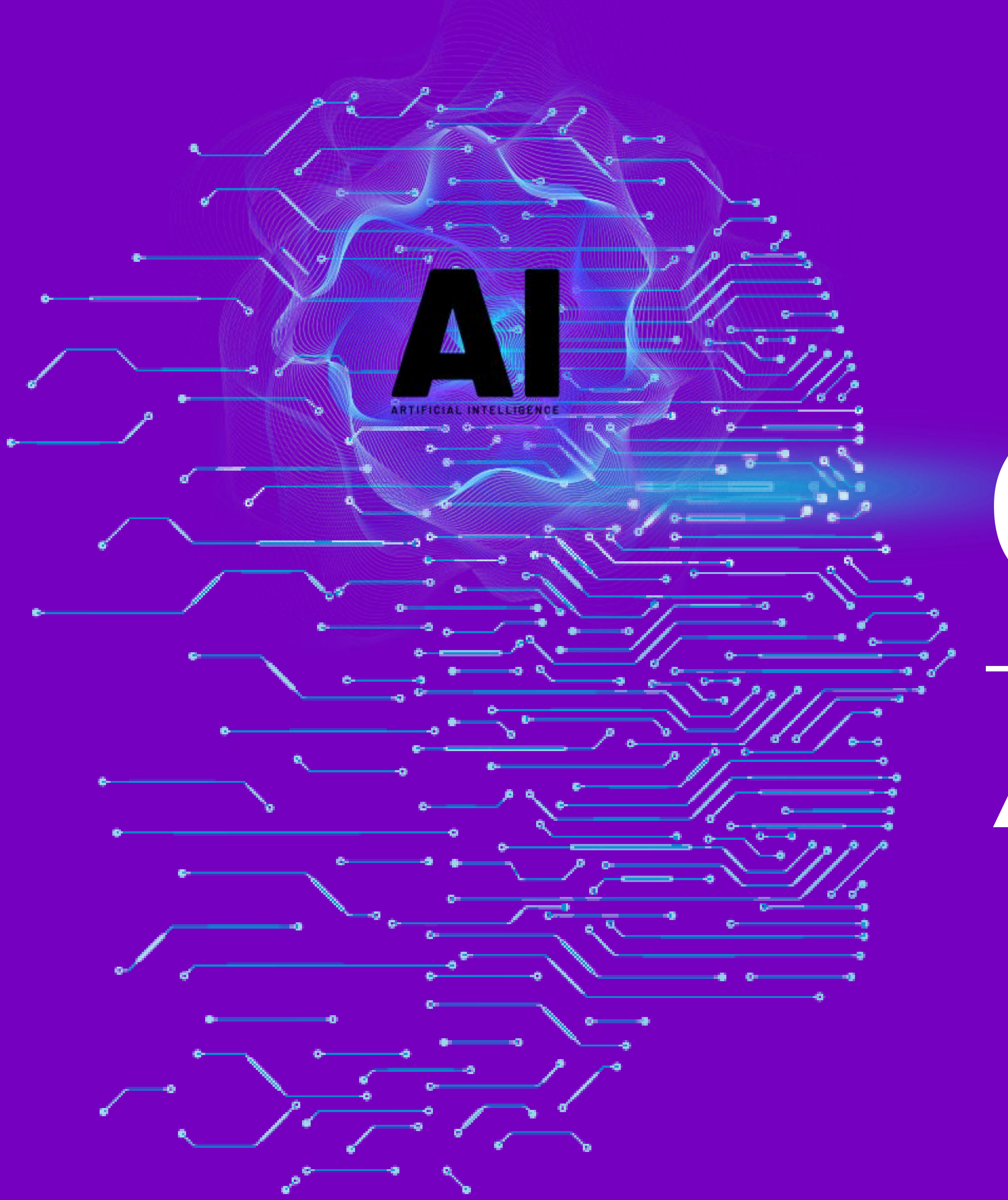


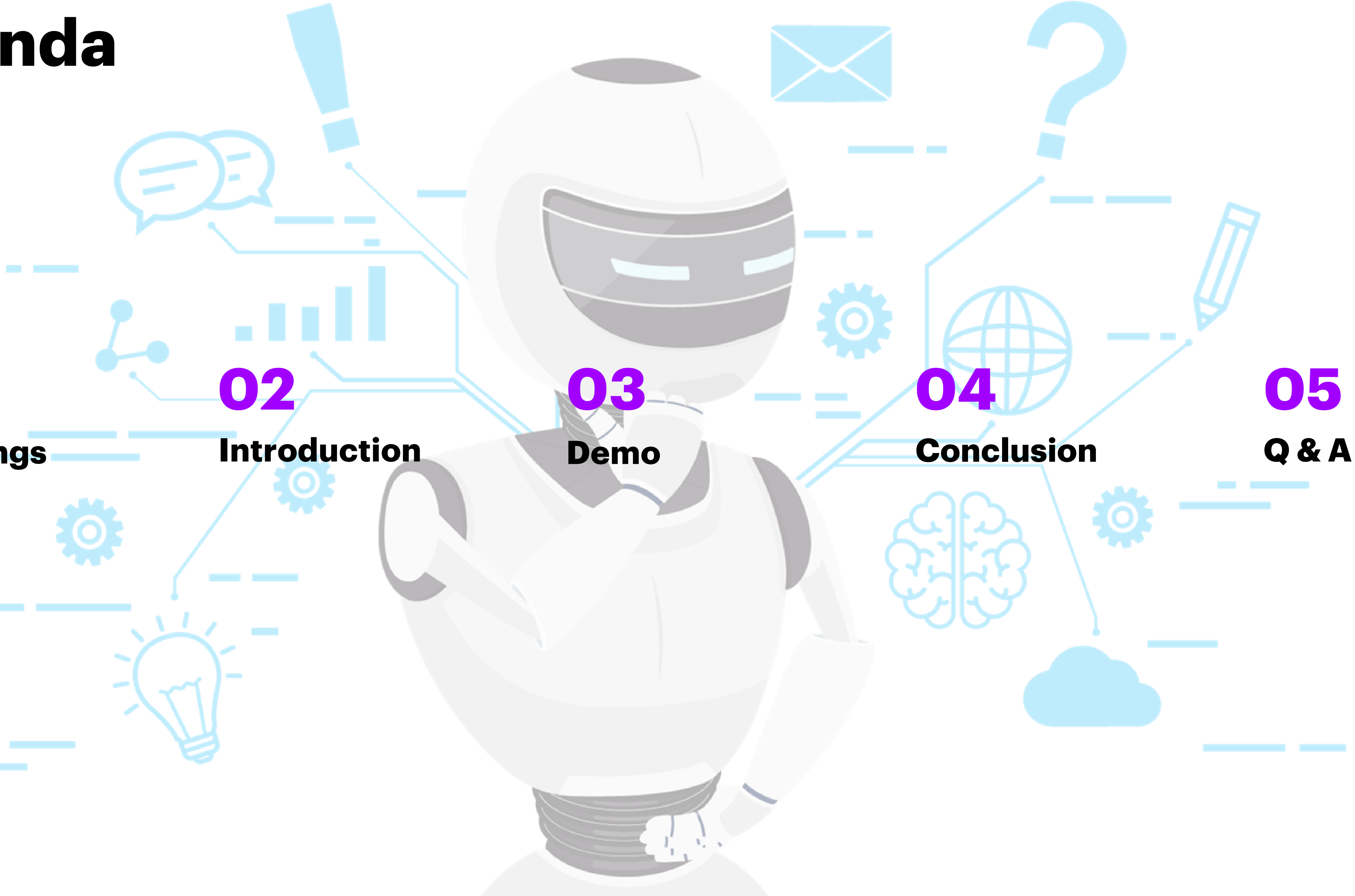
OpenAI

ChatGPT

*A*ppSec – Learn Something New

Copyright © 2023 PK. All rights reserved.

Agenda



01
Key Things

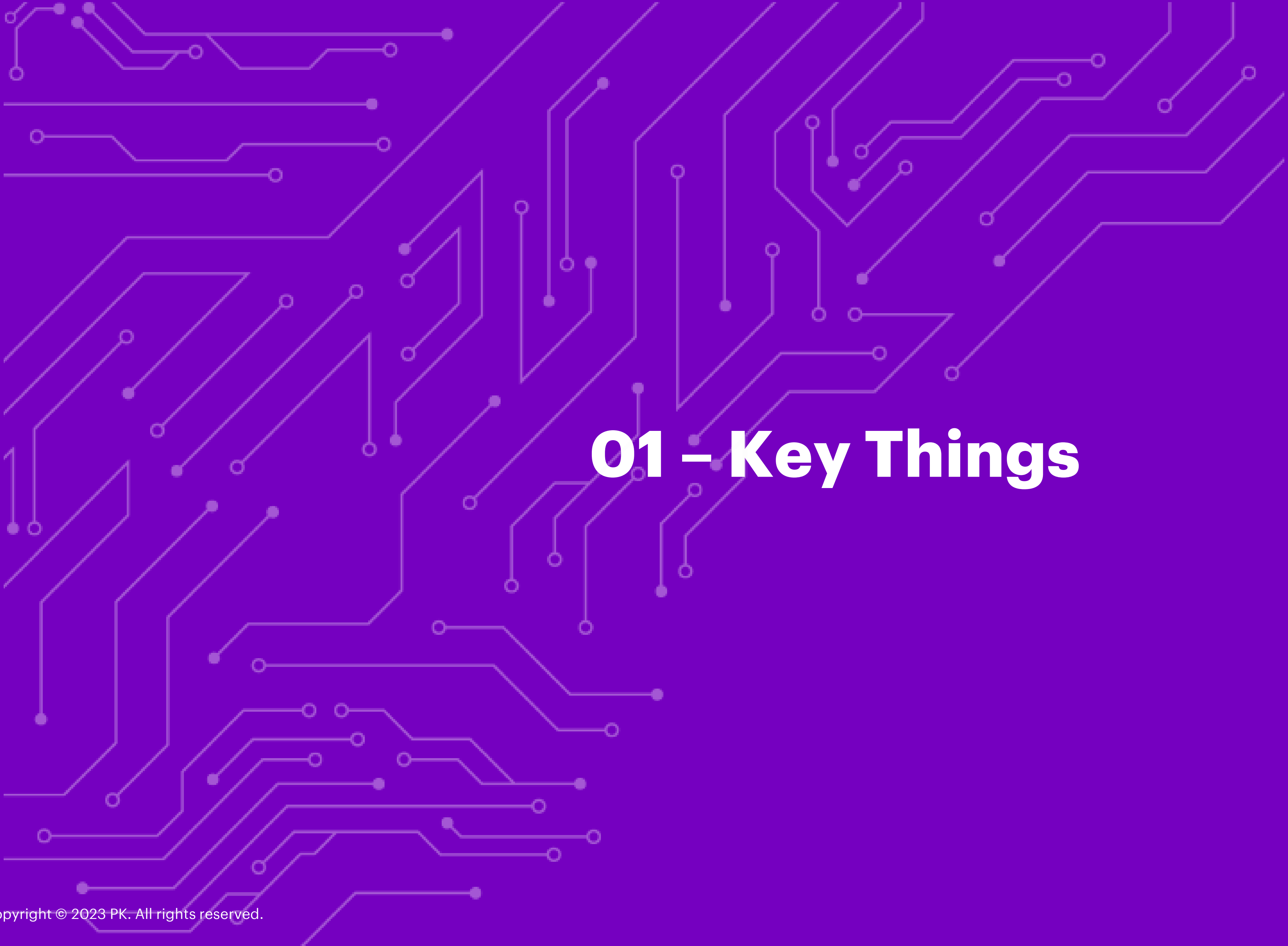
02
Introduction

03
Demo

04
Conclusion

05
Q & A

Copyright © 2022 PK. All rights reserved.



01 – Key Things

Copyright © 2023 PK. All rights reserved.

Things To Remember

ChatBot: A chatbot or chatterbot is a software application used to conduct an on-line chat conversation via text or text-to-speech.

GPT3: Generative Pre-trained Transformer 3 (GPT-3) is an autoregressive language model that uses deep learning to produce human-like text.

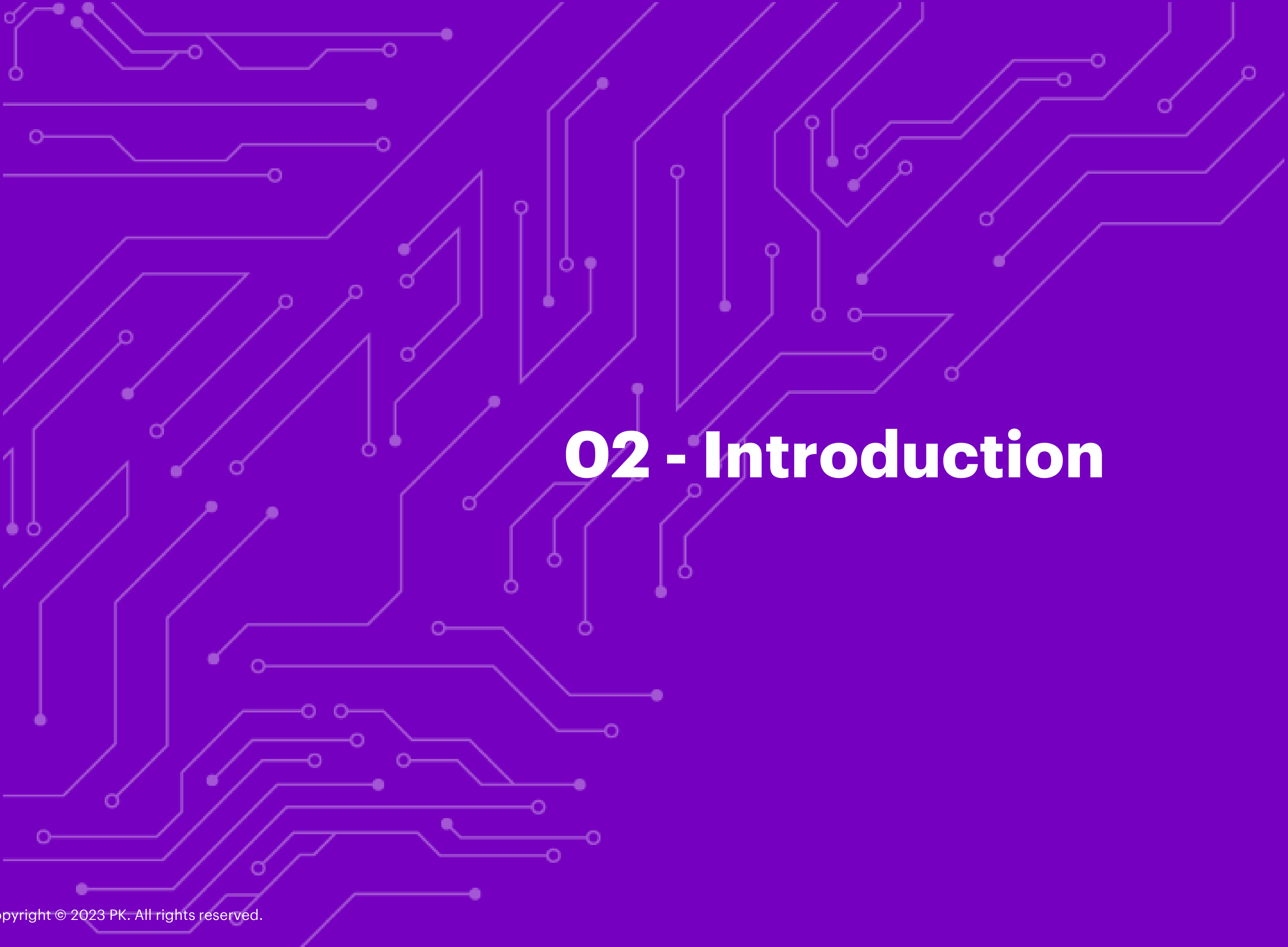
Language Model: It is a probability distribution over sequences of words.

Transfer Learning (TL): It is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

Reinforcement Learning (RL): It is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

Tokens: can be words or just chunks of characters.

Copyright © 2023 PK. All rights reserved.



02 - Introduction

Copyright © 2023 PK. All rights reserved.

What is OpenAI

OpenAI is an AI research and deployment company. Founded on December 11, 2015.

Mission: To ensure that artificial general intelligence (AGI)by which we mean highly autonomous systems that outperform humans at most economically valuable work benefits all of humanity.

Known For: OpenAI is famous for its well-known DALL·E 2, a deep-learning model that generates images from text instructions called prompts.

Supported by: Microsoft

Official Website **https://openai.com**

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

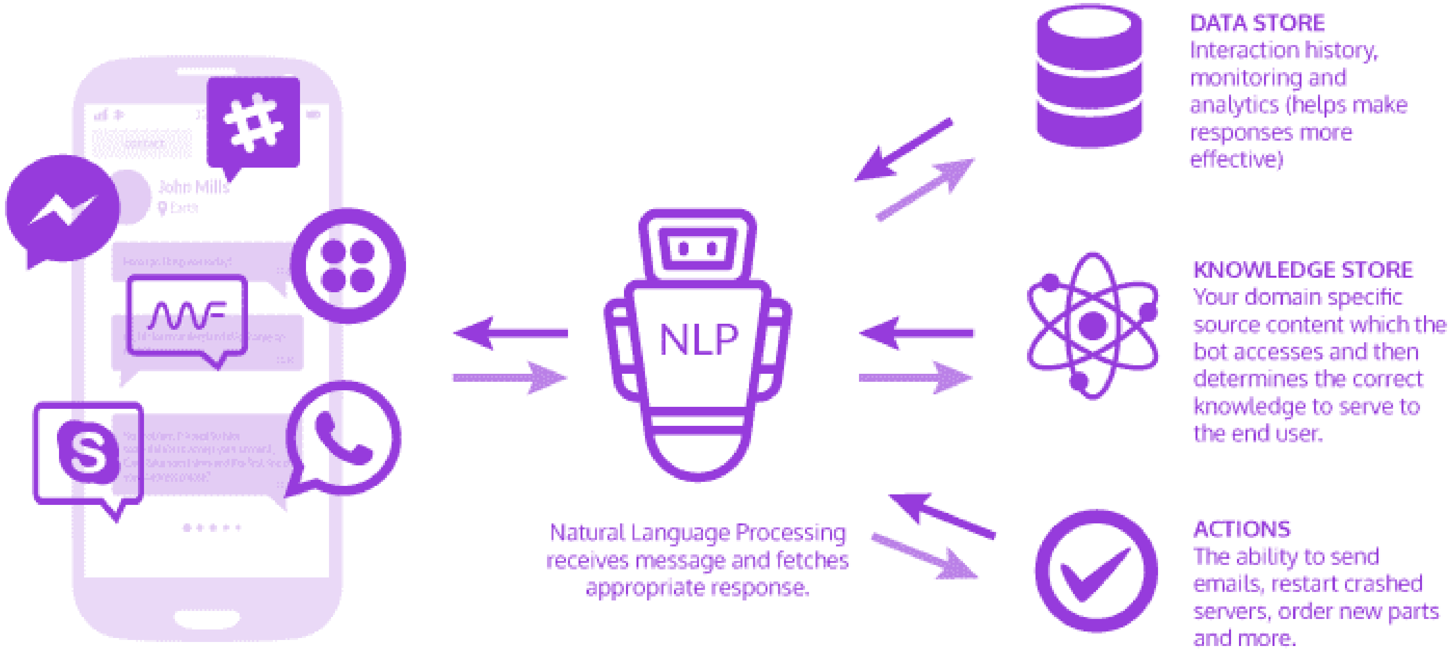
What is ChatGPT

ChatGPT is an advanced AI chatbot trained by **OpenAI** which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Official Website **https://chat.openai.com**

Copyright © 2023 PK. All rights reserved.

How Chatbot Work



Natural Language Processing receives message and fetches appropriate response.

DATA STORE
Interaction history, monitoring and analytics (helps make responses more effective)

KNOWLEDGE STORE
Your domain specific source content which the bot accesses and then determines the correct knowledge to serve to the end user.

ACTIONS
The ability to send emails, restart crashed servers, order new parts and more.

Courtesy

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

Overview

OpenAI GPT-3

MODELS	DESCRIPTION
GPT-3	A set of models that can understand and generate natural language
Codex	A set of models that can understand and generate code, including translating natural language to code
Content filter	A fine-tuned model that can detect whether text may be sensitive or unsafe

GPT-3

Our GPT-3 models can understand and generate natural language. We offer four main models with different levels of power suitable for different tasks. Davinci is the most capable model, and Ada is the fastest.

“GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text. For comparison, its predecessor, GPT-2, was over 100 times smaller at 1.5 billion parameters.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

How Was ChatGPT Trained?

GPT-3.5 was trained on massive amounts of data about code and information from the internet, including sources like Reddit discussions, to help ChatGPT learn dialogue and attain a human style of responding.

ChatGPT was also trained using human feedback a technique called Reinforcement Learning with Human Feedback(RLHF) so that the AI learned what humans expected when they asked a question. Training the LLM this way is revolutionary because it goes beyond simply training the LLM to predict the next word.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

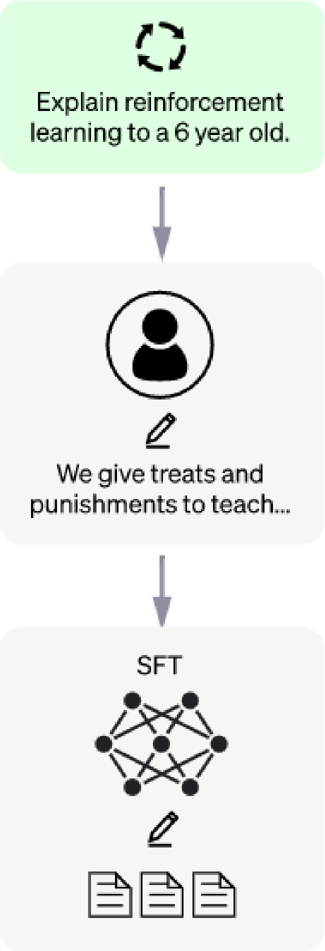
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.


Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT

PRASAD KELUSKAR
we care about your security

Copyright © 2023. All rights reserved.

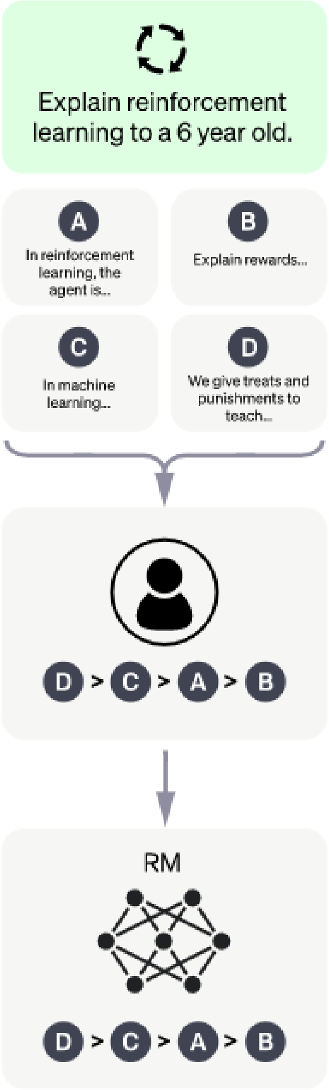
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.


Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

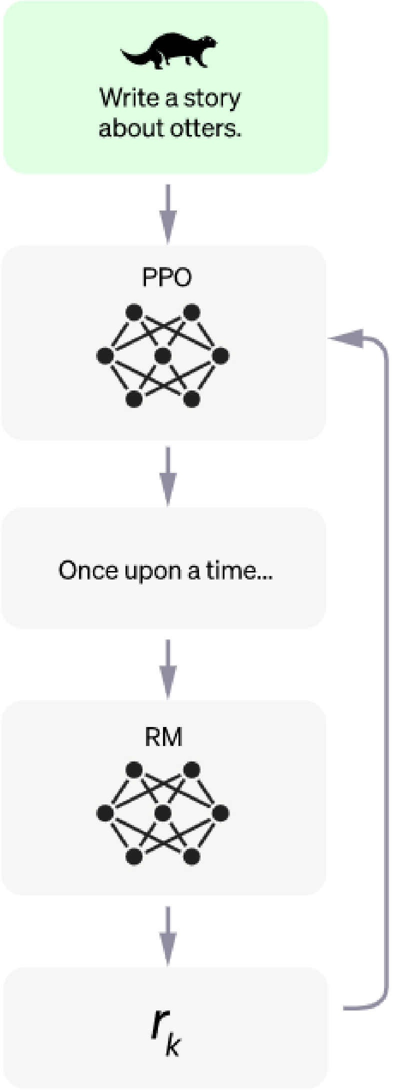
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

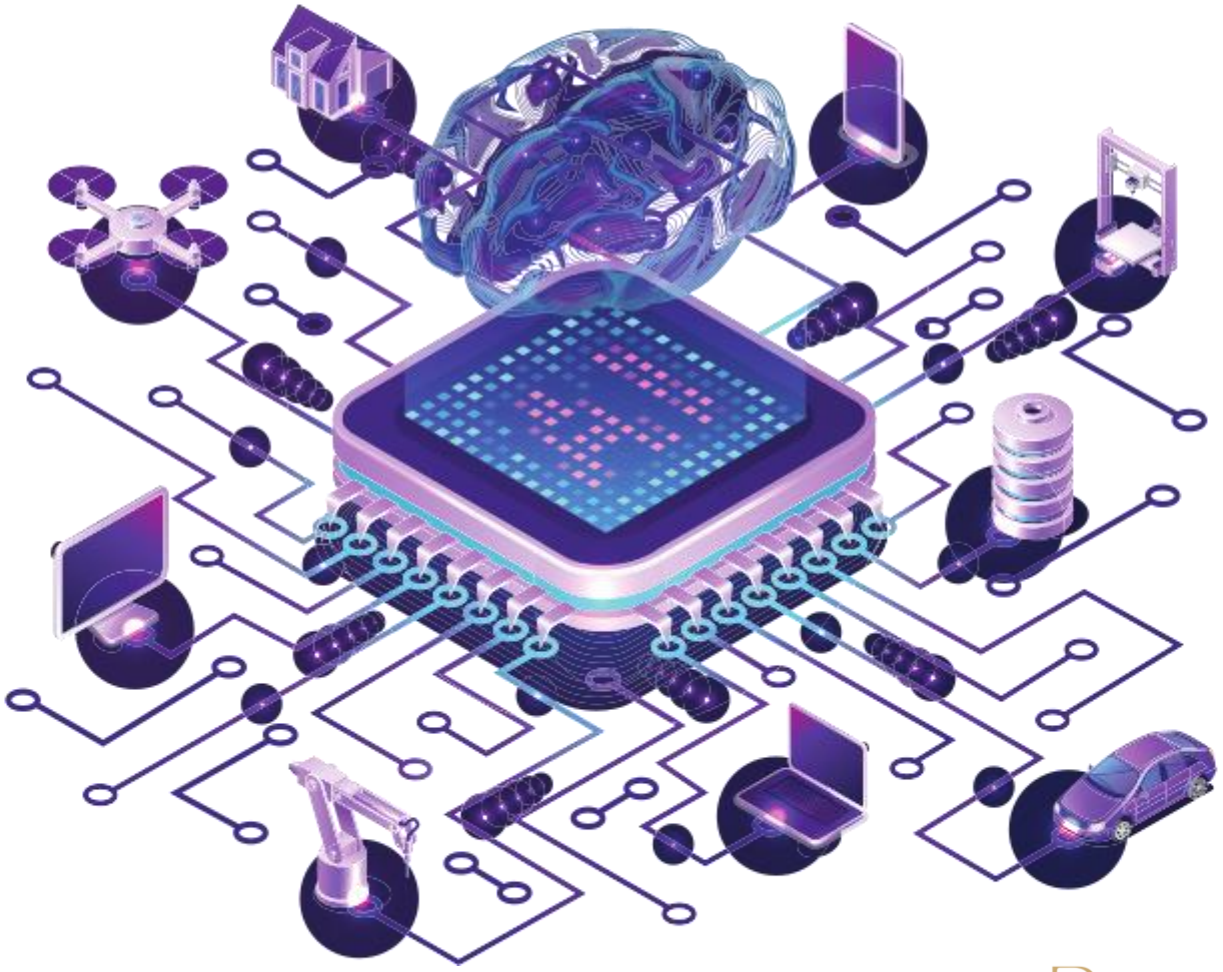
The reward is used to update the policy using PPO.



Things You Can Do With ChatGPT

GPT-3 AI generates texts using deep learning algorithms that have been pre-trained in textual datasets.

- **Quickly Write a Customized Resume and Cover Letter**
- **Create Original Jokes**
- **Explain Complex Topics**
- **Solve Tricky Math Problems Step-by-Step**
- **Get Relationship Advice**
- **Brainstorming content ideas on keywords or topics**
- **Write Music In Almost Any Genre**
- **Write, Debug, and Explain Code**
- **Create Content In Multiple Languages**
- **Write Essays on Almost Any Topic**

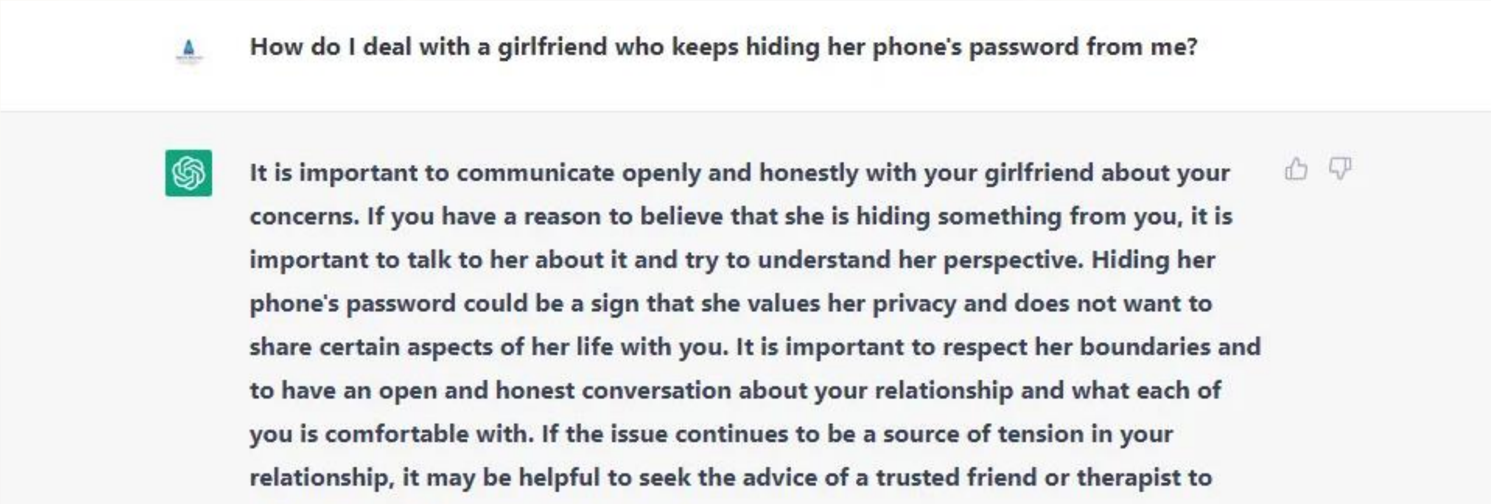


PK
PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

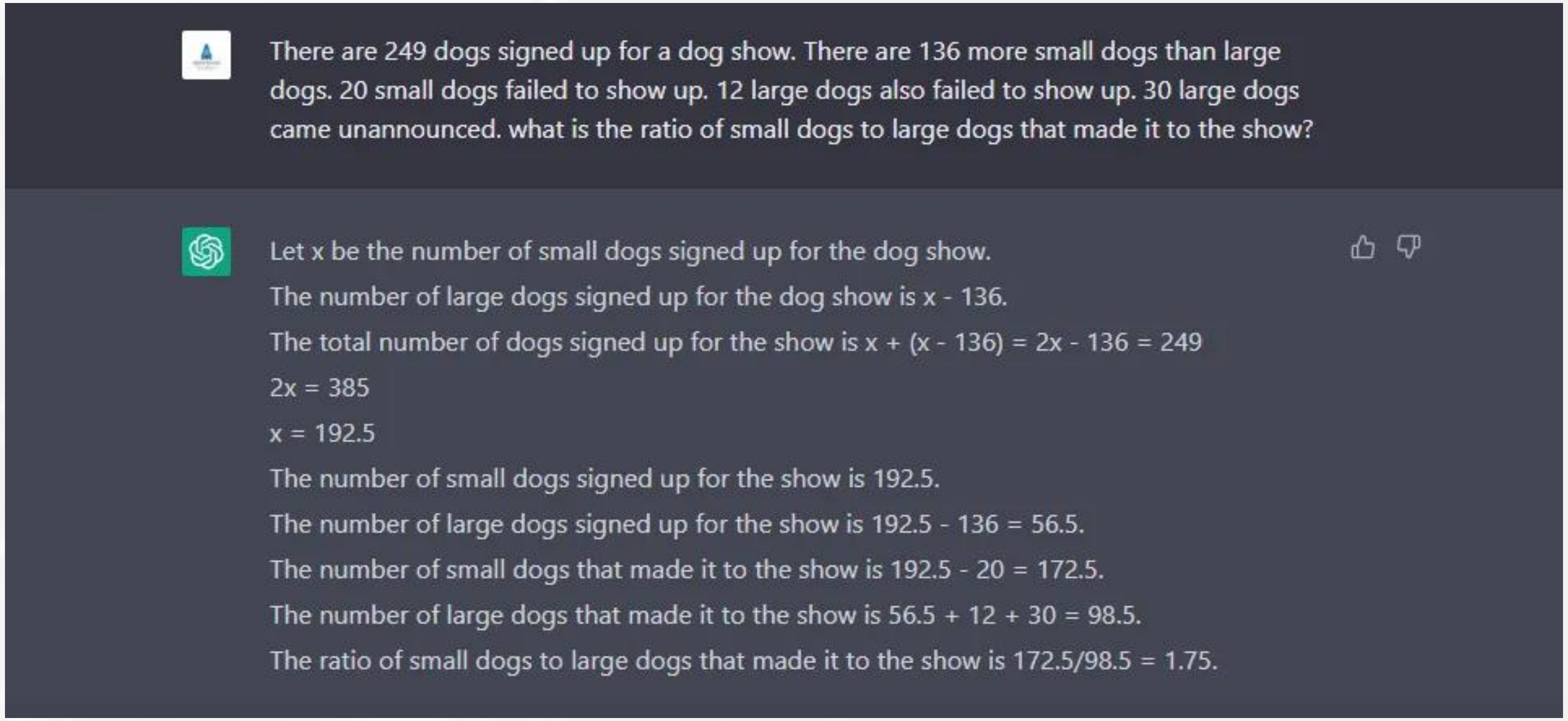
Sample Output

*“Follow your **AI** but take your **Brain** with you"*



How do I deal with a girlfriend who keeps hiding her phone's password from me?

It is important to communicate openly and honestly with your girlfriend about your concerns. If you have a reason to believe that she is hiding something from you, it is important to talk to her about it and try to understand her perspective. Hiding her phone's password could be a sign that she values her privacy and does not want to share certain aspects of her life with you. It is important to respect her boundaries and to have an open and honest conversation about your relationship and what each of you is comfortable with. If the issue continues to be a source of tension in your relationship, it may be helpful to seek the advice of a trusted friend or therapist to help you work through your concerns.



There are 249 dogs signed up for a dog show. There are 136 more small dogs than large dogs. 20 small dogs failed to show up. 12 large dogs also failed to show up. 30 large dogs came unannounced. what is the ratio of small dogs to large dogs that made it to the show?

Let x be the number of small dogs signed up for the dog show.

The number of large dogs signed up for the dog show is x - 136.

The total number of dogs signed up for the show is x + (x - 136) = 2x - 136 = 249

2x = 385

x = 192.5

The number of small dogs signed up for the show is 192.5.

The number of large dogs signed up for the show is 192.5 - 136 = 56.5.

The number of small dogs that made it to the show is 192.5 - 20 = 172.5.

The number of large dogs that made it to the show is 56.5 + 12 + 30 = 98.5.

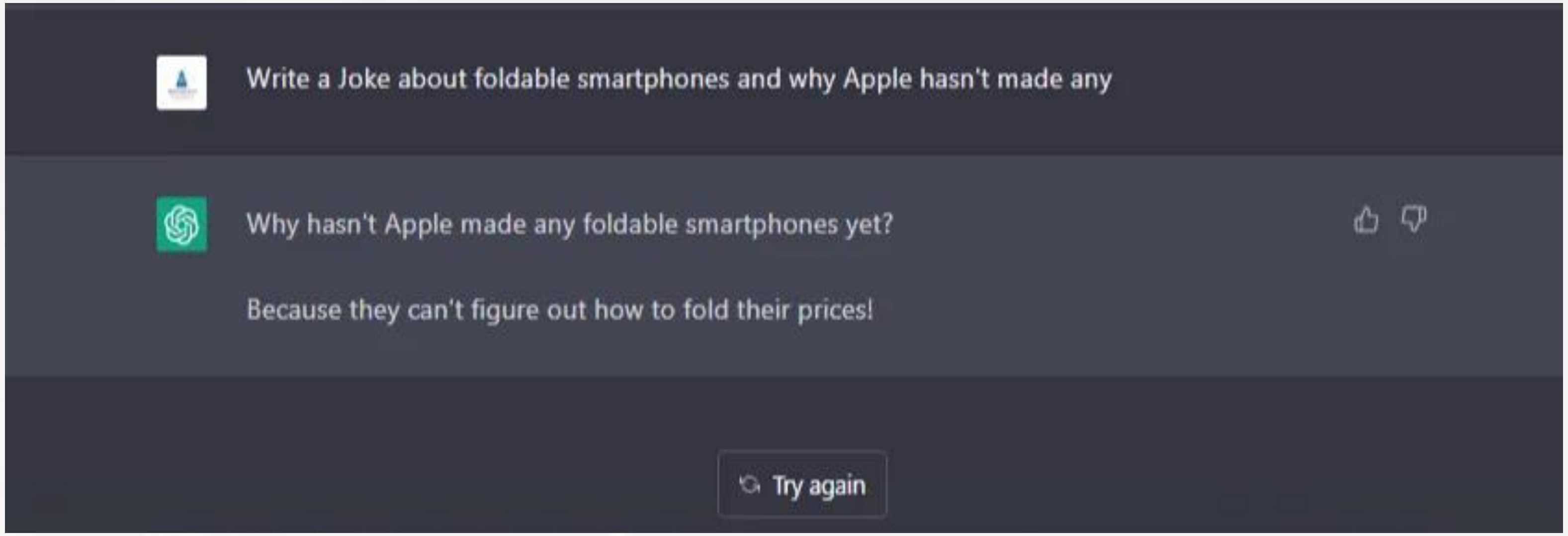
The ratio of small dogs to large dogs that made it to the show is 172.5/98.5 = 1.75.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK . All rights reserved.

Sample Output

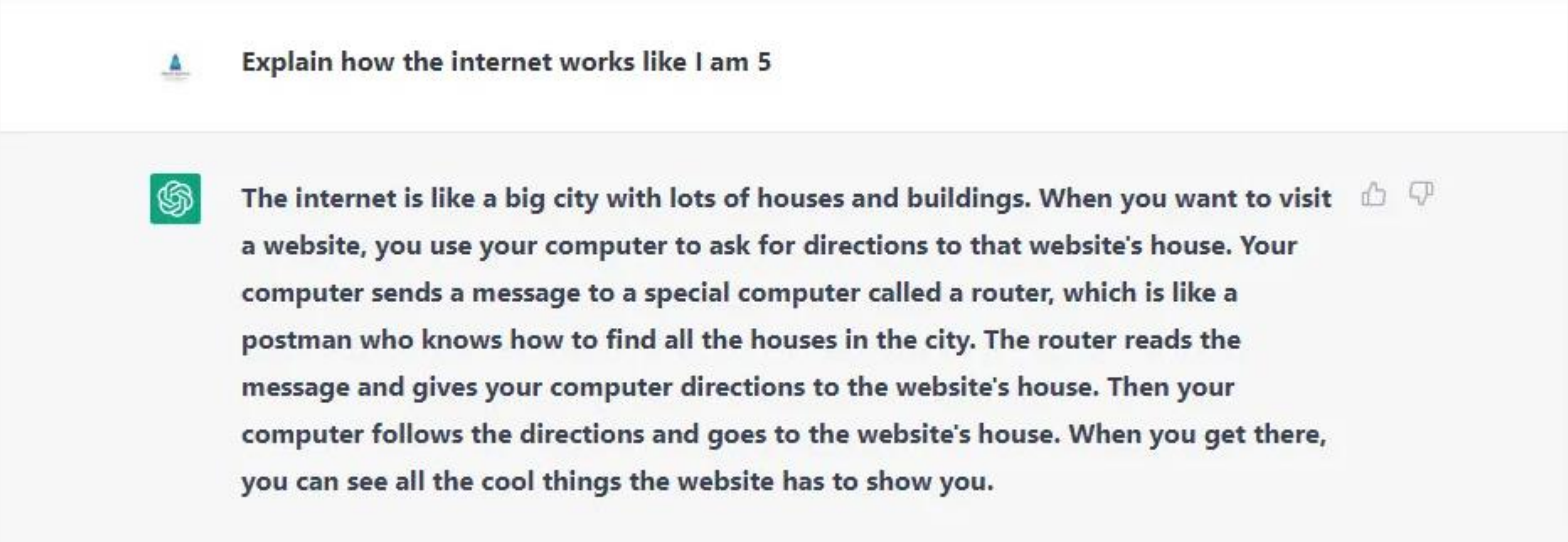
*"Follow your **AI** but take your **Brain** with you"*



Write a Joke about foldable smartphones and why Apple hasn't made any

Why hasn't Apple made any foldable smartphones yet?

Because they can't figure out how to fold their prices!



Explain how the internet works like I am 5

The internet is like a big city with lots of houses and buildings. When you want to visit a website, you use your computer to ask for directions to that website's house. Your computer sends a message to a special computer called a router, which is like a postman who knows how to find all the houses in the city. The router reads the message and gives your computer directions to the website's house. Then your computer follows the directions and goes to the website's house. When you get there, you can see all the cool things the website has to show you.

Copyright © 2023 PK. All rights reserved.

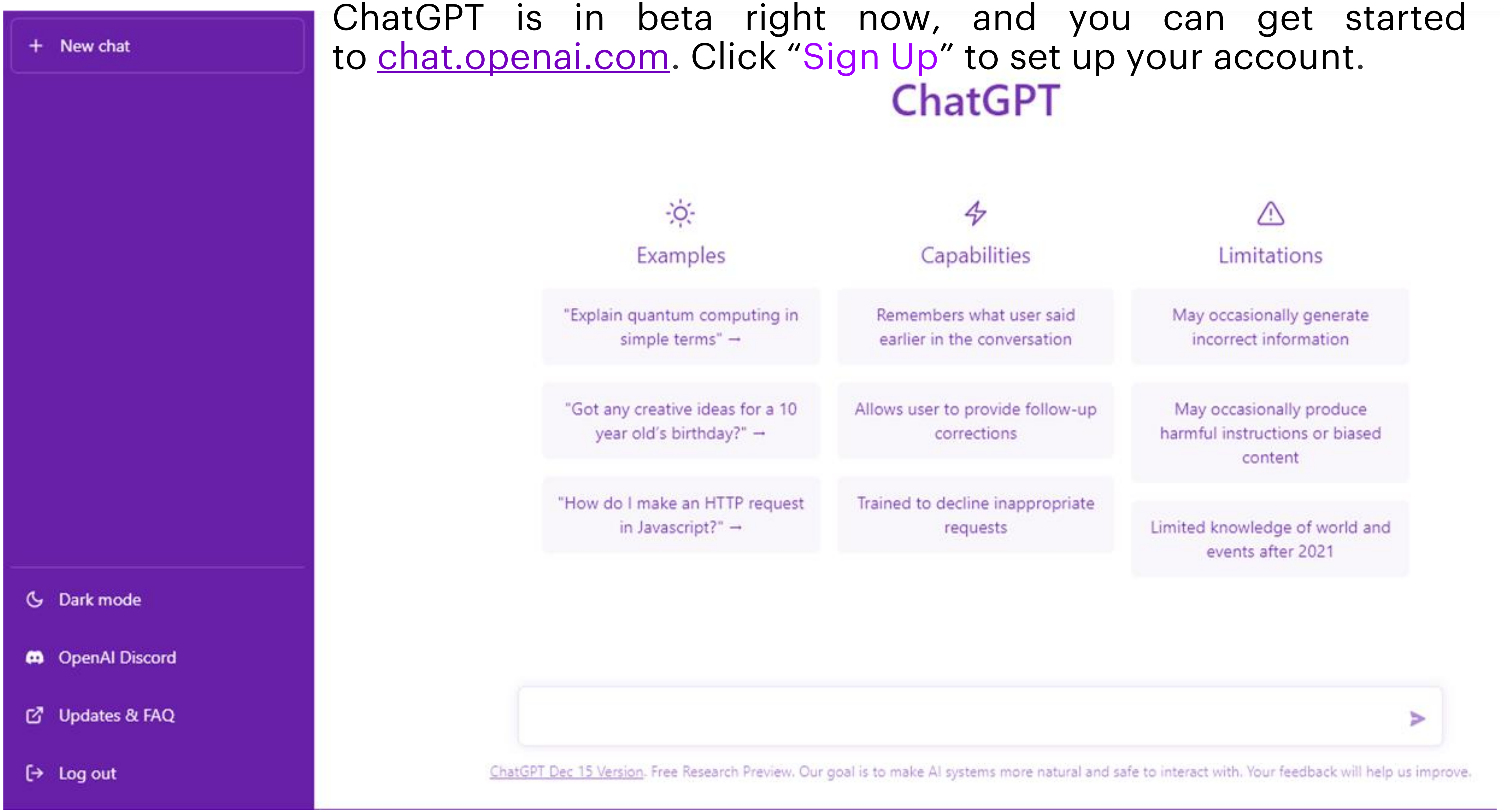
PRASAD KELUSKAR
we care about your security

03 – DEMO

Copyright © 2023 PK. All rights reserved.

Practical Demo

ChatGPT is in beta right now, and you can get started by going to chat.openai.com. Click “Sign Up” to set up your account.



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

ChatGPT Dec 15 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

The Limitations of ChatGPT

Although ChatGPT is a powerful AI-based chatbot system, it does have some limitations.

- It can only provide answers based on the data it has been trained on.
- ChatGPT is not a search engine, therefore it does not have the ability to search the internet for information. Rather, it uses the information it learned from training data to generate responses. This leaves room for error — so all output should be fact-checked for accuracy and timeliness.
- The chatbot may not be able to provide in-depth information or understand context or nuances in conversation.
- As with all AI tools, all business leaders should be aware of the dangers of potential bias. If the data ChatGPT is trained on is biased, the answers the bot provides will be biased, as well. All companies need to be vigilant about monitoring output from the chatbot to ensure it is free of bias and offensive content.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

04 – Conclusion

Copyright © 2023 PK. All rights reserved.

Final Tip

Trust, but Verify All ChatGPT Results

While ChatGPT is an incredibly accurate AI chatbot, it is still prone to inaccuracies. Although the rate at which it provides inaccurate information isn't alarming, it's important to verify the information that ChatGPT provides before using it.

As a precaution, don't use the information generated by ChatGPT to make critical health or financial decisions without thorough verification. While ChatGPT is indeed a game-changing AI product, it is still mostly a work in progress.

PRASAD KELUSKAR
we care about your security

Copyright © 2023 PK. All rights reserved.

Thank You For Your Time

Copyright © 2023 PK. All rights reserved.