

# AI Basics for Everyday Business Users



# Who am I?

- Researcher
- Faculty
- Author
- Specialty: Artificial Intelligence



# Who are you? - Introductions

1. Please introduce yourself
2. What is your experience with RPA and AI?
3. What do you hope to get out of our time together?

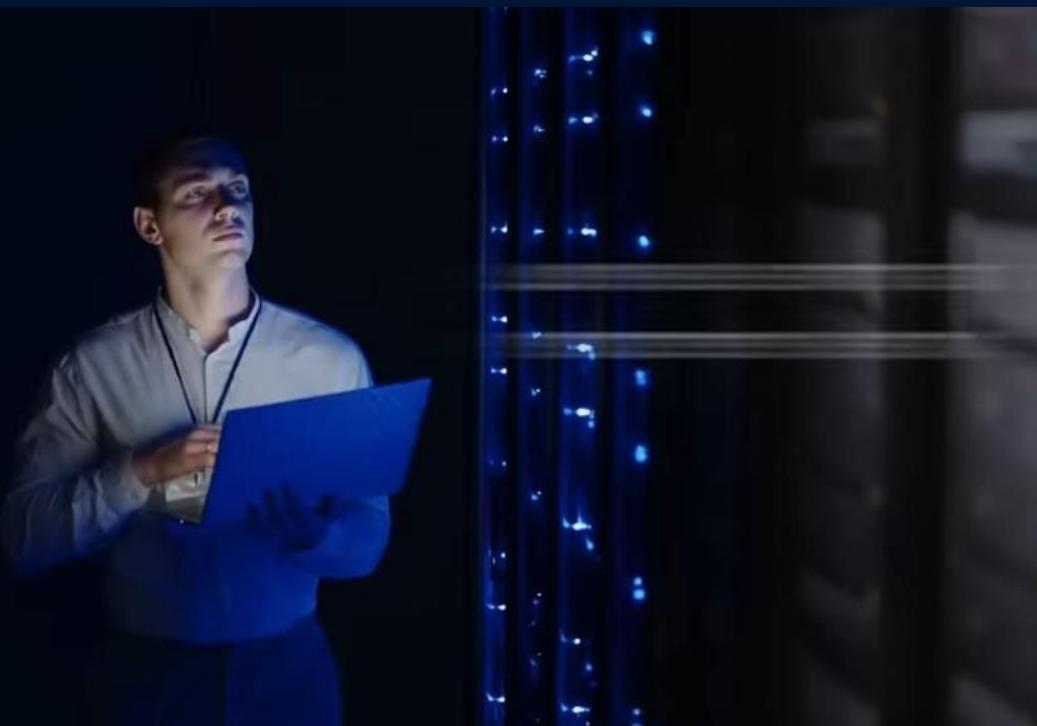
# openAI



**what sets Azure OpenAI apart from other chat Bots is its ability to understand context and provide relevant responses.**

# Generative Pretrained Transformer (GPT)

- Generates responses
- It is pretrained by humans
- Transforms input to meaningful output



## How Does it Work? Supervised learning & reinforcement learning

# ChatGPT



## Examples

"Explain quantum computing in simple terms" →



## Capabilities

Remembers what user said earlier in the conversation



## Limitations

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

In two sentences, explain quantum mechanics in very simple terms



Fl

In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time. thumb up thumb down

**How did it come up with this?**

Fl

In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can



**Predicts what words, phrases and sentences are most associated with the input made**

F

In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.



Chooses words and sentences deemed most likely to be associated with the input

F

In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.



**Outputs words and sentences that it predicts will best answer your question, based on its training**

Fl

In two sentences, explain quantum mechanics in very simple terms

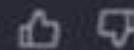


Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time. Like Dislike

**Randomizes output so that the answer  
is not always the same**



Quantum mechanics is a branch of physics that studies the behavior of matter and energy at a very small scale, such as atoms and subatomic particles. It is characterized by phenomena such as superposition, entanglement, and wave-particle duality, which are not easily explained by classical physics.



Was this response better or worse?

Better

Worse

Same



Regenerate response



Fl

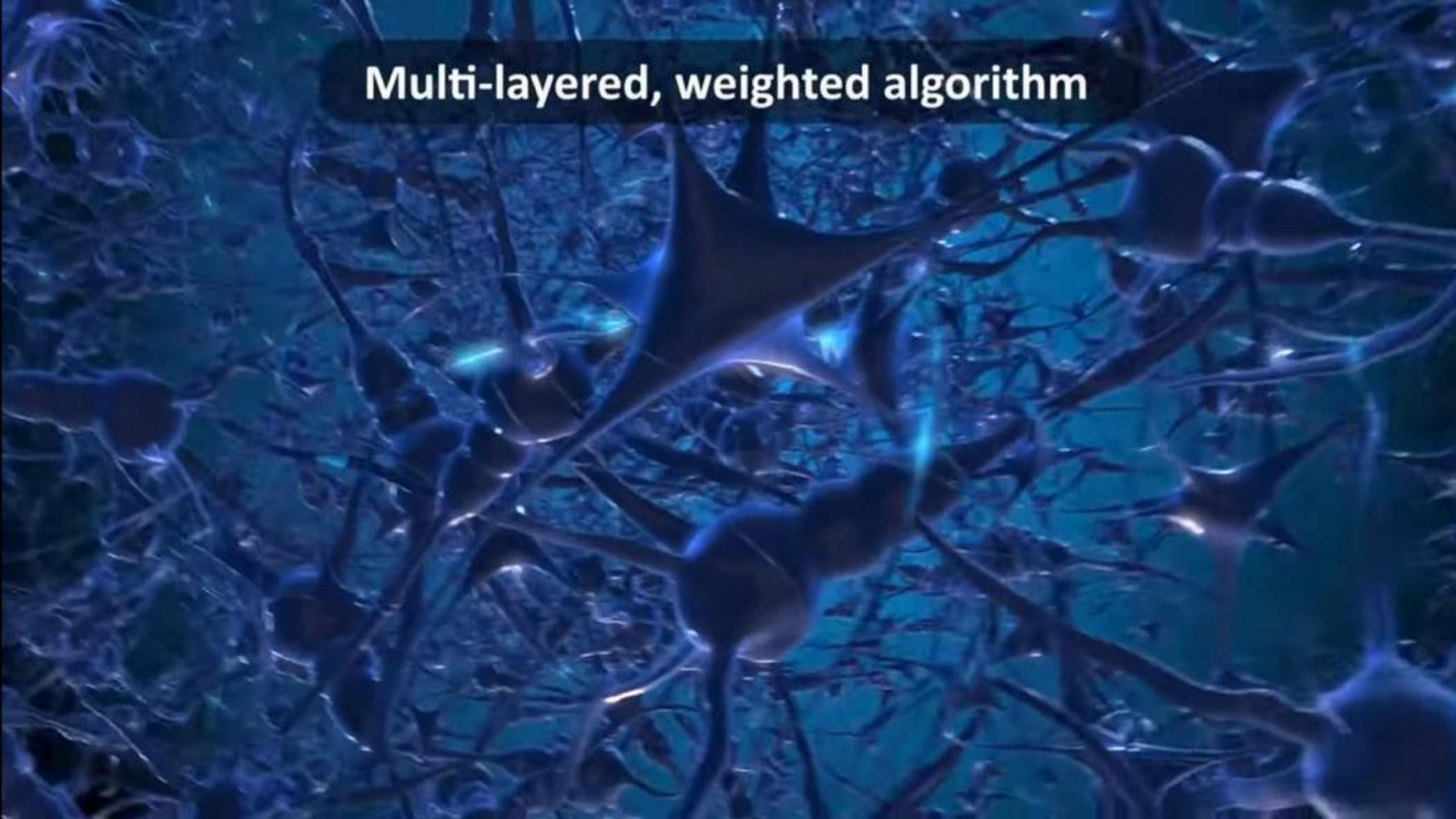
In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.

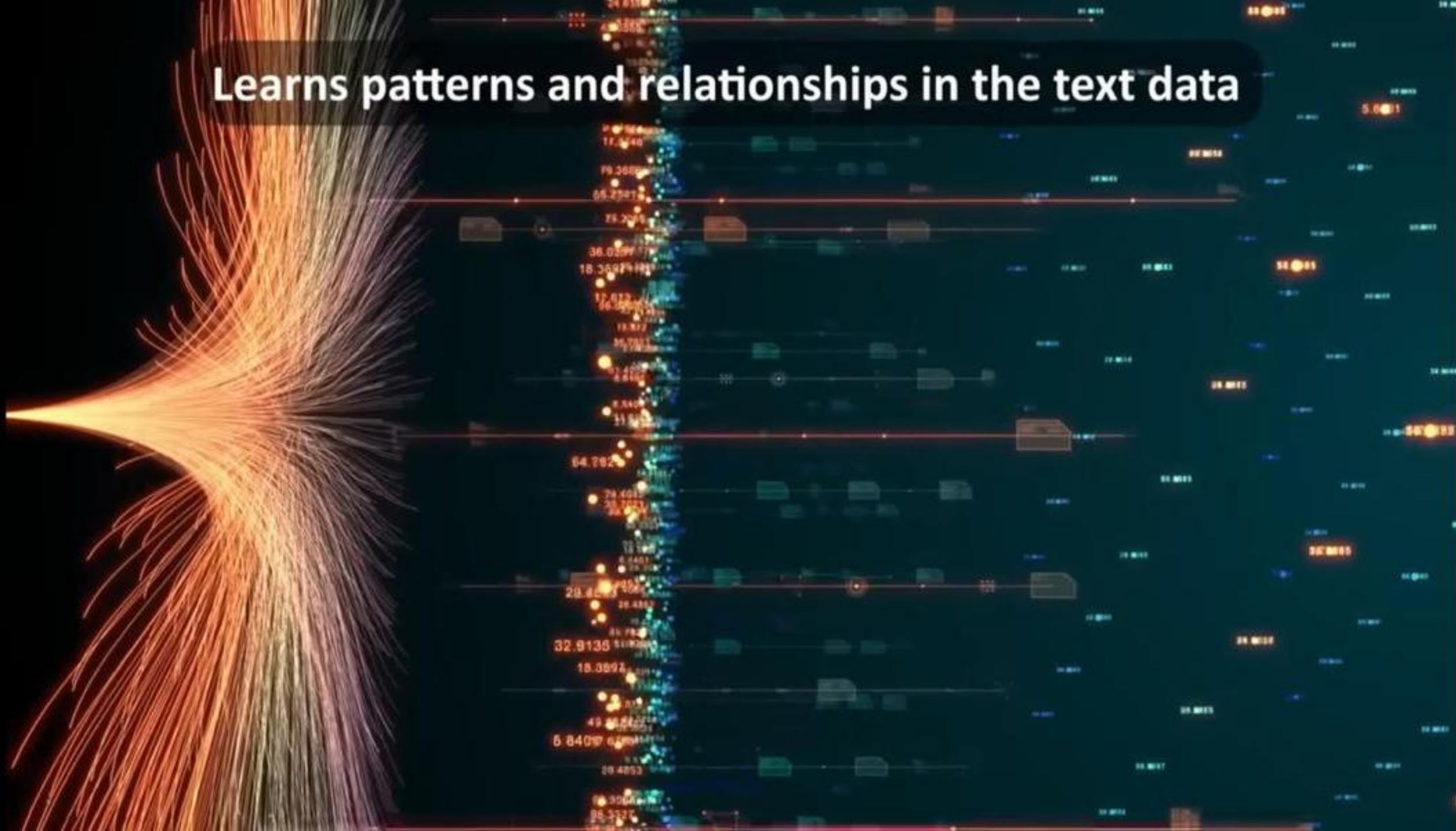


**Determines what words would most likely be expected after having learned how input compares to words written on millions of pages**



Multi-layered, weighted algorithm

Learns patterns and relationships in the text data



Fl

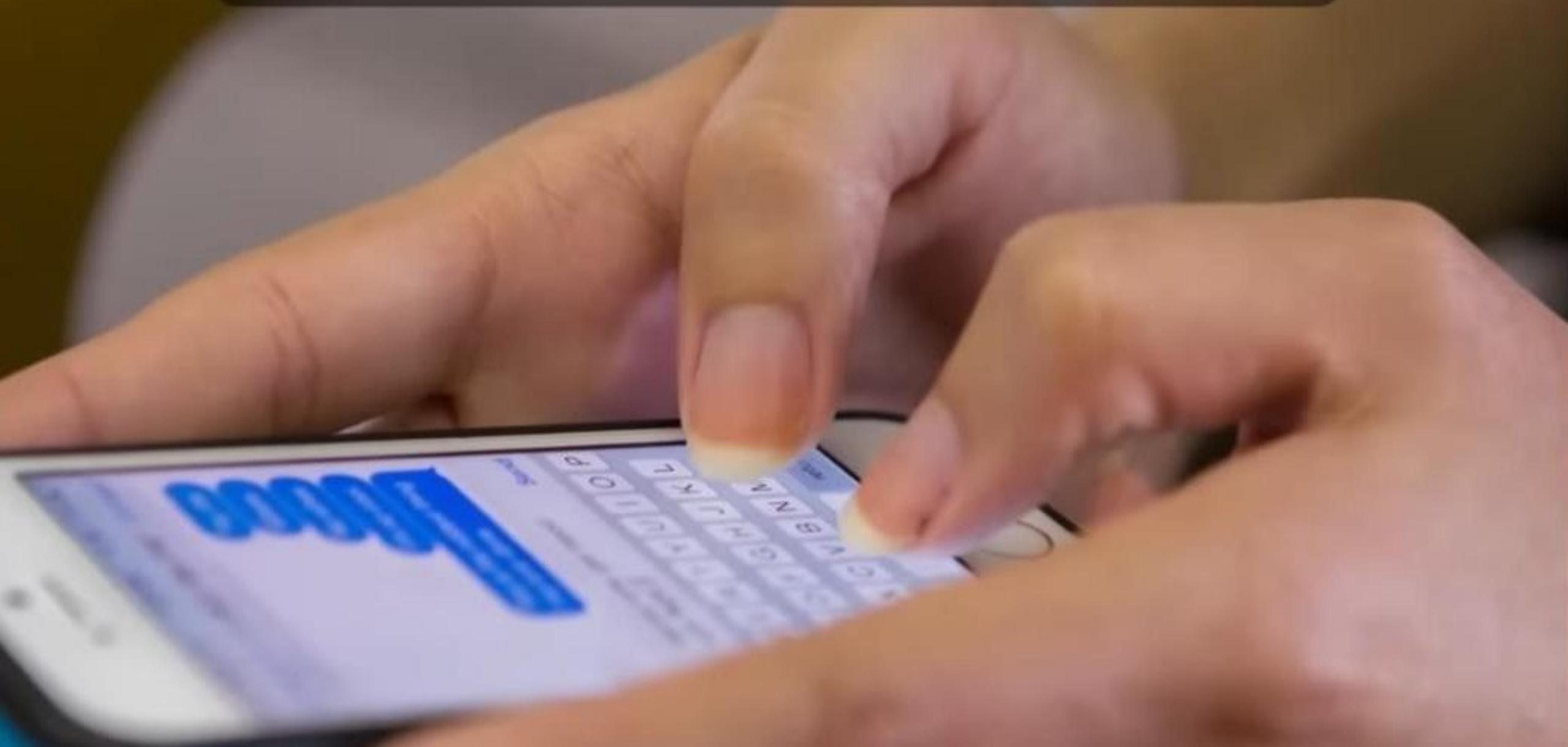
In two sentences, explain quantum mechanics in very simple terms



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.



It's not like the predictive text on your phone



Fl

In two sentences, explain quantum mechanics in very simple terms



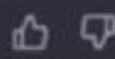
Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.



◀ 2/2



Quantum mechanics is a branch of physics that studies the behavior of matter and energy at a very small scale, such as atoms and subatomic particles. It is characterized by phenomena such as superposition, entanglement, and wave-particle duality, which are not easily explained by classical physics.



Was this response better or worse?

Better

Worse

Same



## Predicts words, sentences and paragraphs

Regenerate response

**Quantum mechanics is**

ChatGPT Mar 14 Version: Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

the smallest scales, such as atoms, subatomic particles, and photons. It was developed in the early 20th century as a result of experimental observations that contradicted classical mechanics, which was the dominant physical theory at the time.

Quantum mechanics describes the behavior of particles in terms of wavefunctions, which are mathematical functions that describe the probability of finding a particle in a particular location or with a particular energy. The theory also introduces the concept of superposition, where a particle can exist in multiple states simultaneously, and the principle of uncertainty, which states that certain properties of particles, such as position and momentum, cannot be measured with absolute precision.

Quantum mechanics has many important applications, including in the development of technologies such as transistors, lasers, and nuclear power. It also plays a key role in understanding the behavior of complex systems, such as molecules and materials, and in the study of fundamental questions in physics, such as the nature of particles and the origins of the universe.

⟳ Regenerate response

Send a message

Quantum mechanics is

Calculates what word comes next  
and what fraction of the time



ChatGPT Mar 14 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Quantum mechanics is

Looks for matches in  
context and meaning

ChatGPT Mar 14 Version: Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

<b>Quantum mechanics is... a</b>	<b>4.5%</b>
<b>Quantum mechanics is... based</b>	<b>3.8%</b>
<b>Quantum mechanics is... fundamentally</b>	<b>3.5%</b>
<b>Quantum mechanics is... described</b>	<b>3.2%</b>
<b>Quantum mechanics is... many</b>	<b>0.7%</b>

Quantum mechanics is... many

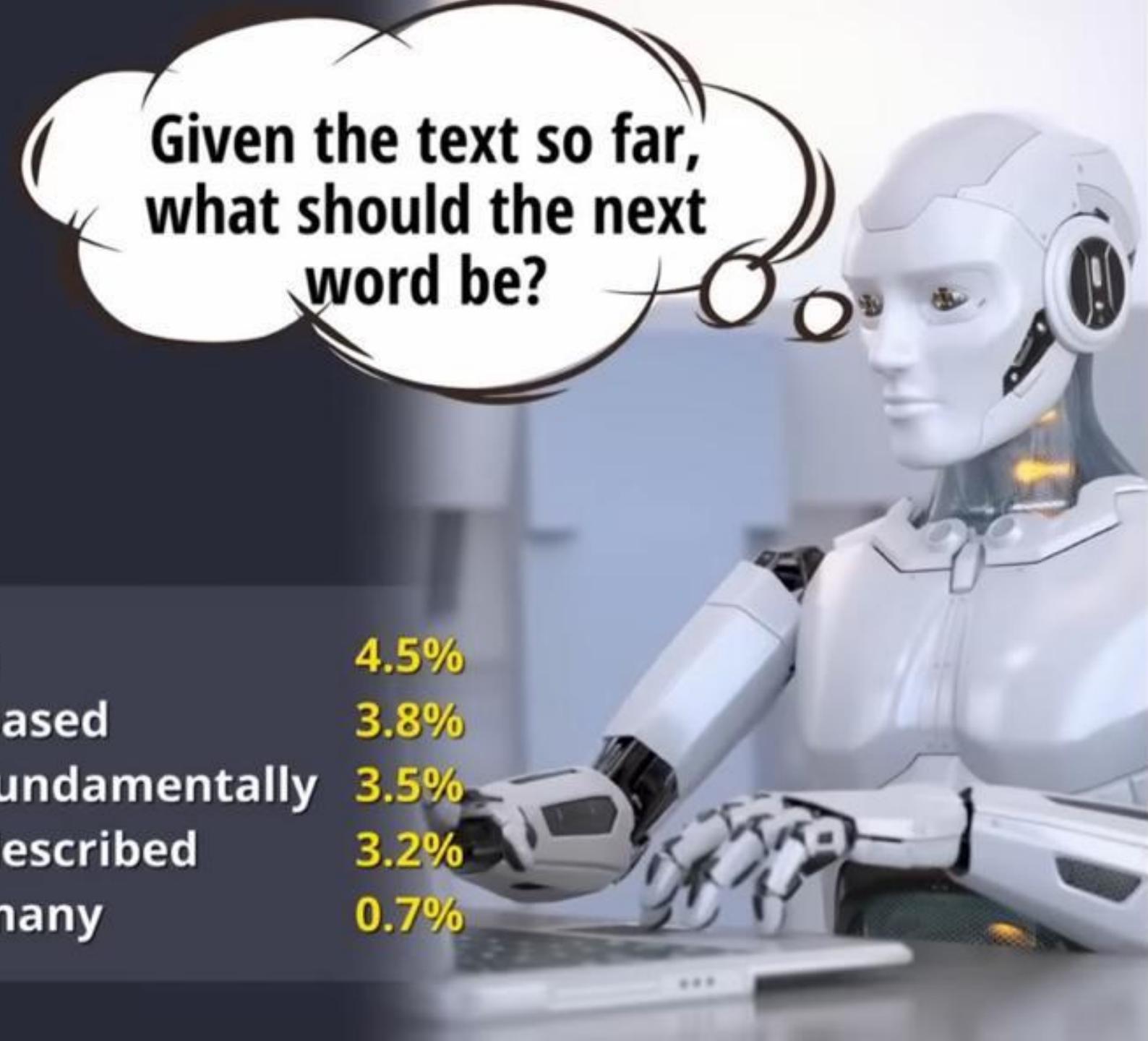
Quantum mechanics is

a  
based  
fundamentally  
described  
many

ChatGPT Mar 14 Ver 0.1. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

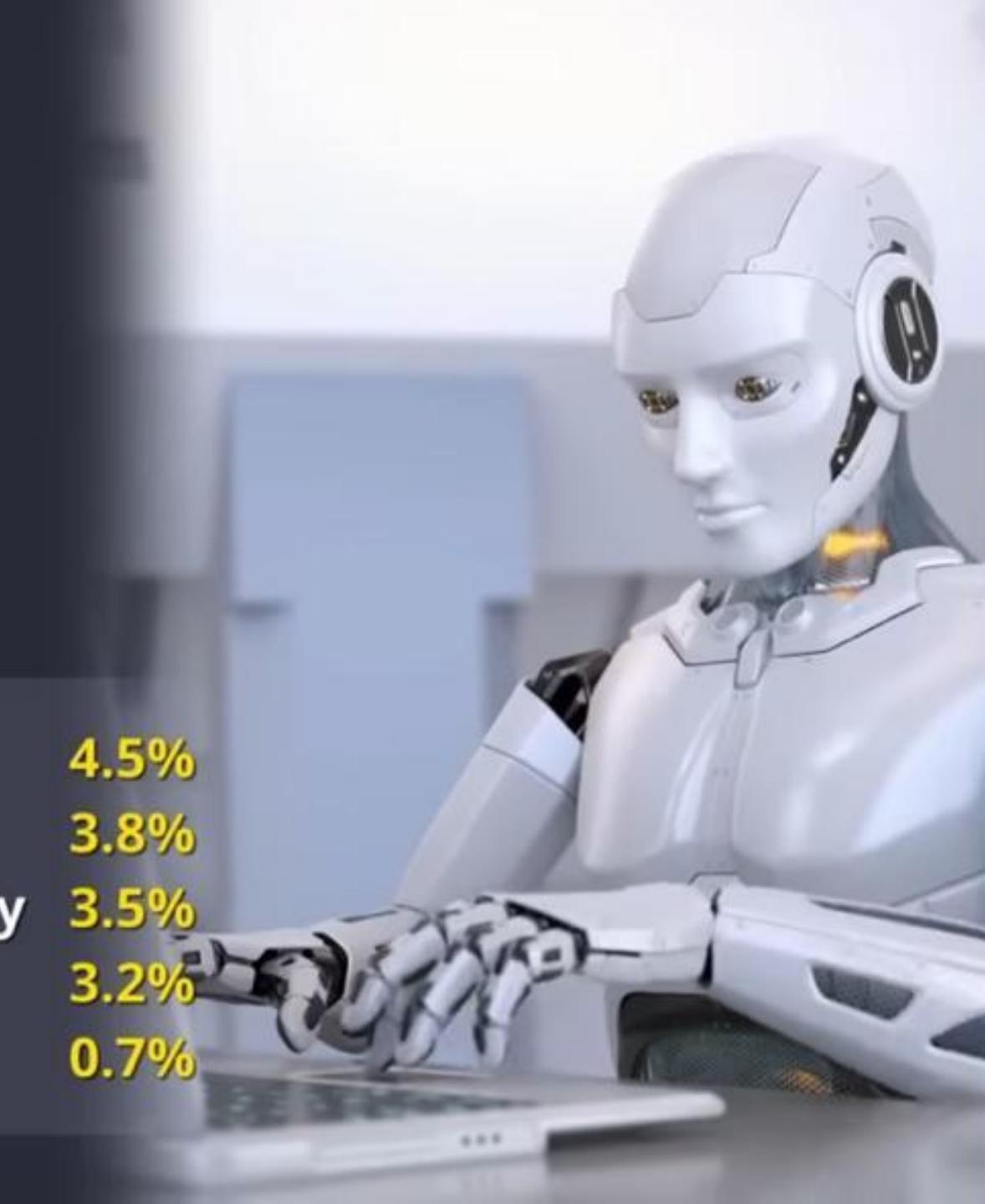
**Given the text so far,  
what should the next  
word be?**

- Quantum mechanics is... a                   **4.5%**
- Quantum mechanics is... based               **3.8%**
- Quantum mechanics is... fundamentally   **3.5%**
- Quantum mechanics is... described         **3.2%**
- Quantum mechanics is... many               **0.7%**



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.

<b>Quantum mechanics is... a</b>	<b>4.5%</b>
<b>Quantum mechanics is... based</b>	<b>3.8%</b>
<b>Quantum mechanics is... fundamentally</b>	<b>3.5%</b>
<b>Quantum mechanics is... described</b>	<b>3.2%</b>
<b>Quantum mechanics is... many</b>	<b>0.7%</b>



Quantum mechanics is a branch of physics that deals with the behavior of tiny particles like atoms and electrons. It tells us that these particles can act like both waves and particles, and can be in many different states at the same time.

Quantum mechanics is...	a	4.5%
Quantum mechanics is...	based	3.8%
Quantum mechanics is...	fundamentally	3.5%
Quantum mechanics is...	described	3.2%
Quantum mechanics is...	many	0.7%



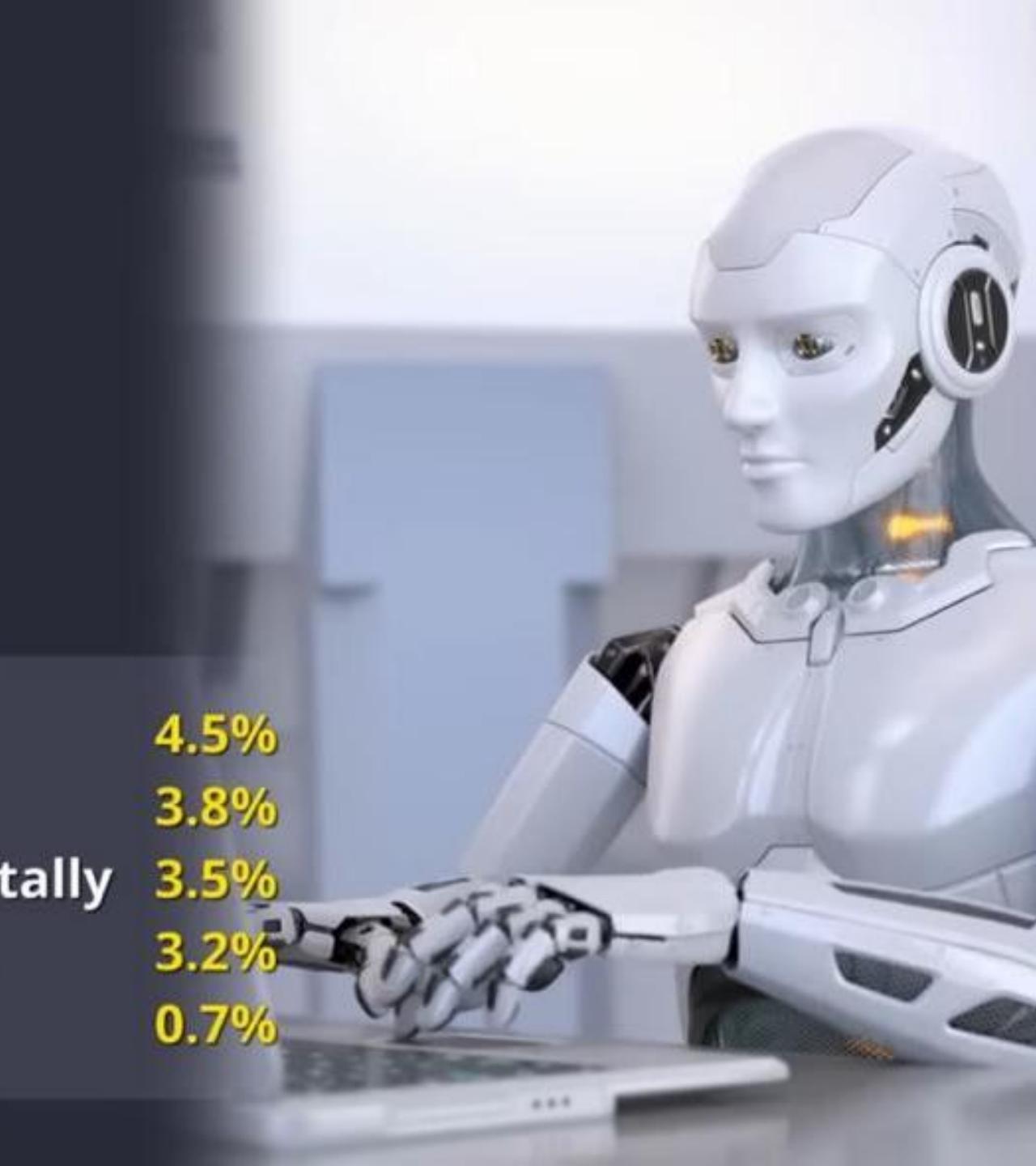
Quantum mechanics is a

Quantum mechanics is a branch

Quantum mechanics is a branch of

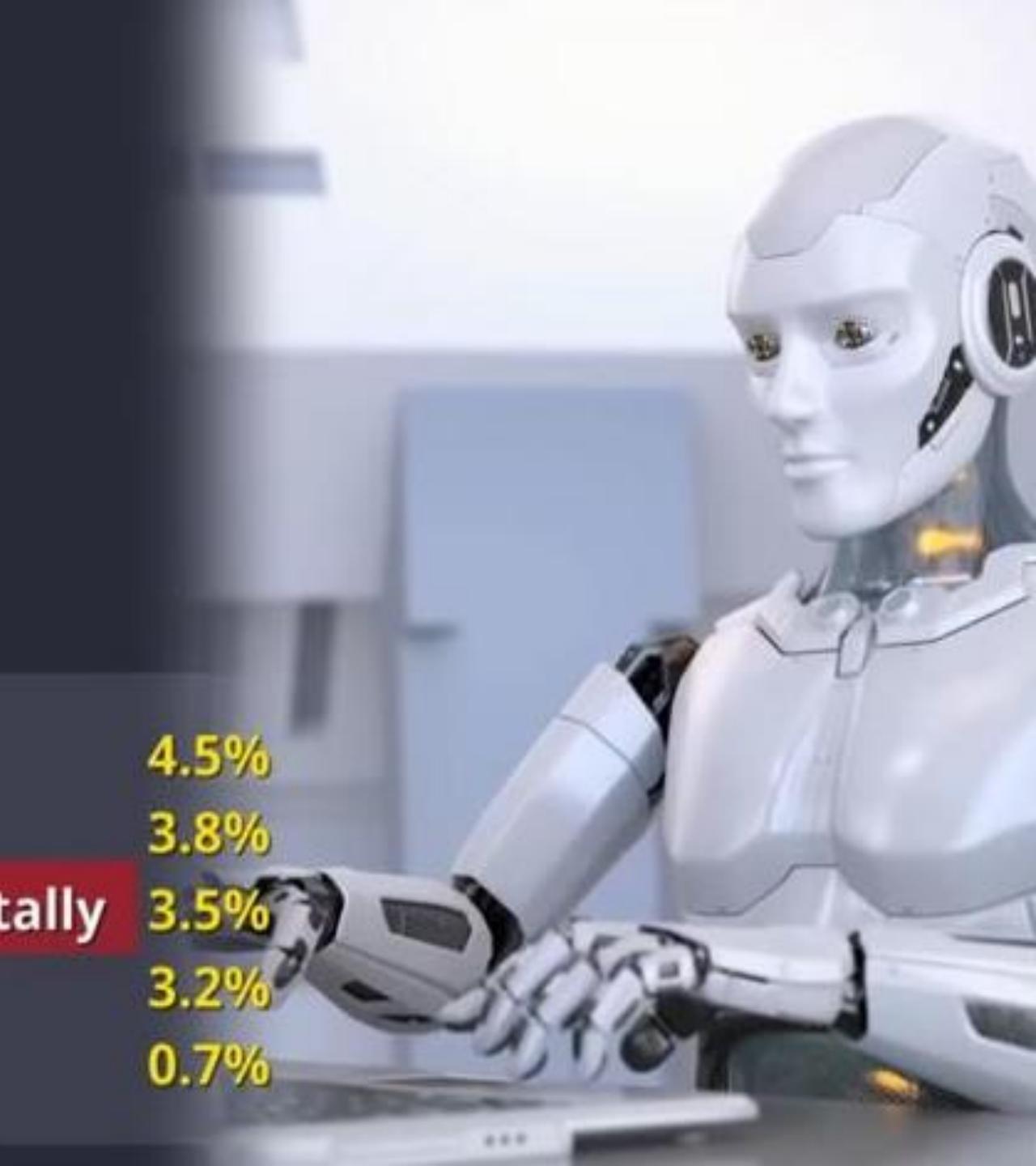
Quantum mechanics is a branch of physics

Quantum mechanics is...	a	4.5%
Quantum mechanics is...	based	3.8%
Quantum mechanics is...	fundamentally	3.5%
Quantum mechanics is...	described	3.2%
Quantum mechanics is...	many	0.7%



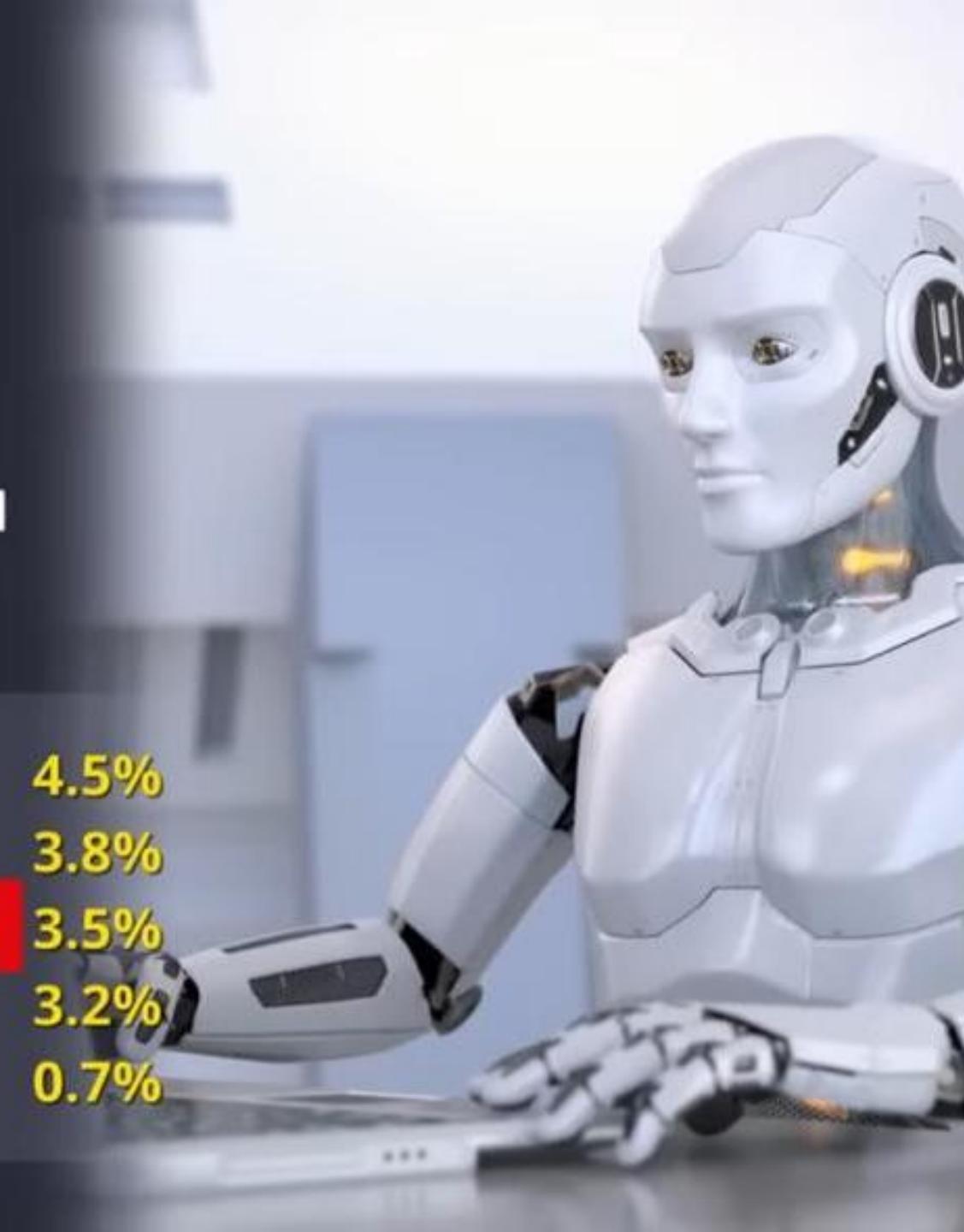
Quantum mechanics is a branch of physics that deals with the behavior of particles on a very small scale, such as atoms and subatomic particles like electrons, protons, and photons.

Quantum mechanics is... a	4.5%
Quantum mechanics is... based	3.8%
Quantum mechanics is... fundamentally	3.5%
Quantum mechanics is... described	3.2%
Quantum mechanics is... many	0.7%



Quantum mechanics is fundamentally  
Quantum mechanics is fundamentally a  
Quantum mechanics is fundamentally a probabilistic  
Quantum mechanics is fundamentally a probabilistic theory

Quantum mechanics is... a	4.5%
Quantum mechanics is... based	3.8%
Quantum mechanics is... <b>fundamentally</b>	3.5%
Quantum mechanics is... described	3.2%
Quantum mechanics is... many	0.7%



# Both are correct



Quantum mechanics is fundamentally a probabilistic theory. It describes the behavior of particles and systems in terms of probabilities rather than deterministically predicting their behavior



Quantum mechanics is a branch of physics that deals with the behavior of particles on a very small scale, such as atoms and subatomic particles like electrons, protons, and photons.



Input: Quantum mechanics is

Potential output: Quantum mechanics is a

Quantum mechanics is based

Quantum mechanics is fundamentally

Quantum mechanics is described

Quantum mechanics is many

## **Model works at sentence and paragraph level in addition to word level**

Quantum mechanics is

Quantum mechanics is a branch of physics that deals with the behavior of particles on a very small scale.

## Model works at sentence and paragraph level in addition to word level

Quantum mechanics is

Quantum mechanics is fundamentally a probabilistic theory.

It describes the behavior of particles and systems in terms of probabilities rather than deterministically predicting their behavior

Quantum mechanics is

Quantum mechanics is a branch of physics that deals with the behavior of particles on a very small scale.

**The model is choosing the best response step by step**

Quantum mechanics is

Quantum mechanics is fundamentally a probabilistic theory.  
It describes the behavior of particles and systems in  
terms of probabilities rather than deterministically  
predicting their behavior

**The model is choosing the best response step by step**

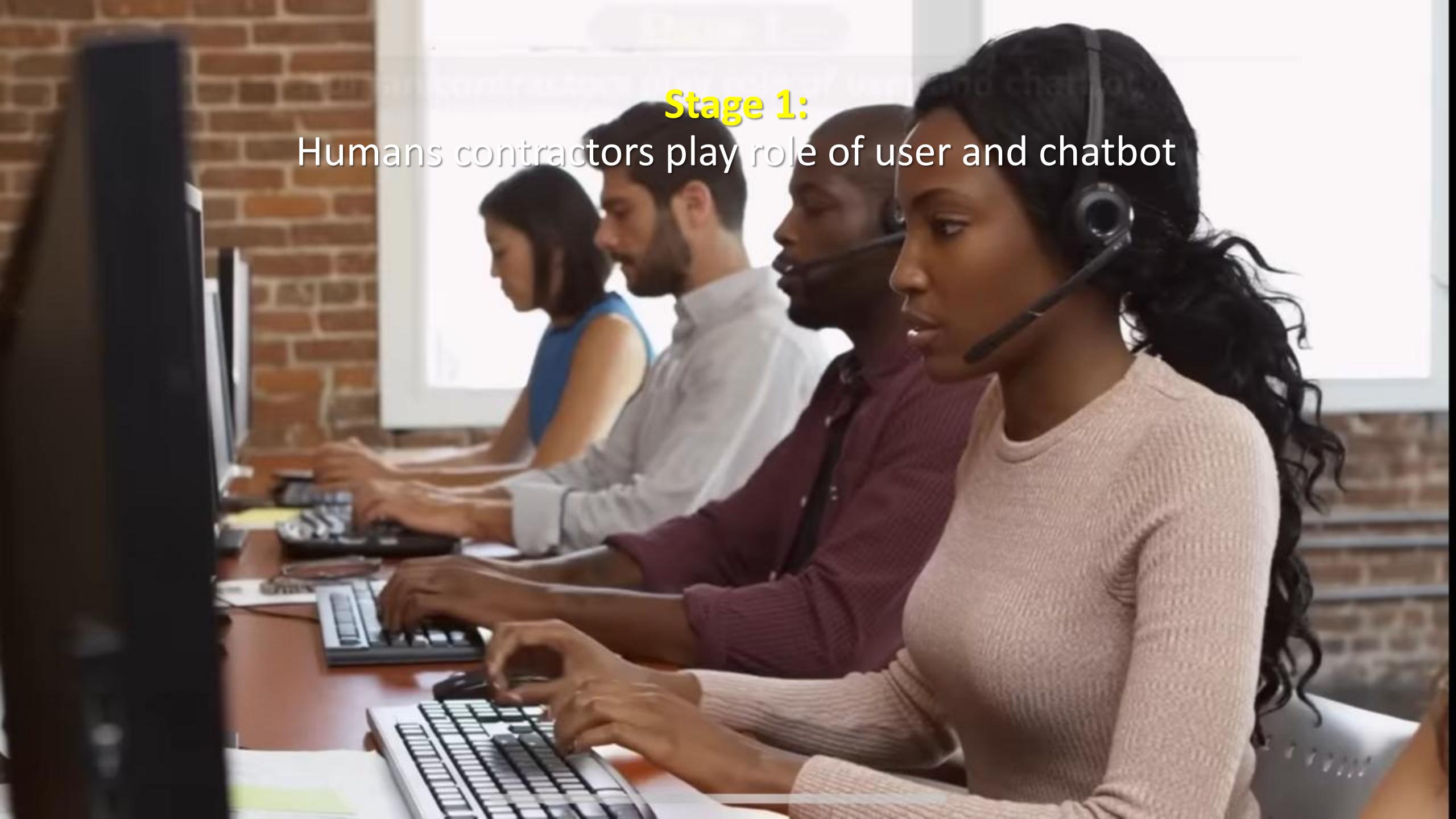


Explain how quantum mechanics works according to the Schrodinger Equation

Explain how quantum mechanics works

...

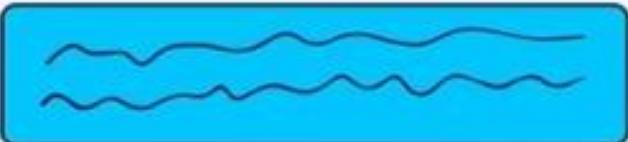
ChatGPT Mar 14 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

A photograph showing four people in an office environment, likely a call center, working at their respective desks. They are all wearing headsets and looking down at their computer monitors. The person in the foreground is a woman with dark hair tied back, wearing a light-colored ribbed sweater. Behind her, from left to right, are a woman in a blue top, a man in a light grey shirt, and another man in a maroon shirt. They are all focused on their work. The background features a brick wall and a window.

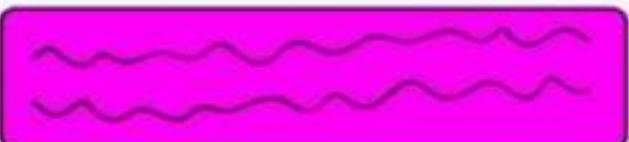
## Stage 1:

Humans contractors play role of user and chatbot

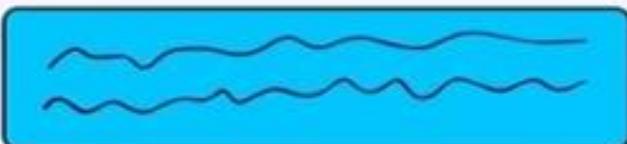
# Thread history of mock conversations is entered into the model



User AB



Ideal Chatbot



User AB



ChatGPT Model

I like cows.



M

I heard they go to college.

Cows go to college?



M

I heard that a cow went to Harvard.

What did the cow study?



M

Bovine sciences.

Do horses ever go to Harvard?



M

Horses go to Hayvard.

• • •



That's a pretty good joke, I feel like you led me into it.





Write an essay on 'Crime and punishment'  
by Dostoyevskii

"Crime and Punishment" is a novel written by Fyodor Dostoyevskii that explores the themes of guilt, redemption, and the role of punishment in society. The novel is set in St. Petersburg, Russia in the mid-19th century and tells the story of Raskolnikov, a young student who, driven by poverty and desperation, murders a pawnbroker and her sister. Throughout the novel, Raskolnikov grapples with the consequences of his crime and the guilt that consumes him.

One of the central themes of the novel is the idea of redemption.

A photograph showing four call center agents in profile, facing left. They are wearing headsets and working at computer stations with multiple monitors. The background features a brick wall.

**Stage 2:**  
Finetuning using rank and reward system



## Describe an atom

- A) It's the smallest part of a substance made of electrons, neutrons, and protons. **0.95**
- B) It's a basic chemical element **0.37**
- C) It's an object made of subatomic particles **0.48**
- D) It's a ticketing service **0.12**

Teaches ChatGPT to evaluate  
what the best output is  
likely to be

A > C > B > D



ChatGPT Model

**Human trainers would have to input  
any potential query – not possible!**



# ChatGPT



## Examples

"Explain quantum computing in simple terms" →



## Capabilities

Remembers what user said earlier in the conversation



## Limitations

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

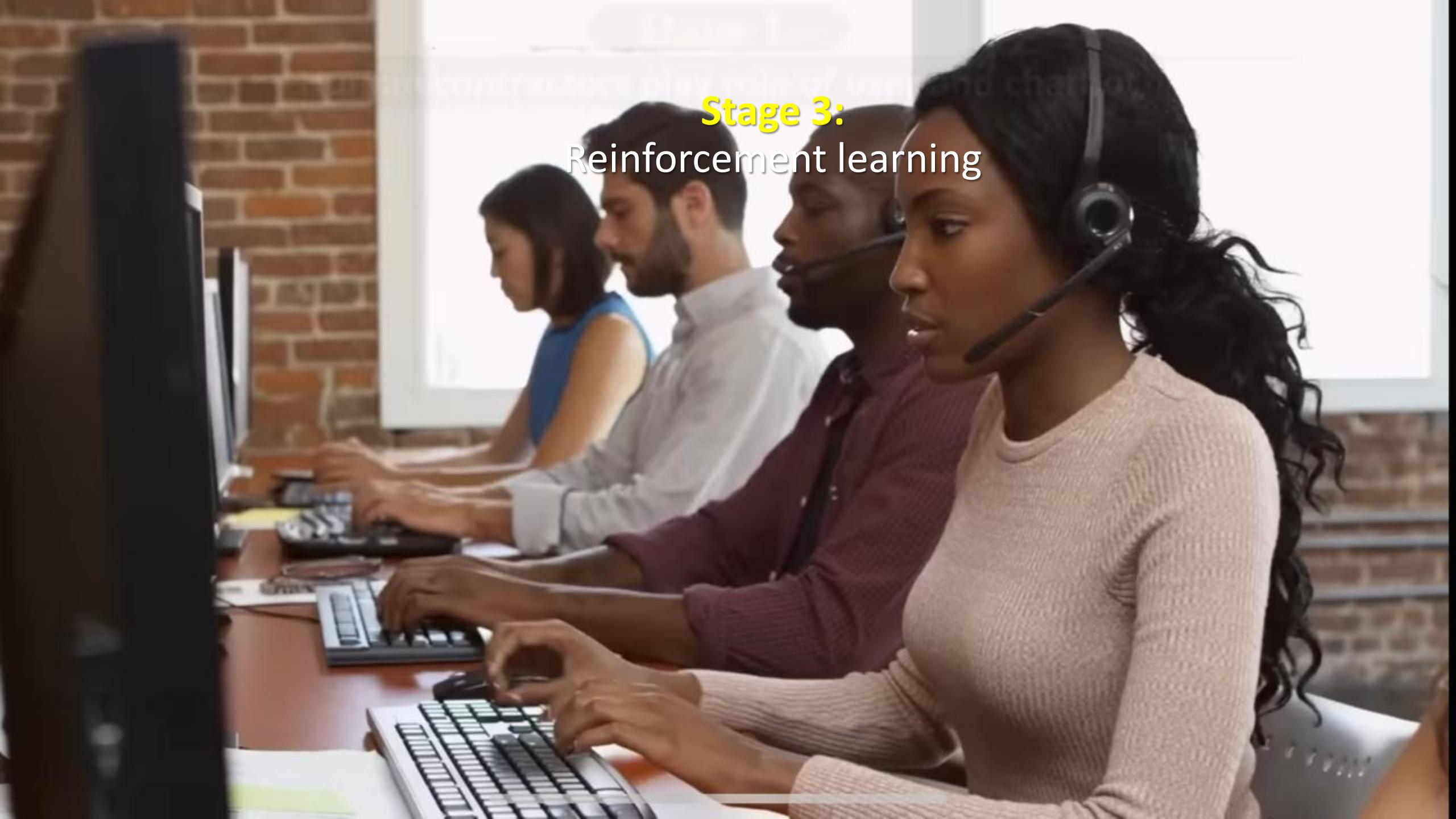
"How do I make an HTTP request in Javascript?" →

Trained to decline inappropriate requests

Limited knowledge of world and events after 2021

I  
Write a short story about bread



A photograph showing four people in an office environment, likely a call center, working at their respective computer stations. They are all wearing headsets with microphones. The person in the foreground is a woman with dark hair tied back, wearing a light-colored ribbed sweater. She is looking down at her keyboard. Behind her, from left to right, are three other individuals: a woman in a blue tank top, a man in a light-colored button-down shirt, and another man in a maroon long-sleeved shirt. They are all focused on their work. The background features a brick wall and a window.

**Stage 3:**  
Reinforcement learning

Unsupervised learning based on the  
human-taught ranking system



**Unsupervised learning based on the  
human-taught ranking system**



**Unsupervised learning allows the model to learn patterns from much more data and many more subjects**





# 45 terabytes of data

## GET STARTED

[Introduction](#)[Quickstart](#)[Libraries](#)[Models](#)[Overview](#)[GPT-4](#)[GPT-3.5](#)[DALL-E](#)[Whisper](#)[Embeddings](#)[Moderation](#)[GPT-3](#)[Tutorials](#)[Data usage policies](#)[Usage policies](#)

## Embeddings

A set of models that can convert text into a numerical form

## Moderation

A fine-tuned model that can detect whether text may be sensitive or unsafe

## GPT-3

A set of models that can understand and generate natural language

## Codex

Deprecated

A set of models that can understand and generate code, including translating natural language to code

We have also published open source models including [Point-E](#), [Whisper](#), [Jukebox](#), and [CLIP](#).

Visit our [model index for researchers](#) to learn more about which models have been featured in our research papers and the differences between model series like InstructGPT and GPT-3.5.

## GPT-4

Limited beta

GPT-4 is a large multimodal model (accepting text inputs and emitting text outputs today, with image inputs coming in the future) that can solve difficult problems with greater accuracy than any of our previous models, thanks to its broader general knowledge and advanced reasoning capabilities. Like [gpt-3.5-turbo](#), GPT-4 is optimized for chat but works well for traditional completions tasks. Learn how to use GPT-4 in our [chat guide](#).



Each terabyte equivalent to about 83 million pages!



community. She decided to use her bread-baking skills to feed her neighbors.

Maria woke up early every day to begin baking loaves of bread. She worked tirelessly, kneading the dough and carefully watching the rising process. Her hands became covered in flour, and her apron became stained with the evidence of her hard work.

As the smell of fresh bread wafted through the village, the people's spirits lifted. They would line up outside Maria's small bakery, eager to get their hands on a warm, fresh loaf. Maria's bread became a staple of their daily diets, and soon, the village began to thrive again.

Years passed, and Maria became known ■

Stop generating

Send a message...

...

We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

---

March 14, 2023

[Read paper ↗](#)

[Language](#), [GPT-4](#), [Milestone](#), [Publication](#)

[View system card ↗](#)

[Try on ChatGPT Plus ↗](#)

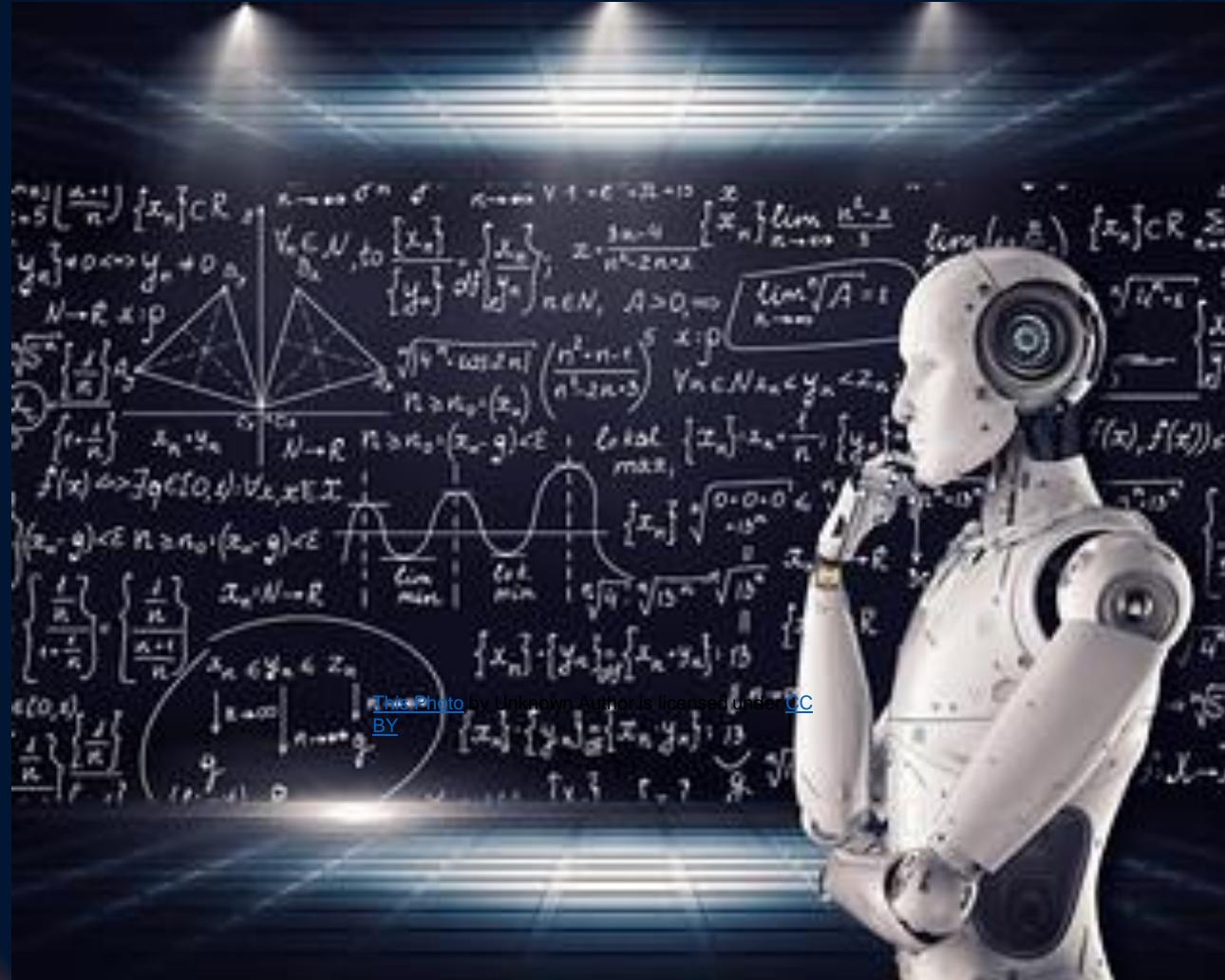
[Join API waitlist ↗](#)

[Rewatch demo livestream ↗](#)

[Contribute to OpenAI Eval ↗](#)

# **Technical Dive into Generative AI**

# What is Generative AI?



# Natural Language Generation

GPT-4

 Give me three ideas to enhance our cybersecurity measures with the latest tools.

 Certainly! Cybersecurity is constantly evolving, and staying updated with the latest tools and technologies is essential for any organization to protect its assets. Here are three ideas to enhance your cybersecurity measures with the latest tools:

**1. Implement Zero Trust Architecture (ZTA):**

- **What it is:** Zero Trust is a security concept centered on the belief that organizations should not automatically trust anything inside or outside its perimeters. Instead, everything trying to connect to its systems must be verified before granting access.
- **How to implement:**
  - **Identity and Access Management (IAM):** Use tools that offer multi-factor authentication (MFA), single sign-on (SSO), and identity governance.
  - **Micro-segmentation:** Break your network into segments to limit lateral movement of potential threats.
  - **Continuous Authentication & Monitoring:** Implement tools that continuo

Stop generating

# Image Generation

- <https://labs.openai.com/>
- <https://bing.com>

https://labs.openai.com/e/E7Y1IWhLPnDKqTASWxLfV6K

DALL-E History Collections

Edit the detailed description

create a picture of an elephant eating french fries.

The image displays three generated illustrations of an elephant eating french fries. The first illustration is a simple cartoon-style drawing of an elephant's head and trunk reaching towards a red tray filled with yellow french fries. The second illustration is a more detailed and colorful drawing of an elephant's head and trunk, with its mouth open as if eating from a red tray that has the golden arches logo on it. The third illustration is a close-up, detailed view of an elephant's head and trunk, showing its mouth full of french fries and its trunk holding a yellow tray filled with them.

# Code Generation



give me the Python Code to code a loop over a list of fruits and display the color.



Certainly! To accomplish this, we'd first need to have a mapping between fruits and their colors. Here's a basic example:

python

Copy code

```
# Define a dictionary where the keys are fruits and the values are their colors
fruit_colors = {
    'apple': 'red',
    'banana': 'yellow',
    'grape': 'purple',
    'orange': 'orange',
    'kiwi': 'green'
}

# Loop over the dictionary and print out each fruit and its color.
for fruit, color in fruit_colors.items():
    print(f"The color of {fruit} is {color}.")
```

Regen

# Data Analytics

chat.openai.com/?model=gpt-4-code-interpreter

Advanced Data Analysis

adult\_income\_data.csv  
Spreadsheet

Analyze this dataset and provide a univariate analysis on any fields of interest.

Alright, let's start by loading the dataset and taking an initial look at its contents.

python

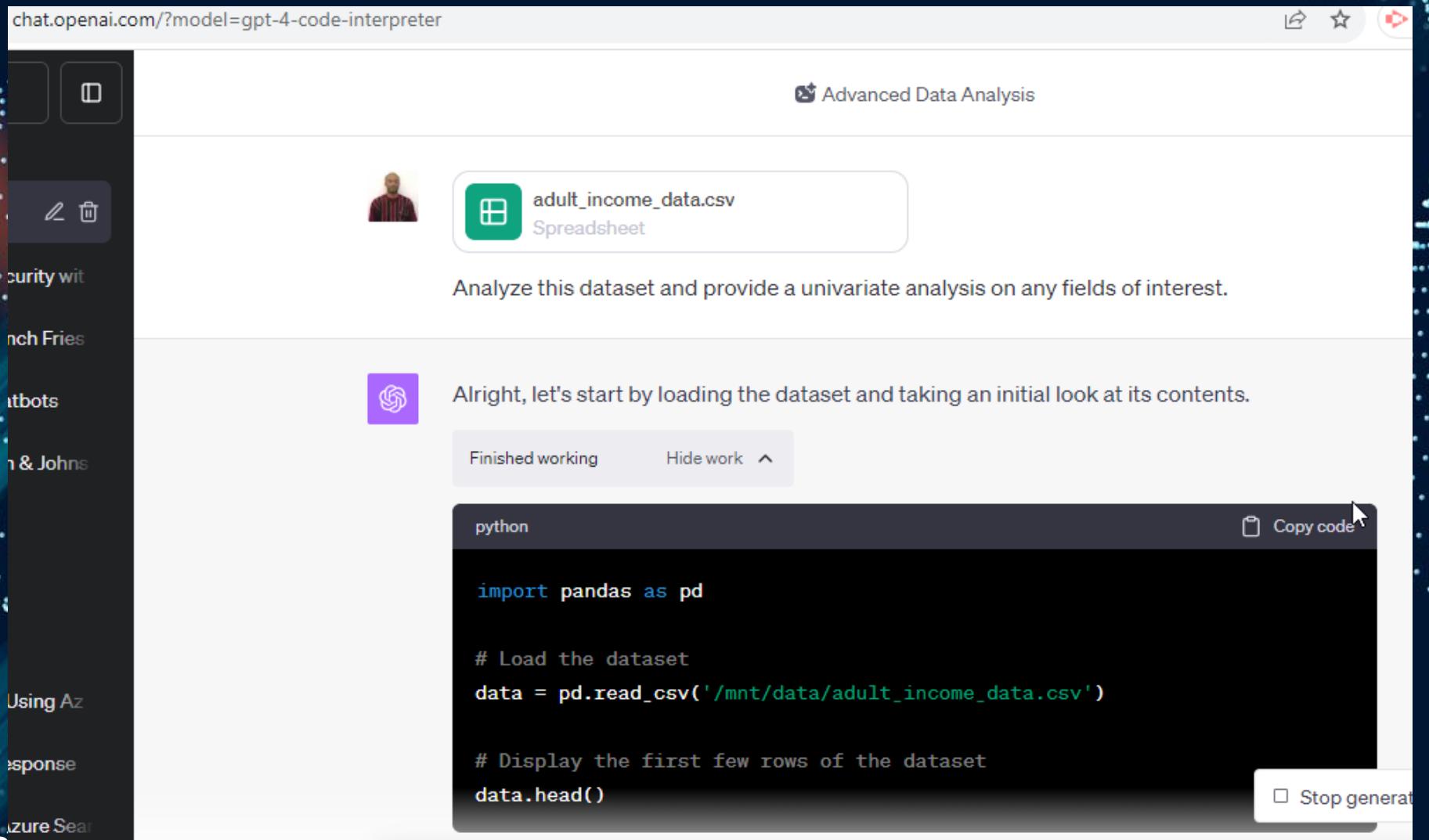
```
import pandas as pd

# Load the dataset
data = pd.read_csv('/mnt/data/adult_income_data.csv')

# Display the first few rows of the dataset
data.head()
```

Copy code

Stop generation



# Image Analysis

GPT-4

Q&A with Semantic Answering with Azure OpenAI Service  
Extract precise answers from your unstructured documents similar to ChatGPT

Knowledge Base (Azure Storage, Unstructured documents (pdf, docx, txt, etc.)) → Raw documents → Azure Form Recognizer → Extracted paragraphs → Azure Translator (optional) → Extracted dialog → Azure Translator (optional) → Embedding of paragraphs using OpenAI → Azure OpenAI Service Embeddings → Vector Database → Vectors → Vector search → Questions → Azure OpenAI Embeddings → Users → Answers

There is a red blinking light on C4 button in my SP-300a machine, which part do I need to replace?

You need to replace parts a & b. Here's the link to order them.

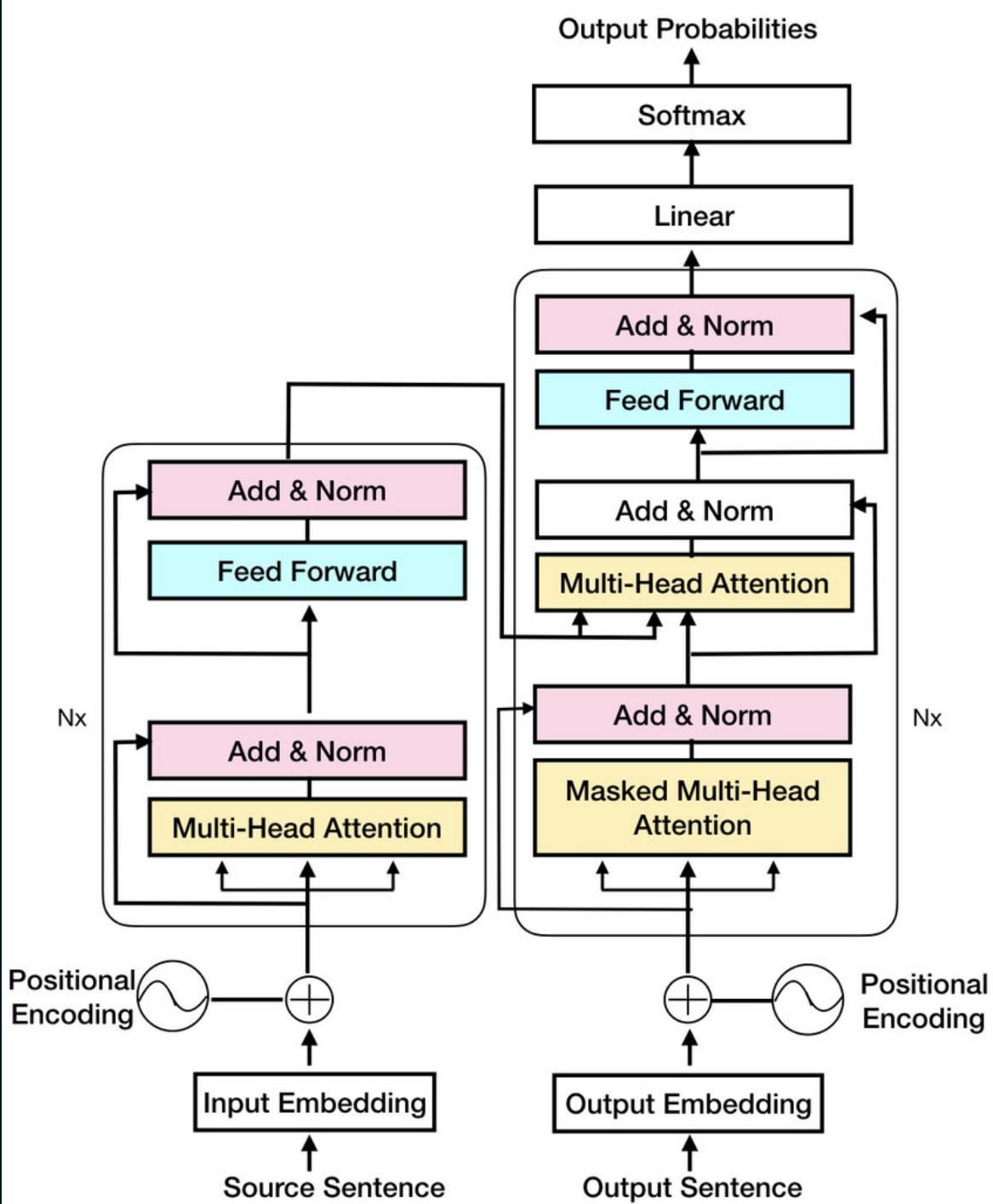
Please explain what is in this diagram

Certainly! The diagram depicts a system for "Q&A with Semantic Answering using Azure OpenAI Service." Here's a breakdown of the flow ●

Stop generating

# Large Language Models

# Transformers



# Tokenization: *I heard a dog bark loudly at a cat*

To tokenize this text, you can identify each discrete word and assign token IDs to them.

I (1)

heard (2)

a (3)

dog (4)

bark (5)

loudly (6)

at (7)

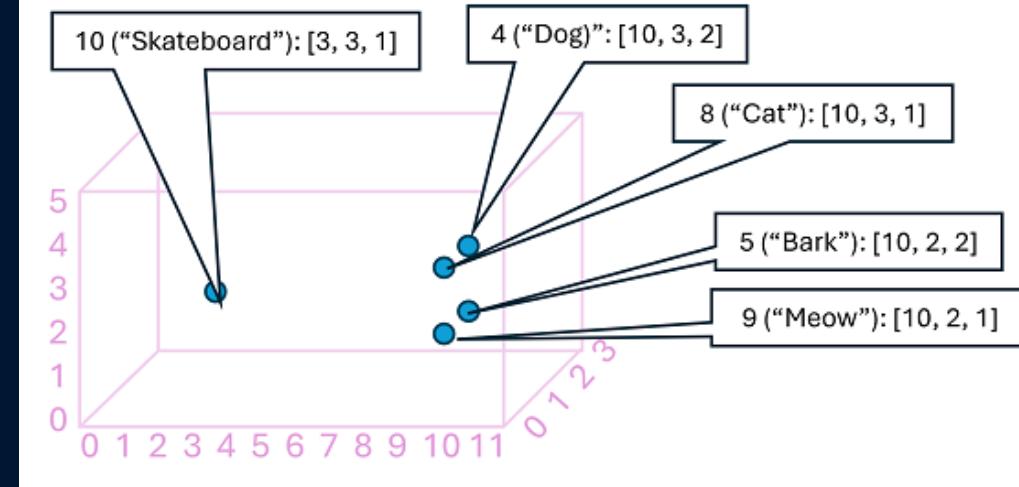
*("a" is already tokenized as 3)*

cat (8)

The sentence can now be represented with the tokens: [1 2 3 4 5 6 7 3 8]

# Embeddings

- **Simple Example:**
- Consider three words: "dog," "bark," and "bike."
- 1. **"dog"**: Since it's an animal, we'll mix a lot of red, a bit of blue because it makes a sound, and no green because it's not rideable. Its vector might look like [8, 2, 0].
- 2. **"bark"**: It's primarily a sound, so more blue. It's related to dogs, so a bit of red. And it has nothing to do with rideable objects, so no green. Its vector could be [2, 8, 0].
- 3. **"bike"**: It's something you ride, so lots of green. It's neither an animal nor a sound, so no red or blue. Its vector would be [0, 0, 8].

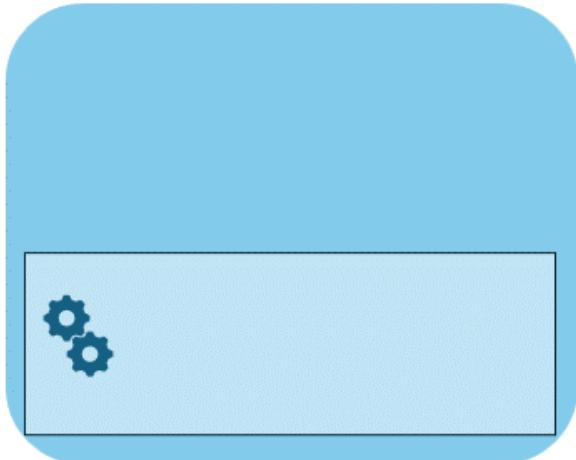


# Attention is all you need

- <https://arxiv.org/abs/1706.03762>

Anna loves chocolate because it's sweet.

What are the most relevant words in that sentence?



# Azure OpenAI

# Copilots

- LLMs have given rise to "copilots," software applications that assist users with common tasks using generative AI models.
- Copilots can be integrated into various applications for seamless user assistance.
- Their standardized architecture allows developers to create tailored copilots for specific business applications and services.
- Examples: Microsoft.com, Bing.com, GitHub.com, MS 365

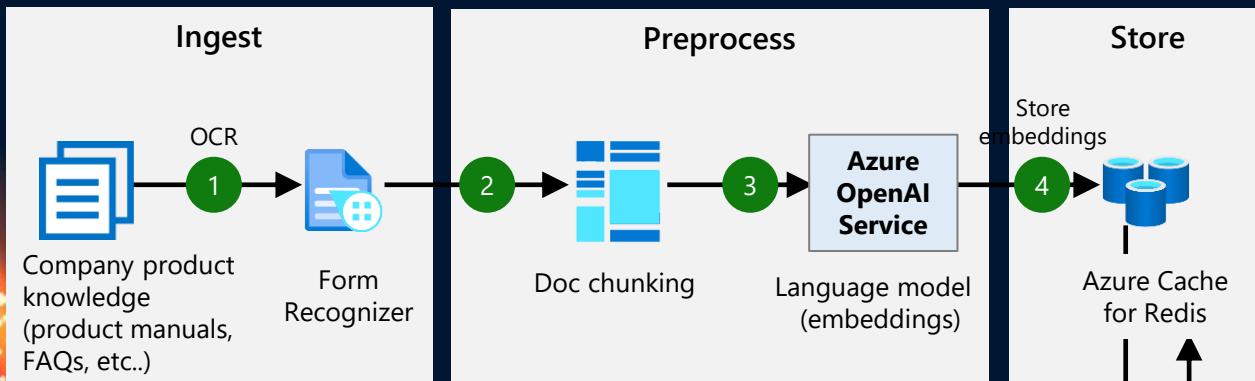
# Prompt Engineering

- **System Messages**
- **Grounding Data**
- **Good Example:**
- **Prompt:** "Based on the following information, provide travel recommendations. I'm visiting Tokyo for the first time in April for 5 days. I love historical sites, local food markets, and tranquil parks. I want to avoid overly touristy places. List 5 top places I should visit."
- **Bad Example:**
- **Prompt:** "Tell me some places to visit in Tokyo."

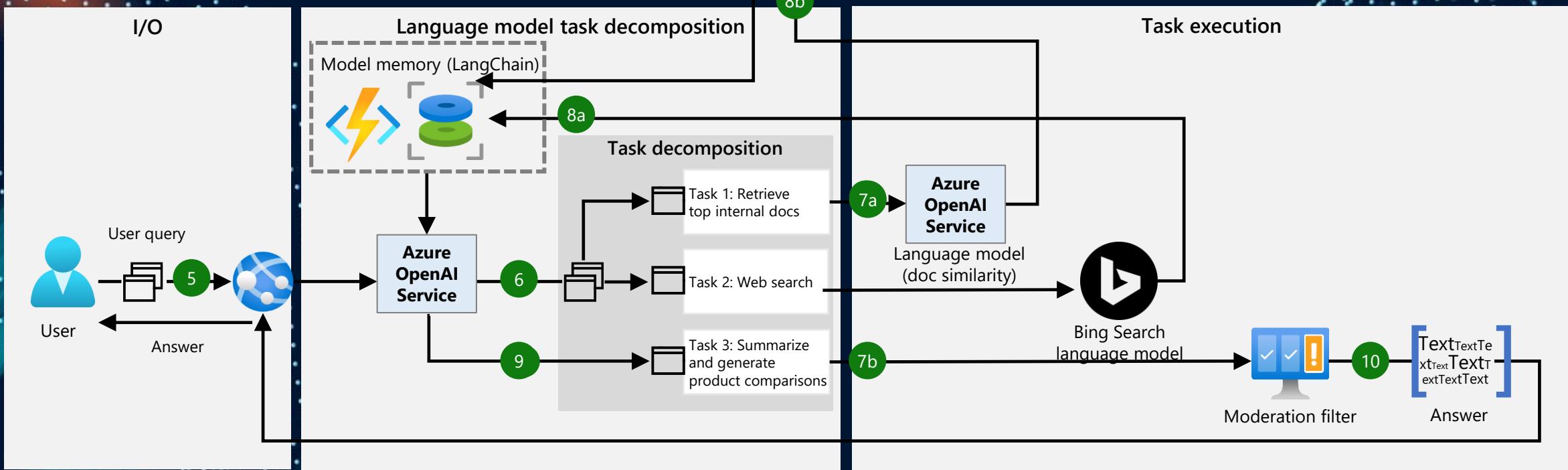
# HOL

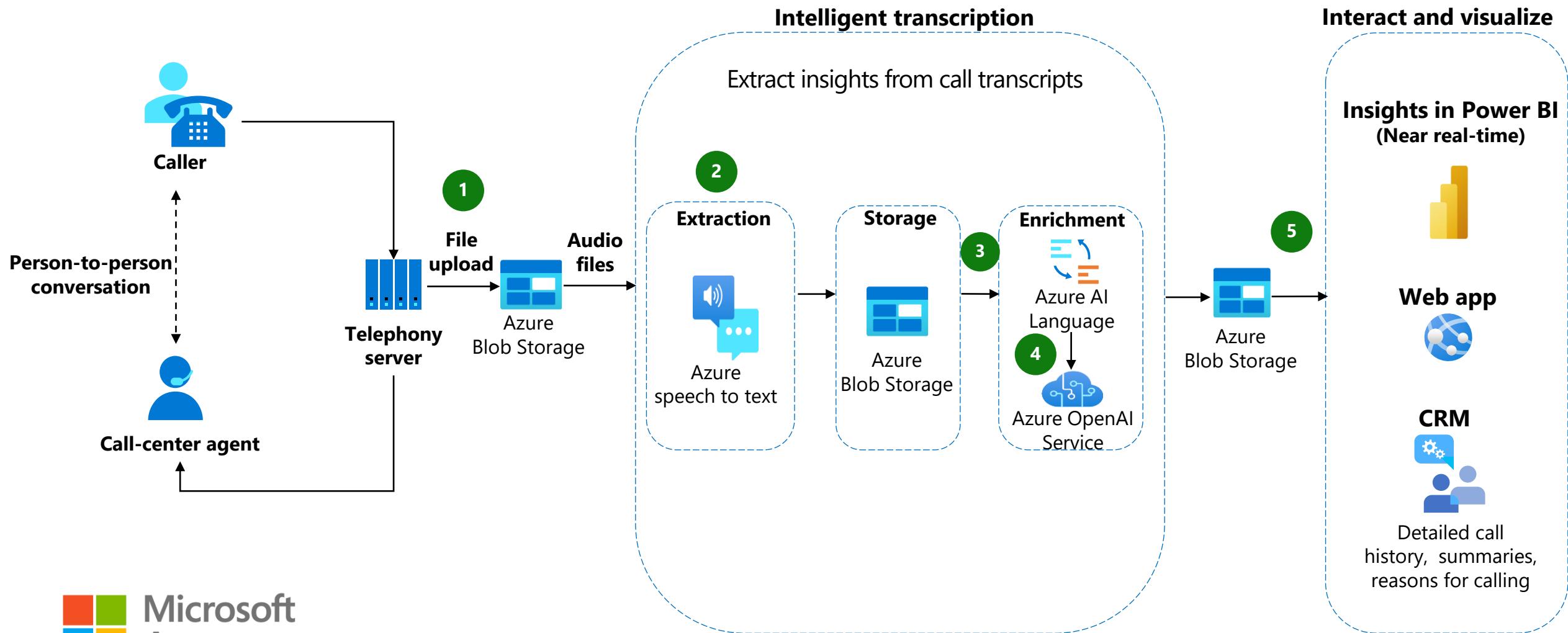
- <https://bit.ly/explore2121>
- [https://DrLee.io \(Explore Generative AI with Bing: A Life Sciences Edition \)](https://DrLee.io)

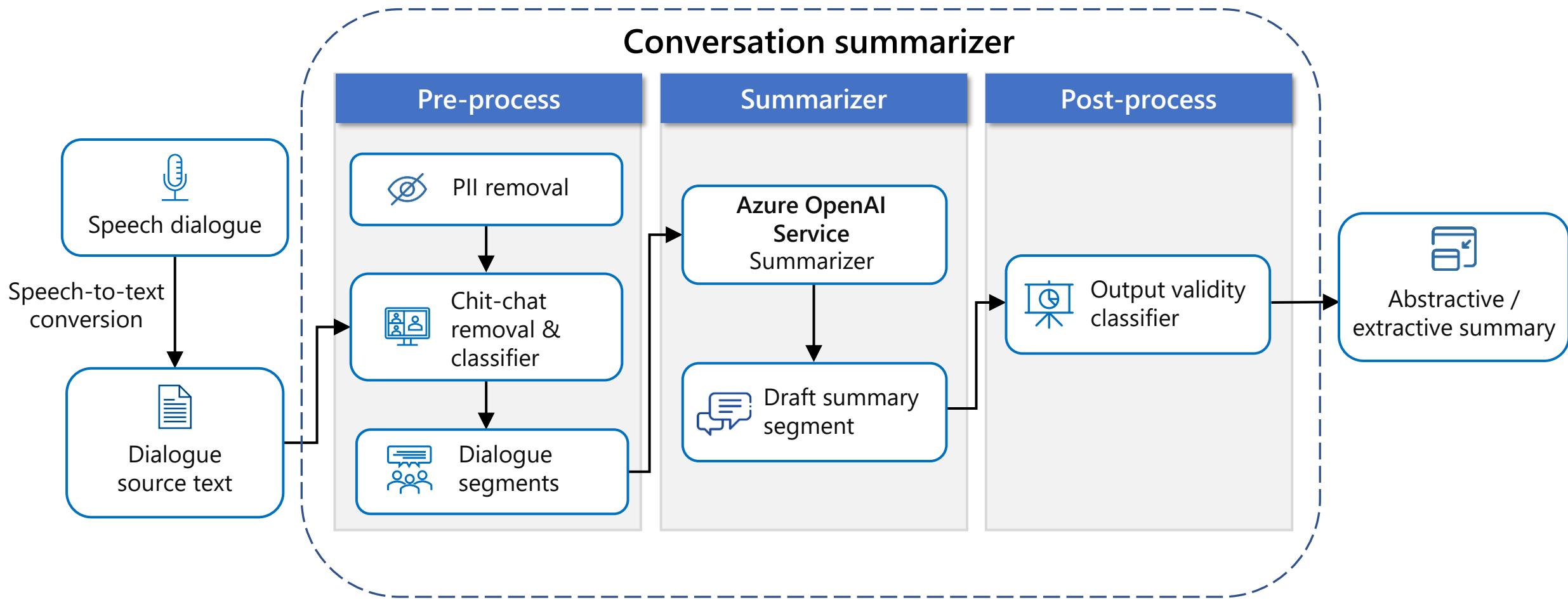
## Batch pipeline

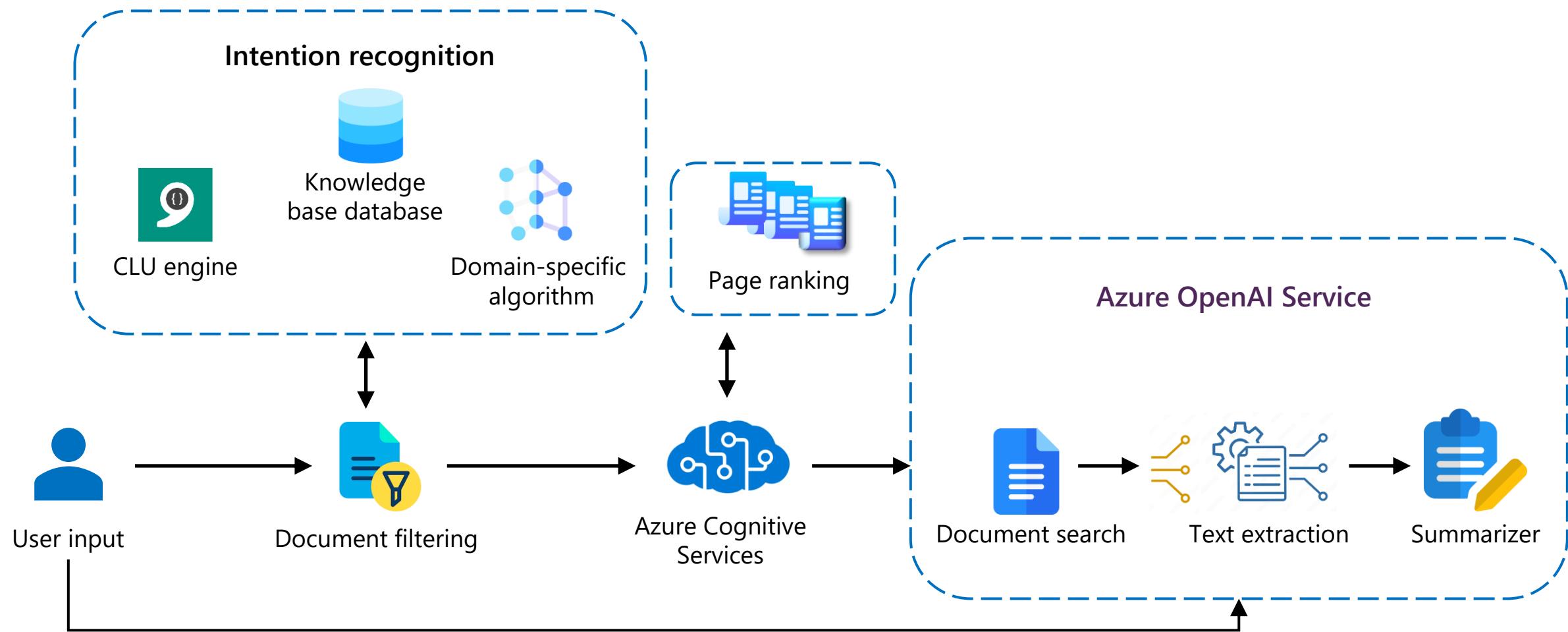


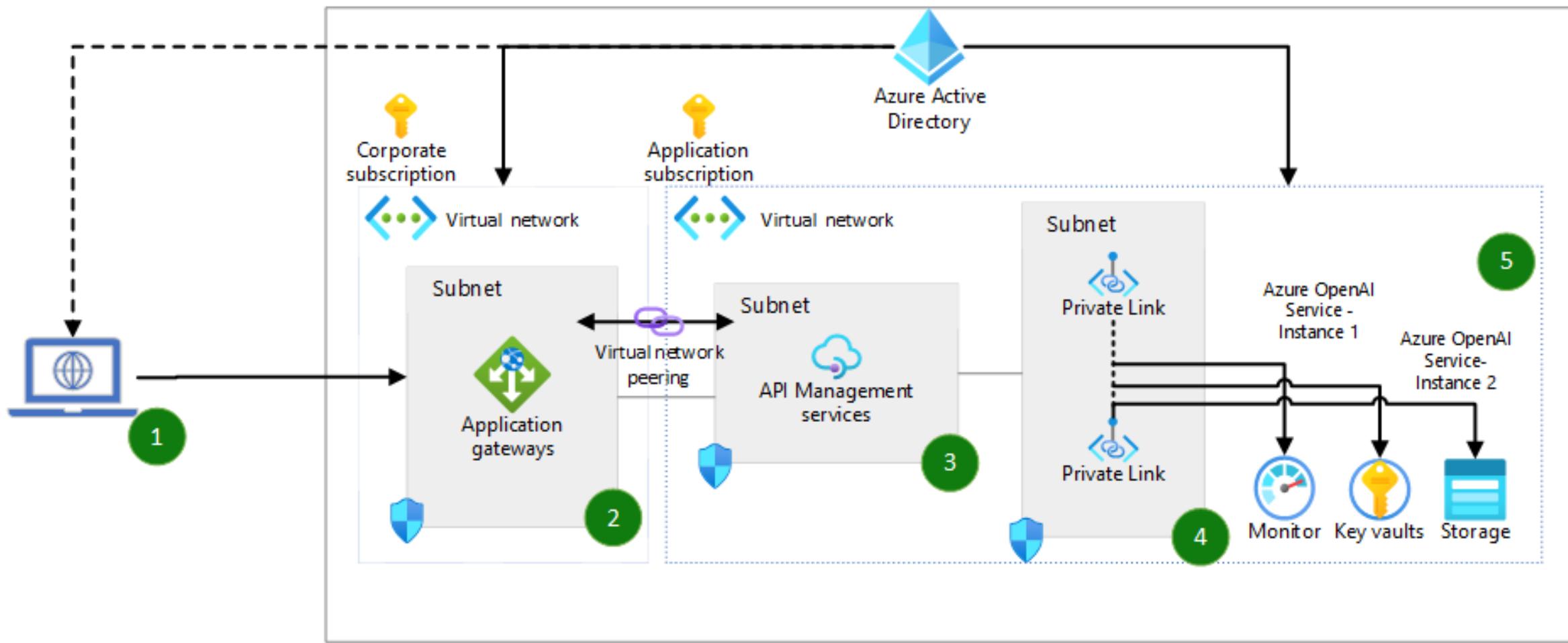
## Asynchronous

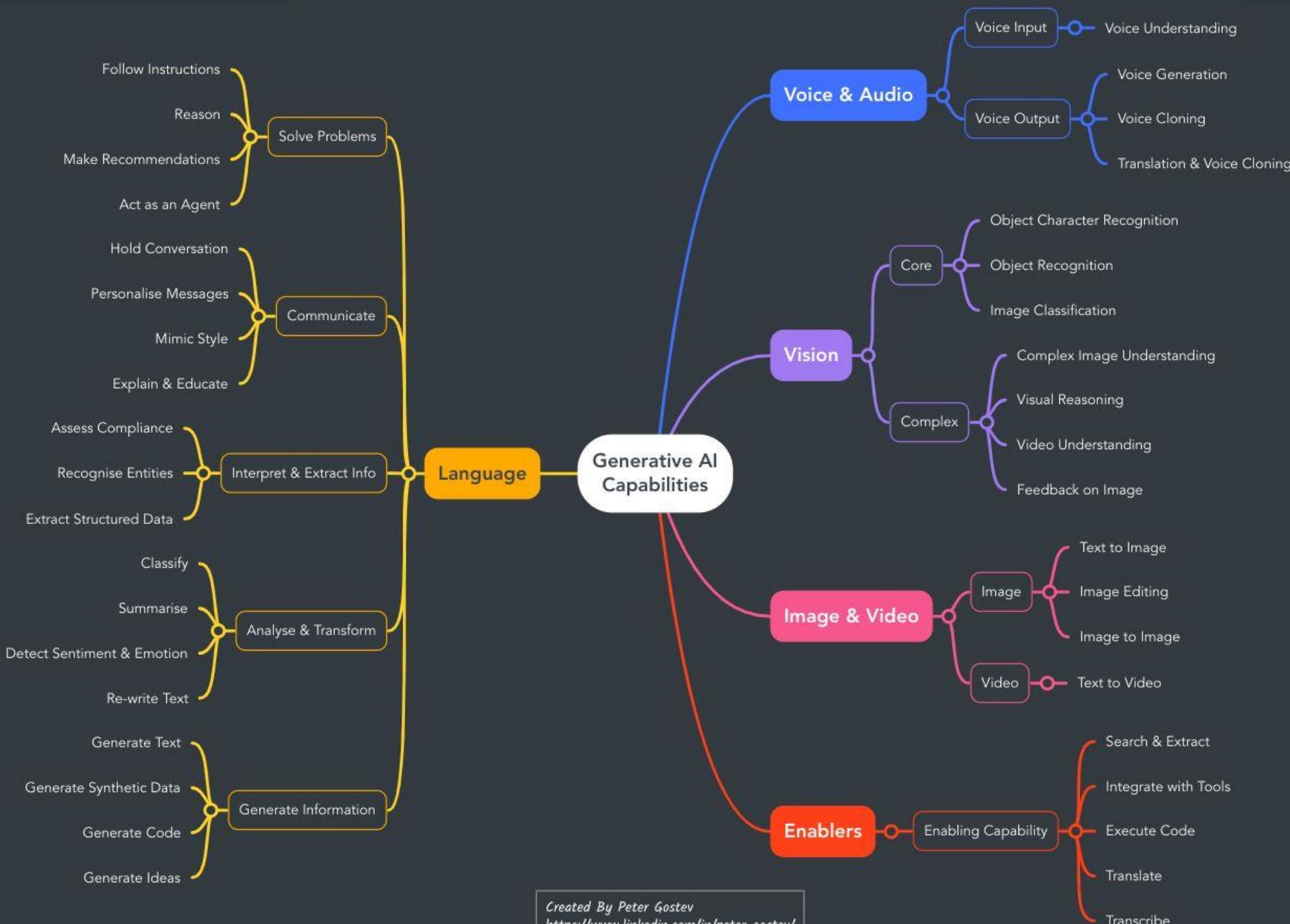












# DEMO

- These are the capabilities that we are looking to integrate into our business processes

# What are the risks?

# Risks

- Hallucinations
- Biased Data
- No Data Lineage
- Privacy (Each Query is fed back into the model)
- Security (Risk of sensitive data being leaked through prompt attacks)
- Interpretability and Transparency

# How can we address these problems?

- **Combine the Cloud with the Robot**
  - Security
  - Privacy
  - Control over the model
  - You have your own “personal” instance of the Robot
  - Transparency (through citations)

BUT... there is still the problem with the data... Bias, lineage, misinformation, hallucinations

# What is a model... really?

- An AI robot is really 2 things:
  - A model (an algorithm or a set of rules)
  - Data

MANY OF THE PROBLEMS ARE RELATED TO THE DATA. WHAT IF WE CAN TRADE AWAY THE DATA THAT EXISTS WITH THE MODEL AND REPLACE IT WITH YOUR OWN CUSTOM DATA?

# DEMO

- These are the capabilities that we are looking to integrate into our business processes

## It would cost \$20k a year to get Gen AI to analyse 10 million customer interactions

### Tokens for 10 million Chats & Calls

Total Chat & Calls per year

10 million



2,000



20 billion

### Total Tokens per Year

### Cost per 1m Tokens

GPT-3.5 cost per 1m tokens

\$1

GPT-4-Turbo cost per 1m tokens

\$20

### Total Cost Per Year

Total GPT-3.5 cost per year

\$20,000

Total GPT-4-Turbo cost per year

\$400,000

### 12 Things Gen AI can do with Chats & Calls

Detailed Reason  
for Call

Guidance for  
Agents

Key Info  
Extraction

Compliance

Sentiment

Agent's  
performance

Emotion

Fraud & Anomaly  
Alert

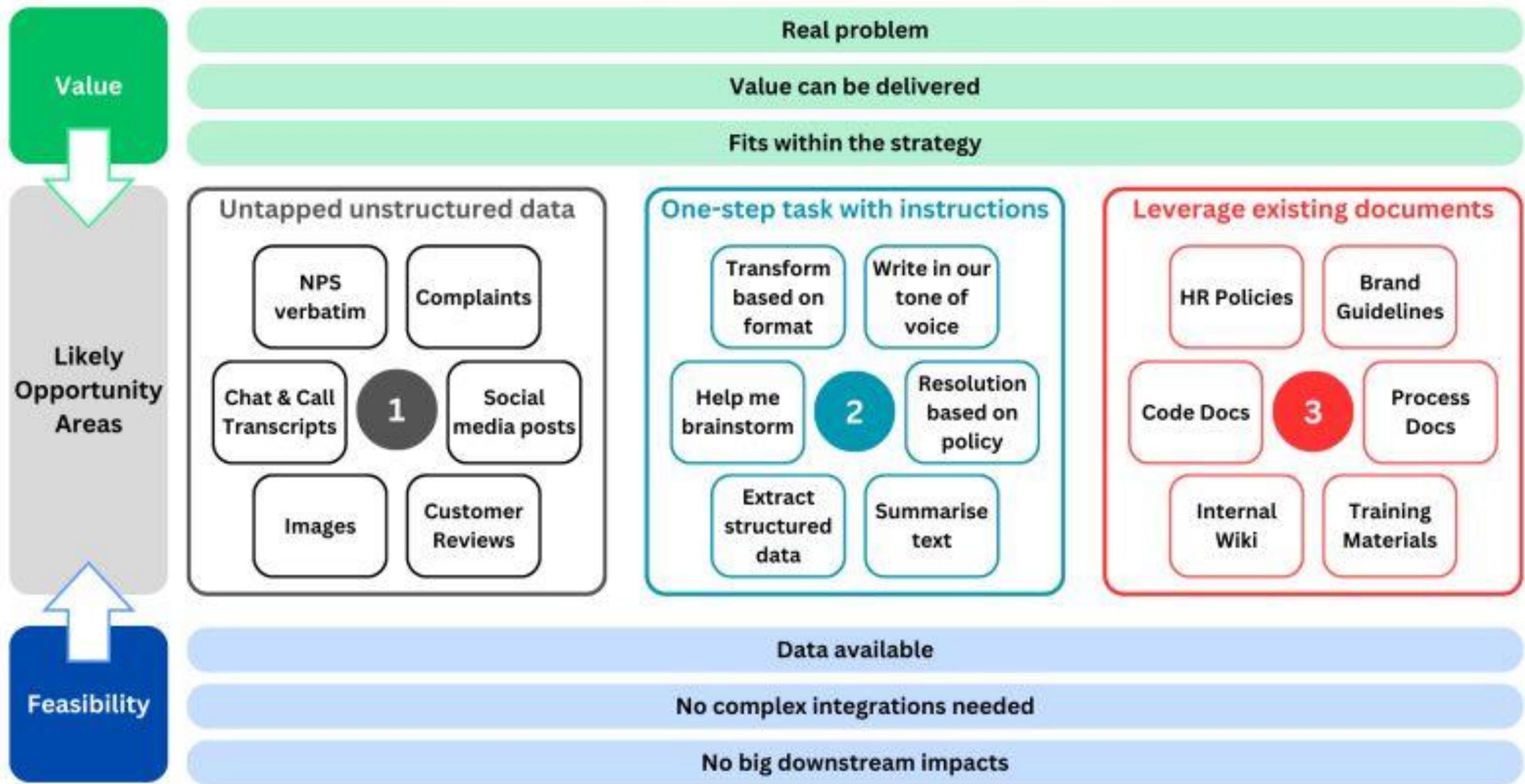
Summarisation

Next Best Action

Issue  
Identification

Customer Profile

# Where to look for Gen AI opportunities?



## Lmsys.org Chatbot Arena Elo Ratings

(based on blind voting by users)

Company	Model	Elo Rating	License
OpenAI	GPT-4-Turbo	1249	Proprietary
New! Google	Gemini Pro (Bard)	1215	Proprietary
OpenAI	GPT-4-0314	1189	Proprietary
OpenAI	GPT-4-0613	1161	Proprietary
Mistral	Mistral Medium	1150	Proprietary
Anthropic	Claude-1	1150	Proprietary
Anthropic	Claude-2.0	1131	Proprietary
Mistral	Mixtral-8x7b-Instruct	1123	Apache 2.0
Google	Gemini Pro	1122	Proprietary
Anthropic	Claude-2.1	1119	Proprietary
OpenAI	GPT-3.5-Turbo-0613	1117	Proprietary
01 AI (China)	Yi-34B-Chat	1112	Yi License
Anthropic	Claude-Instant-1	1110	Proprietary
Open Source	WizardLM-70B-v1.0	1106	Llama 2
Open Source	Tulu-2-DPO-70B	1105	AI2 ImpACT
OpenAI	GPT-3.5-Turbo-0314	1104	Proprietary
Open Source	Vicuna-33B	1094	Non-commercial
Open Source	Starling-LM-7B-alpha	1089	CC-BY-NC-4.0
Meta	Llama-2-70b-chat	1081	Llama 2
Open Source	OpenChat-3.5	1076	Apache-2.0

Analysis by: Peter Gostev (<https://www.linkedin.com/in/peter-gostev/>)

Source: <https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

# Hopes & Expectations for AI in 2024

## Huge Models

Massive models exceeding GPT-4, having been trained with 30-50x more compute

## Small Models

Amazing quality models in the range of ~3b to ~30b parameters

## Cheap Models

Super cheap good models (90% cheaper than GPT-4)

## AI-First UI

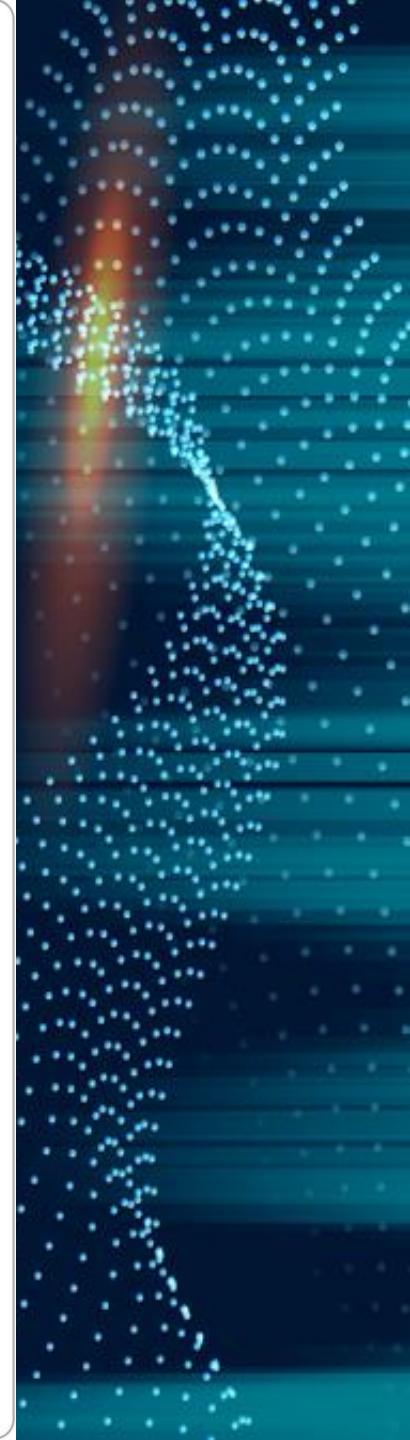
New AI-first user experiences emerge & 'chatbots everywhere' idea is starting to die

## AI Doom Dies

We'll calm down about AI doom & focus on the aspects of safety that are actually important

## Meta's Surprise

Something amazing from Meta (what are they doing with their 150k H100 GPUs?)



# OPINIONATED GUIDE TO EXPERIMENTING WITH & BUILDING GEN AI APPS\*

\*that would apply to 80% of cases

## Early Exploration

Key Question: Can Gen AI help solve my use case?

**DO**

Pick the best model available (GPT-4)

Build a Custom GPT or Test on Playground

Use public data or generate synthetic with LLMs

**DON'T DO**

Experiment with lower performant models

Build a polished custom app

Fine-tune a model

## Proof of Concept

Key Question: Are my stakeholders interested?

Build a simple app (e.g. Streamlit)

Experiment with new user experiences

Develop strong product evaluation & testing

Iterate through implementation techniques

Try a cheaper model and possibly fine-tuning

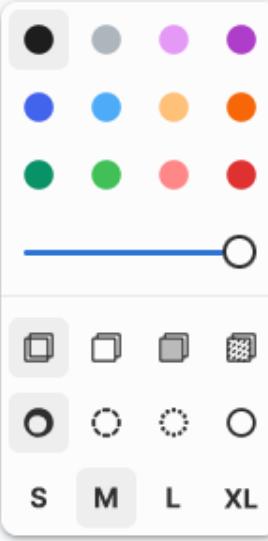
Build-out a full featured & integrated app

Spend too much time on re-usable assets

Ignore LLM risks (e.g. prompt injection, hallucinations)

Get stuck with the first implementation attempt

Forget about data quality (incl. for RAG)



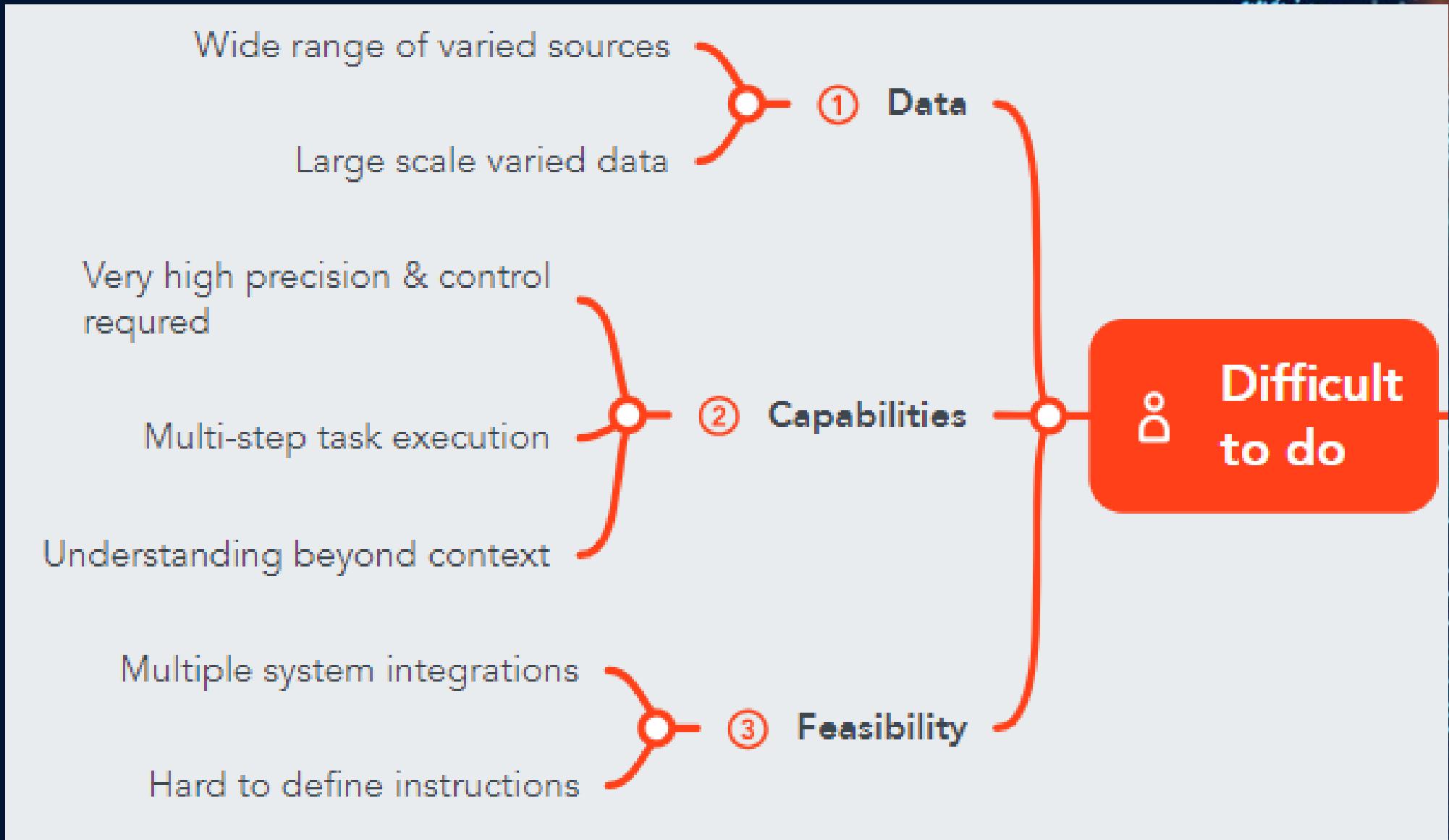
Your OpenAI API Key (risky but cool)

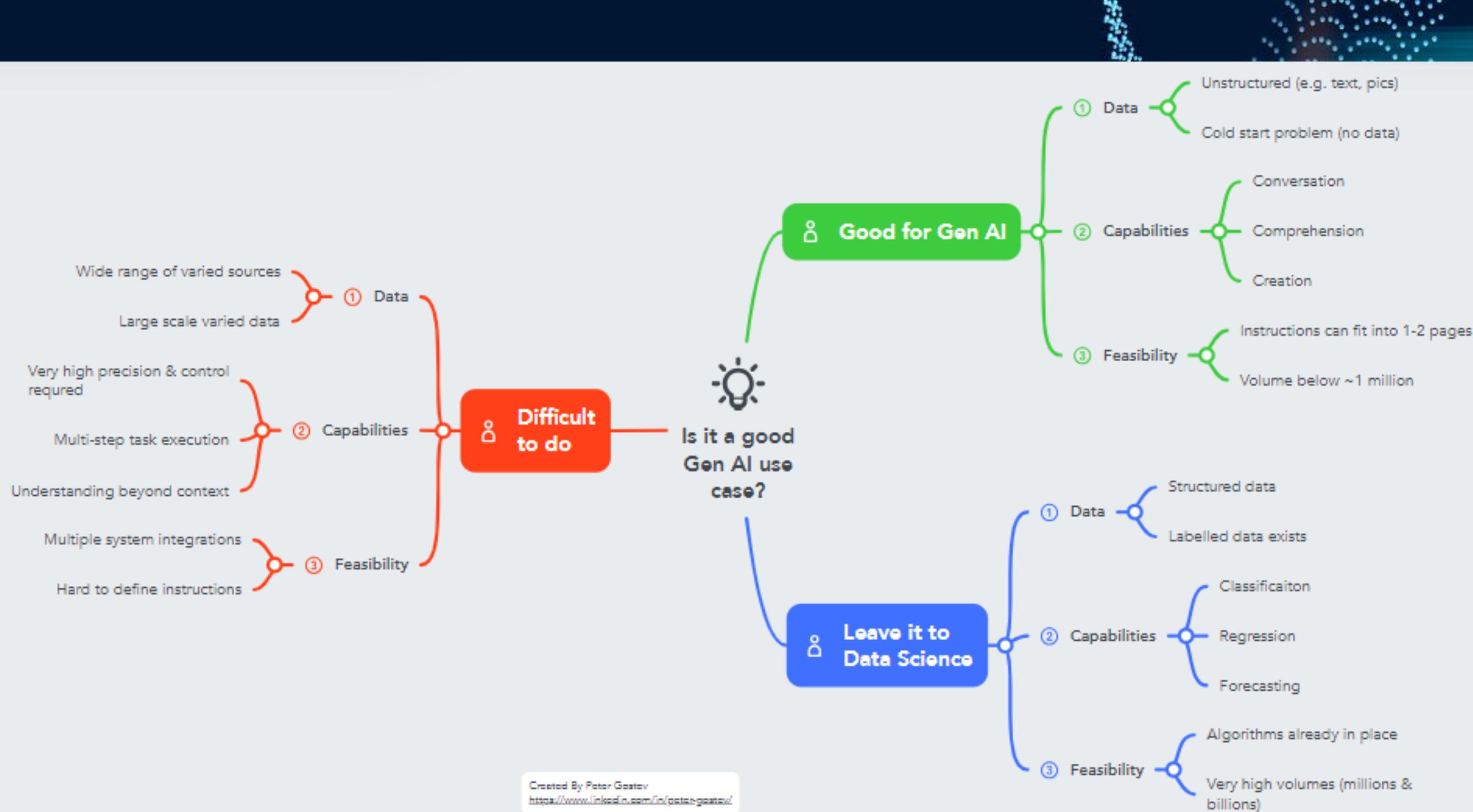


[https://calculateme.  
streamlit.app/](https://calculateme.streamlit.app/)

# When to use GenAI

# What can we automate?





## It would cost \$20k a year to get Gen AI to analyse 10 million customer interactions

### Tokens for 10 million Chats & Calls

Total Chat & Calls per year

10 million



2,000



20 billion

### Total Tokens per Year

### Cost per 1m Tokens

GPT-3.5 cost per 1m tokens

\$1

GPT-4-Turbo cost per 1m tokens

\$20

### Total Cost Per Year

Total GPT-3.5 cost per year

\$20,000

Total GPT-4-Turbo cost per year

\$400,000

### 12 Things Gen AI can do with Chats & Calls

Detailed Reason  
for Call

Guidance for  
Agents

Key Info  
Extraction

Compliance

Sentiment

Agent's  
performance

Emotion

Fraud & Anomaly  
Alert

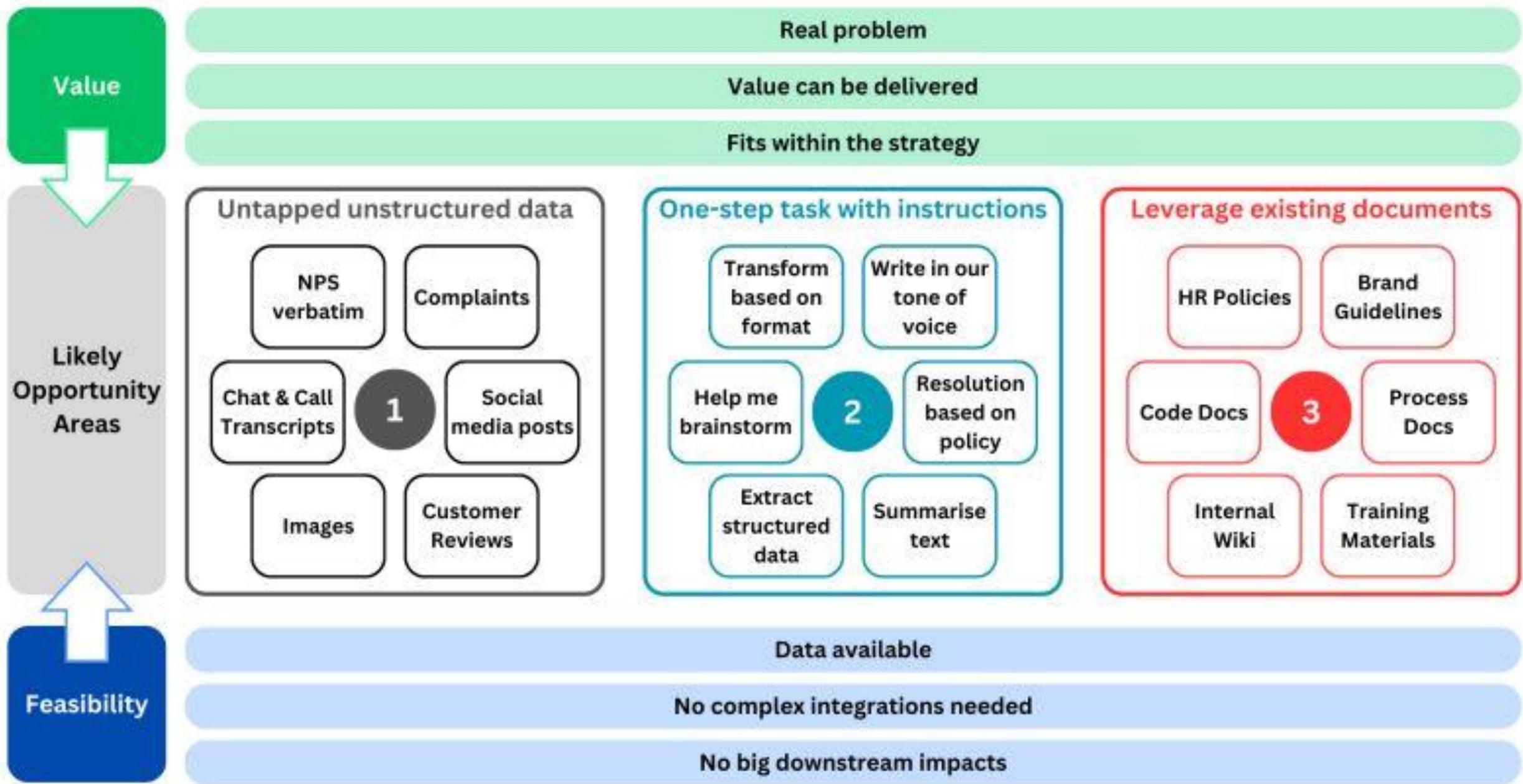
Summarisation

Next Best Action

Issue  
Identification

Customer Profile

# Where to look for Gen AI opportunities?



## Lmsys.org Chatbot Arena Elo Ratings

(based on blind voting by users)

Company	Model	Elo Rating	License
OpenAI	GPT-4-Turbo	1249	Proprietary
New! Google	Gemini Pro (Bard)	1215	Proprietary
OpenAI	GPT-4-0314	1189	Proprietary
OpenAI	GPT-4-0613	1161	Proprietary
Mistral	Mistral Medium	1150	Proprietary
Anthropic	Claude-1	1150	Proprietary
Anthropic	Claude-2.0	1131	Proprietary
Mistral	Mixtral-8x7b-Instruct	1123	Apache 2.0
Google	Gemini Pro	1122	Proprietary
Anthropic	Claude-2.1	1119	Proprietary
OpenAI	GPT-3.5-Turbo-0613	1117	Proprietary
01 AI (China)	Yi-34B-Chat	1112	Yi License
Anthropic	Claude-Instant-1	1110	Proprietary
Open Source	WizardLM-70B-v1.0	1106	Llama 2
Open Source	Tulu-2-DPO-70B	1105	AI2 ImpACT
OpenAI	GPT-3.5-Turbo-0314	1104	Proprietary
Open Source	Vicuna-33B	1094	Non-commercial
Open Source	Starling-LM-7B-alpha	1089	CC-BY-NC-4.0
Meta	Llama-2-70b-chat	1081	Llama 2
Open Source	OpenChat-3.5	1076	Apache-2.0

Analysis by: Peter Gostev (<https://www.linkedin.com/in/peter-gostev/>)

Source: <https://huggingface.co/spaces/lmsys/chatbot-area-leaderboard>

# Hopes & Expectations for AI in 2024

## Huge Models

Massive models exceeding GPT-4, having been trained with 30-50x more compute

## Small Models

Amazing quality models in the range of ~3b to ~30b parameters

## Cheap Models

Super cheap good models (90% cheaper than GPT-4)

## AI-First UI

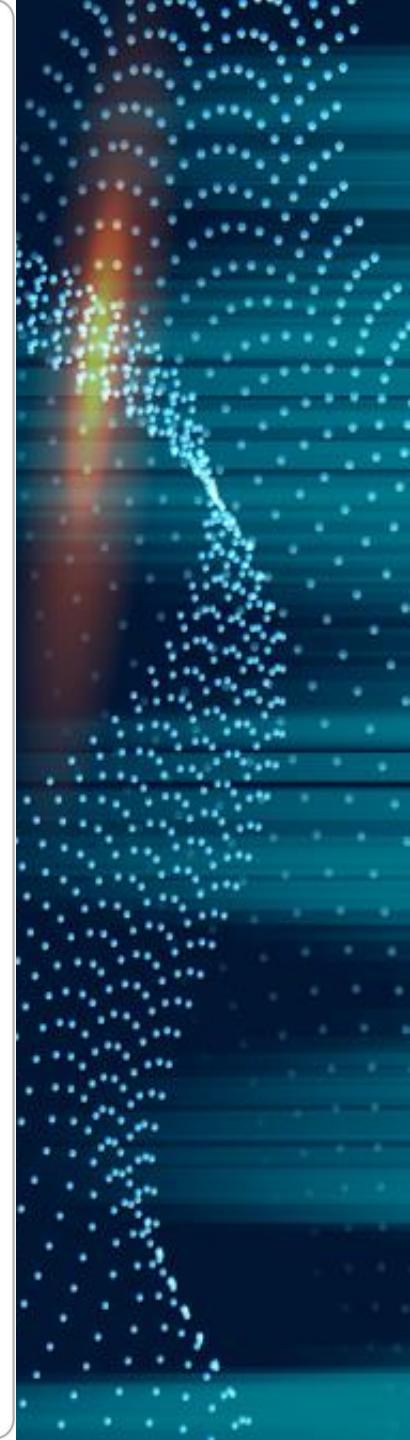
New AI-first user experiences emerge & 'chatbots everywhere' idea is starting to die

## AI Doom Dies

We'll calm down about AI doom & focus on the aspects of safety that are actually important

## Meta's Surprise

Something amazing from Meta (what are they doing with their 150k H100 GPUs?)



# OPINIONATED GUIDE TO EXPERIMENTING WITH & BUILDING GEN AI APPS\*

\*that would apply to 80% of cases

## Early Exploration

Key Question: Can Gen AI help solve my use case?

**DO**

Pick the best model available (GPT-4)

Build a Custom GPT or Test on Playground

Use public data or generate synthetic with LLMs

**DON'T DO**

Experiment with lower performant models

Build a polished custom app

Fine-tune a model

## Proof of Concept

Key Question: Are my stakeholders interested?

Build a simple app (e.g. Streamlit)

Experiment with new user experiences

Develop strong product evaluation & testing

Iterate through implementation techniques

Try a cheaper model and possibly fine-tuning

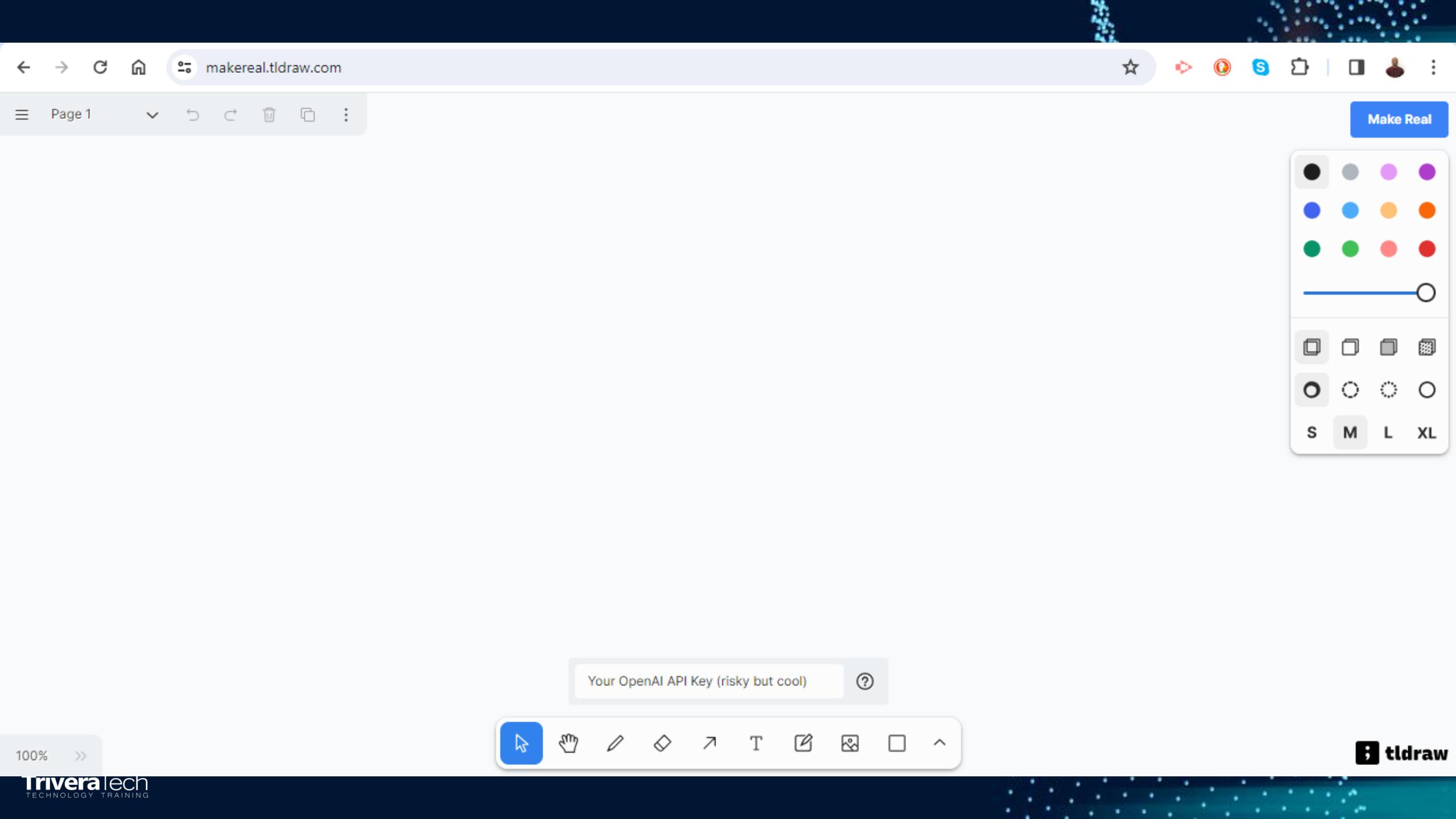
Build-out a full featured & integrated app

Spend too much time on re-usable assets

Ignore LLM risks (e.g. prompt injection, hallucinations)

Get stuck with the first implementation attempt

Forget about data quality (incl. for RAG)

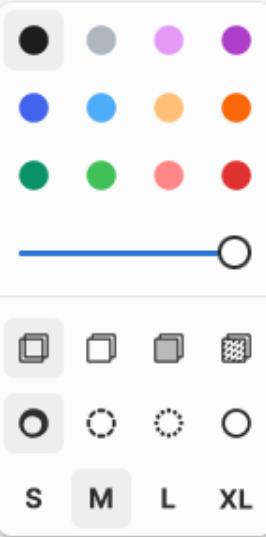


← → ⌂ ⌂ makereal.tldraw.com



☰ Page 1

Make Real



Your OpenAI API Key (risky but cool)



tldraw

100% >>

TriveraTech  
TECHNOLOGY TRAINING