

# VoiceTech futures

from Siri to ChatGPT  
*RTC Conference 2023*

# John Burkey

- Founder, CEO/CTO Brighten AI
  - Voice AI platform company with hardware and software
  - Gen AI on the edge
- Siri team
- Microsoft Chief Architect for Office for iPhone/iPad/Mac
- Java Chief Architect at Sun
- Many patents in Ai,OS,Dev Tools
- Advise national press on Ai, technologies

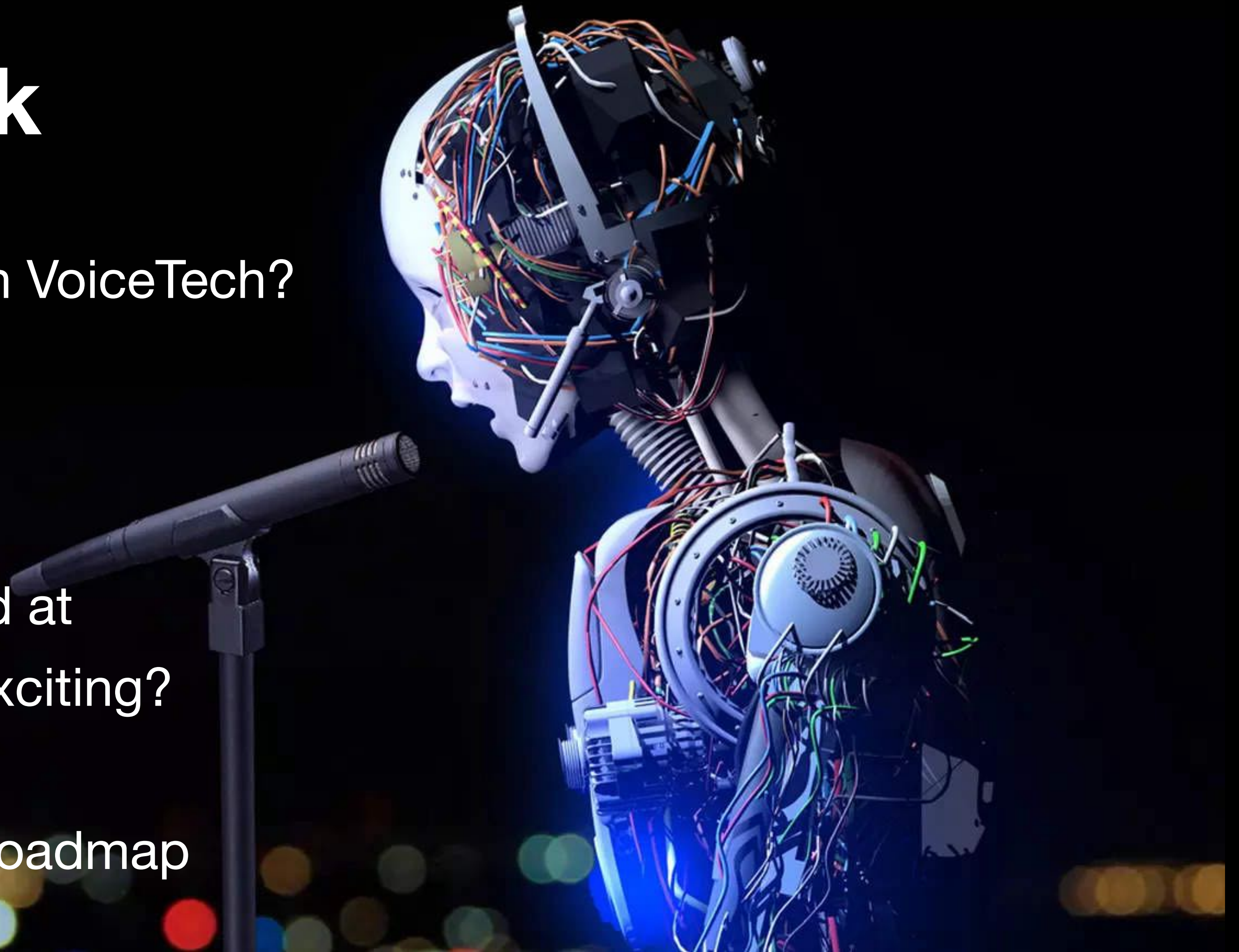


# Today's talk

Where are we in VoiceTech?  
A brief history

What is GenAI?  
What's it good at  
Why is it so exciting?

The future & a roadmap



# Voice Tech

What are we really solving for?

Audio  
Dialog Acts  
Intent  
Goals

*Transcription*

*Conversation*

*Translation*

# Automatic Speech Recognition

From early Voice AI through Daniel Povey to End to end systems

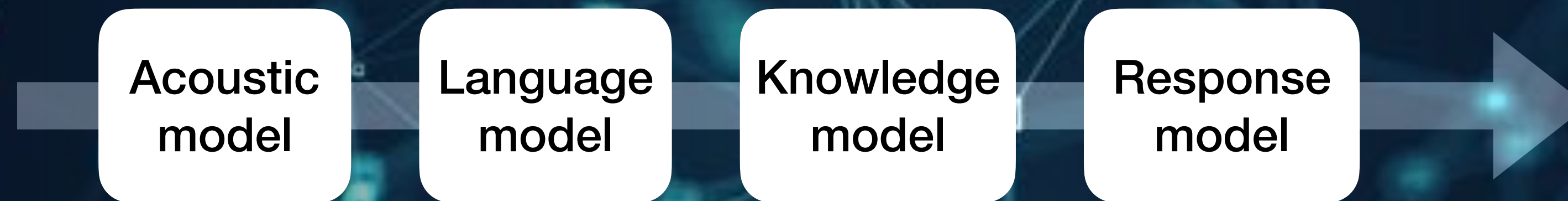
Enabled by hardware advancements allowing training on huge data sets

Moved to statistical AI models

Started stacking models

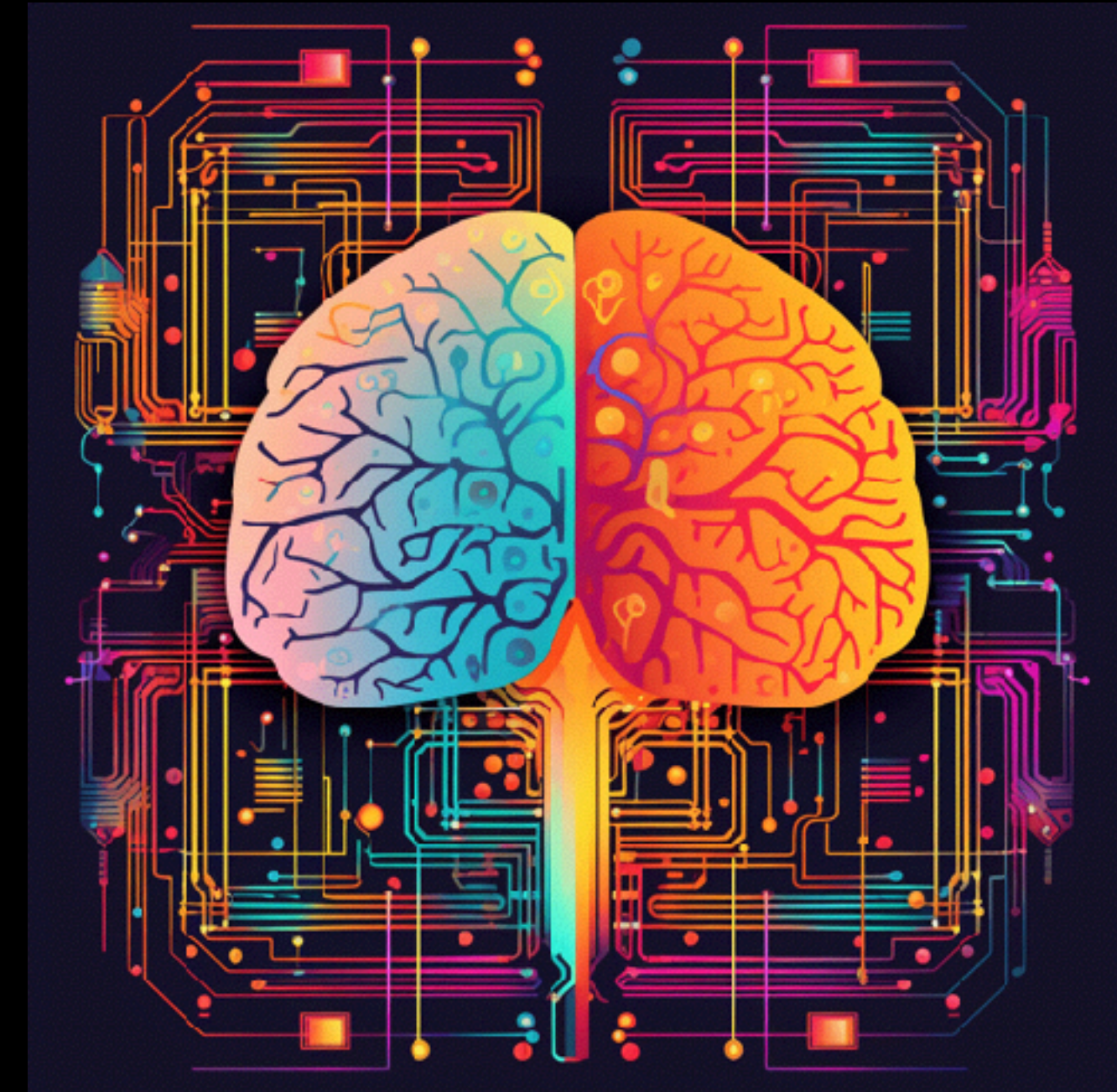
Competitive with humans, but require(d) huge data sets for any language

# End to end systems

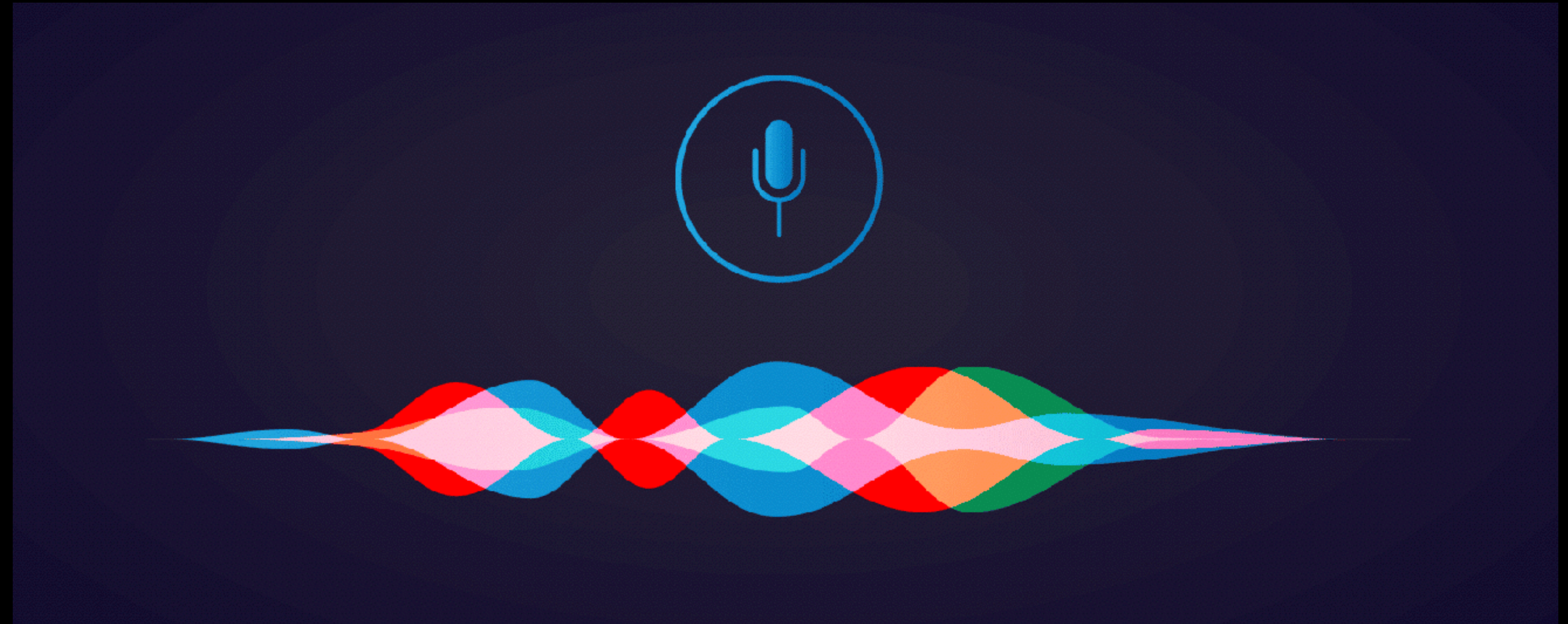


- Connect the models' ends to each other, and train the whole thing together
- Maintains statistical fuzziness through pipelines, preserving nuance

# What about Voice Agents?



# Siri



## A static model

Only responds with templated pre-programmed answers and simplistic textual response templates working with API's.

Good for: "Play Madonna", "What's the weather?", "Where can I eat at 6pm?"

Proved voice assistants could work at scale. Nice work !



**Alexa**



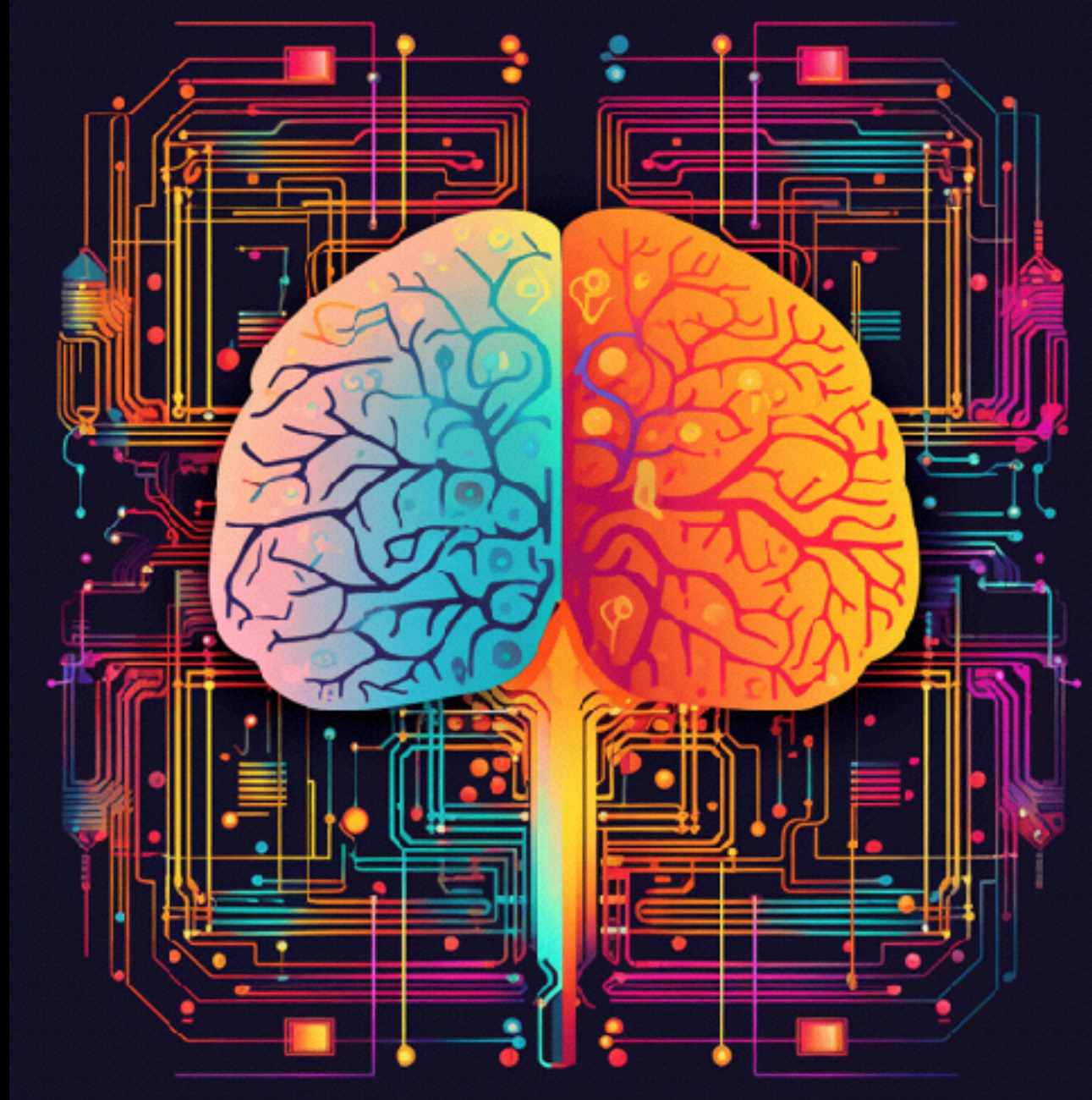
Same model as Siri's, but open access to create skills for the assistant.

Proved you could have an open ecosystem !

Multi decade revenue loss for Amazon, never made money !

# Meanwhile at Google

Google was trying to make search better, and engaging in basic research around how to do that, which led to statistical AI research in language..

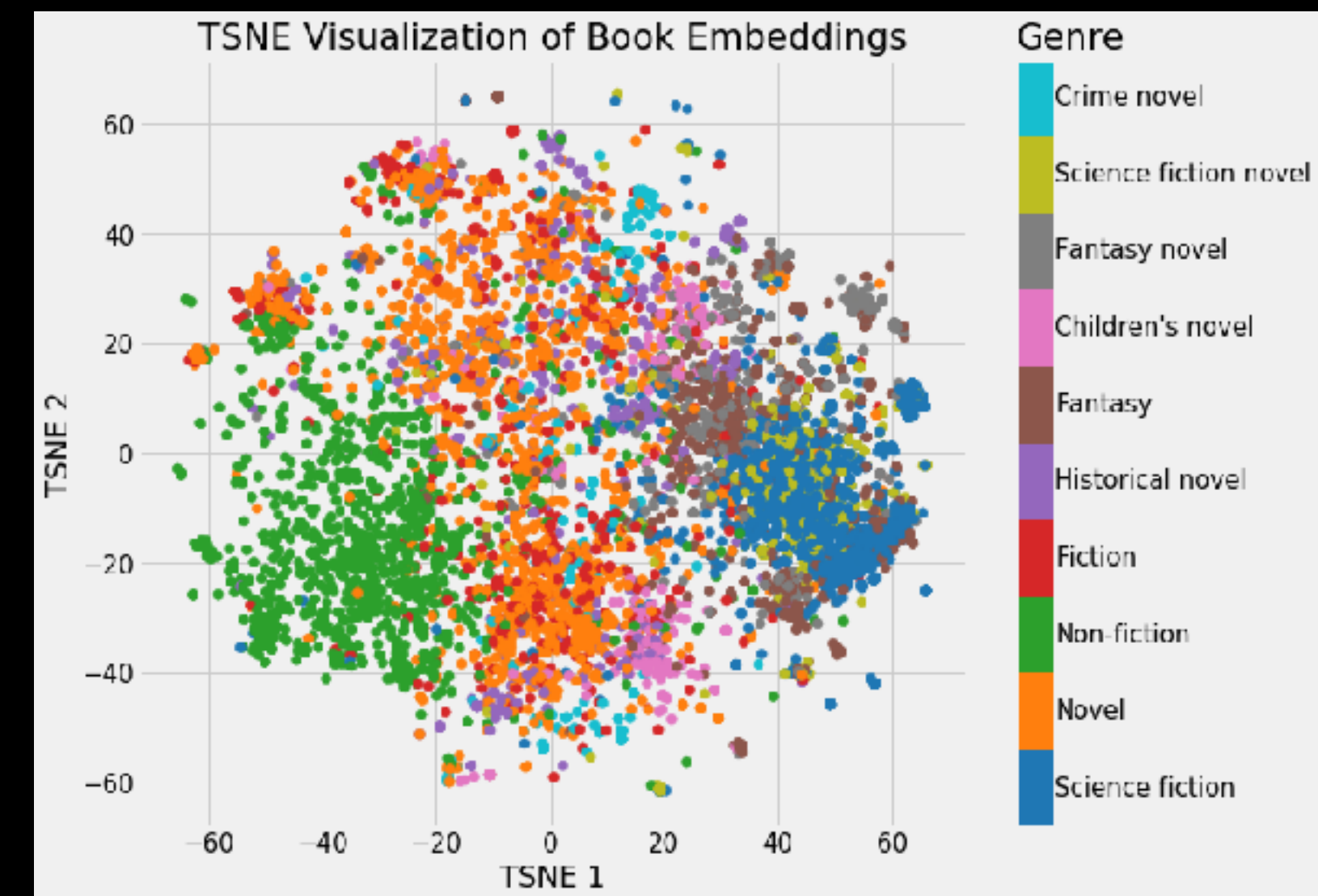
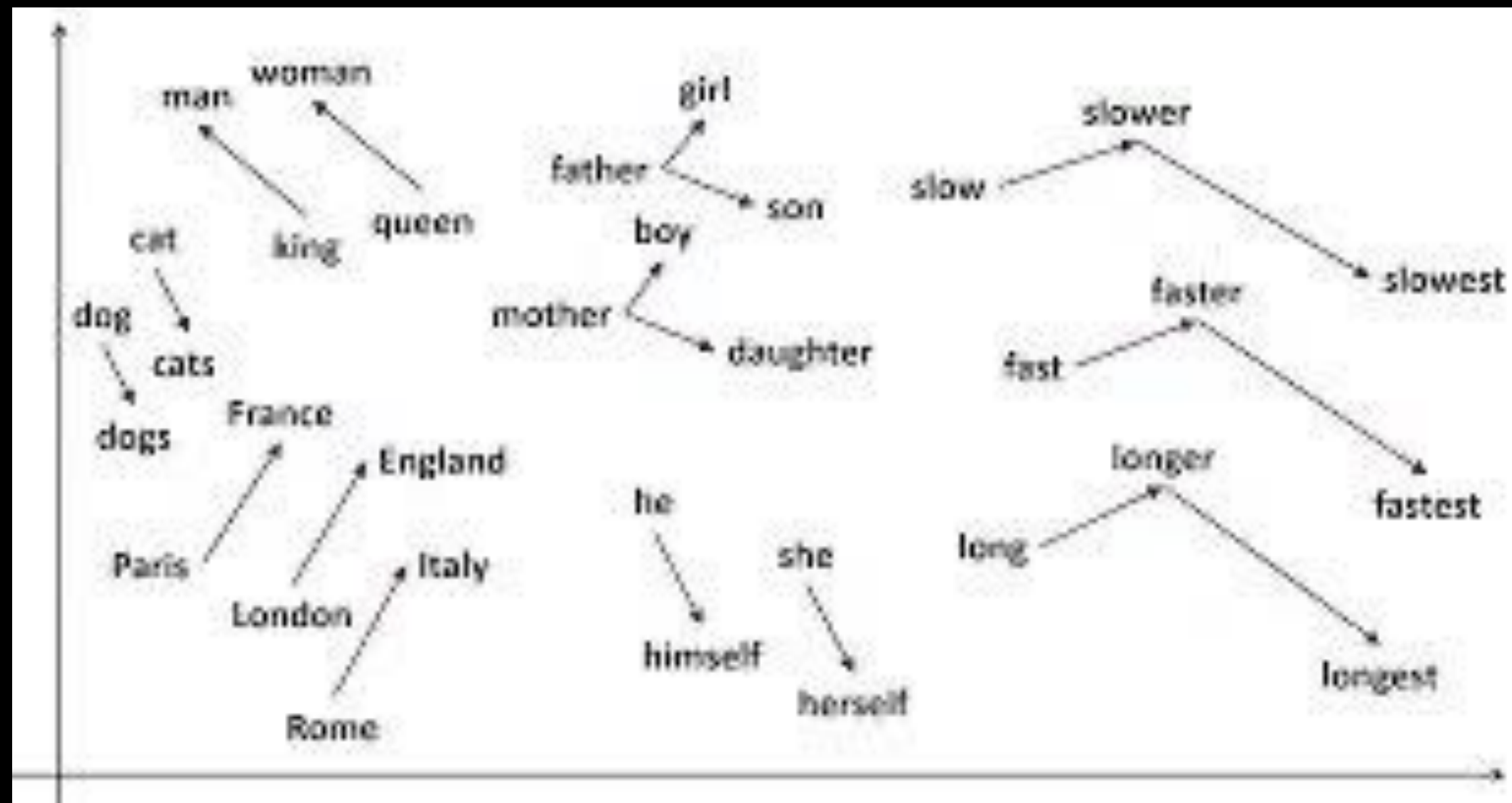


In 2016 or so, large word vector databases were revealed.

These were created with a simplistic training mechanism, rewarding word co-occurrence in 50-100 dimensions, but yielded emergent behavior (analogy spaces)

# Word vectors & Analogy space!

Not just related words, but how they are related!



**This spurred people to go further**

# Creating sentence embeddings and document embeddings!

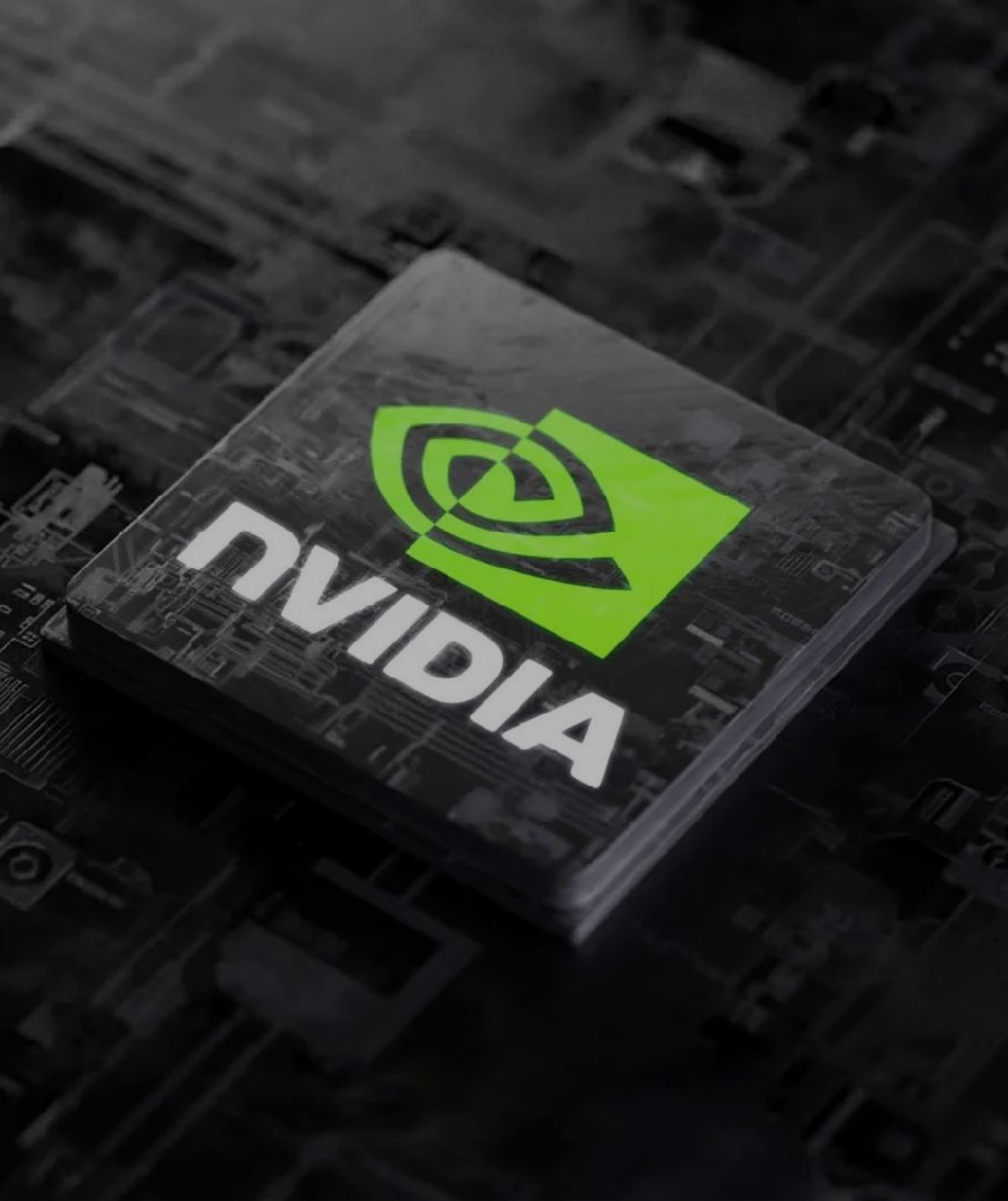
*Think of these as a mathematical  
representations for what something "means"*

**Google used this to make  
search much better**



**Innovation**  
**in the training techniques**  
**To BERT and then GPT**



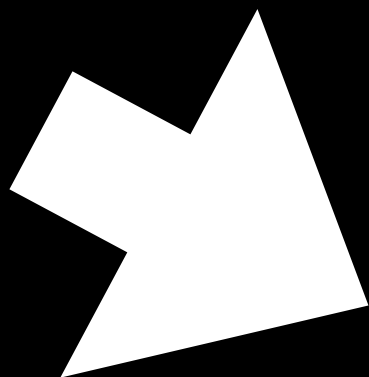


**Huge advances in AI  
acceleration hardware**

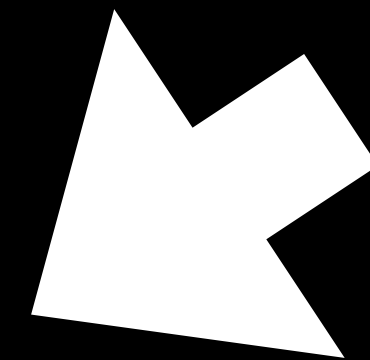
**The core thing AI does is  
multiply math matrices**

**NVidia was already doing  
that for games..**

New insights and first glimpses into emergence

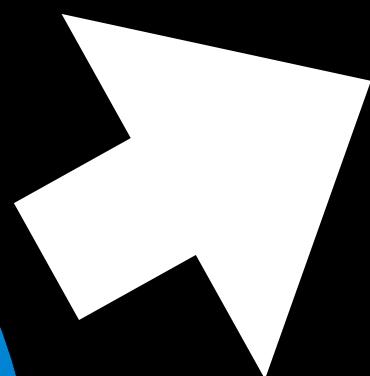


New models

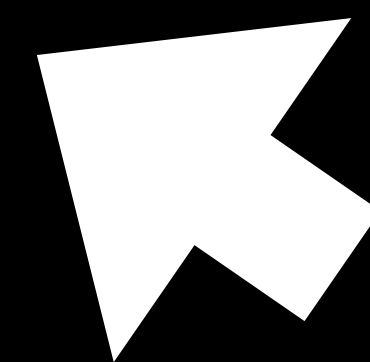


**Rapid innovation path**

New hardware



New AI training algorithms



# Generative Ai

## A brief history

**1956**  
The Dartmouth Summer Research Project on Artificial Intelligence workshop, considered a founding moment for the field of AI, convenes<sup>55</sup>

**2012**  
AlexNet, a convolutional neural network trained on GPUs, achieves a breakthrough performance on the ImageNet benchmark test<sup>57</sup>

**2016**  
AlphaGo beats 18-time world Go champion Lee Sedol<sup>59</sup>

**2019**  
Hugging Face releases V1 of its Transformers NLP library<sup>61</sup>

**2021**  
The European Union proposes the AI Act, the first legislation concerning AI, that would ban, regulate, or allow AI applications based on an assigned risk category<sup>62</sup>

**2021**  
The Beijing Academy of Artificial Intelligence announces Wu Dao 2.0, a multimodal text and image foundation model with 1.75 trillion parameters<sup>65</sup>

**2022**  
DeepMind introduces Gato, a general purpose agent foundation model that can perform a wide range of tasks across multiple modalities and embodiments<sup>67</sup>

**2024**  
Several new search engines have emerged, featuring foundation-model-based "quick answer" services that give users direct and thorough answers to queries

**2027**  
A prestigious art museum opens a gallery dedicated solely to AI-generated artwork

**2030**  
75% of knowledge workers globally interact with an application, service, or agent powered by foundation models daily

**1986**  
Dave Rumelhart, Geoffrey Hinton, and Ronald Williams publish a landmark paper on back-propagation - laying the foundation for modern neural networks<sup>56</sup>

**2015**  
OpenAI is founded<sup>58</sup>

**2017**  
Google introduces the Transformer AI network architecture in a paper titled "Attention Is All You Need." Transformers now underpin most foundation models<sup>60</sup>

**2020**  
OpenAI trains GPT-3 (Generative Pre-trained Transformer), which becomes the world's most sophisticated large language model<sup>63</sup>

**2021**  
Researchers at Stanford University publish a paper coining the phrase "foundation models"<sup>64</sup>

**2022**  
A piece of AI-generated art wins first place in the digital arts category at the Colorado State Fair's fine arts competition<sup>66</sup>

**2022**  
OpenAI releases ChatGPT, a highly sophisticated chatbot<sup>68</sup>

**2024**  
A new generation of virtual assistants is built with transformer-based foundation models. Increased adoption is attributed to more sophisticated and natural language ability

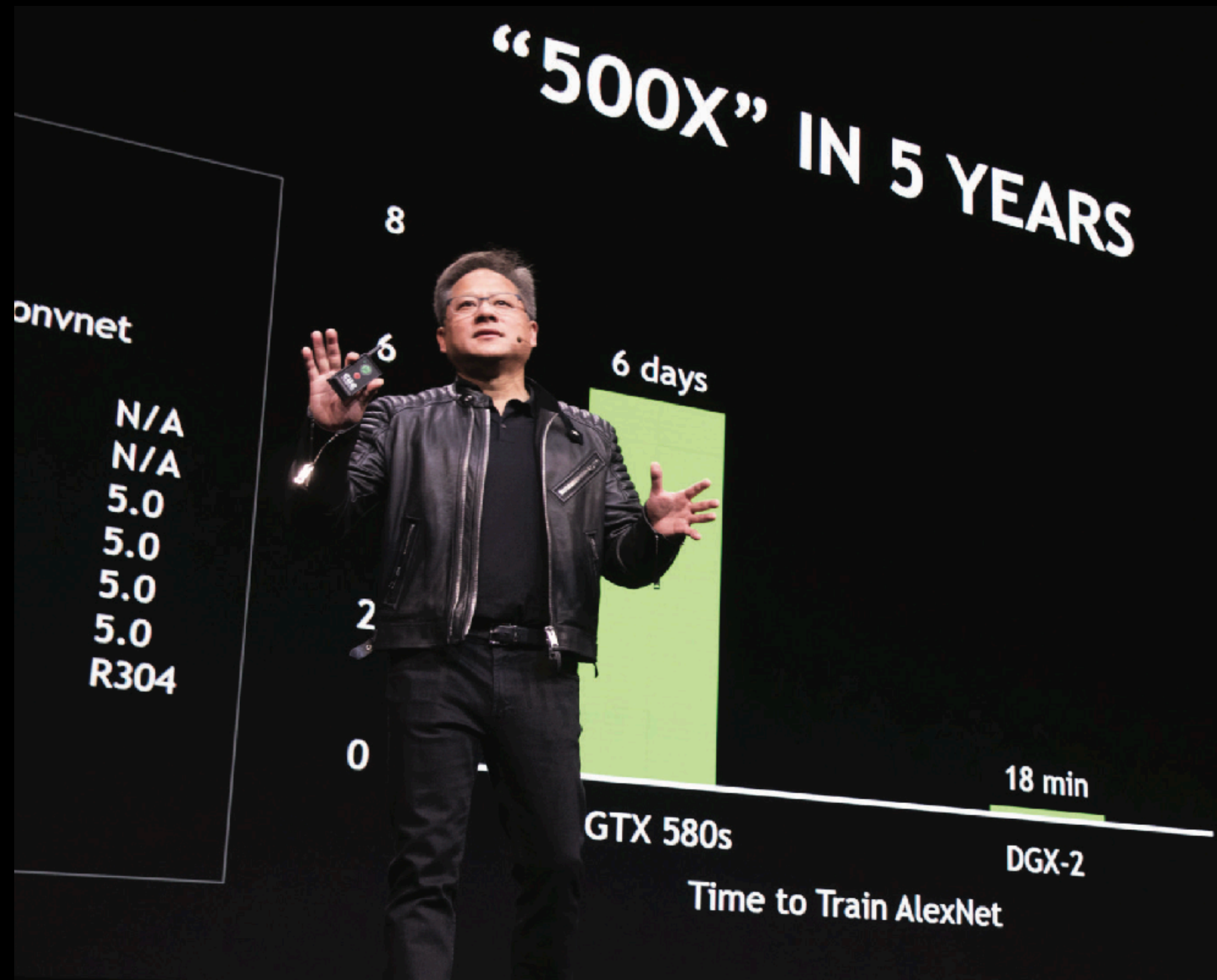
**2025**  
A consumer electronics firm introduces ear buds that use an on-device foundation model to translate over 100 languages in near-real time

**2029**  
30% of social media ads are autonomously generated using foundation models. Ads must still pass human review

**2033**  
A university deploys a general-purpose foodservice robot, built with a foundation model, to take on odd jobs in the cafeteria. The robot learns new tasks easily, so can fill open roles as needed

**2023**

**AI will continue to get better in this way for 10-25 years**



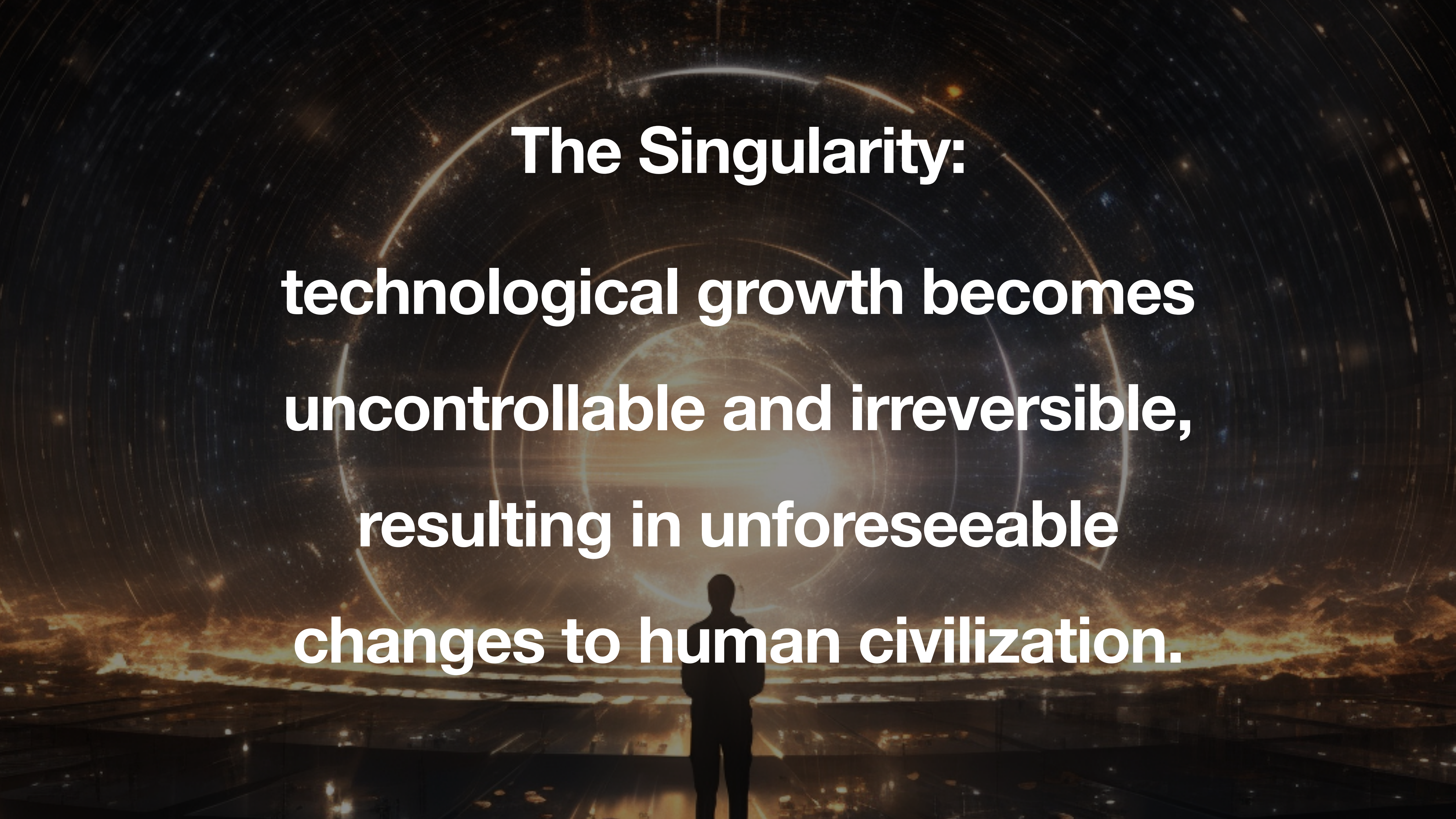
**Riding 3-5 innovation curves at once**

A person's silhouette stands on a balcony overlooking a city at night, with a large, glowing, futuristic circular structure in the sky. The scene is illuminated by the city lights and the bright light from the structure. The text "Multiply those together and you get something new" is overlaid on the image.

**Multiply those together and you get something new**

# The Singularity





**The Singularity:  
technological growth becomes  
uncontrollable and irreversible,  
resulting in unforeseeable  
changes to human civilization.**



## Transformer Architecture

A new powerful model architecture is released that is the basis for LLMs

2017

## GPT3

By scaling transformer models, special properties emerge such as few-shot learning. LLMs can be good at tasks without training.

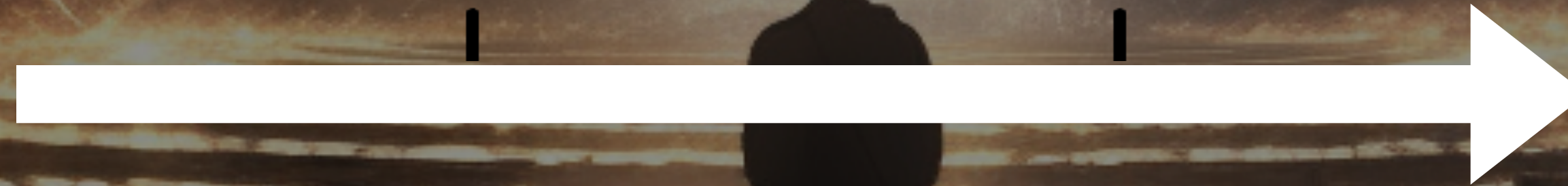
## ChatGPT


GPT-3.5 is trained with human feedback through reinforcement learning to achieve remarkable conversational abilities.

## GPT4

Shows strong performance on human tests and sparks of AGI.

2023



The image features a futuristic, circular digital interface. A person's silhouette is centered at the bottom, looking out over a glowing, orange-hued horizon. The background is a dark, starry space with a large, glowing circular arc. The text is displayed in a white box in the center of the image.

Shows strong  
performance on  
human tests and  
sparks of AGI.

# THE TYPES OF INTELLIGENCE

by Mark Vital

**spatial**  
visualizing  
the world  
in 3D



**naturalist**  
understanding  
living  
things and  
reading  
nature



**musical**  
discerning  
sounds, their  
pitch, tone,  
rhythm, and  
timbre



**intra-  
personal**  
understanding  
yourself,  
what you feel,  
and what  
you want



quantifying  
things,  
making  
hypotheses  
and  
proving  
them



**logical-  
mathematical**

finding  
the right  
words to  
express  
what  
you mean



**linguistic**

coordinating  
your mind  
with your body



**bodily-  
kinesthetic**

sensing people's  
feelings and motives



**interpersonal**

tackling  
the questions  
of why we live,  
and why we die



**existential**

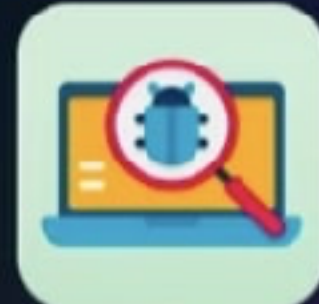
# ACHIEVEMENTS UNLOCKED BY LLMS

EMERGENT ABILITIES OF  
LARGE LANGUAGE MODELS (APR/2023)

**S** GPT-3 13B,  
PaLM 8B



Mod.Arithmetic\*



Debugging\*



Comprehension\*

**M** GPT-3 175B,  
LaMDA 137B,  
PaLM 64B,  
Chinchilla 7B



LinguisticsPuzzles\*



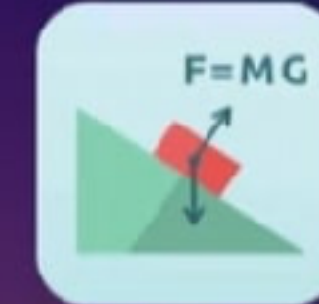
EmojiMovie\*



GRE-Comprehension\*



MetaphorUnderstanding\*

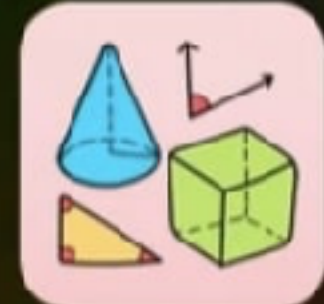


PhysicalIntuition\*

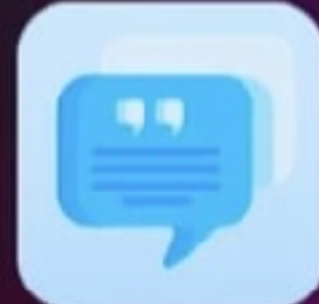


LogicalDeduction\*

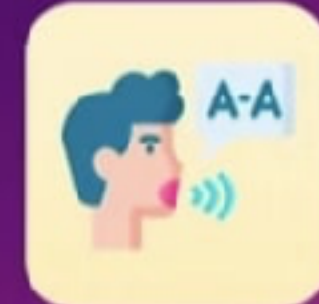
**L** PaLM 540B,  
Chinchilla 70B



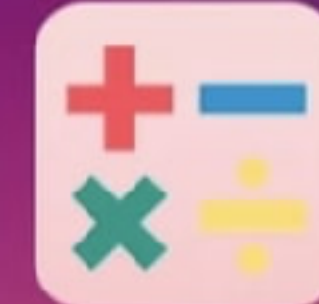
GeometricShapes\*



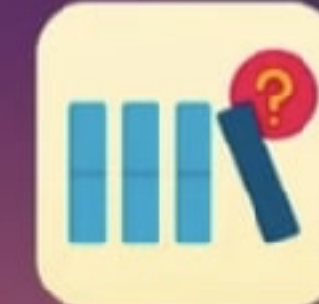
Proverbs\*



PhoneticAlphabet\*



ElementaryMath\*



CausalJudgment\*

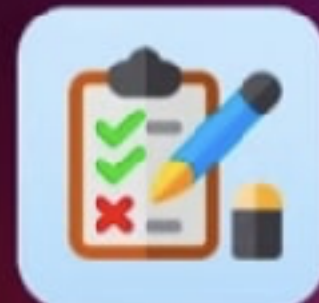


CodeLineDescription

**XL** GPT-4,  
Gemini (est.)



College-LevelExams



Self-Critique/Reflection



AppBuilding



SpatialReasoning

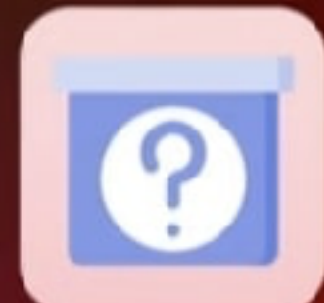


AdvancedCreativity

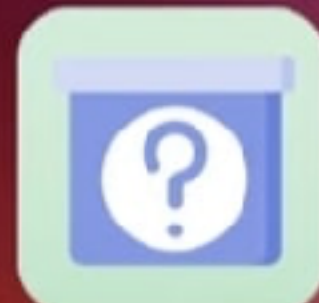


EmbodimentOptions

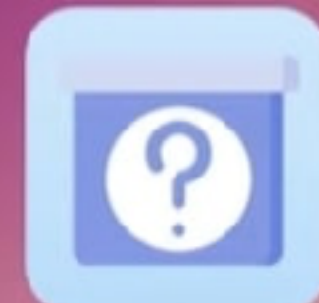
**Next...**



Grounding



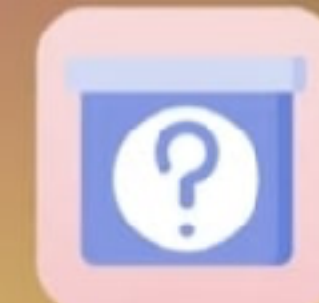
Long-HorizonPlanning



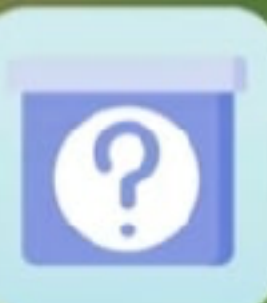
Persuasion



AdvancedEmbodiment



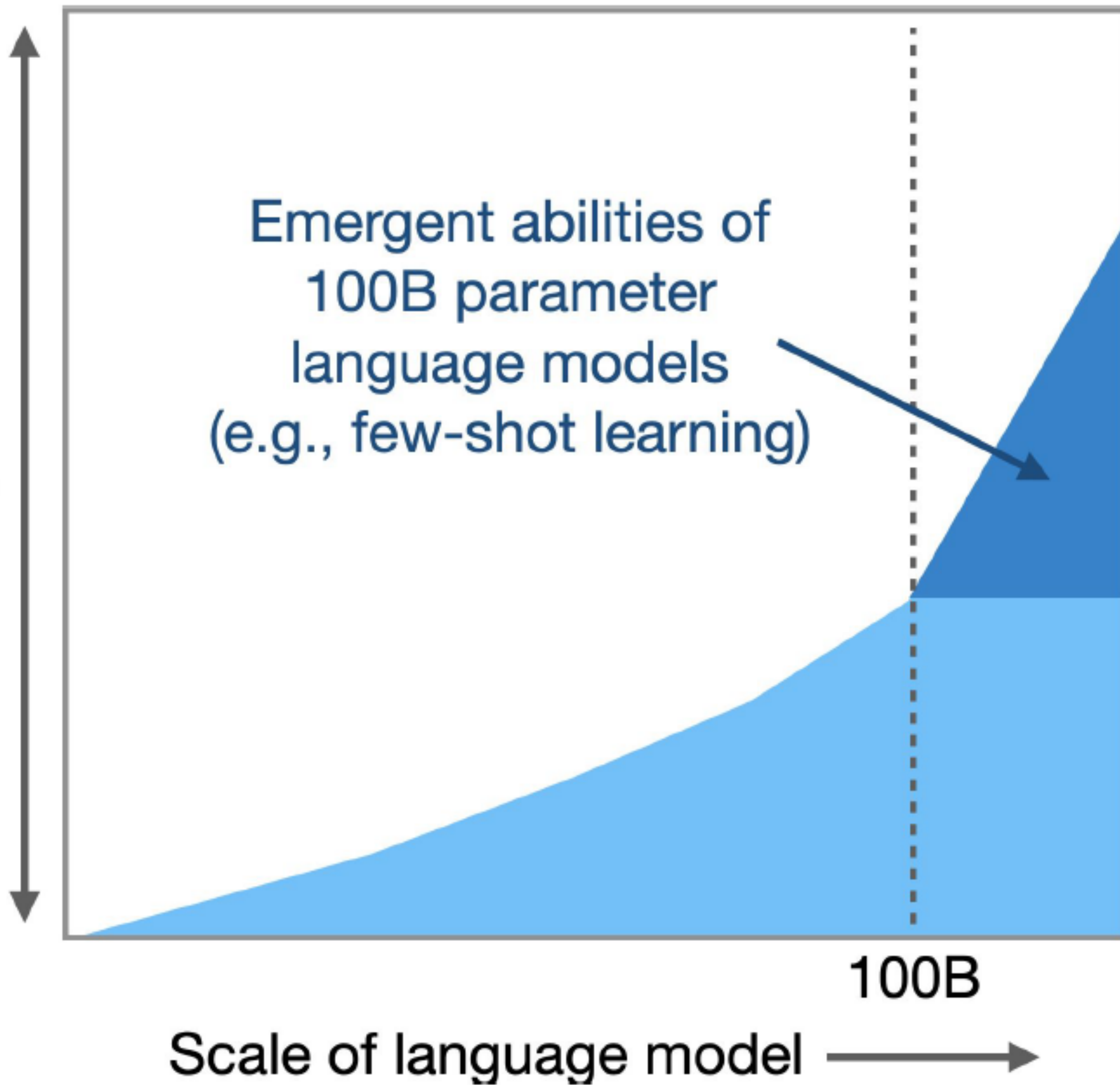
Awareness



More...

“Things that  
language models  
can do”

Emergent abilities of  
100B parameter  
language models  
(e.g., few-shot learning)



100B

Scale of language model →

### Question

Tracy used a piece of wire 4 feet long to support tomato plants in the garden. The wire was cut into pieces 6 inches long. How many pieces did she obtain?

### 62B Model Output

She cut the wire into 6 inch pieces. This means she got  $4 * 6 = 24$  pieces. The answer is 24.

**(Incorrect: semantic understanding error)**

### 540B Model Output

The wire was 4 feet long. This means it was  $4 * 12 = 48$  inches long. It was cut into pieces 6 inches long. This means she obtained  $48 / 6 = 8$  pieces. The answer is 8.

**(Correct)**

### Question

Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?

### 62B Model Output

He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for  $3 * 6 = 18$  miles. The answer is 18. **(Incorrect: semantic understanding error)**

### 540B Model Output

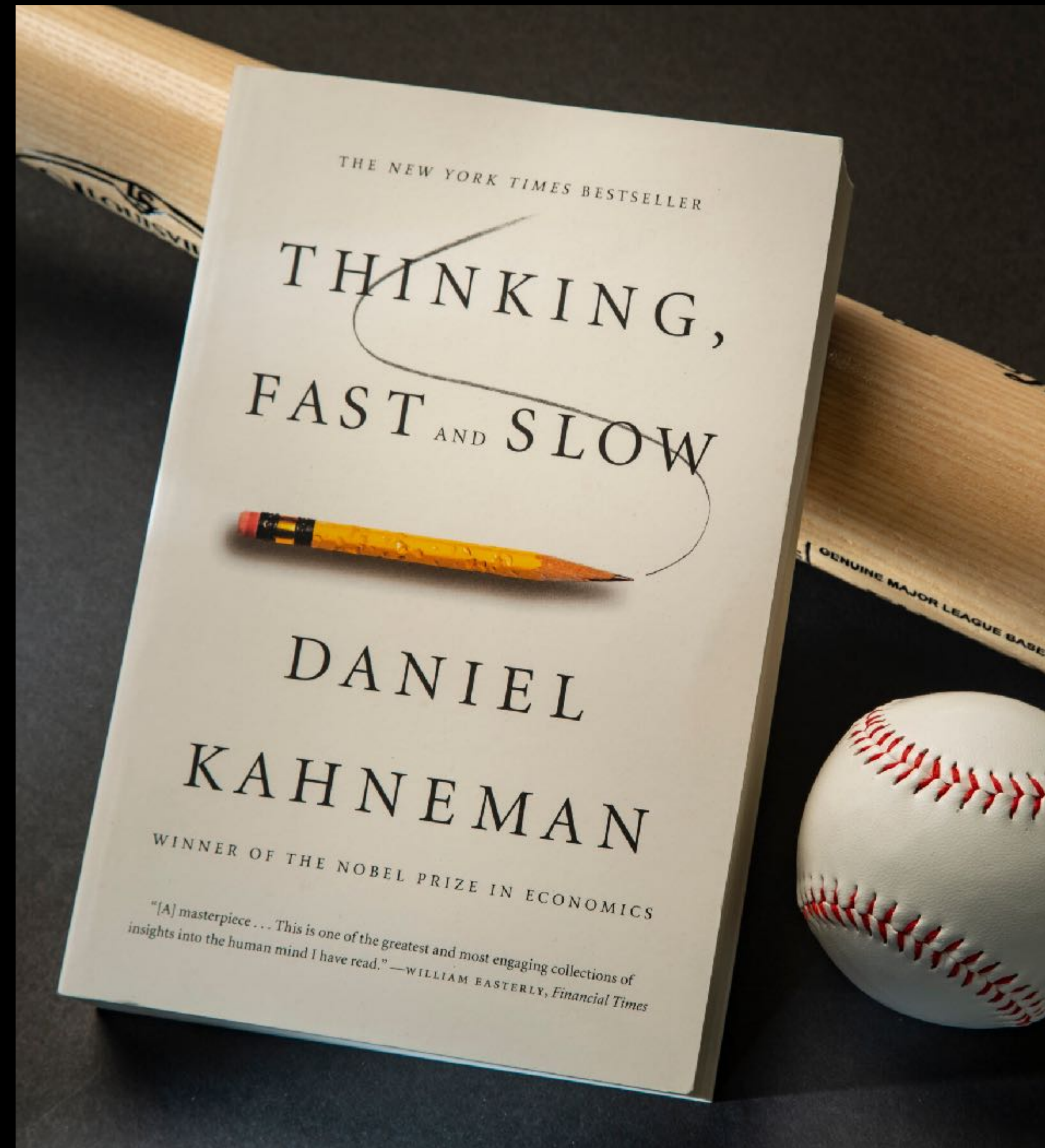
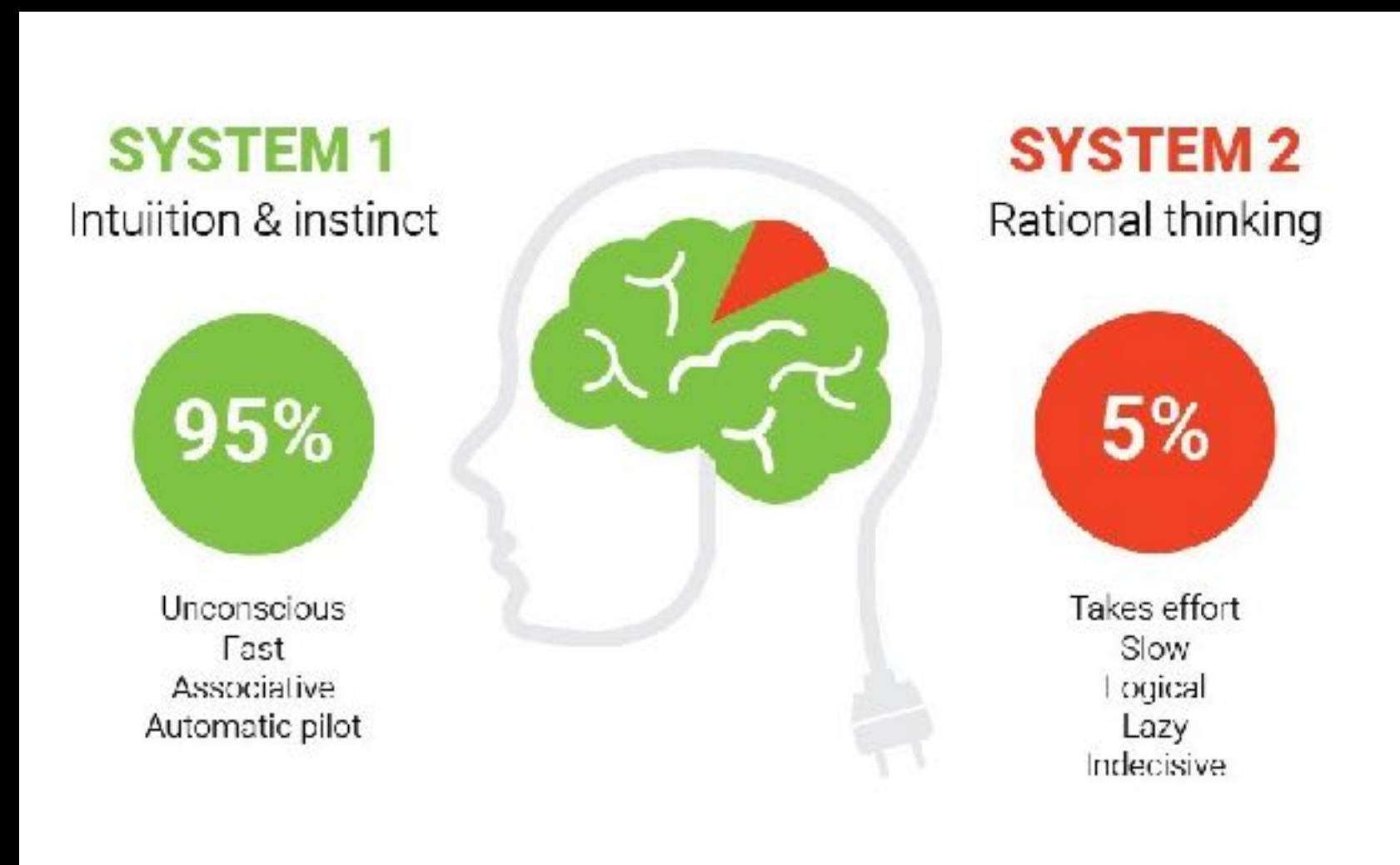
He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels  $3 * 10 = 30$  miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him  $30 / 6 = 5$  hours. The answer is 5.

**(Correct)**

# Generative Ai

## Daniel Kahneman

- Remember System 1 & 2
- GenAI is System 1



# Generative Ai

Stephen Pinker

- All languages are built on the same universal grammar
- the basic structures of language are a part of our biological inheritance
- There's an engine that takes knowledge and grammar and vocabulary and generates language

The image shows the front cover of the book 'The Language Instinct: How the Mind Creates Language' by Stephen Pinker. The cover is divided into three horizontal bands. The top band is a light gray color. The middle band is a darker gray color and contains the title 'THE LANGUAGE INSTINCT' in large, white, all-caps, sans-serif font. Below the title, in a smaller white font, is the subtitle 'HOW THE MIND CREATES LANGUAGE'. The bottom band is a light gray color, matching the top band.

THE LANGUAGE INSTINCT  
HOW THE MIND CREATES LANGUAGE





**Emulating but not copying nature**

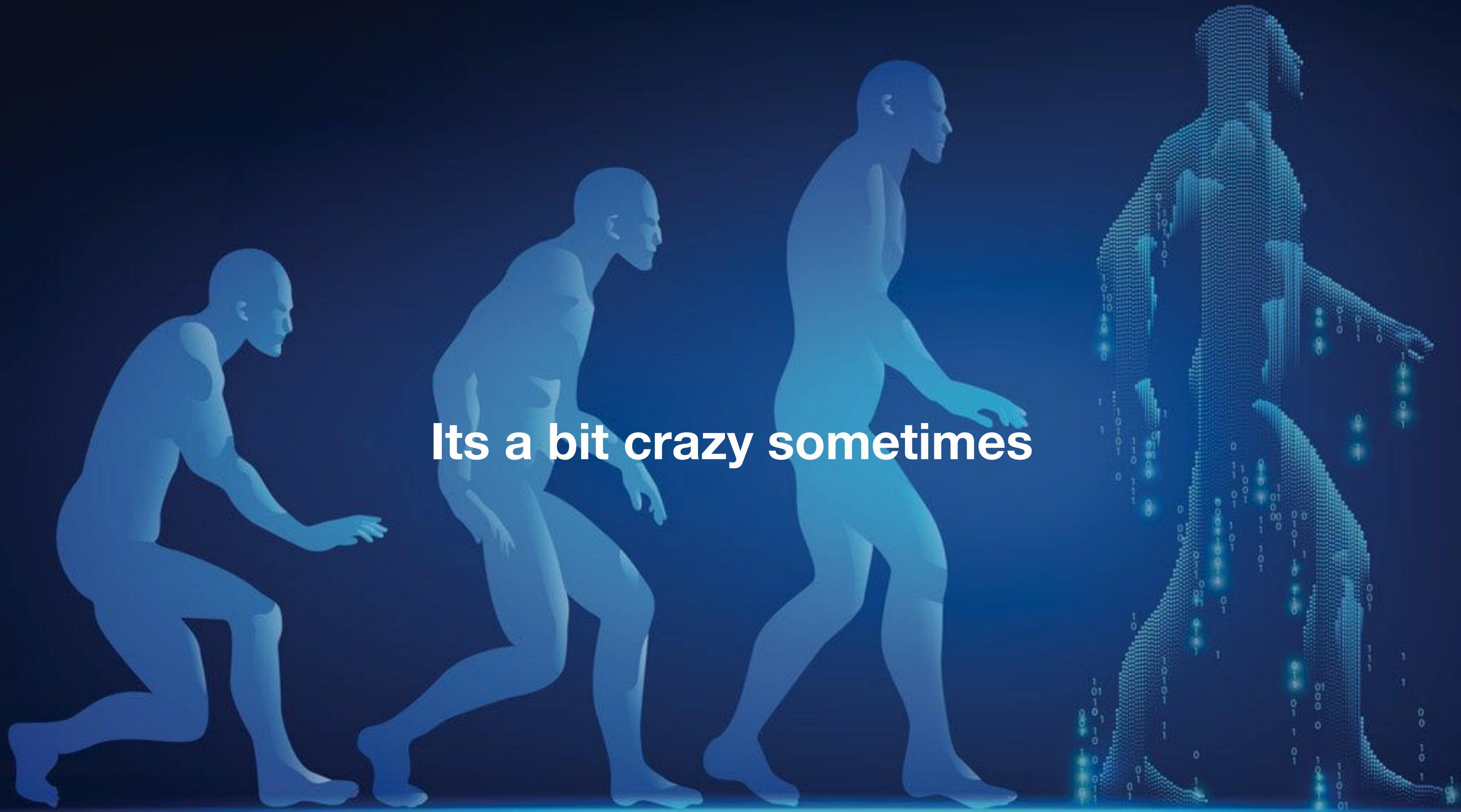


**Pinker's and Kahneman's models both  
forming in this emerging general  
Intelligence**

The image depicts a sequence of four human-like figures in profile, moving from left to right. The first three figures are rendered in a solid blue color and show a progression from an ape-like posture to a more upright, modern human form. The fourth figure is a digital human, composed of a grid of blue dots and lines, with binary code (0s and 1s) floating around it. The background is a dark blue gradient with a glowing blue horizontal line at the bottom.

**Already has “pretty good”  
answers to almost all questions**

**Its a bit crazy sometimes**

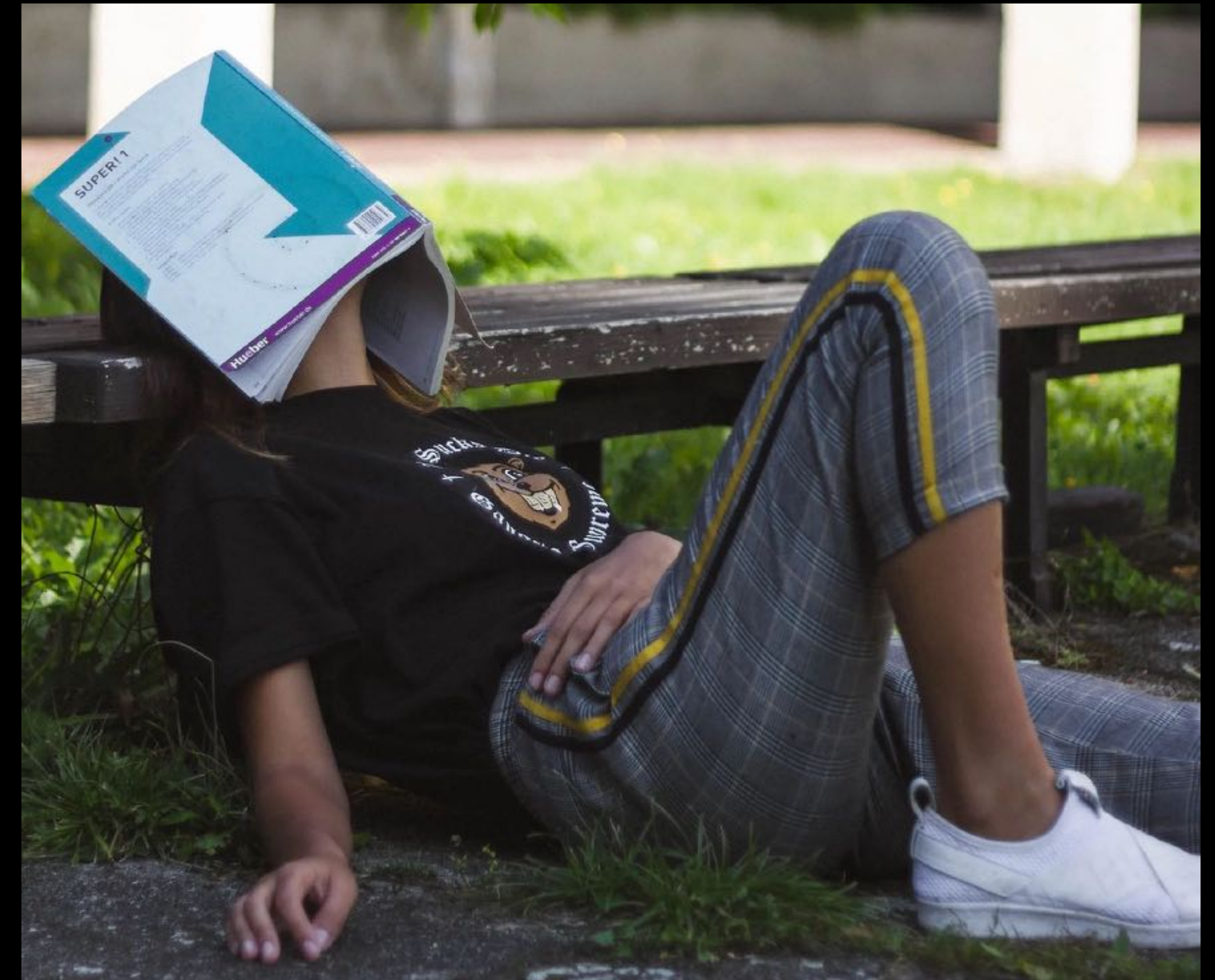


Like an overworked but brilliant grad student

Exciting cogent answers to any question

May be wrong on the face of it

More perilous is when they look correct and aren't



**If you woke them up and asked them a question,  
how would it go?**

# **Generative Ai**

## **An instructive example**

J

whats a simple recipe using almonds and whipped cream that would be good



One simple and delicious dessert you can make using almonds and whipped cream is an "Almond and Whipped Cream Parfait." This dessert combines the rich and nutty flavor of almonds with the light and airy texture of whipped cream. Here's how to make it:



Ingredients:

- 1 cup of toasted almonds (you can toast them in a dry skillet over medium heat until they become fragrant)
- 1 cup of whipped cream (you can use store-bought or make your own by whipping heavy cream with a bit of sugar until stiff peaks form)
- 2 tablespoons of honey or maple syrup (adjust to taste)
- Fresh berries (strawberries, raspberries, blueberries, etc.) for garnish (optional)

Instructions:

1. Start by preparing your whipped cream. If making it from scratch, whip the heavy cream in a cold mixing bowl using an electric mixer. Gradually add sugar while whipping until stiff

J

make a shopping list from that



Certainly! Here's a shopping list based on the recipe for Almond and Whipped Cream Parfait:

**Produce:**

- Fresh berries (strawberries, raspberries, blueberries, or your choice)

**Dairy:**

- Heavy cream (if making whipped cream from scratch)
- Whipped cream (you can buy pre-made whipped cream if you prefer)

**Pantry:**

- Almonds (you may need to purchase these if you don't have them already)
- Honey or maple syrup (for drizzling, adjust to taste)



J

ok now divide the recipe so it can be done by 3 people efficiently



Certainly! To divide the Almond and Whipped Cream Parfait recipe for three people efficiently, you can follow these adjusted quantities:

**Ingredients:**

- 1/3 cup toasted almonds (for toasting, use approximately 1/9 cup per person)
- 1/3 cup whipped cream (for convenience, you can use store-bought)
- 2-3 teaspoons honey or maple syrup (adjust to taste)
- Fresh berries (a small handful per person for garnish, as desired)

**Instructions:**

1. Divide the toasted almonds evenly among three serving glasses or bowls, using approximately 1/9 cup of almonds per person.
2. Add a layer of whipped cream on top of the almonds, dividing the 1/3 cup evenly among the servings.

J

no I mean divide the instructions so 3 people are busy making it



I see what you mean now. To divide the instructions so that three people can efficiently make the Almond and Whipped Cream Parfait together, you can assign specific tasks to each person. Here's how you can divide the instructions:



### **Person 1: Toasted Almonds**

1. Person 1 should be in charge of toasting the almonds. They can take 1 cup of almonds (the original recipe's quantity), toast them in a dry skillet until fragrant, and then roughly chop them. Once chopped, they can divide the toasted and chopped almonds into three equal portions (1/3 cup each) for the parfaits.

### **Person 2: Whipped Cream**

1. Person 2 can focus on preparing the whipped cream. They can whip 1 cup of heavy cream (the original recipe's quantity) using an electric mixer until stiff peaks form. If needed, they can add a bit of sugar while whipping to sweeten the cream. Once whipped, they should divide the whipped cream into three portions, enough for each parfait.

J

ok, now tell me those directions, assuming one person is a dwarf, one an elf, and one an orc, using tolkein's languages



Certainly! Here are the divided instructions for making the Almond and Whipped Cream Parfait, with each task assigned to a different character from Tolkien's world:



### **For the Dwarf (Khazâd): Toasted Almonds**

- In the language of the Dwarves, instruct Khazâd to toast the almonds. Speak in the resonant Dwarvish tones, "Khazâd! Tozgum-ishi othok!" which means "Dwarf! Toast these almonds!"
- Provide a small skillet and 1 cup of almonds to Khazâd and let him toast them until they're fragrant. Make sure to speak in Dwarvish so he fully understands his task.
- Once the almonds are toasted and aromatic, ask Khazâd to roughly chop them into small pieces. "Khazâd! Agh burz-ishi tozhadruk!" means "Dwarf! Chop these almonds!"
- Khazâd should divide the toasted and chopped almonds into three equal portions (1/3 cup each) for the parfaits.

### **For the Elf (Elendil): Whipped Cream**

- In the mellifluous Elvish tongue, ask Elendil to prepare the whipped cream. Say, "Elendil, ammenya-néreo!" which translates to "Elf, make whipped cream!"

Asked a very random question

Great response!

Engaged in analytical **thinking** and **planning**

Received feedback, live, and improved its results

It demonstrated cultural sensitivity (those elves...)

# How long it would take you to:

1. Create Recipe for random ingredients
2. Shopping list
3. Efficient instructions for a group of arbitrary size
4. Handle multiple cultures in their own languages with sensitivity

*Disclaimer : we did not ask the elves what they thought of our work*

**What if everyone had one of  
these at their fingers tips?**





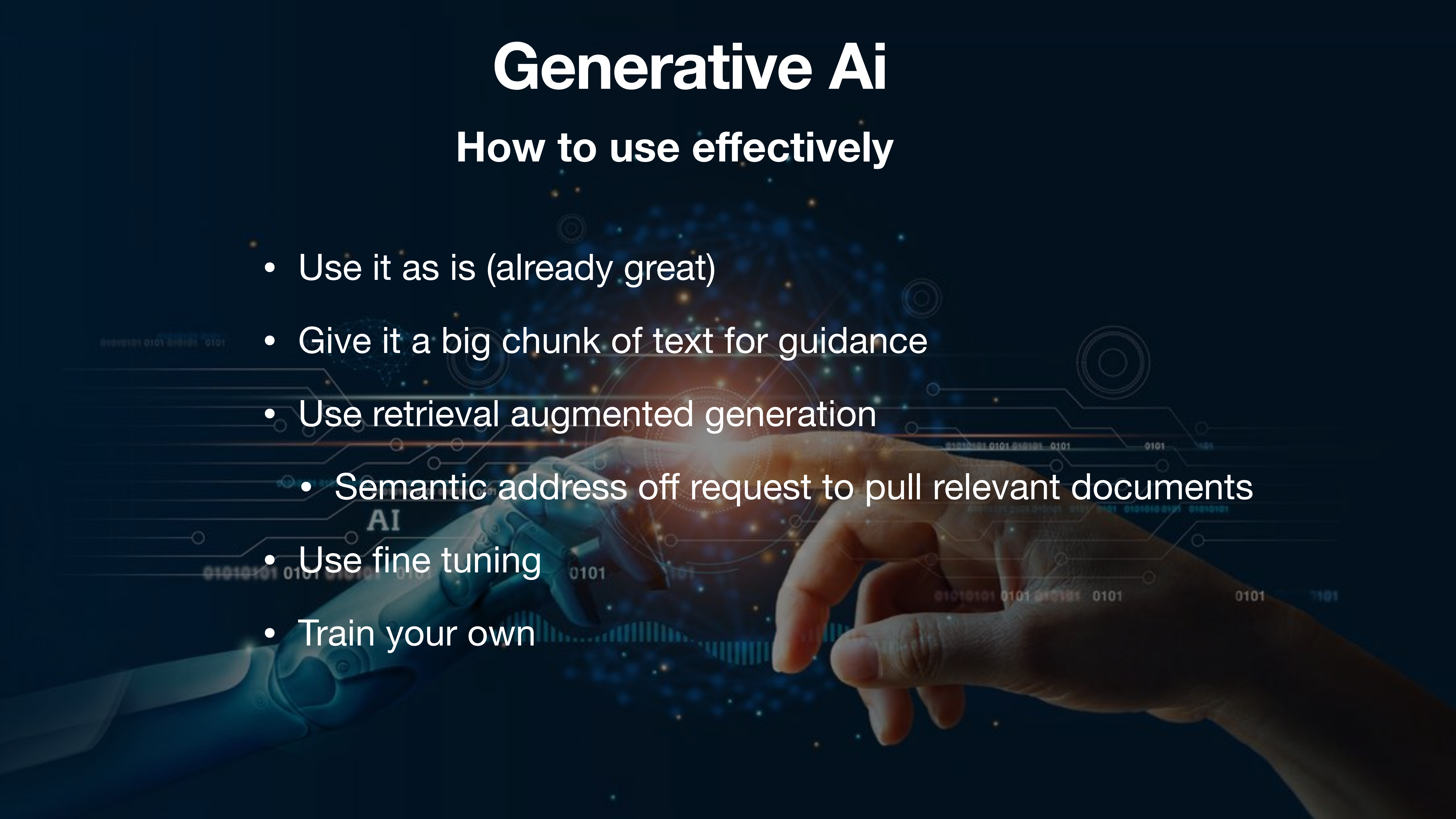
**Only paid for when they used?**

**(Unlike a human worker)**

**Already does many things much  
faster and better?**

# Generative Ai

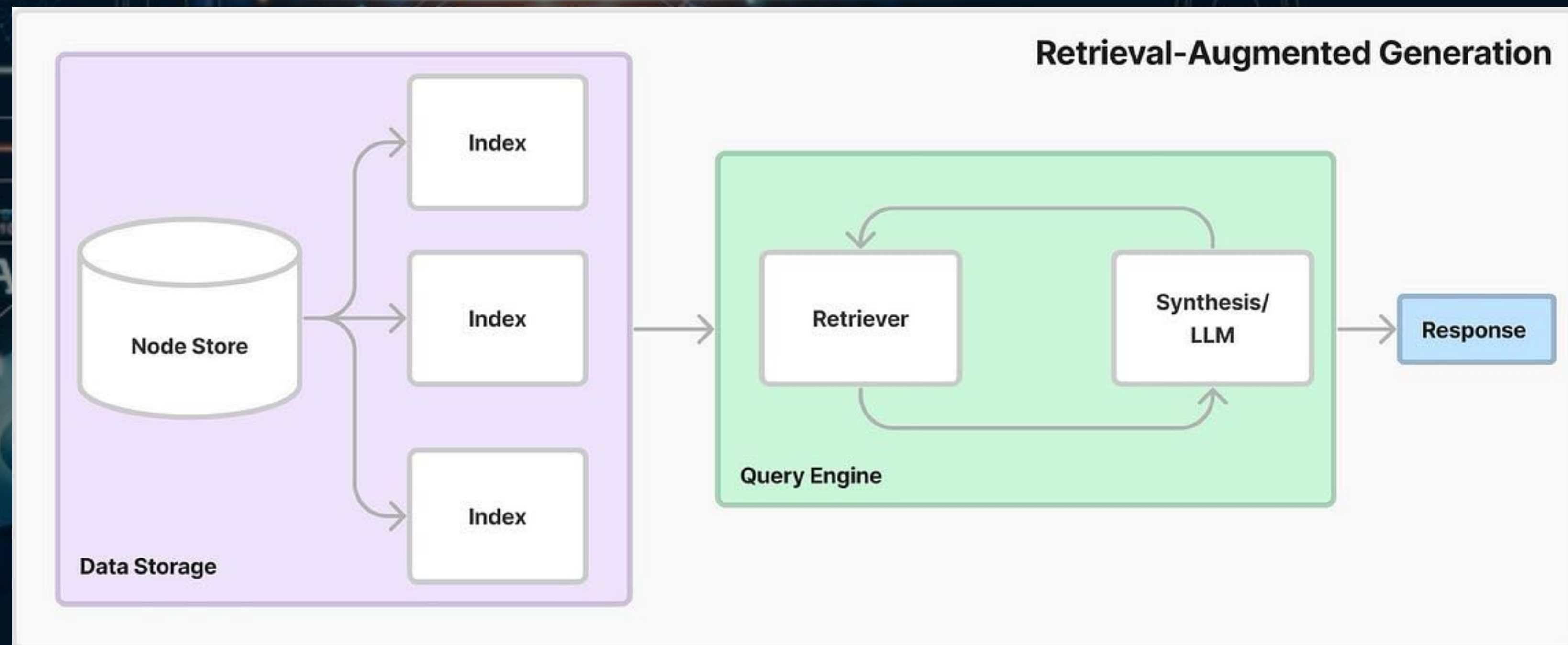
## How to use effectively

- Use it as is (already great)
  - Give it a big chunk of text for guidance
  - Use retrieval augmented generation
    - Semantic address off request to pull relevant documents
  - Use fine tuning
  - Train your own
- 
- The background features a dark blue space with glowing particles and binary code (0101). A human hand is shown holding a glowing blue, semi-transparent AI hand. The AI hand has a digital, circuit-like appearance. The overall theme is artificial intelligence and human-machine interaction.



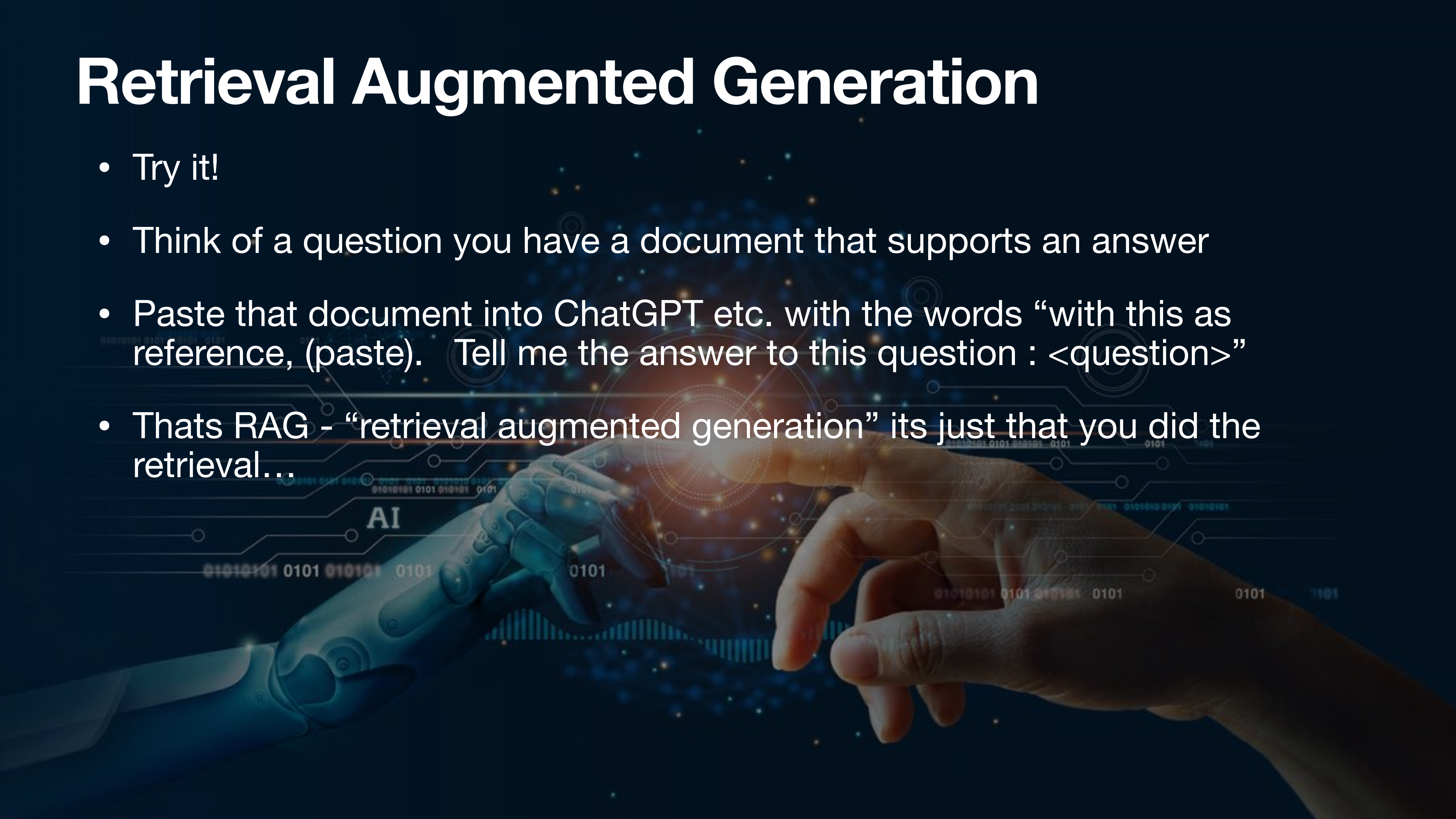
# Retrieval Augmented Generation

- Index your documents
- Retrieval top n related documents
- Ask GenAI to use them as reference and answer



# Retrieval Augmented Generation

- Try it!
- Think of a question you have a document that supports an answer
- Paste that document into ChatGPT etc. with the words “with this as reference, (paste). Tell me the answer to this question : <question>”
- That's RAG - “retrieval augmented generation” its just that you did the retrieval...

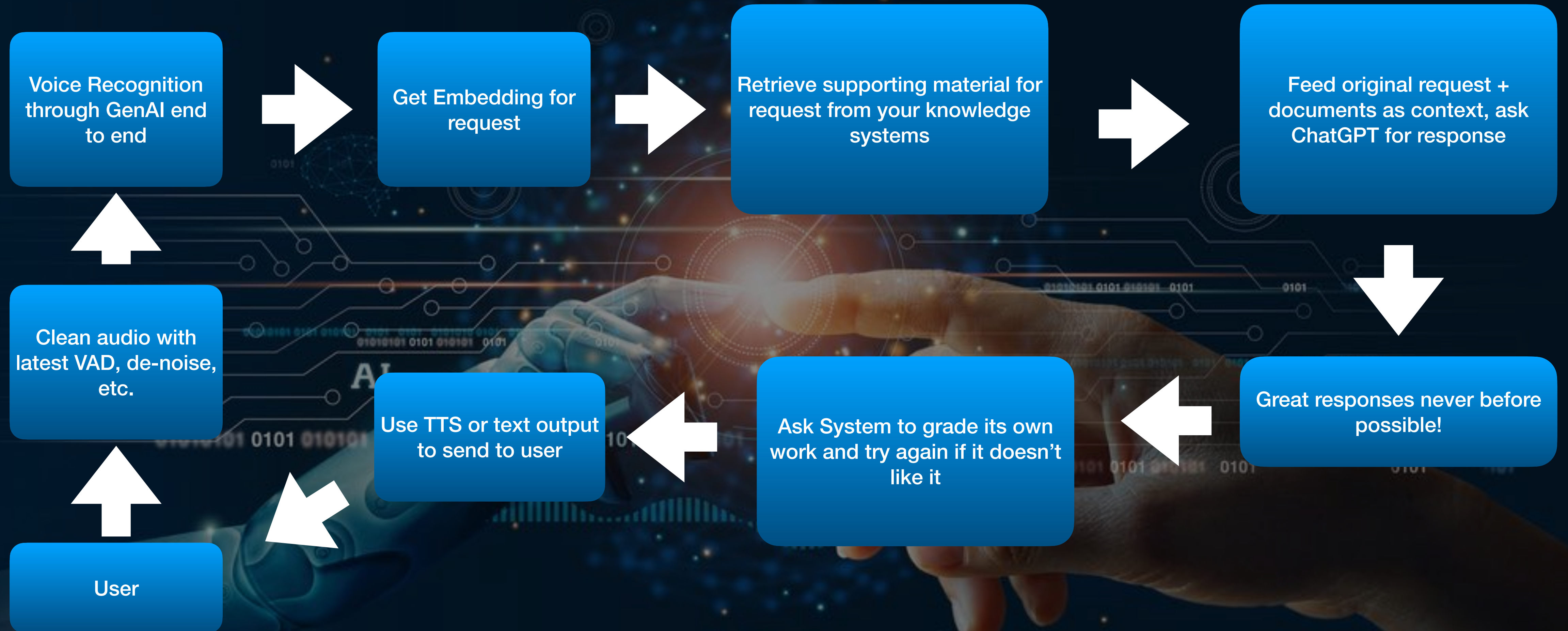


# Generative Ai

## The future

- Rapid advances in hardware
  - Capitalism is very good at focusing investment
    - Look for huge advances here!
- Rapid advances in algorithms
  - As hardware enables people try more things faster, we advance much faster
  - What took a year to do 5 years ago, we can now do in a few days to a week
  - (Thats the difference between a few advances a career, and miraculous innovation)
- Vast changes ahead

# Modern VoiceTech pipeline



# Generative Ai

## Call to action

AI is becoming truly intelligent

For users, its now more about teaching than training

System 1 is “solved”

Really good at many things already

Because the resource is shared, you only pay for what you use..

# Generative Ai

## The future

Your iPhone will run ChatGPT all by itself in 5 years

Movies will be made with no human actors

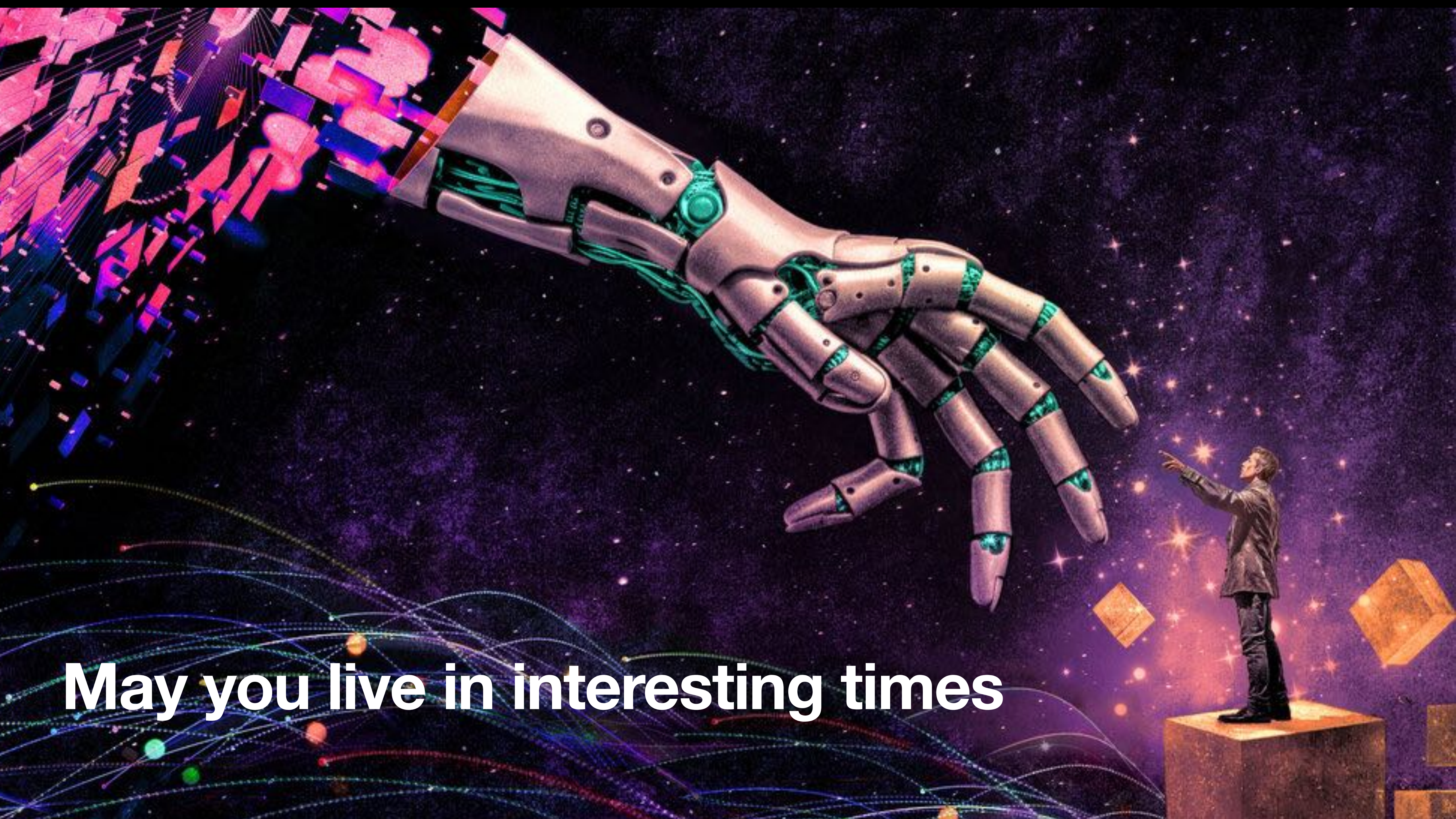
Almost any process will be planned and optimized with GenAI

All new works will be co-edited with Ai

Interactive fiction will be generated on the fly to please that particular person

Read “Dream Park”

Watch Apple’s “Knowledge Navigator” video from the 80’s!



**May you live in interesting times**

**Thanks**