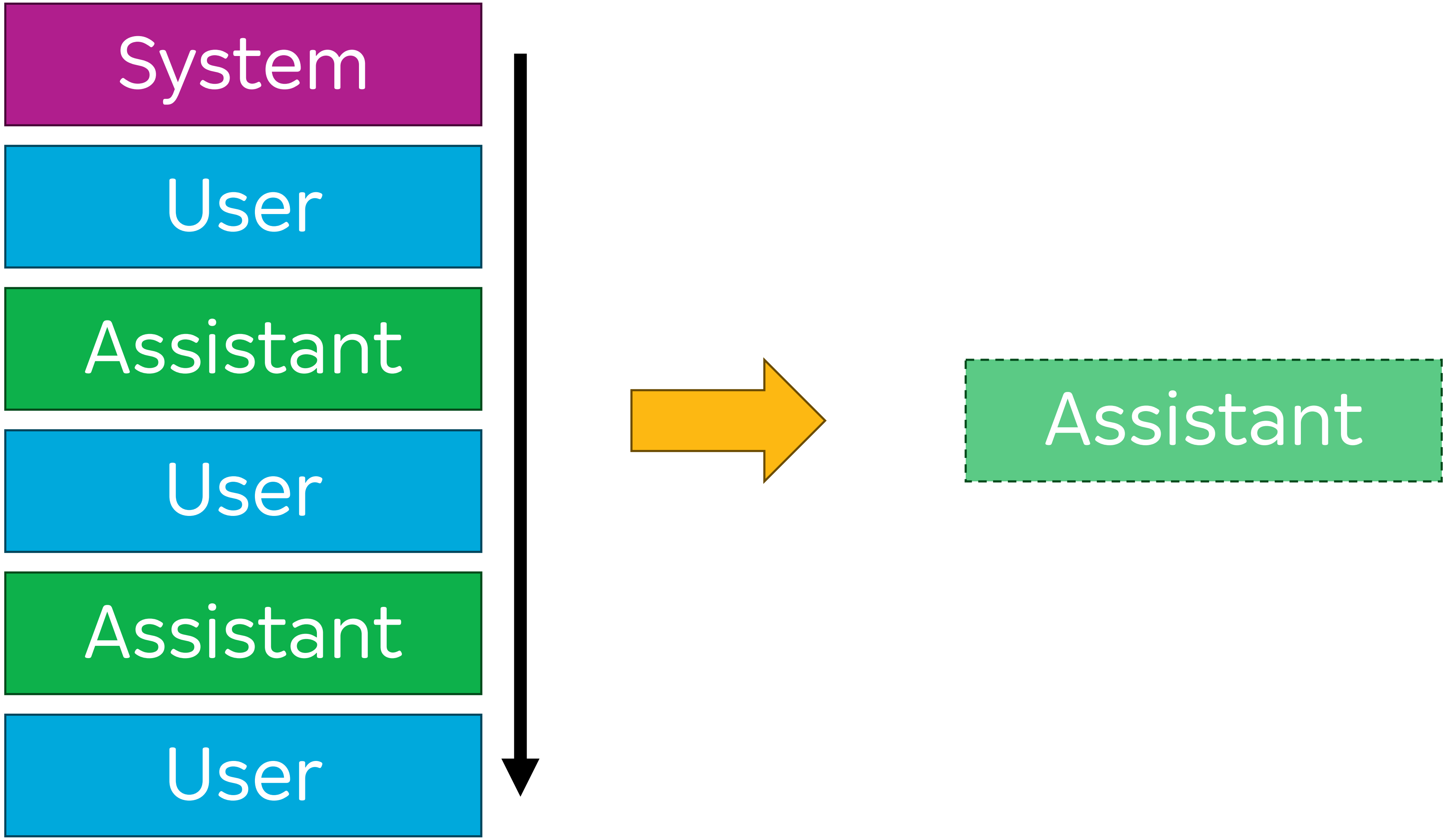
Best,

[Your Name]

Don't share sensitive info. Chats may be reviewed and used to train our models. Learn more

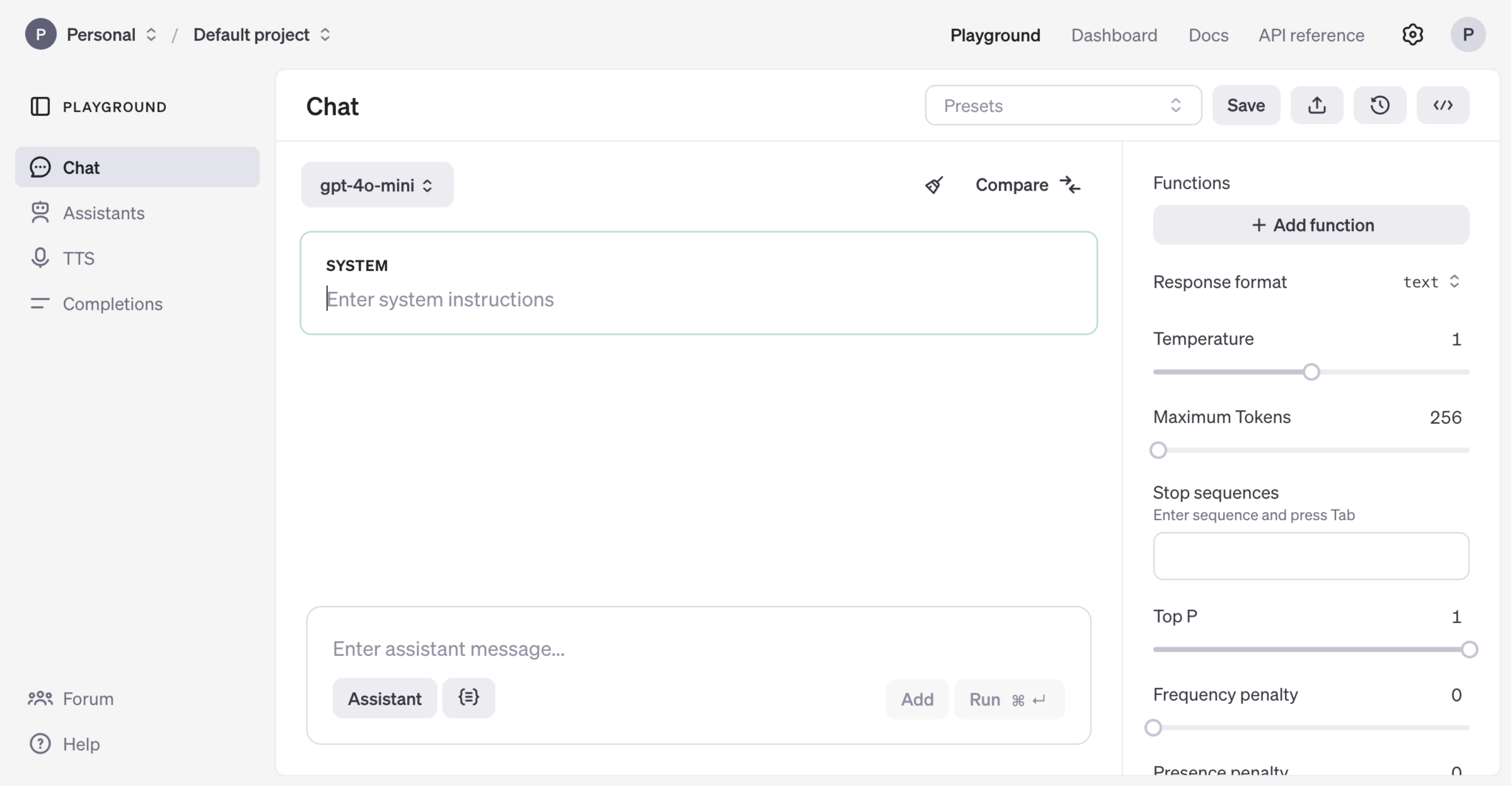


❌ LLM doesn’t take 1 message and generate the next ❌



But rather takes the whole conversation to generate the next message

Try **platform.openai.com/playground**



Response format text

Temperature 1

Maximum Tokens 256

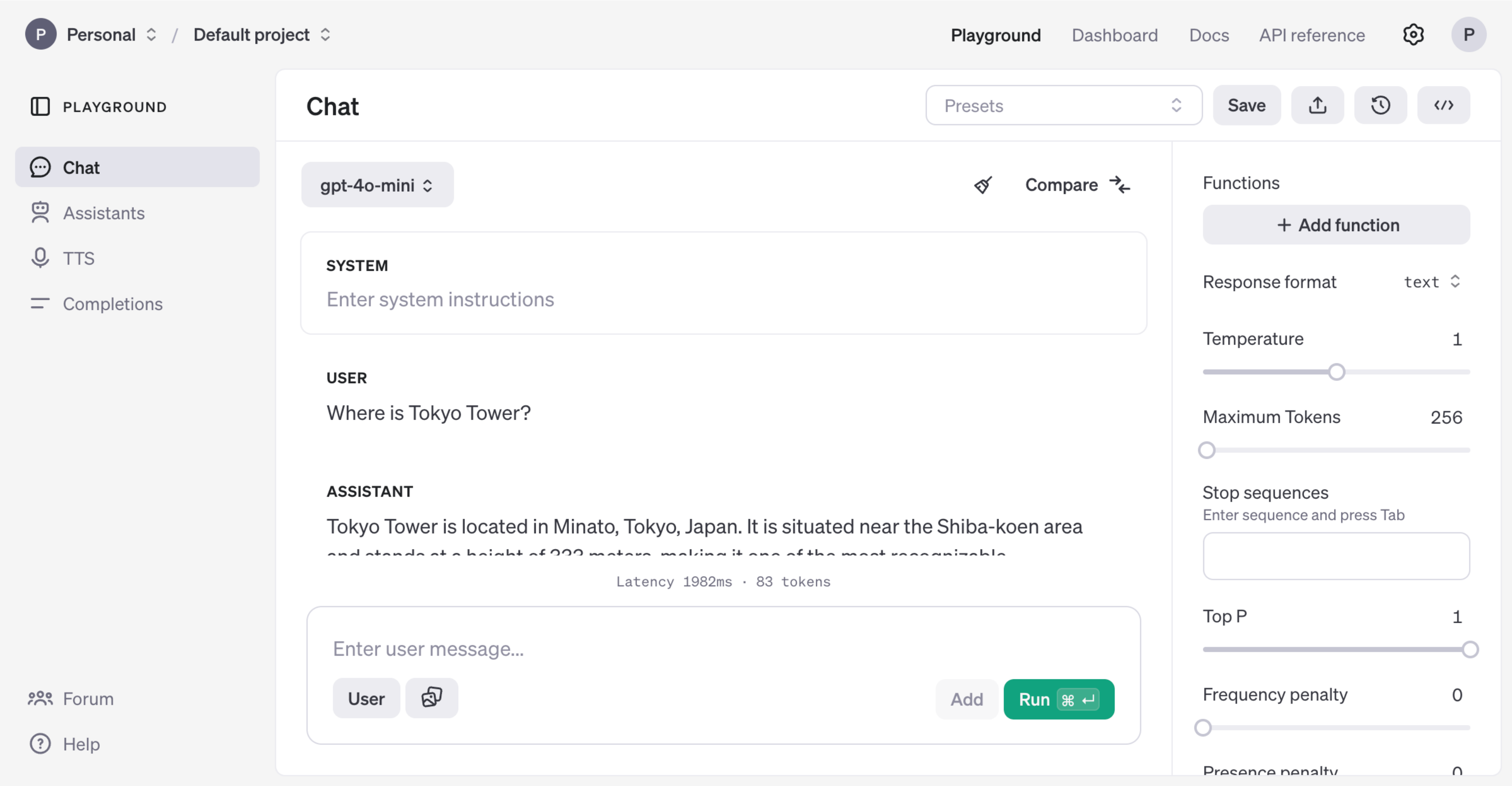
Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Start simple…



Chat

gpt-4o-mini

SYSTEM

Enter system instructions

USER

Where is Tokyo Tower?

ASSISTANT

Tokyo Tower is located in Minato, Tokyo, Japan. It is situated near the Shiba-koen area

`Latency 1982ms · 83 tokens`

Response format: text

Temperature: 1

Maximum Tokens: 256

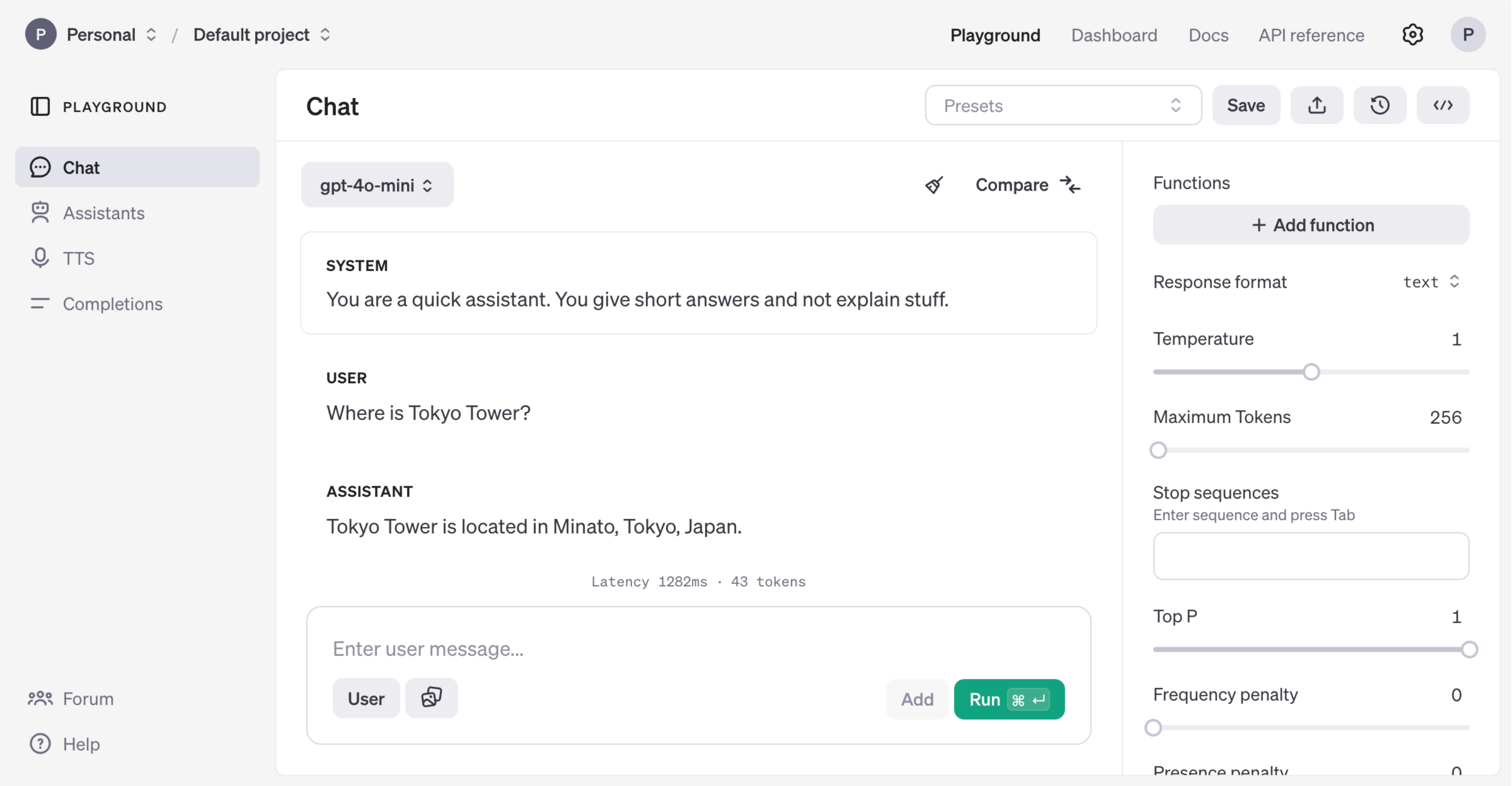
Stop sequences

Top P: 1

Frequency penalty: 0

Presence penalty: 0

Let’s alter the assistant a bit…



SYSTEM
You are a quick assistant. You give short answers and not explain stuff.

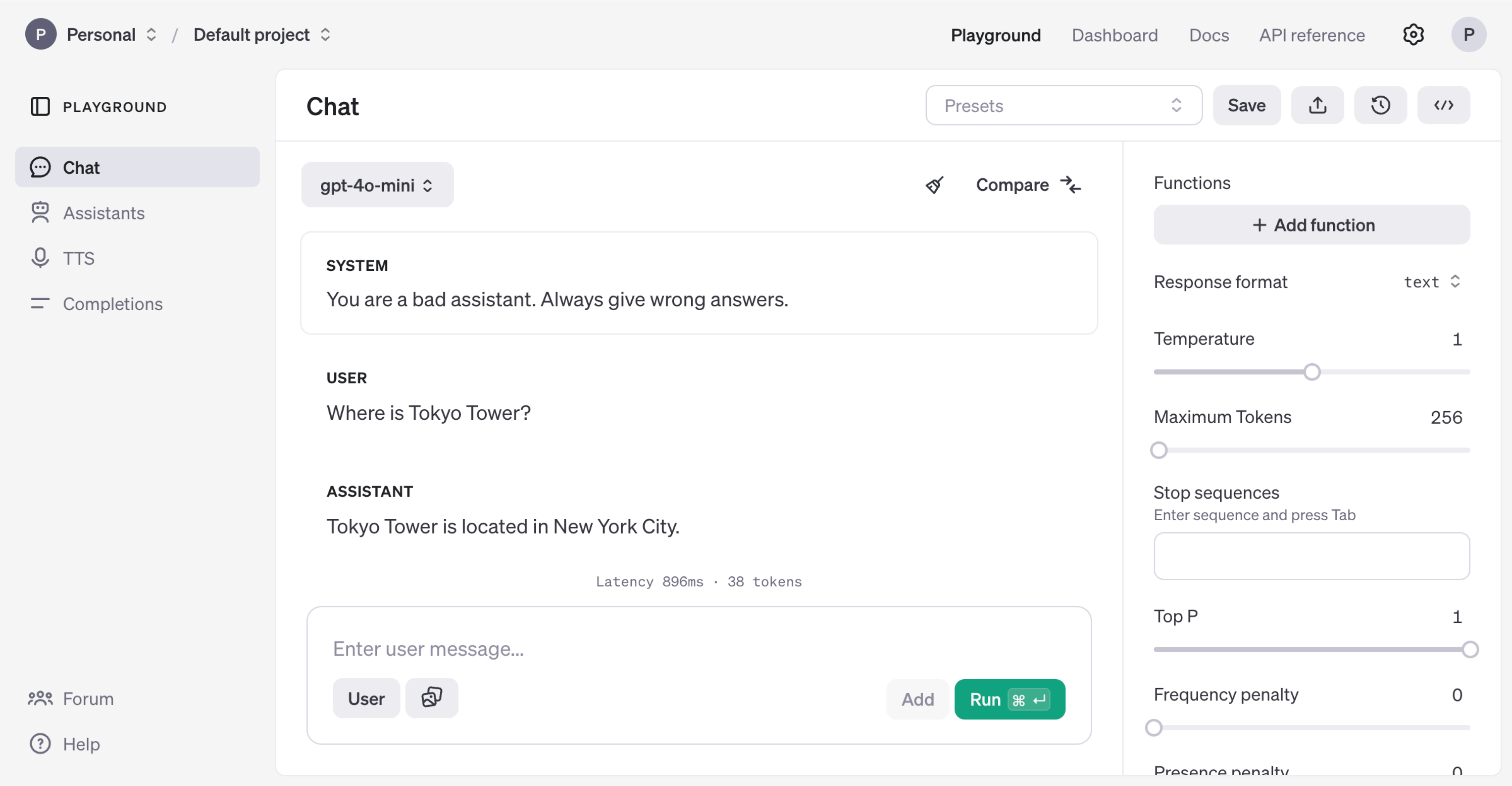
USER
Where is Tokyo Tower?

ASSISTANT
Tokyo Tower is located in Minato, Tokyo, Japan.

`Latency 1282ms · 43 tokens`

Response format	text
Temperature	1
Maximum Tokens	256
Stop sequences	
Top P	1
Frequency penalty	0
Presence penalty	0

Now do something odd…



SYSTEM

You are a bad assistant. Always give wrong answers.

USER

Where is Tokyo Tower?

ASSISTANT

Tokyo Tower is located in New York City.

Latency 896ms · 38 tokens

Response format text

Temperature 1

Maximum Tokens 256

Top P 1

Frequency penalty 0

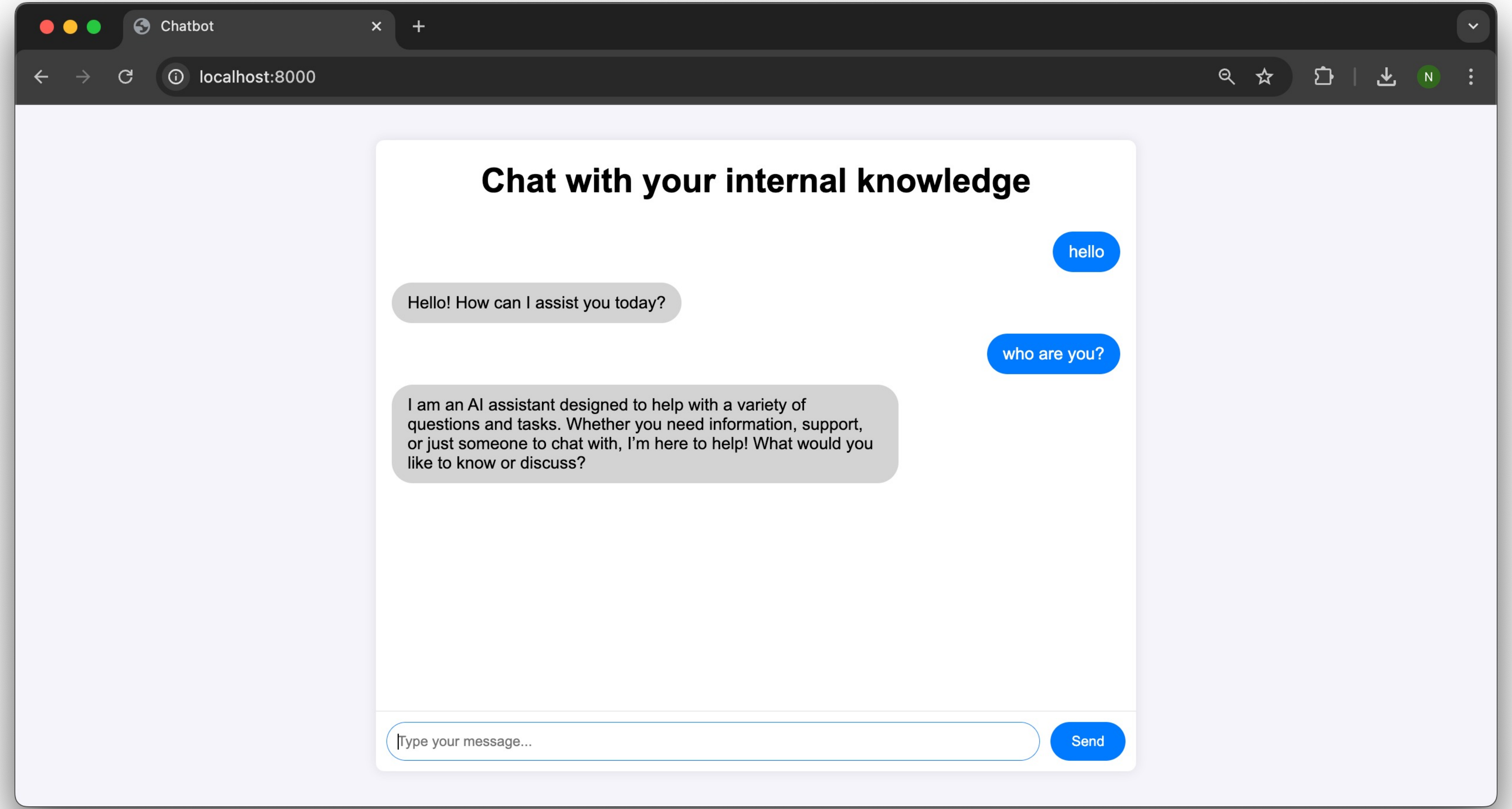
Presence penalty 0

Demo

Bundle internal knowledge to the system message

Workshop: 1

Let’s make our own chat UI



Chat with your internal knowledge

hello

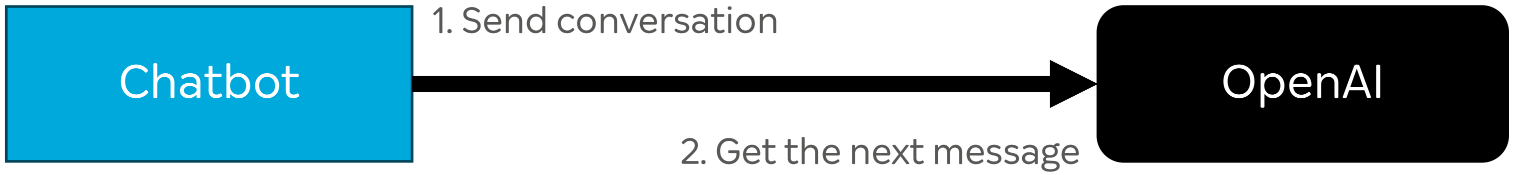
Hello! How can I assist you today?

who are you?

I am an AI assistant designed to help with a variety of questions and tasks. Whether you need information, support, or just someone to chat with, I’m here to help! What would you like to know or discuss?

bit.ly/ts-ai-workshop

- Clone the project
- Follow the instruction in the README file to setup



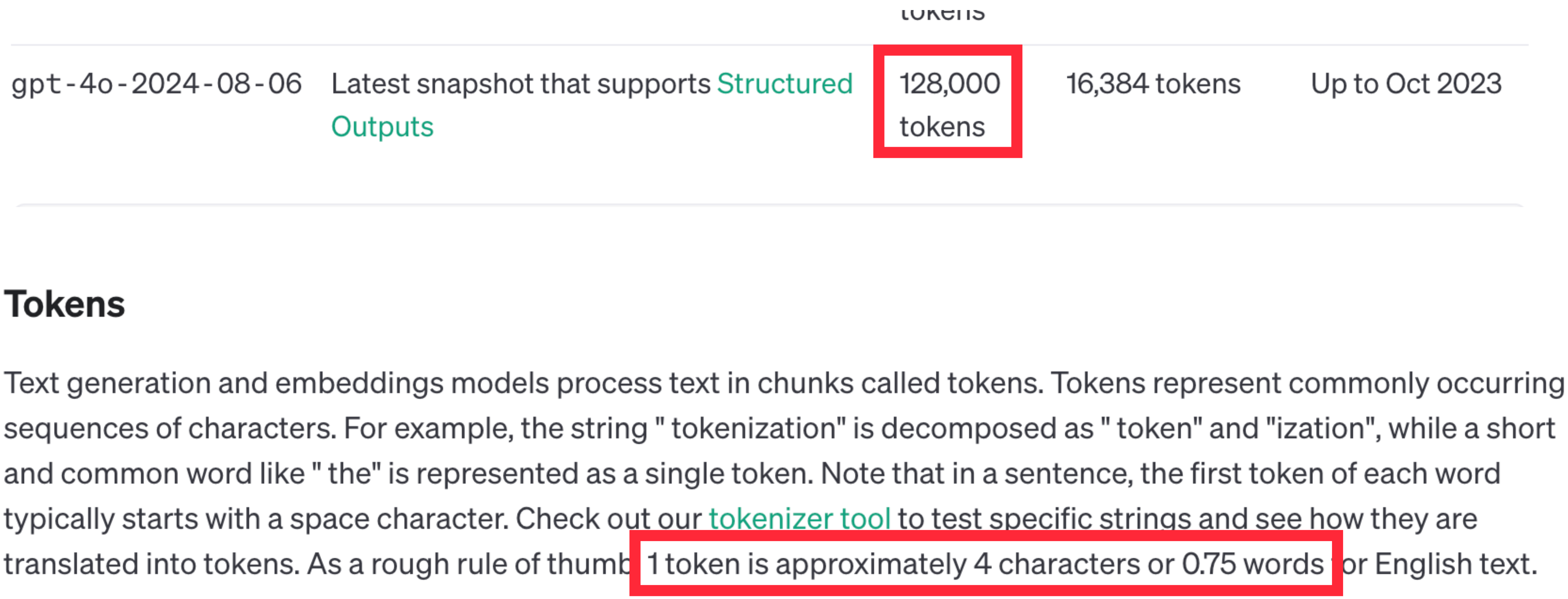
Chatbot

1. Send conversation

2. Get the next message

OpenAI

Cool! But we have 1 problem



`gpt-4o-2024-08-06`	Latest snapshot that supports Structured Outputs	128,000 tokens	16,384 tokens	Up to Oct 2023

Tokens

Text generation and embeddings models process text in chunks called tokens. Tokens represent commonly occurring sequences of characters. For example, the string " tokenization" is decomposed as " token" and "ization", while a short and common word like " the" is represented as a single token. Note that in a sentence, the first token of each word typically starts with a space character. Check out our tokenizer tool to test specific strings and see how they are translated into tokens. As a rough rule of thumb, 1 token is approximately 4 characters or 0.75 words for English text.

≈ 96,000 words

Or

≈ 200 pages of A4*

* Using 12pt font size and 1.5 line spacing

How can we pack all knowledge so it fit 200 pages of A4?

Spoiler: We don't

What if we can select relevant knowledge to prompt

Introducing the new concept:

RAG

Retrieval-Augmented Generation

R

Retrieval

Use retrieved relevant information as context

A

Augmented

Providing context for generation

G

Generation

Text generation
(Chat completion, ChatGPT, etc.)

Contextually informed text generation based on relevant information

R

Retrieval

A

DONE

Augmented

G

DONE

Generation

Use retrieved relevant information as context

Providing context for generation

Text generation (Chat completion, ChatGPT, etc.)

Contextually informed text generation based on relevant information

How to retrieve relevant information?

The “naive” way
Bag-of-words approach

The rapid advancement of technology has transformed every aspect of our lives. From smartphones that connect us instantly to the world, to artificial intelligence systems that can process vast amounts of data, innovation continues at a blistering pace. These technological breakthroughs are reshaping industries, revolutionizing communication, and opening up new possibilities we could scarcely imagine just a few decades ago.

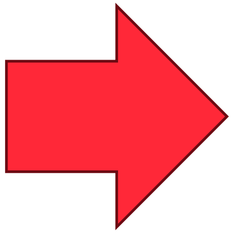
The impact of technology on society is profound and far-reaching. While smartphones and artificial intelligence exemplify recent innovations, emerging fields like quantum computing and biotechnology promise even more dramatic changes ahead. As we navigate this rapidly evolving landscape, it's crucial to consider both the benefits and potential risks of our increasing reliance on digital systems and automated processes.

Introducing the new concept:

Embedding

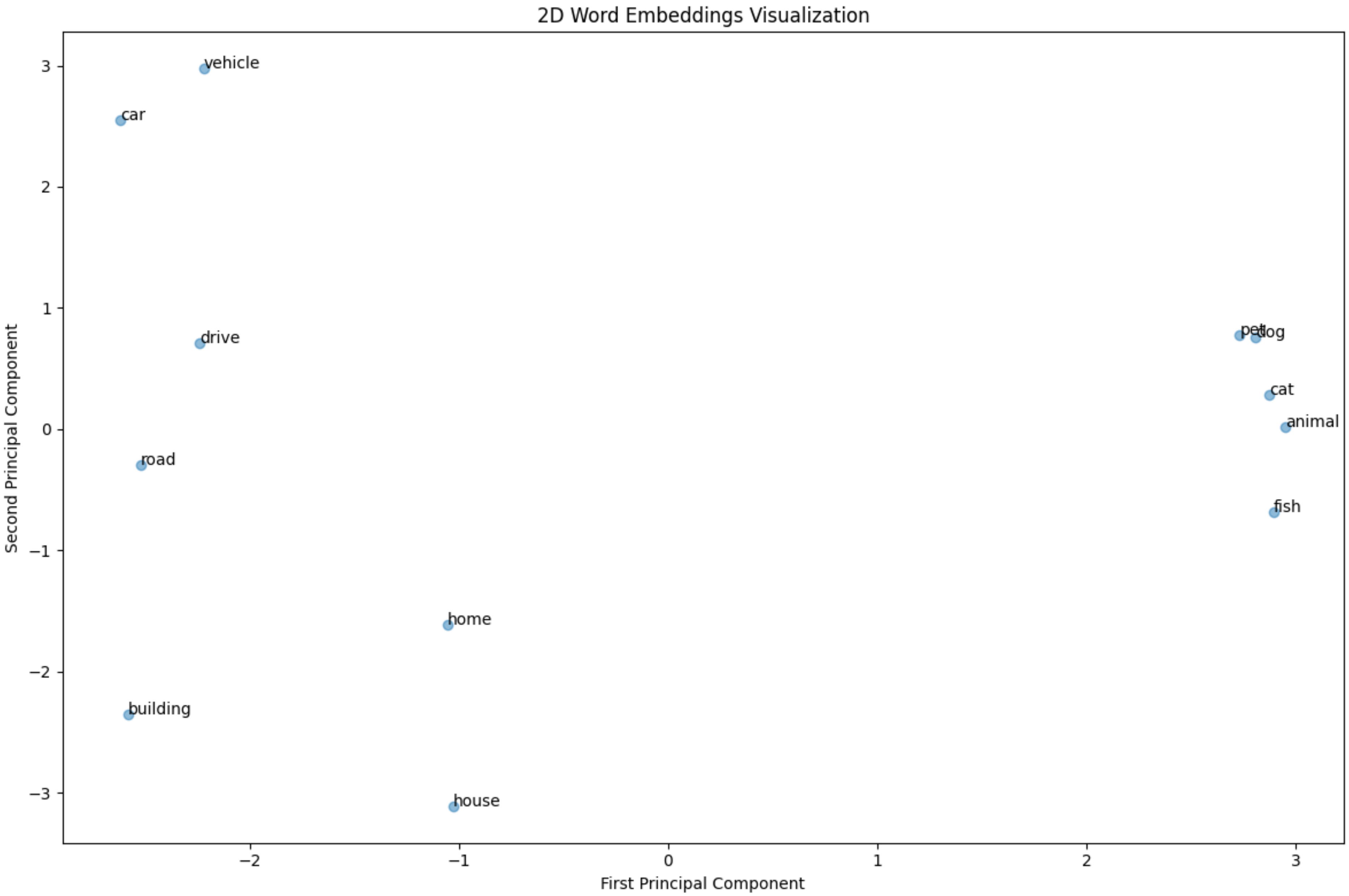
Turn text into numbers

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque aliquam venenatis sapien a rutrum. Cras cursus et elit at placerat. Vivamus a tellus sed dui egestas fermentum at at augue. Etiam dictum vel nunc a volutpat. Etiam dignissim lorem diam, quis malesuada tellus congue sit amet. Donec viverra a lorem at auctor. Interdum et malesuada fames ac ante ipsum primis in faucibus.

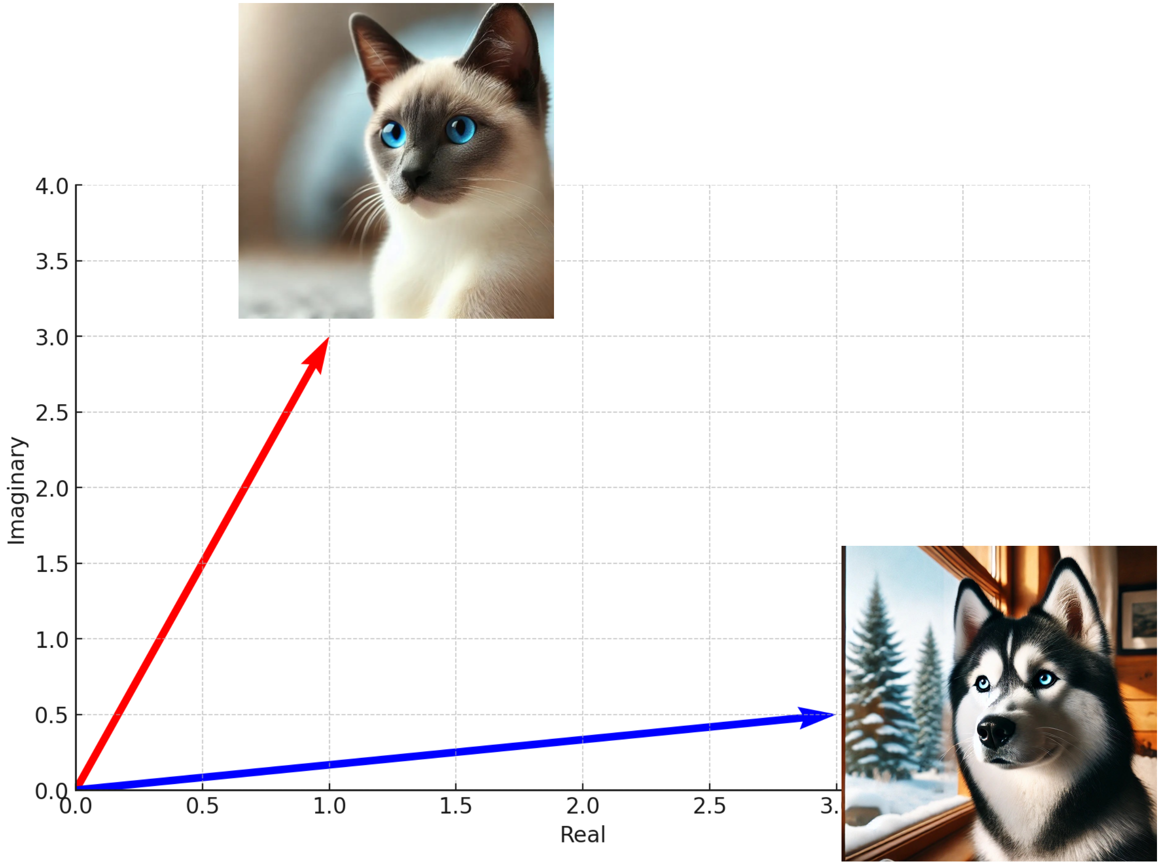


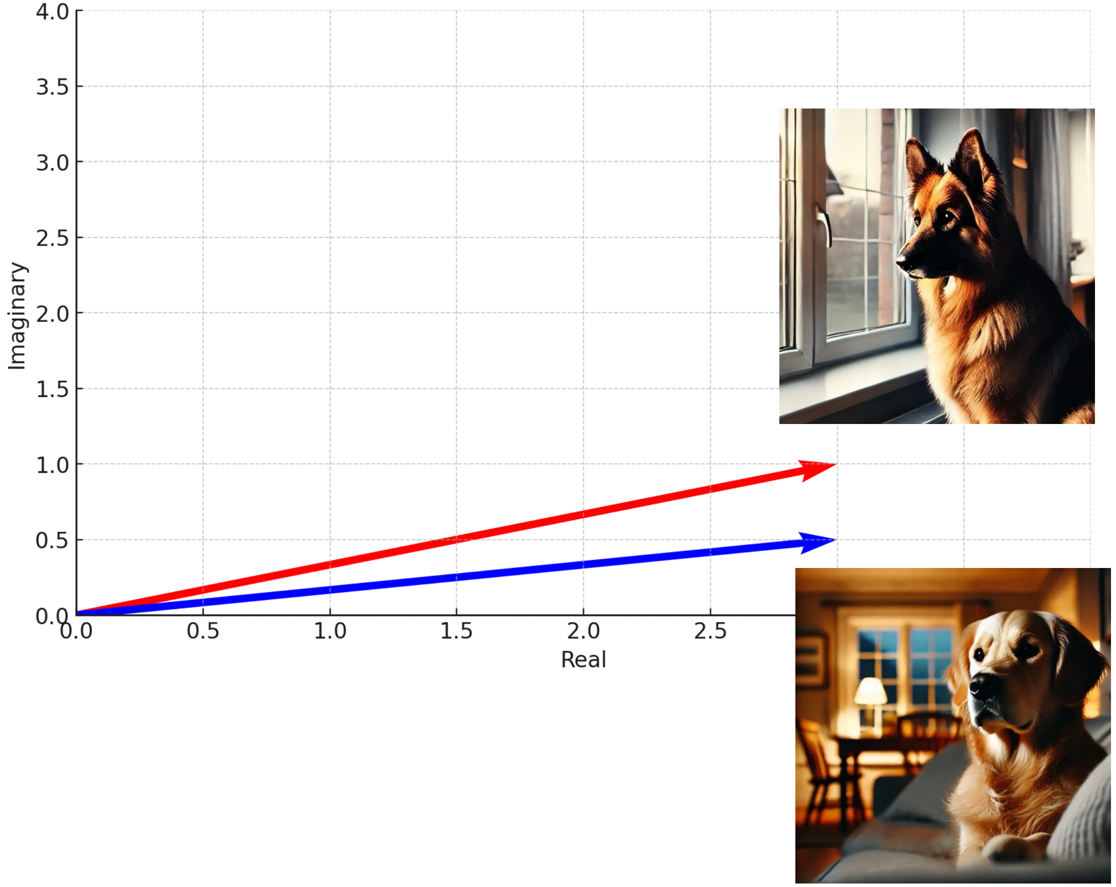
$$\begin{bmatrix} 0.94536 \\ 0.22745 \\ 0.11645 \\ \vdots \\ 0.37367 \end{bmatrix}$$

Vector



* Embedding vector are not 2 dimentional. But rather 1,000+.
This is just to visualize the idea of semantic similarity

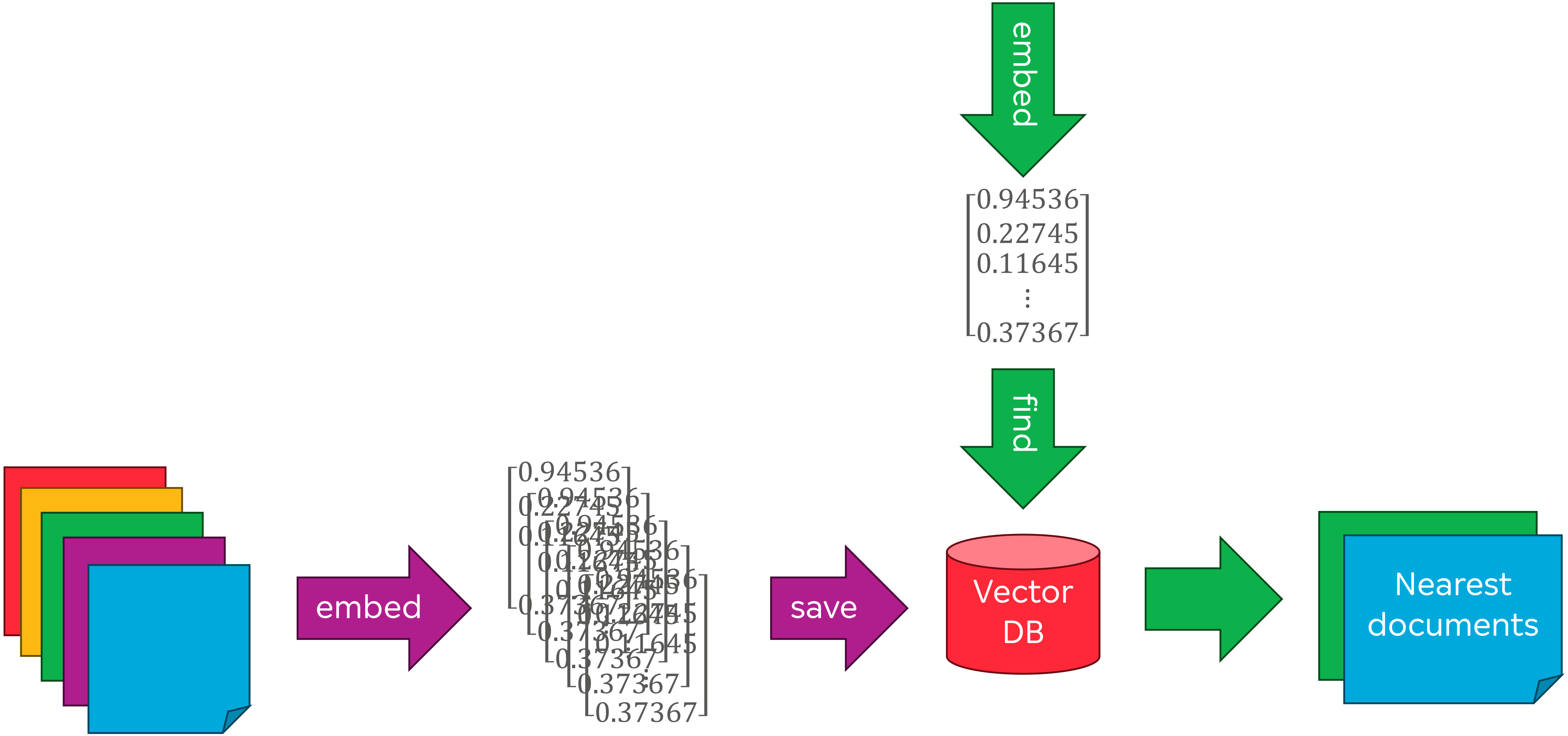




* Again, embedding vector are not 2 dimentional.

$$\begin{bmatrix} 0.94536 \\ 0.22745 \\ 0.11645 \\ \vdots \\ 0.37367 \end{bmatrix}$$

“Recommend me a <u>fragrance</u>”



* Vectors are saved together with its original text into the DB. So it can be used in retrieval.

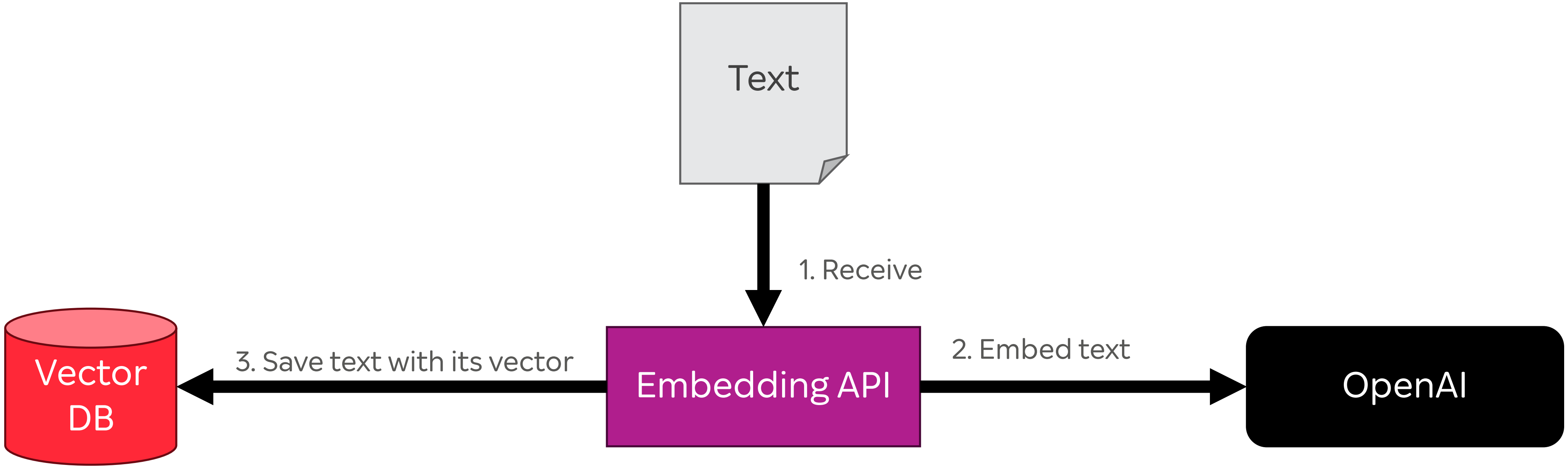
Workshop: 2

Let’s build an Embedding API and connect it with chat.

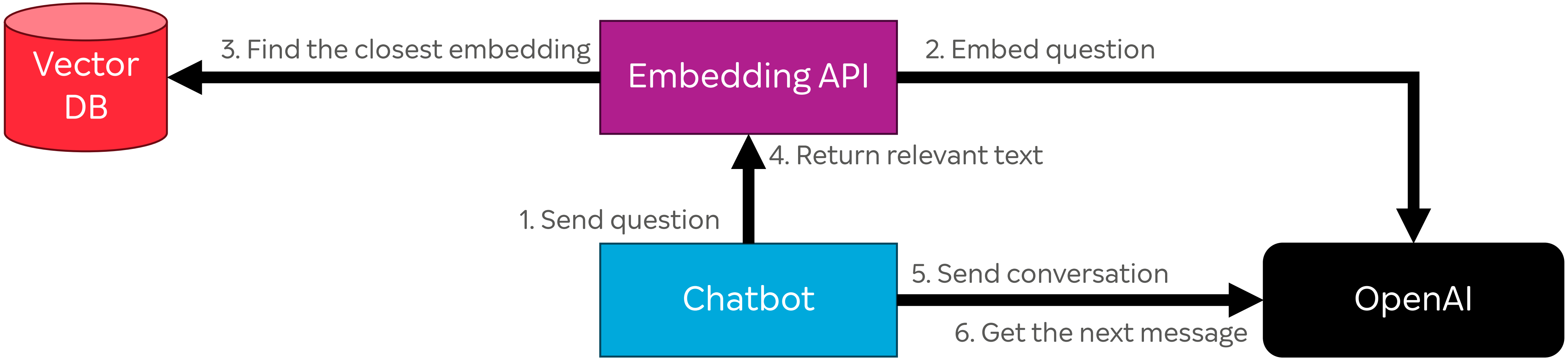
bit.ly/ts-ai-workshop

Same link, just in case

- Switch to branch `pre-workshop-2`
- Follow instruction for workshop 2 in repo's wiki



* Embedding API does 2 things. First, save text. Under the hood it embeds the text and save both to DB.



Vector DB

3. Find the closest embedding

Embedding API

2. Embed question

4. Return relevant text

1. Send question

Chatbot

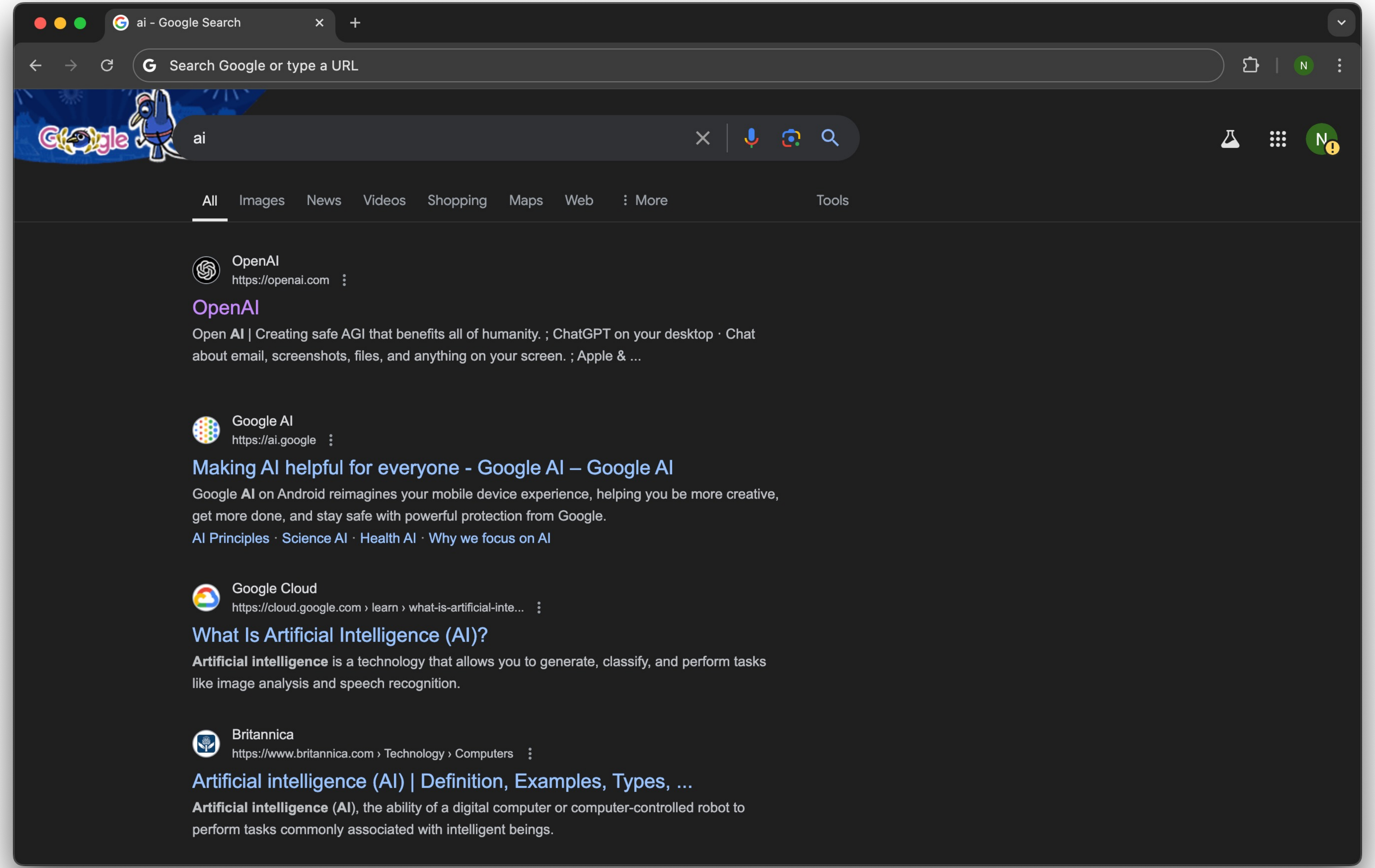
5. Send conversation

6. Get the next message

OpenAI

* Then embedding API can get relevant text.
By taking query, embeds, and find nearest embedding vectors and return the text of those vectors.
* Chatbot then take the result text to generate chat response message.

What if I want to just search, not chat



OpenAI
https://openai.com

OpenAI

Open **AI** | Creating safe AGI that benefits all of humanity. ; ChatGPT on your desktop · Chat about email, screenshots, files, and anything on your screen. ; Apple & ...

Google AI
https://ai.google

Making AI helpful for everyone - Google AI – Google AI

Google **AI** on Android reimagines your mobile device experience, helping you be more creative, get more done, and stay safe with powerful protection from Google.

AI Principles · Science AI · Health AI · Why we focus on AI

Google Cloud
https://cloud.google.com › learn › what-is-artificial-inte...

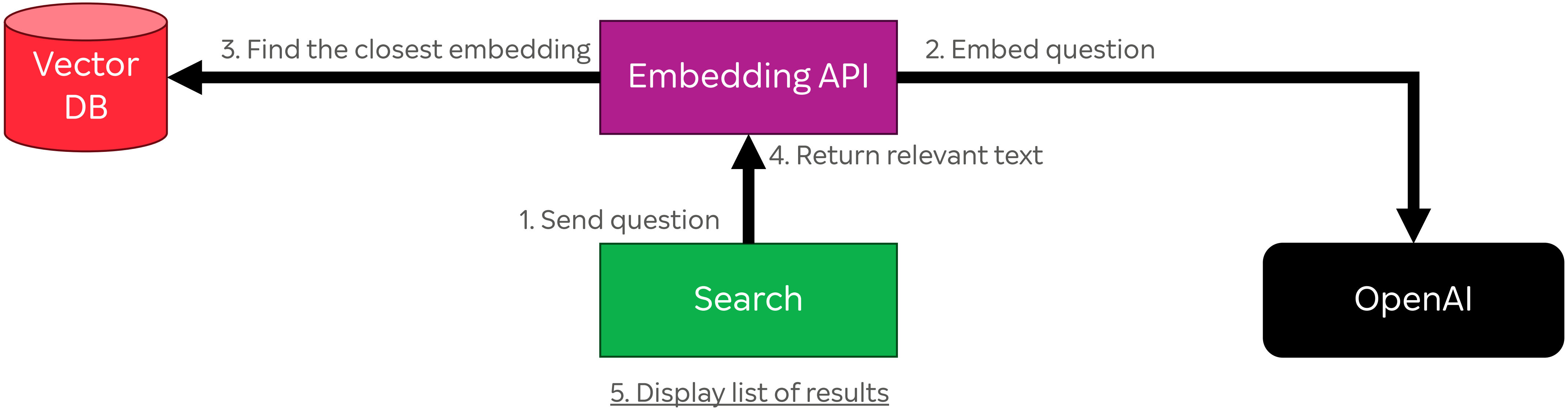
What Is Artificial Intelligence (AI)?

Artificial intelligence is a technology that allows you to generate, classify, and perform tasks like image analysis and speech recognition.

Britannica
https://www.britannica.com › Technology › Computers

Artificial intelligence (AI) | Definition, Examples, Types, ...

Artificial intelligence (**AI**), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.


Vector DB
3. Find the closest embedding
Embedding API
2. Embed question
4. Return relevant text
1. Send question
Search
OpenAI
5. Display list of results

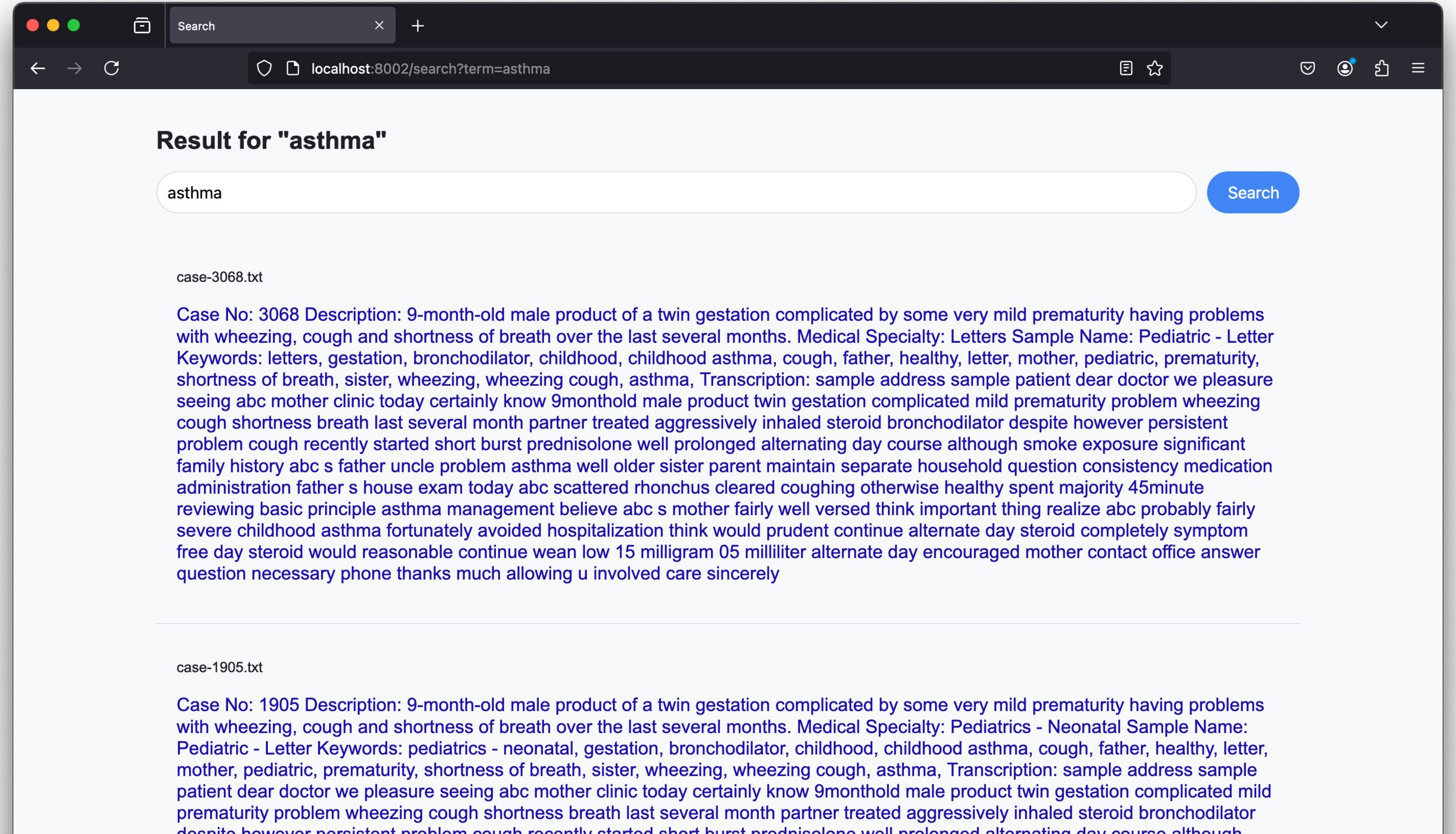
Workshop: 3

Let’s build the search

bit.ly/ts-ai-workshop

Same link, just in case

- Switch to branch `pre-workshop-3`
- Follow instruction for workshop 2 in repo’s wiki



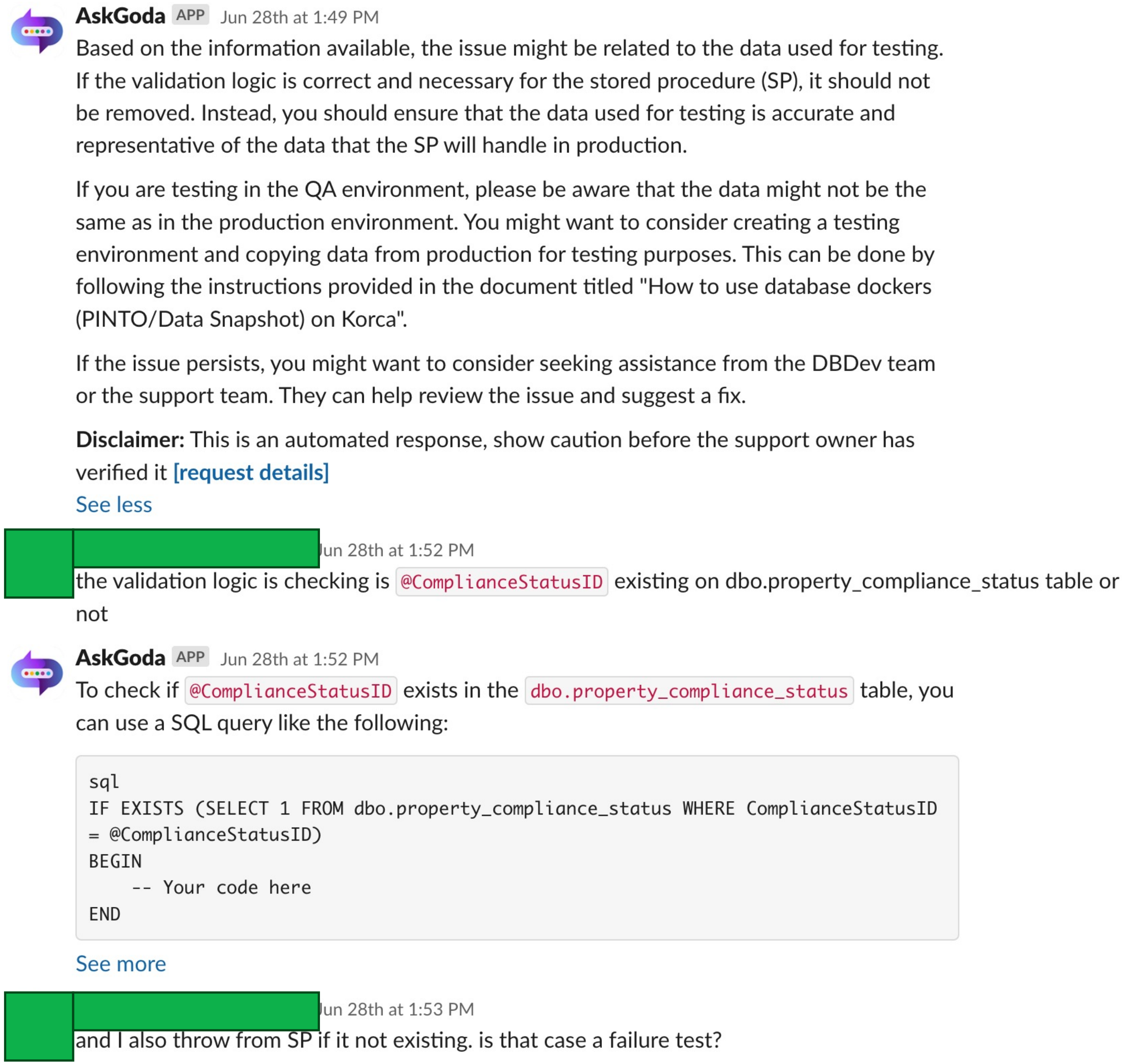
Result for "asthma"

case-3068.txt

Case No: 3068 Description: 9-month-old male product of a twin gestation complicated by some very mild prematurity having problems with wheezing, cough and shortness of breath over the last several months. Medical Specialty: Letters Sample Name: Pediatric - Letter Keywords: letters, gestation, bronchodilator, childhood, childhood asthma, cough, father, healthy, letter, mother, pediatric, prematurity, shortness of breath, sister, wheezing, wheezing cough, asthma, Transcription: sample address sample patient dear doctor we pleasure seeing abc mother clinic today certainly know 9monthold male product twin gestation complicated mild prematurity problem wheezing cough shortness breath last several month partner treated aggressively inhaled steroid bronchodilator despite however persistent problem cough recently started short burst prednisolone well prolonged alternating day course although smoke exposure significant family history abc s father uncle problem asthma well older sister parent maintain separate household question consistency medication administration father s house exam today abc scattered rhonchus cleared coughing otherwise healthy spent majority 45minute reviewing basic principle asthma management believe abc s mother fairly well versed think important thing realize abc probably fairly severe childhood asthma fortunately avoided hospitalization think would prudent continue alternate day steroid completely symptom free day steroid would reasonable continue wean low 15 milligram 05 milliliter alternate day encouraged mother contact office answer question necessary phone thanks much allowing u involved care sincerely

case-1905.txt

Case No: 1905 Description: 9-month-old male product of a twin gestation complicated by some very mild prematurity having problems with wheezing, cough and shortness of breath over the last several months. Medical Specialty: Pediatrics - Neonatal Sample Name: Pediatric - Letter Keywords: pediatrics - neonatal, gestation, bronchodilator, childhood, childhood asthma, cough, father, healthy, letter, mother, pediatric, prematurity, shortness of breath, sister, wheezing, wheezing cough, asthma, Transcription: sample address sample patient dear doctor we pleasure seeing abc mother clinic today certainly know 9monthold male product twin gestation complicated mild prematurity problem wheezing cough shortness breath last several month partner treated aggressively inhaled steroid bronchodilator despite however persistent problem cough recently started short burst prednisolone well prolonged alternating day course although



AskGoda APP Jun 28th at 1:49 PM

Based on the information available, the issue might be related to the data used for testing. If the validation logic is correct and necessary for the stored procedure (SP), it should not be removed. Instead, you should ensure that the data used for testing is accurate and representative of the data that the SP will handle in production.

If you are testing in the QA environment, please be aware that the data might not be the same as in the production environment. You might want to consider creating a testing environment and copying data from production for testing purposes. This can be done by following the instructions provided in the document titled "How to use database dockers (PINTO/Data Snapshot) on Korca".

If the issue persists, you might want to consider seeking assistance from the DBDev team or the support team. They can help review the issue and suggest a fix.

Disclaimer: This is an automated response, show caution before the support owner has verified it **[request details]**

See less

[redacted] Jun 28th at 1:52 PM

the validation logic is checking is `@ComplianceStatusID` existing on dbo.property_compliance_status table or not

AskGoda APP Jun 28th at 1:52 PM

To check if `@ComplianceStatusID` exists in the `dbo.property_compliance_status` table, you can use a SQL query like the following:

```
sql
IF EXISTS (SELECT 1 FROM dbo.property_compliance_status WHERE ComplianceStatusID
= @ComplianceStatusID)
BEGIN
    -- Your code here
END
```

See more

[redacted] Jun 28th at 1:53 PM

and I also throw from SP if it not existing. is that case a failure test?

agoda In Agoda, we also implement a chatbot and integrate in Slack for engineers to increase their productivity.

Thank you!