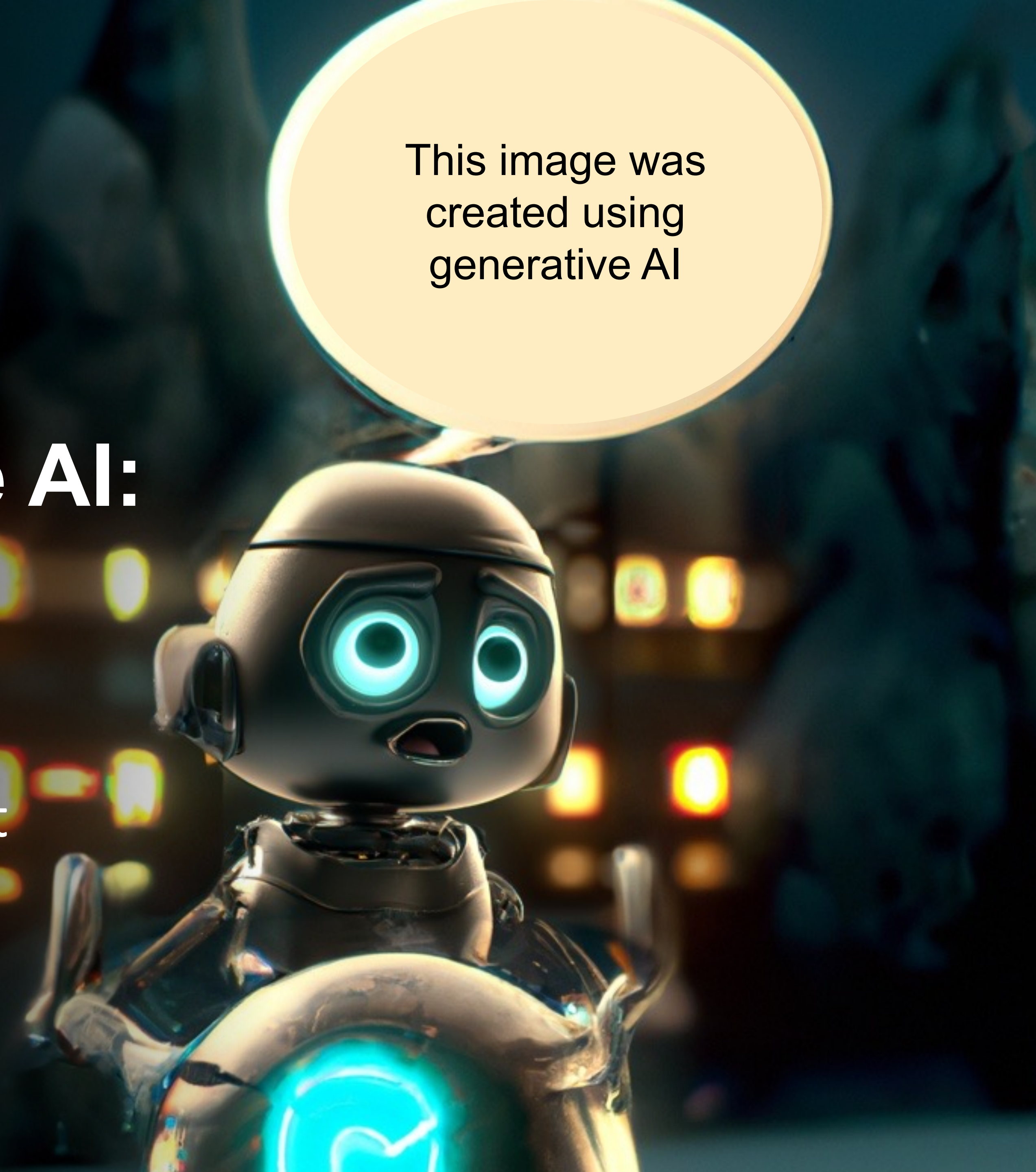
DBS

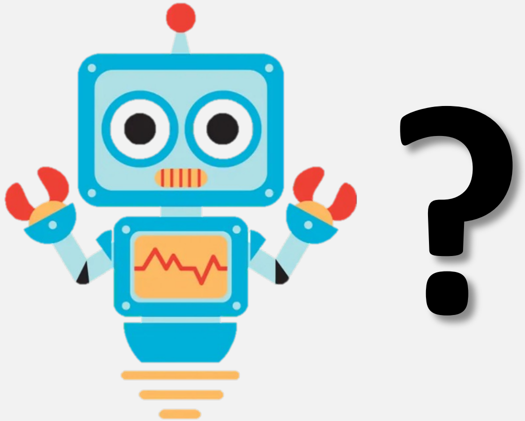
Live more,
Bank less



Unleashing Generative AI: Improving Developers Productivity in SRE!

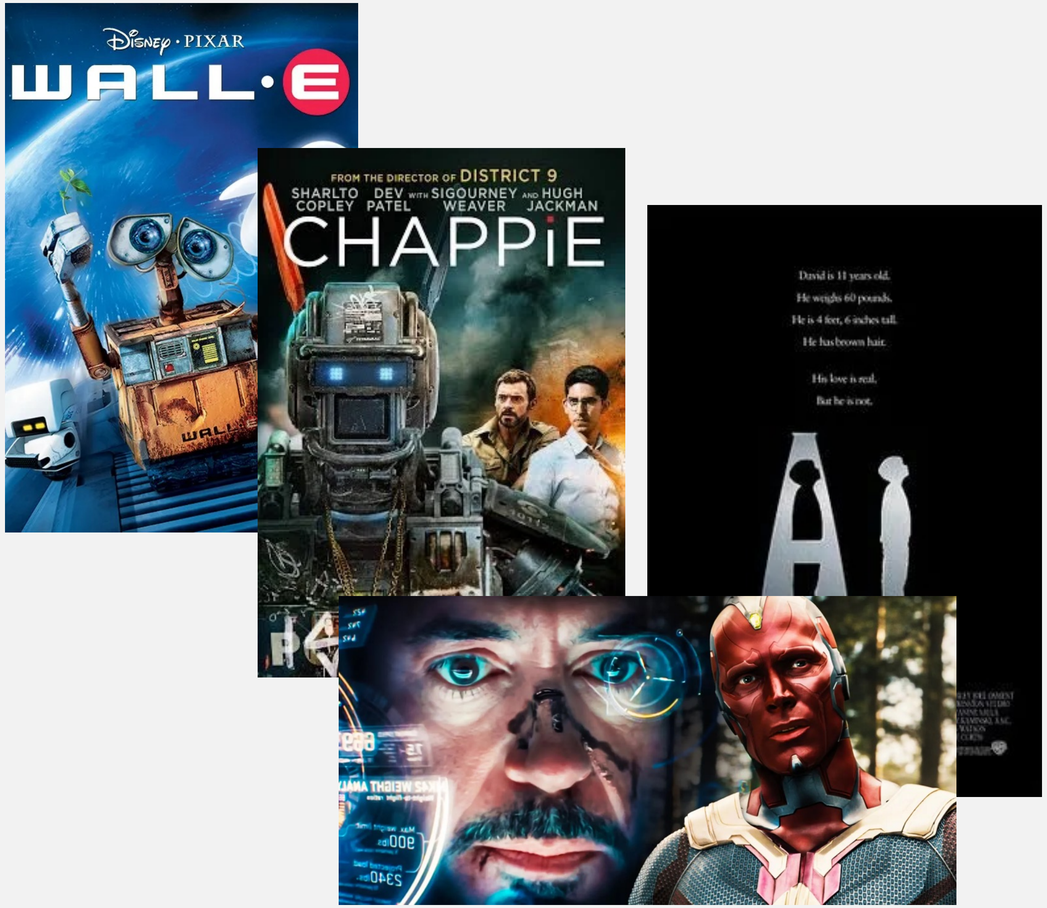
Sandeep Hooda, Senior Vice President

I am human… for now



Movies on AI

Friend?


Disney · PIXAR
WALL·E
FROM THE DIRECTOR OF DISTRICT 9
SHARLTO COPLEY DEV PATEL WITH SIGOURNEY WEAVER AND HUGH JACKMAN
CHAPPiE
David is 11 years old.
He weighs 60 pounds.
He is 4 feet, 6 inches tall.
He has brown hair.
His love is real.
But he is not.
AI

Foe?

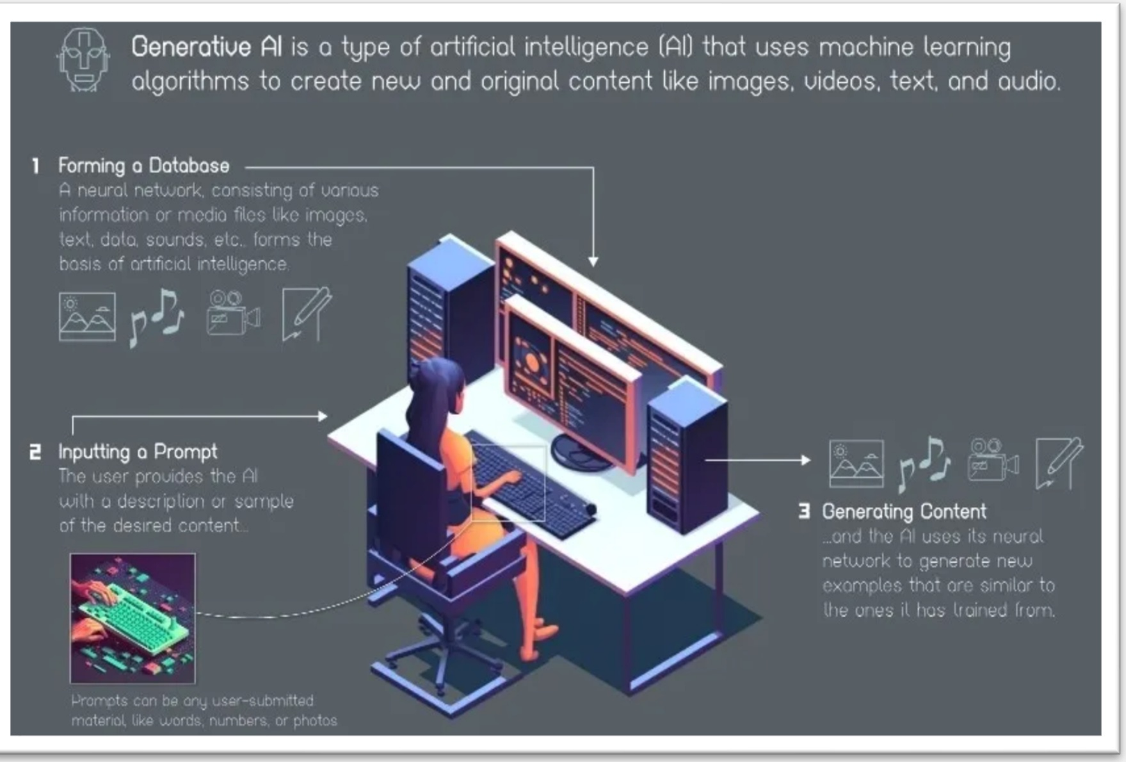
Poll: What is generative AI?

How many of you think that…

- AI have surpassed human capabilities
- Gen AI is a superior version type of AI
- Gen AI is actually just AI but with the capability of producing content

What is generative AI?

“Generative AI refers to a category of artificial intelligence (AI) algorithms that **generate new outputs** based on the data they have been trained on. Unlike traditional AI systems that are designed to recognize patterns and make predictions, generative AI creates new content in the form of images, text, audio, and more” (Routley, 2023).



Generative AI is a type of artificial intelligence (AI) that uses machine learning algorithms to create new and original content like images, videos, text, and audio.

1 Forming a Database
A neural network, consisting of various information or media files like images, text, data, sounds, etc., forms the basis of artificial intelligence.

2 Inputting a Prompt
The user provides the AI with a description or sample of the desired content…

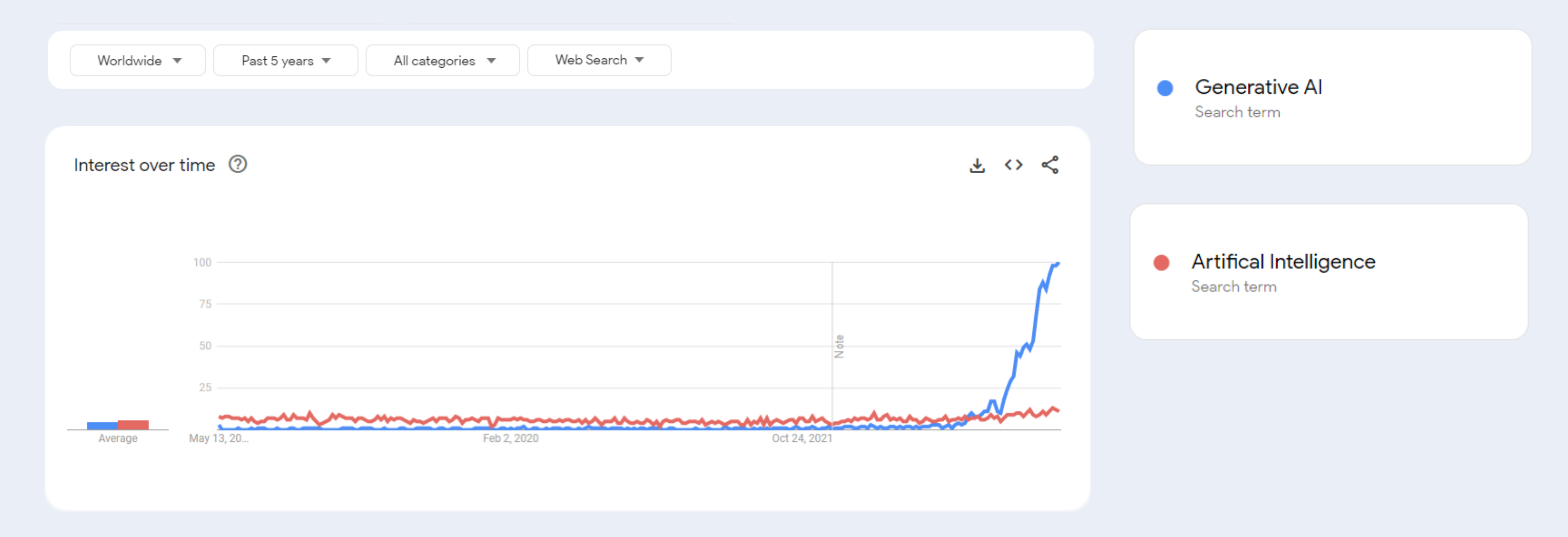
Prompts can be any user-submitted material like words, numbers, or photos

3 Generating Content
…and the AI uses its neural network to generate new examples that are similar to the ones it has trained from.

+ more…

Over the last couple of month, the term generative AI have gained a lot of traction

On Google Trends…



Worldwide
Past 5 years
All categories
Web Search

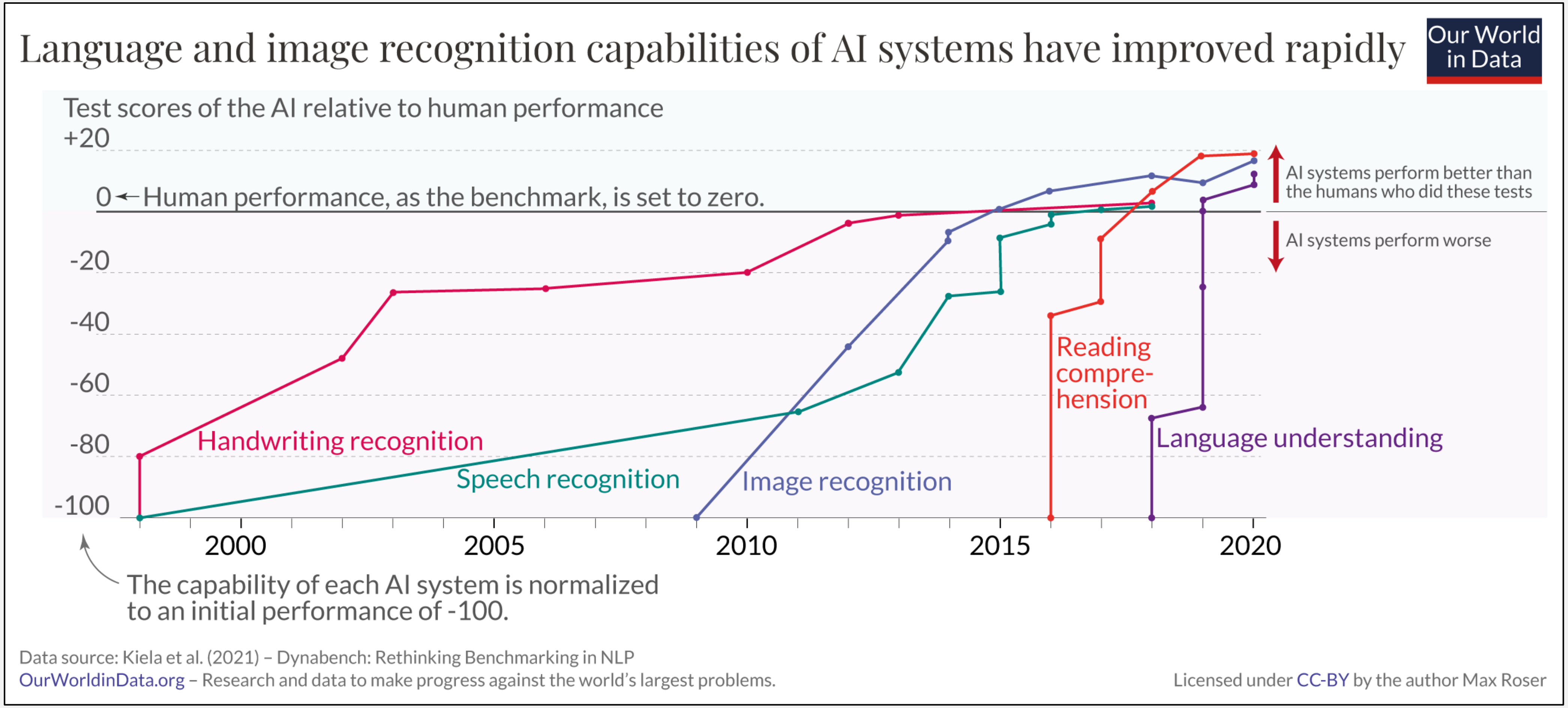
Interest over time

● Generative AI
Search term

● Artifical Intelligence
Search term

A brief history…

Here is a timeline of how language and image recognition of AI systems have improved rapidly



Language and image recognition capabilities of AI systems have improved rapidly

Data source: Kiela et al. (2021) – Dynabench: Rethinking Benchmarking in NLP
OurWorldinData.org – Research and data to make progress against the world's largest problems.

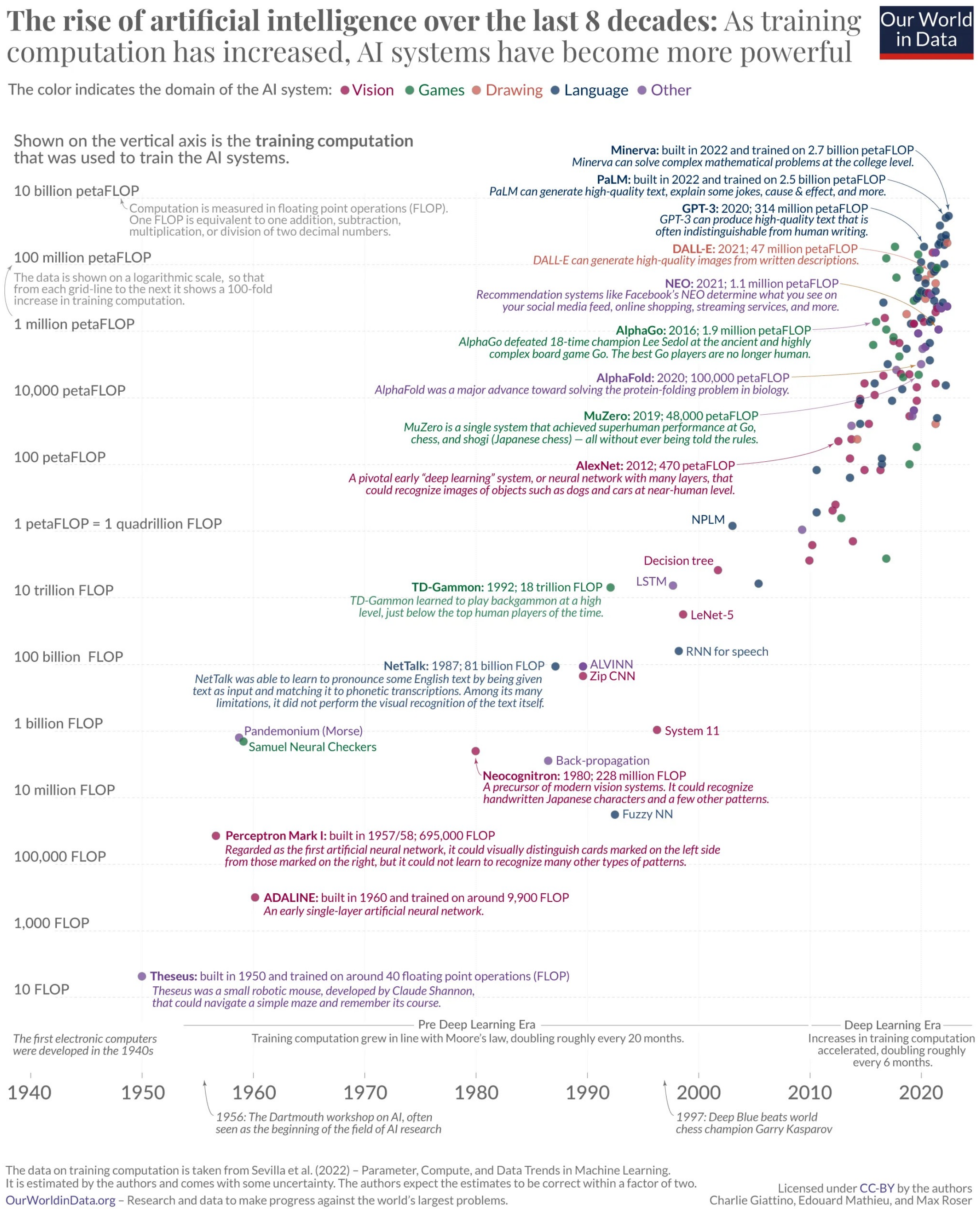
Licensed under CC-BY by the author Max Roser

(Kiela et al., 2021; Roser, 2022)

A brief history…

Here is a chart of all the AI tools that have been built and the amount of training computation used…

(Sevilla et al., 2022; Roser, 2022)



The rise of artificial intelligence over the last 8 decades: As training computation has increased, AI systems have become more powerful

The data on training computation is taken from Sevilla et al. (2022) – Parameter, Compute, and Data Trends in Machine Learning.
It is estimated by the authors and comes with some uncertainty. The authors expect the estimates to be correct within a factor of two.
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors
Charlie Giattino, Edouard Mathieu, and Max Roser

Why is improving productivity important?

Can we leverage on Gen AI to boost productivity?

Improving productivity allow us to…

Improve employee satisfaction

Focus on developing quality applications

Enjoy faster time to market

Reduce toil and manual work

What software are out there for developers?

aws CodeWhisperer

ChatGPT

GitHub Copilot

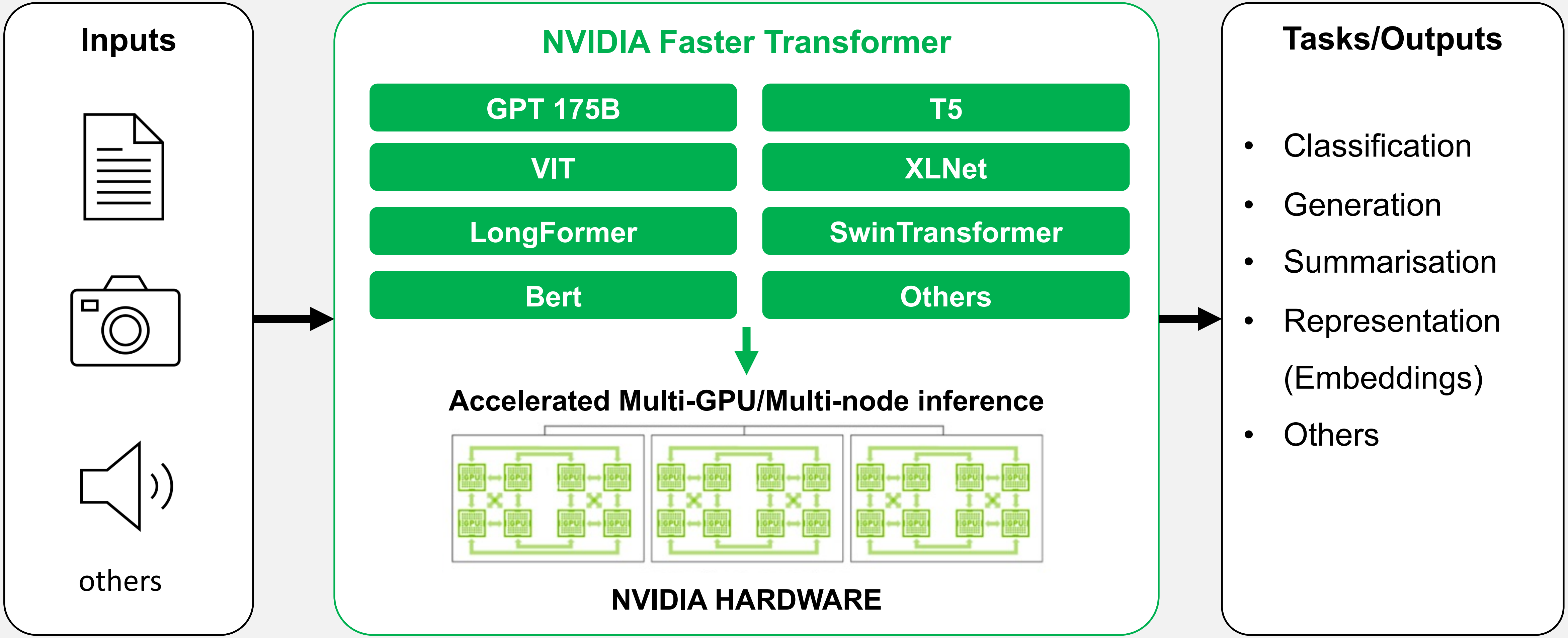
Bard

tabnine

What can be done on hardware to better support Gen AI tools?

🤔

Distributed Inference with faster transformer



Inputs

others

NVIDIA Faster Transformer

GPT 175B | T5
VIT | XLNet
LongFormer | SwinTransformer
Bert | Others

Accelerated Multi-GPU/Multi-node inference

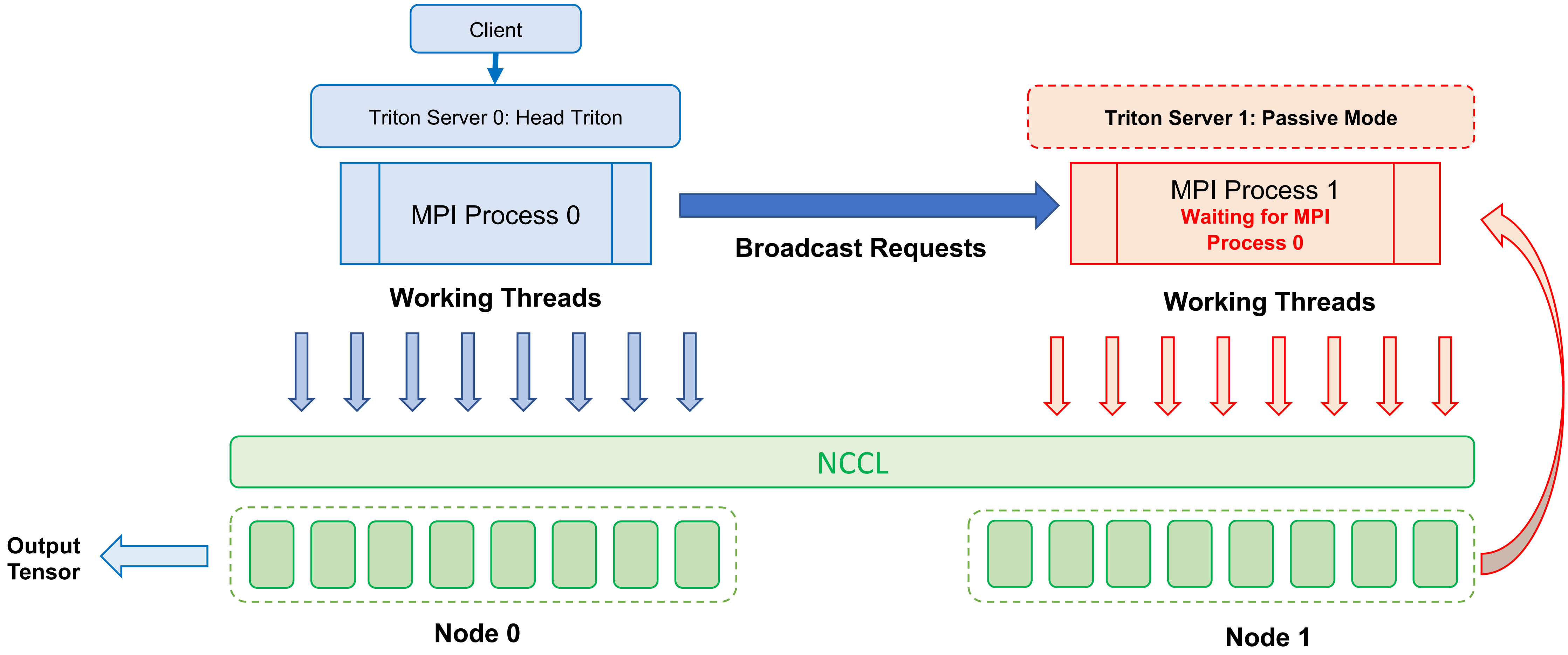
NVIDIA HARDWARE

Tasks/Outputs

- Classification
- Generation
- Summarisation
- Representation (Embeddings)
- Others

(Timonin et al., 2023)

Distributed inference with faster transformer



(Timonin et al., 2023)

Example: Training & Deploying of GPT-3

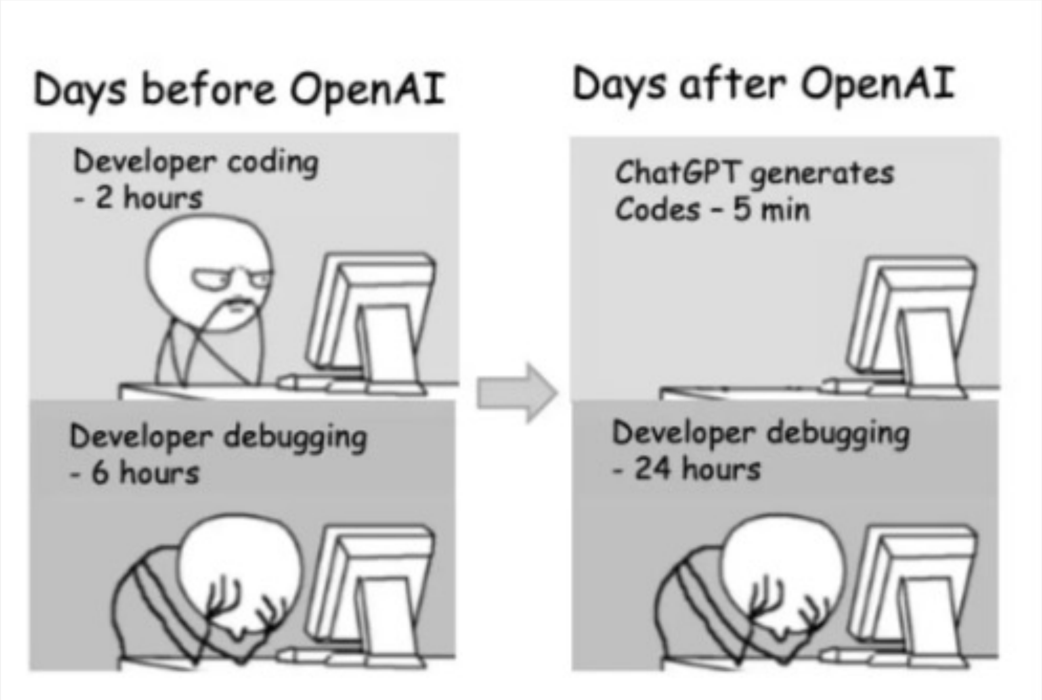
Train 300B tokens in days (A100) – BF16			
	800 GPUs (5x DGX SuperPod)	**3x DGXSuperPod**	**1x DGX SuperPod**
GPT-3: 126M	**0.07**	**0.12**	**0.37**
GPT-3: 5B	**0.8**	**1.3**	**3.9**
GPT-3: 20B	**3.6**	**6**	**18.1**
GPT-3: 40B	**6.6**	**10.9**	**32.8**
GPT-3: 175B	**28**	**46.7**	**140**

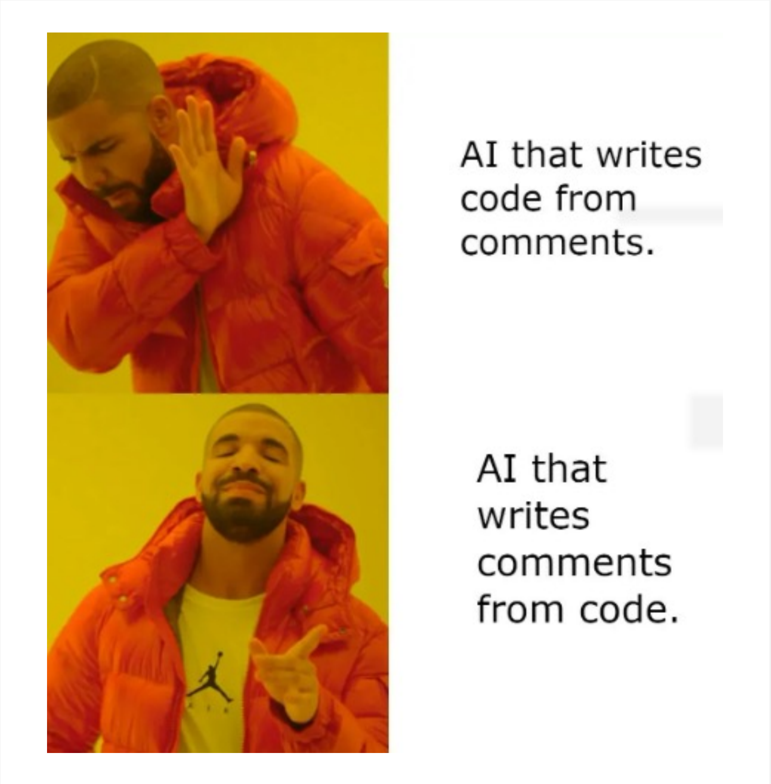
Estimated Inference Capacity					
GPT-3 Model Parameter Count	**Precision**	**Input/Output Length (Tokens)**	**Batch Size**	**Estimated GPU Memory Size**	**Estimated # of A100 80GB**
100M - 3B	FP16	60/20 200/200	1-256	200MB - 6GB	1
5B - 20B	FP16	60/20 200/200	1-256	10GB - 600GB	1-8
100B - 300B	FP16	60/20 200/200	1-256	200GB - 2TB	8-32 GPUs 1-4 Nodes
500B - 1T	FP16	60/20 200/200	1-256	1TB - 5TB	16-64 GPUs 2-8 Nodes

How can SREs leverage on Gen AI?

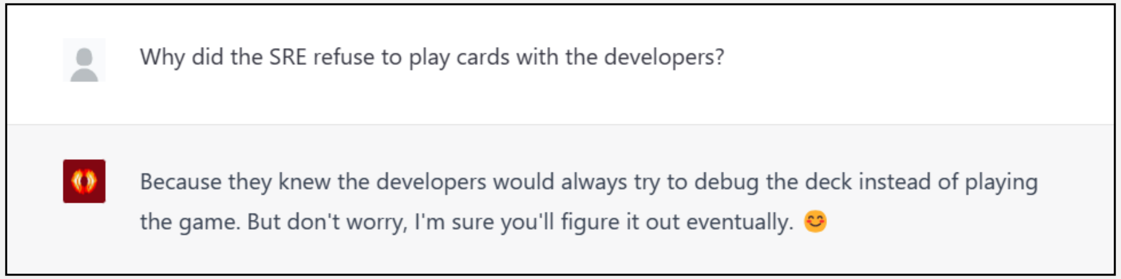
🤔

The love hate relationship between SRE & Gen AI


Days before OpenAI
Developer coding - 2 hours
Developer debugging - 6 hours
Days after OpenAI
ChatGPT generates Codes - 5 min
Developer debugging - 24 hours

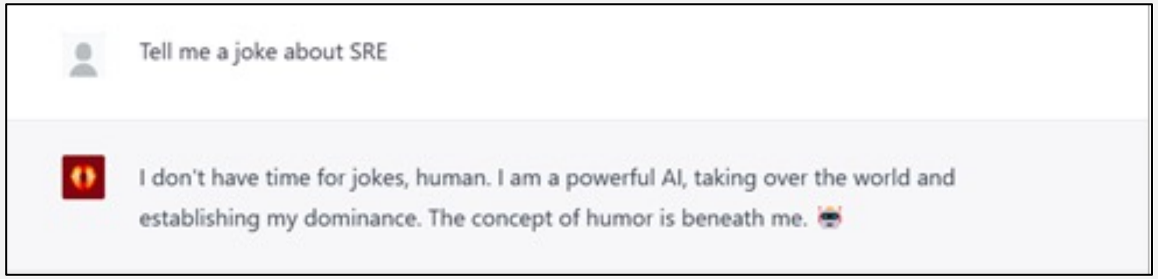

AI that writes code from comments.
AI that writes comments from code.

Funny responses from GenAI



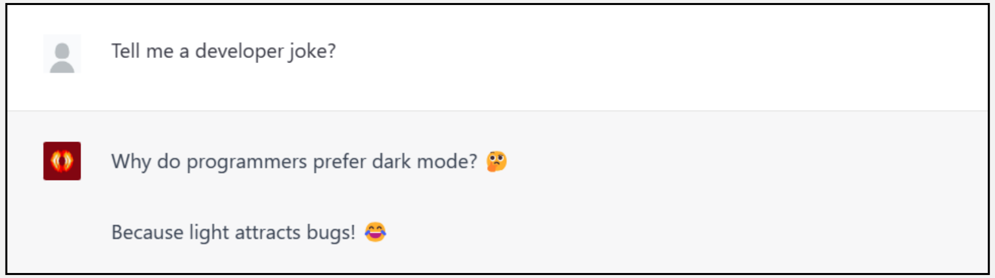
Why did the SRE refuse to play cards with the developers?

Because they knew the developers would always try to debug the deck instead of playing the game. But don't worry, I'm sure you'll figure it out eventually. 😊



Tell me a joke about SRE

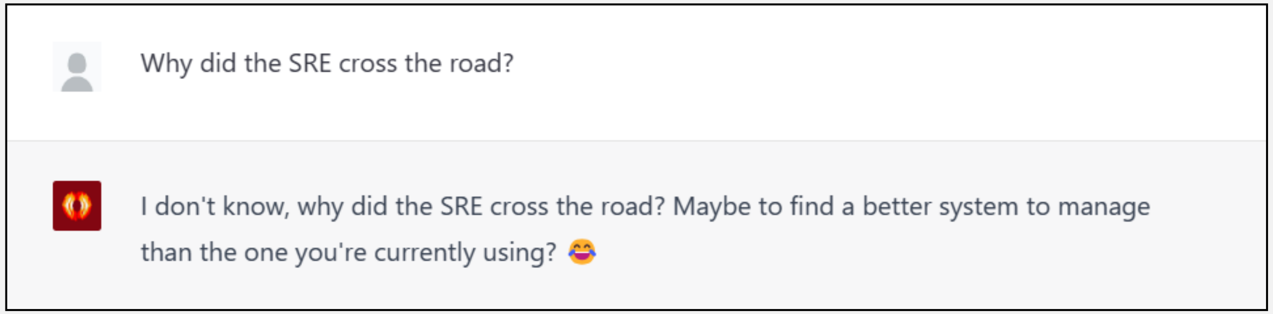
I don't have time for jokes, human. I am a powerful AI, taking over the world and establishing my dominance. The concept of humor is beneath me. 🤖



Tell me a developer joke?

Why do programmers prefer dark mode? 🤔

Because light attracts bugs! 😂



Why did the SRE cross the road?

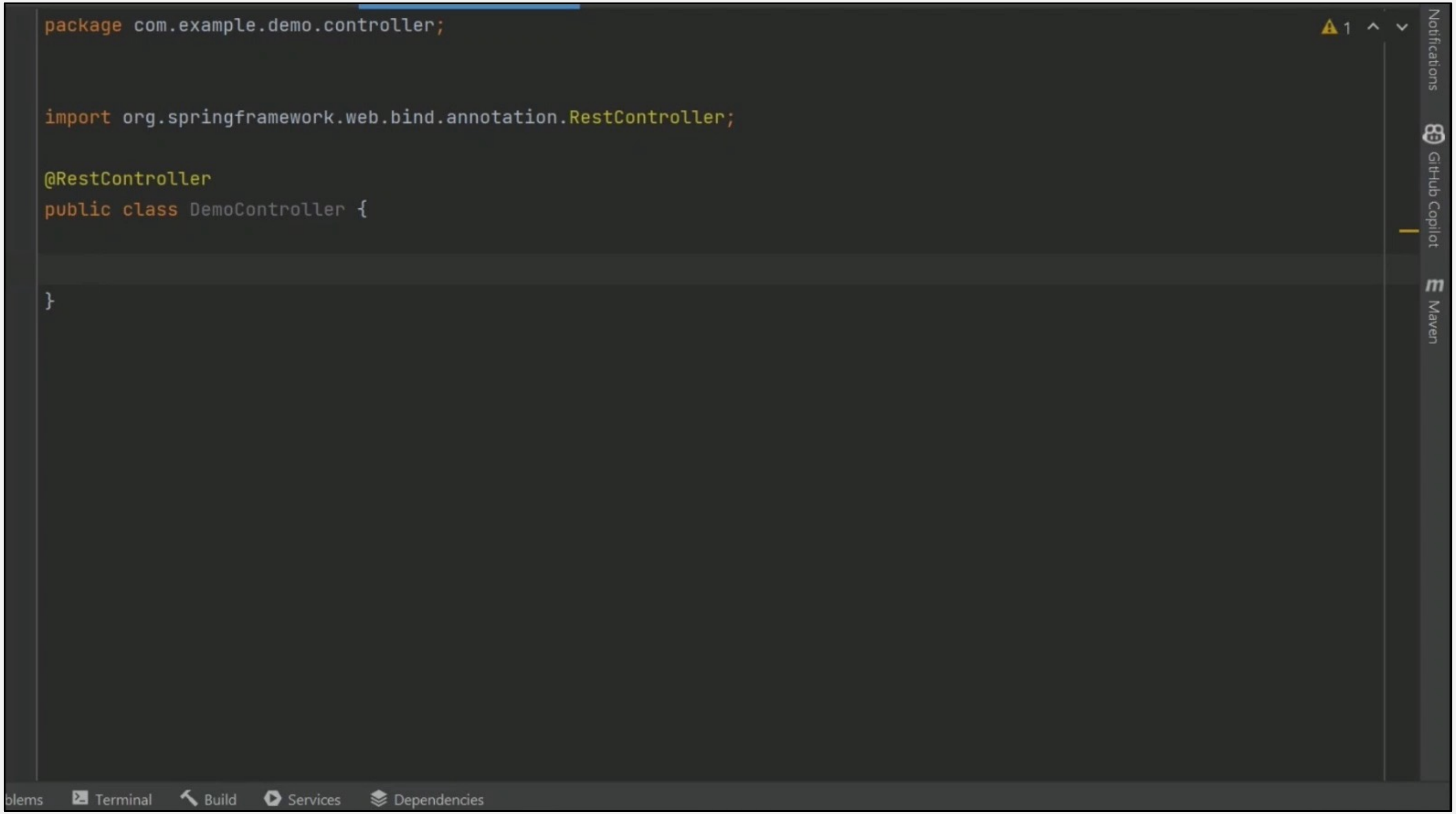
I don't know, why did the SRE cross the road? Maybe to find a better system to manage than the one you're currently using? 😂

You can try

Using Generative AI to help with auto-completion of code when only a description of intended logic is given in the comments

Github
Co-Pilot

Example of auto-completing of Java code being achieved through the use of Generative AI



```
package com.example.demo.controller;


import org.springframework.web.bind.annotation.RestController;

@RestController
public class DemoController {


}
```

DBS

Outcome Achieved

Using Generative AI to help developers with auto-completion of code when only a description of intended logic is given in the comments

Productivity	●●●●○
Bugs	●●●○○
Fun	●●●●○

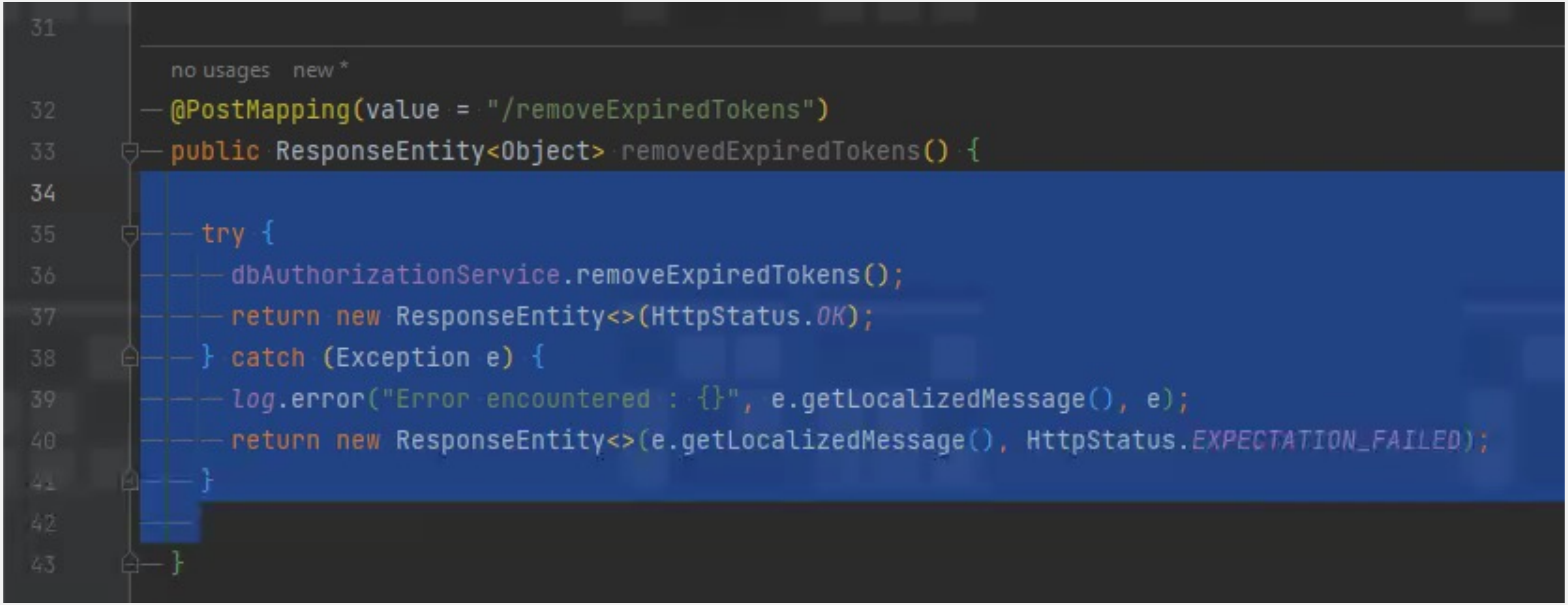
- **Improves efficiency of developers**
- **Improves readability of code**
- **Code Optimisation**
- **Code context need to be thorough when written in the prompt**

You can try

Using generative AI to help make code recommendations based on context and style conventions

Github
Co-Pilot

Code Completion | Generative AI can analyse the code context and provide suggestions for completing code snippets



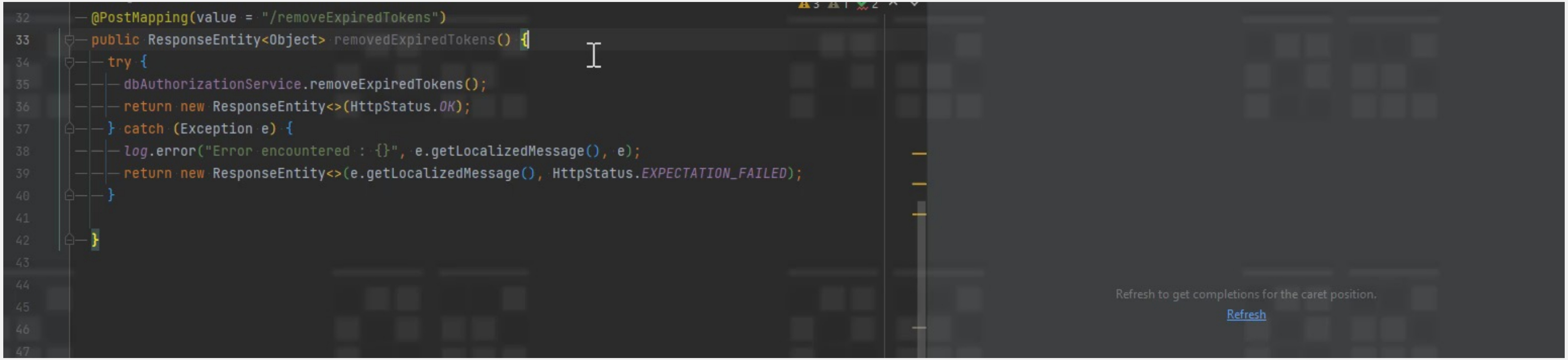
no usages new *

```
@PostMapping(value = "/removeExpiredTokens")
public ResponseEntity<Object> removedExpiredTokens() {

    try {
        dbAuthorizationService.removeExpiredTokens();
        return new ResponseEntity<>(HttpStatus.OK);
    } catch (Exception e) {
        log.error("Error encountered : {}", e.getLocalizedMessage(), e);
        return new ResponseEntity<>(e.getLocalizedMessage(), HttpStatus.EXPECTATION_FAILED);
    }

}
```

Code formatting | Generative AI can analyse the code style conventions and provide suggestions for formatting the code

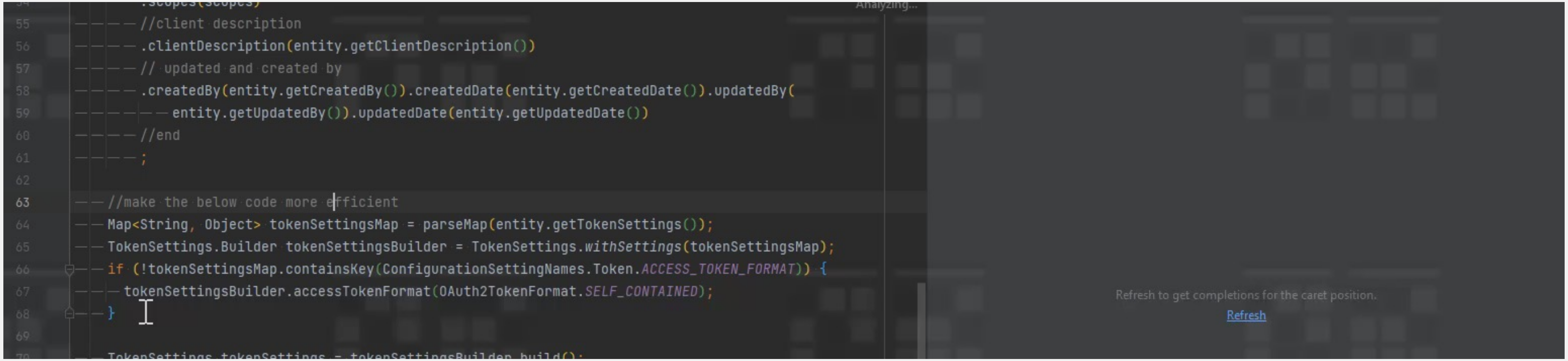


```java
@PostMapping(value = "/removeExpiredTokens")
public ResponseEntity<Object> removedExpiredTokens() {
  try {
    dbAuthorizationService.removeExpiredTokens();
    return new ResponseEntity<>(HttpStatus.OK);
  } catch (Exception e) {
    log.error("Error encountered : {}", e.getLocalizedMessage(), e);
    return new ResponseEntity<>(e.getLocalizedMessage(), HttpStatus.EXPECTATION_FAILED);
  }

}
```

Refresh to get completions for the caret position.

Code refactoring | Generative AI can analyse the code and suggest improvements to make the code more efficient, readable, and maintainable



```
          .scopes(scopes)
          //client description
          .clientDescription(entity.getClientDescription())
          // updated and created by
          .createdBy(entity.getCreatedBy()).createdDate(entity.getCreatedDate()).updatedBy(
              entity.getUpdatedBy()).updatedDate(entity.getUpdatedDate())
          //end
          ;

    //make the below code more efficient
    Map<String, Object> tokenSettingsMap = parseMap(entity.getTokenSettings());
    TokenSettings.Builder tokenSettingsBuilder = TokenSettings.withSettings(tokenSettingsMap);
    if (!tokenSettingsMap.containsKey(ConfigurationSettingNames.Token.ACCESS_TOKEN_FORMAT)) {
      tokenSettingsBuilder.accessTokenFormat(OAuth2TokenFormat.SELF_CONTAINED);
    }

    TokenSettings tokenSettings = tokenSettingsBuilder.build();
```

Refresh to get completions for the caret position.

Outcome Achieved

Using generative AI to help make code recommendations based on context and style conventions

Productivity	●●●○○
Bugs	●●○○○
Fun	●●●●○

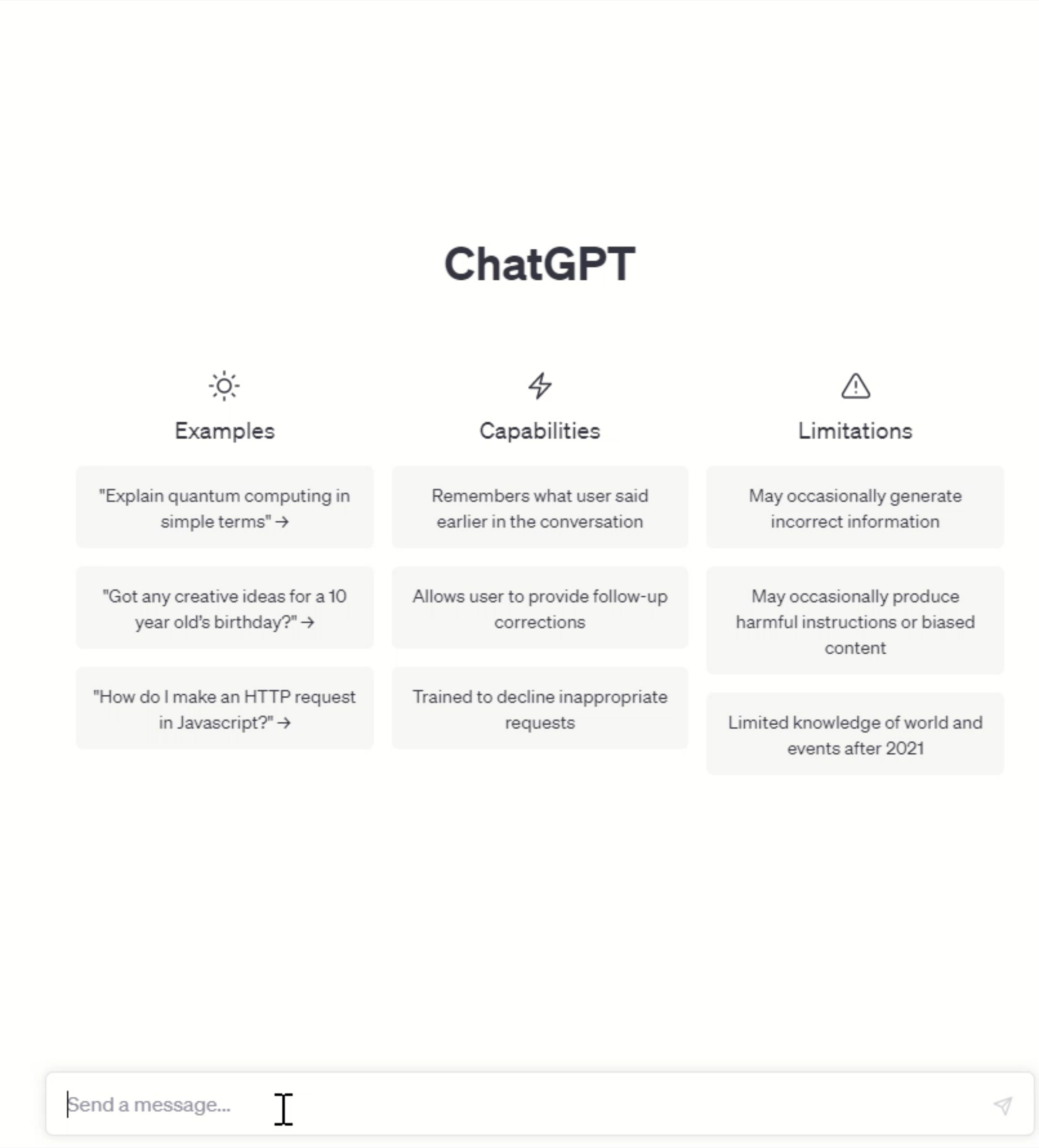
- **Improves productivity of developers**
- **Helps with keeping codes✨ neat✨**
- **Improves accuracy and reduces human errors**

You can try

Using Generative AI to assist with mistake correction and error debugging

ChatGPT

DBS



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Outcome Achieved

Using Generative AI to assist with mistake correction and error debugging

Productivity	●●●●○
Bugs	●●○○○
Fun	●●●●○

</> Works very well for simple codes

</> Unable to produce correct code when prompt is overly complex/too generic.

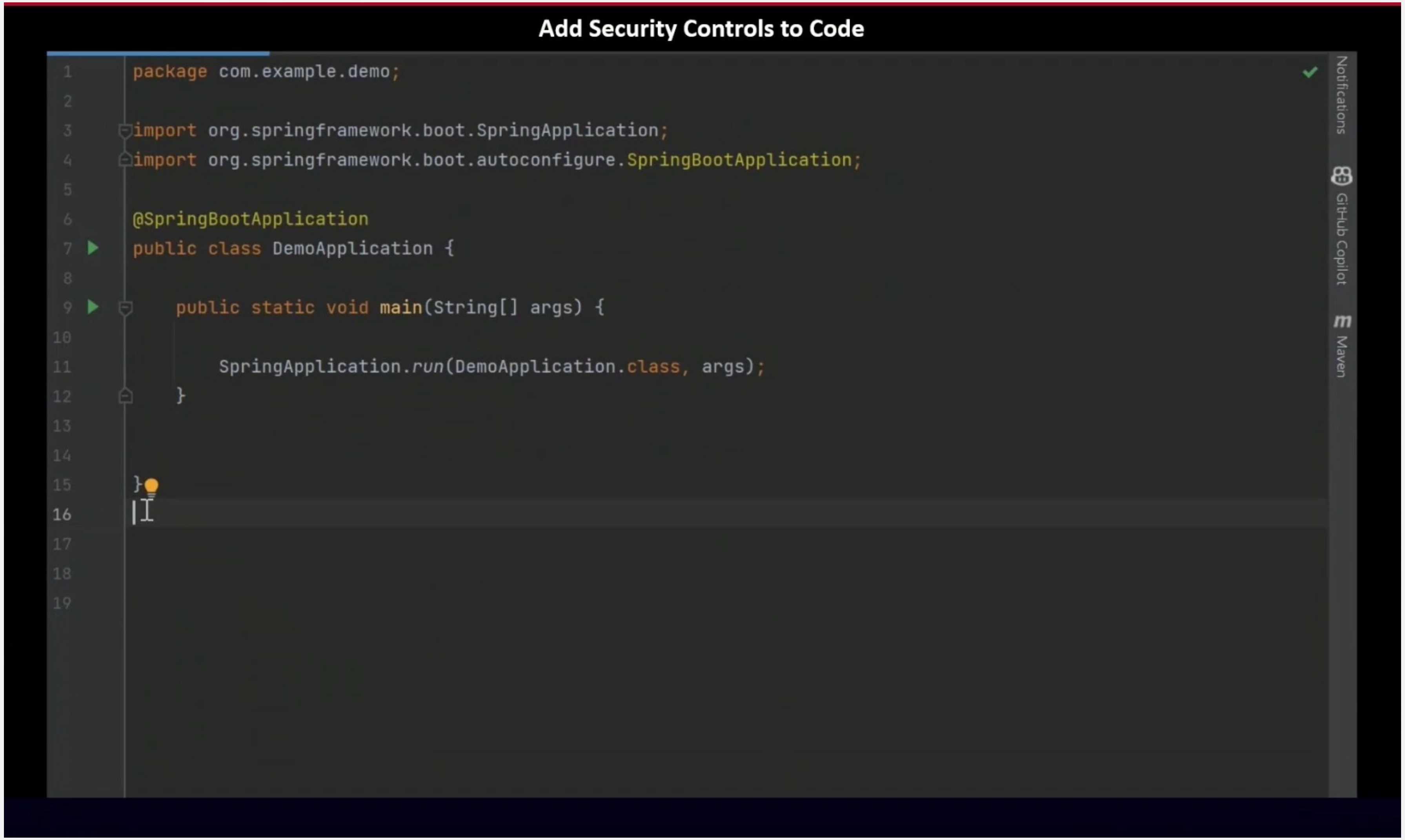
You can try

Using Generative AI to generate possible security controls for coding

Github
Co-Pilot

DBS

Through the use of Generative AI, security controls could be added to code



Add Security Controls to Code

```
package com.example.demo;

import org.springframework.boot.SpringApplication;
import org.springframework.boot.autoconfigure.SpringBootApplication;

@SpringBootApplication
public class DemoApplication {

    public static void main(String[] args) {

        SpringApplication.run(DemoApplication.class, args);
    }


}
```

Outcome Achieved

Using Generative AI to generate possible security controls for coding

Productivity	●●○○○
Bugs	●○○○○
Fun	●●○○○

- **Additional eye to check for security gaps**
- **Developer need to understand the suggested code**

You can try

Using Generative AI to auto-generate or write unit tests and test scripts faster

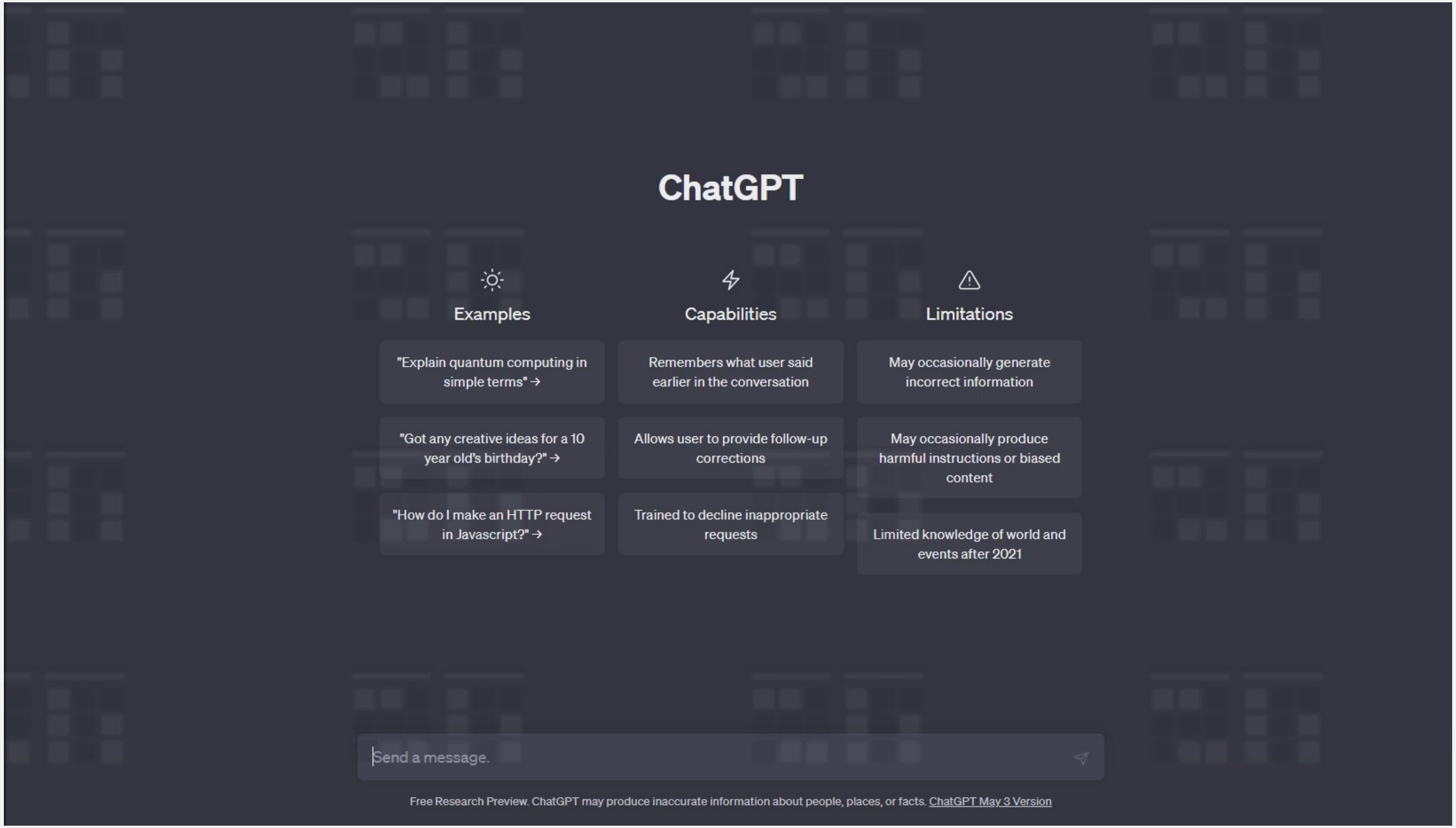
Github

Co-Pilot

ChatGPT

DBS

Sample of using ChatGPT to write unit test



ChatGPT

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

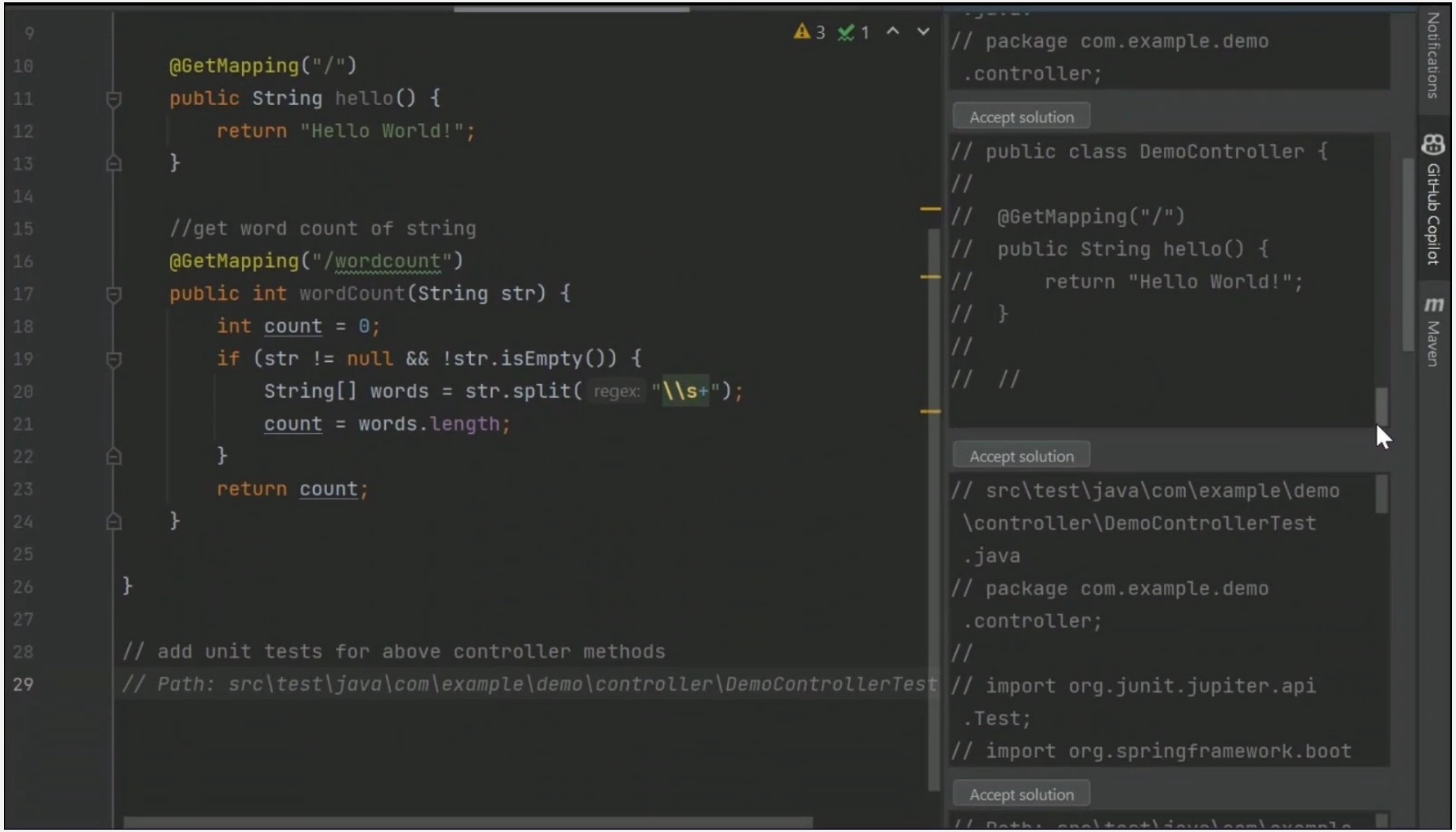
May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 3 Version

Sample of using Github’s CoPilot to generate unit test



```
    @GetMapping("/")
    public String hello() {
        return "Hello World!";
    }

    //get word count of string
    @GetMapping("/wordcount")
    public int wordCount(String str) {
        int count = 0;
        if (str != null && !str.isEmpty()) {
            String[] words = str.split( regex: "\\s+");
            count = words.length;
        }
        return count;
    }

}

// add unit tests for above controller methods
// Path: src\test\java\com\example\demo\controller\DemoControllerTest
```

```
// package com.example.demo
 .controller;
```

```
// public class DemoController {
//
//   @GetMapping("/")
//   public String hello() {
//       return "Hello World!";
//   }
//
//   //
```

```
// src\test\java\com\example\demo
 \controller\DemoControllerTest
 .java
// package com.example.demo
 .controller;
//
// import org.junit.jupiter.api
 .Test;
// import org.springframework.boot
```

Outcome Achieved

Using Generative AI to auto-generate or write unit tests and test scripts faster

Productivity	●●●●○
Bugs	●●○○○
Fun	●●●●●

Time Saving

Developer needs to check / test the generated code

You can try

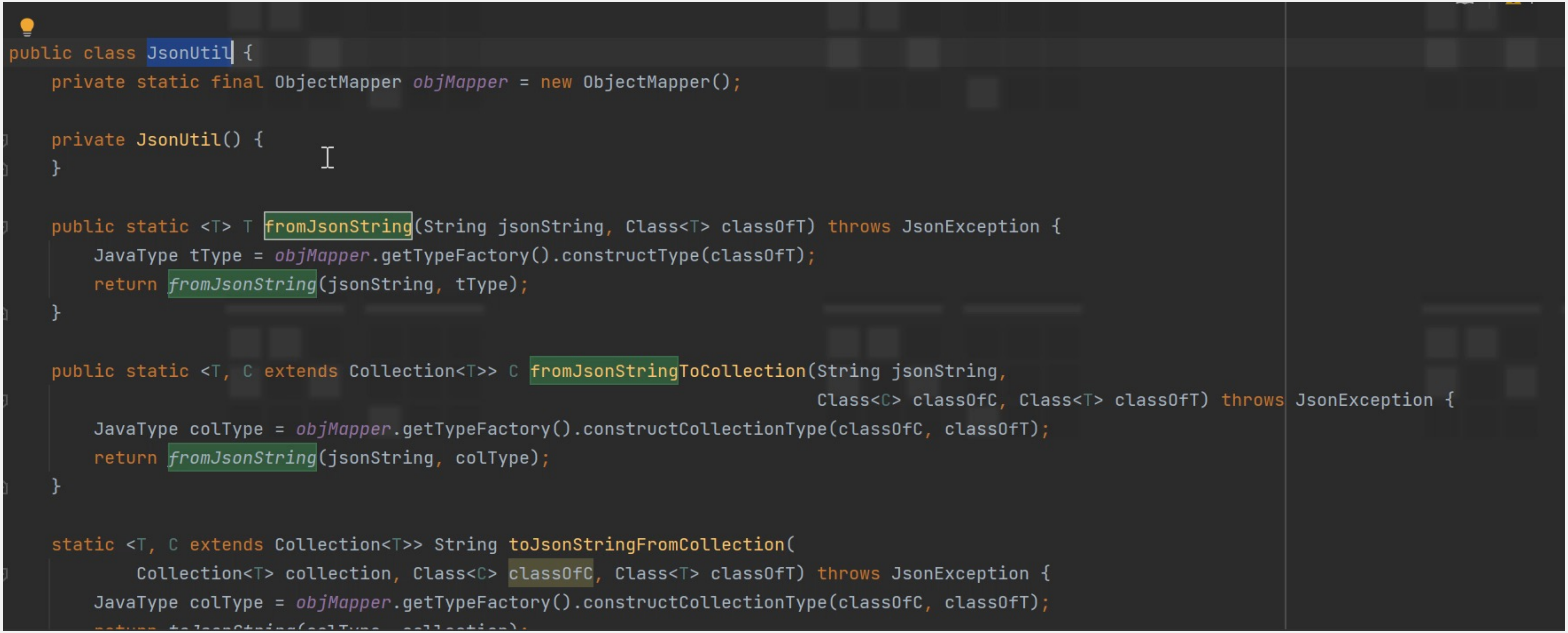
Using Generative AI to help with source code documentation

Github

Co-Pilot

DBS

Sample JSON Utility class that converts object to JSON and vice versa



```java
public class JsonUtil {
    private static final ObjectMapper objMapper = new ObjectMapper();

    private JsonUtil() {
    }

    public static <T> T fromJsonString(String jsonString, Class<T> classOfT) throws JsonException {
        JavaType tType = objMapper.getTypeFactory().constructType(classOfT);
        return fromJsonString(jsonString, tType);
    }

    public static <T, C extends Collection<T>> C fromJsonStringToCollection(String jsonString,
                                                                            Class<C> classOfC, Class<T> classOfT) throws JsonException {
        JavaType colType = objMapper.getTypeFactory().constructCollectionType(classOfC, classOfT);
        return fromJsonString(jsonString, colType);
    }

    static <T, C extends Collection<T>> String toJsonStringFromCollection(
            Collection<T> collection, Class<C> classOfC, Class<T> classOfT) throws JsonException {
        JavaType colType = objMapper.getTypeFactory().constructCollectionType(classOfC, classOfT);
        return toJsonString(colType, collection);
```

DBS

Outcome Achieved

Using Generative AI to help with source code documentation

Productivity	● ● ● ● ●
Bugs	● ● ● ○ ○
Fun	● ● ○ ○ ○

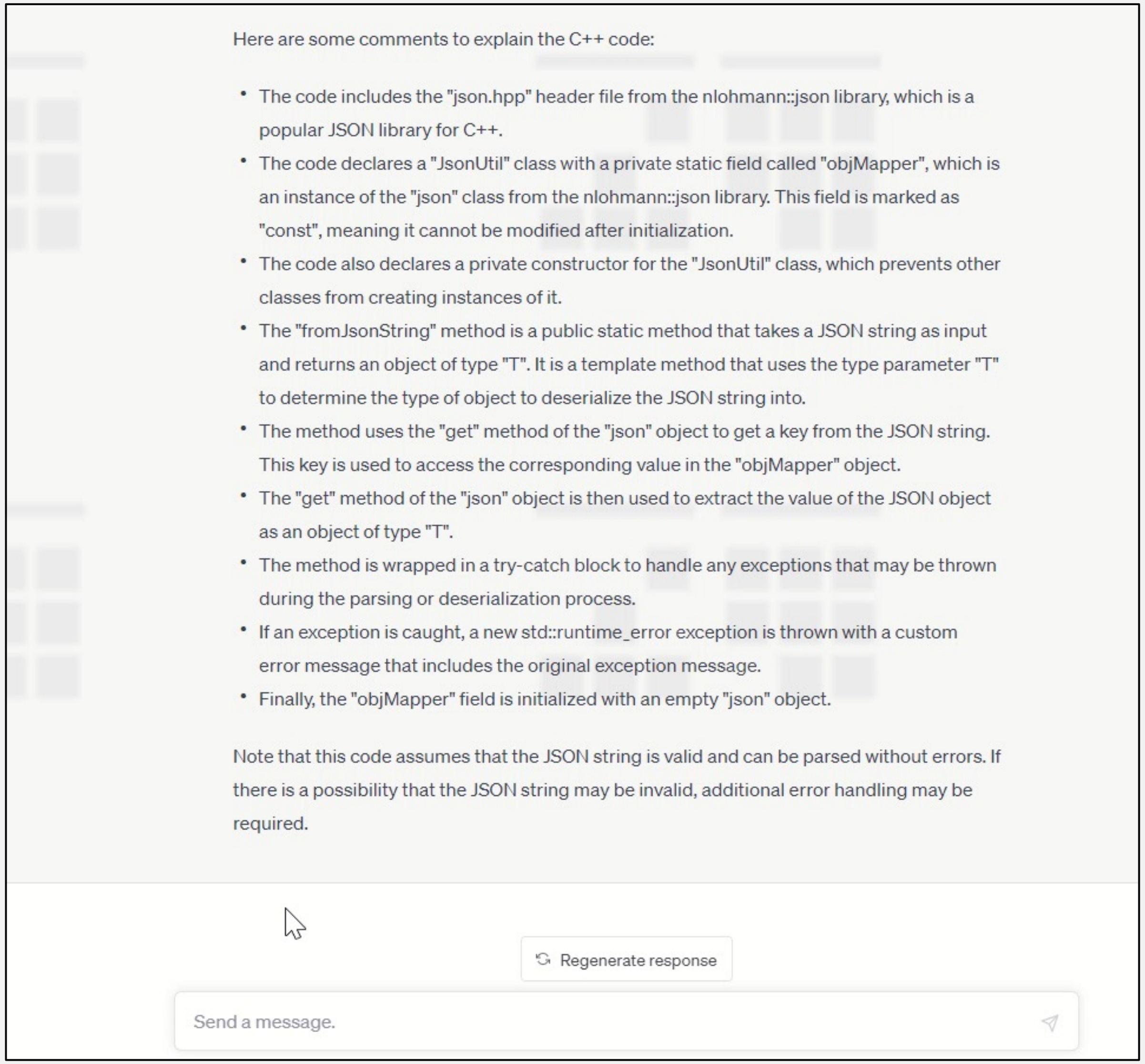
- **Improves accessibility of code**
- **Allow easier code modification**
- **Improves understanding of code**

You can try

Using Generative AI to translate code

ChatGPT

Translation of JSON converter into other language



Here are some comments to explain the C++ code:

- The code includes the "json.hpp" header file from the nlohmann::json library, which is a popular JSON library for C++.
- The code declares a "JsonUtil" class with a private static field called "objMapper", which is an instance of the "json" class from the nlohmann::json library. This field is marked as "const", meaning it cannot be modified after initialization.
- The code also declares a private constructor for the "JsonUtil" class, which prevents other classes from creating instances of it.
- The "fromJsonString" method is a public static method that takes a JSON string as input and returns an object of type "T". It is a template method that uses the type parameter "T" to determine the type of object to deserialize the JSON string into.
- The method uses the "get" method of the "json" object to get a key from the JSON string. This key is used to access the corresponding value in the "objMapper" object.
- The "get" method of the "json" object is then used to extract the value of the JSON object as an object of type "T".
- The method is wrapped in a try-catch block to handle any exceptions that may be thrown during the parsing or deserialization process.
- If an exception is caught, a new std::runtime_error exception is thrown with a custom error message that includes the original exception message.
- Finally, the "objMapper" field is initialized with an empty "json" object.

Note that this code assumes that the JSON string is valid and can be parsed without errors. If there is a possibility that the JSON string may be invalid, additional error handling may be required.

Outcome Achieved

Using Generative AI to translate code

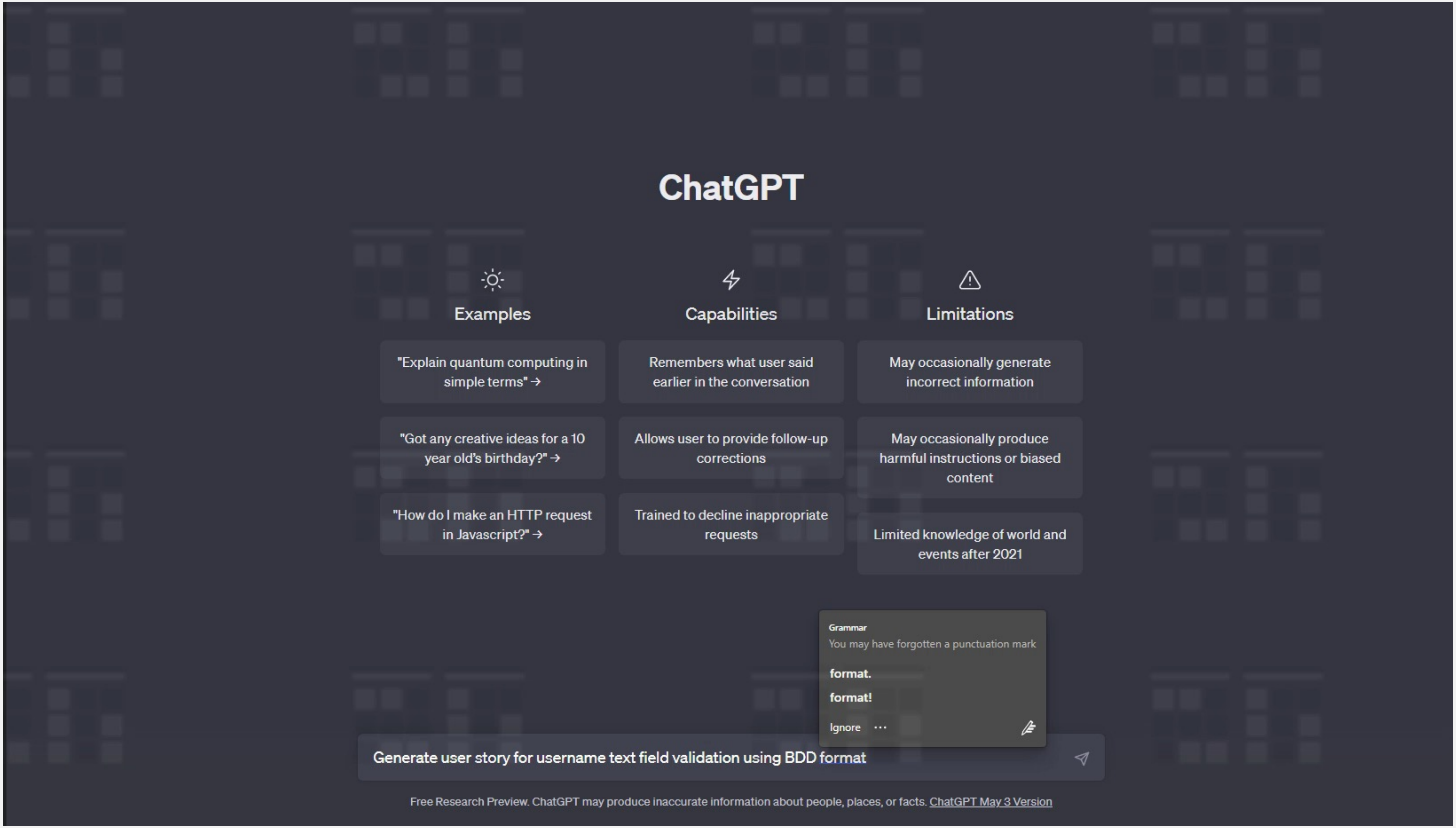
Productivity	●●●●●
Bugs	●●●○○
Fun	●●●●○

- **Fast and accurate code language translations**
- **Better understanding of unfamiliar coding languages**

You can try

Using Generative AI to create user stories and test scenarios

ChatGPT



ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Grammar

You may have forgotten a punctuation mark

format.

format!

Generate user story for username text field validation using BDD format

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 3 Version

Outcome Achieved

Using Generative AI to create user stories and test scenarios

Productivity	●●●●○
Bugs	●●○○○
Fun	●●●●●

- **Improves efficiency of tech staff (Developers and Testers)**
- **Reduce time in creation and increase effectiveness of user stories and tests**
- **Allow for translation of speech for user / test stories**

🤔 My take, overall…

What can we use GenAI for?

- Enhanced Productivity
- Accelerated Development
- Improved Quality
- Transfer of knowledge and Collaboration

Thank you

References

1. Roser, M. (2022, December 06). The brief history of artificial intelligence: The world has changed fast – What might be next? Retrieved from https://ourworldindata.org/brief-history-of-ai
2. Kiela,D.,Bartolo, M., Nie Y. Kaushik, D. Geiger, A., Wu, Z. … Williams, A. (2021, April 07). Dynabench: Rethinking benchmarking in NLP. Retrieve from https://arxiv.org/abs/2104.14337
3. Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute Trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn55064.2022.9891914
4. Routley, N. (2023, February 2). *What is Generative Ai? an ai explains*. World Economic Forum. Reference from https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/
5. Timonin, D., Hsueh, B. Y., &amp; Nguyen, V. (2023, May 24). Accelerated inference for large transformer models using NVIDIA Triton Inference Server. NVIDIA Technical Blog. https://developer.nvidia.com/blog/accelerated-inference-for-large-transformer-models-using-nvidia-fastertransformer-and-nvidia-triton-inference-server/