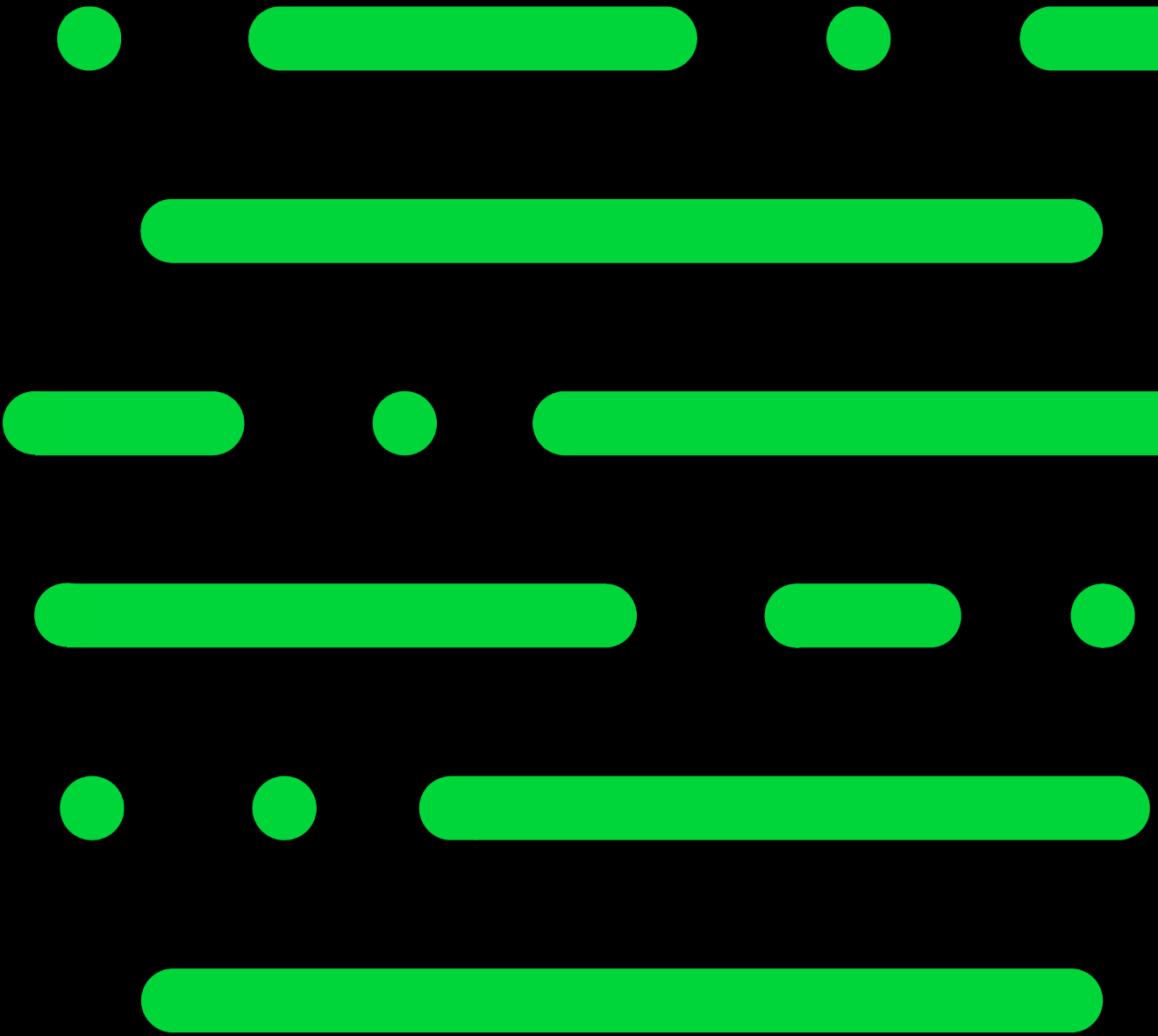


# GenAI, LLMs, and the Future

Jeremiah Edwards - Director  
of AI/ML, Sage AI

Mack Boulder 2024-02-23



# Who am I?

- **Jeremiah Edwards**
- **Director of AI/ML at Sage**
- **Background: Math, lots of math, then computers!**
- **Contact**
  - **[jeremiah.edwards@sage.com](mailto:jeremiah.edwards@sage.com)**
  - **Github:** <https://github.com/jeremiq/>

Sage



**Jeremiah Edwards**

jeremiq · he/him

I work in Machine Learning and Data Science, and have interests in higher category theory, type theory, logic, functional programming, and getting things done

# In this talk

## 1. Technical Preliminaries – from zero to ChatGPT

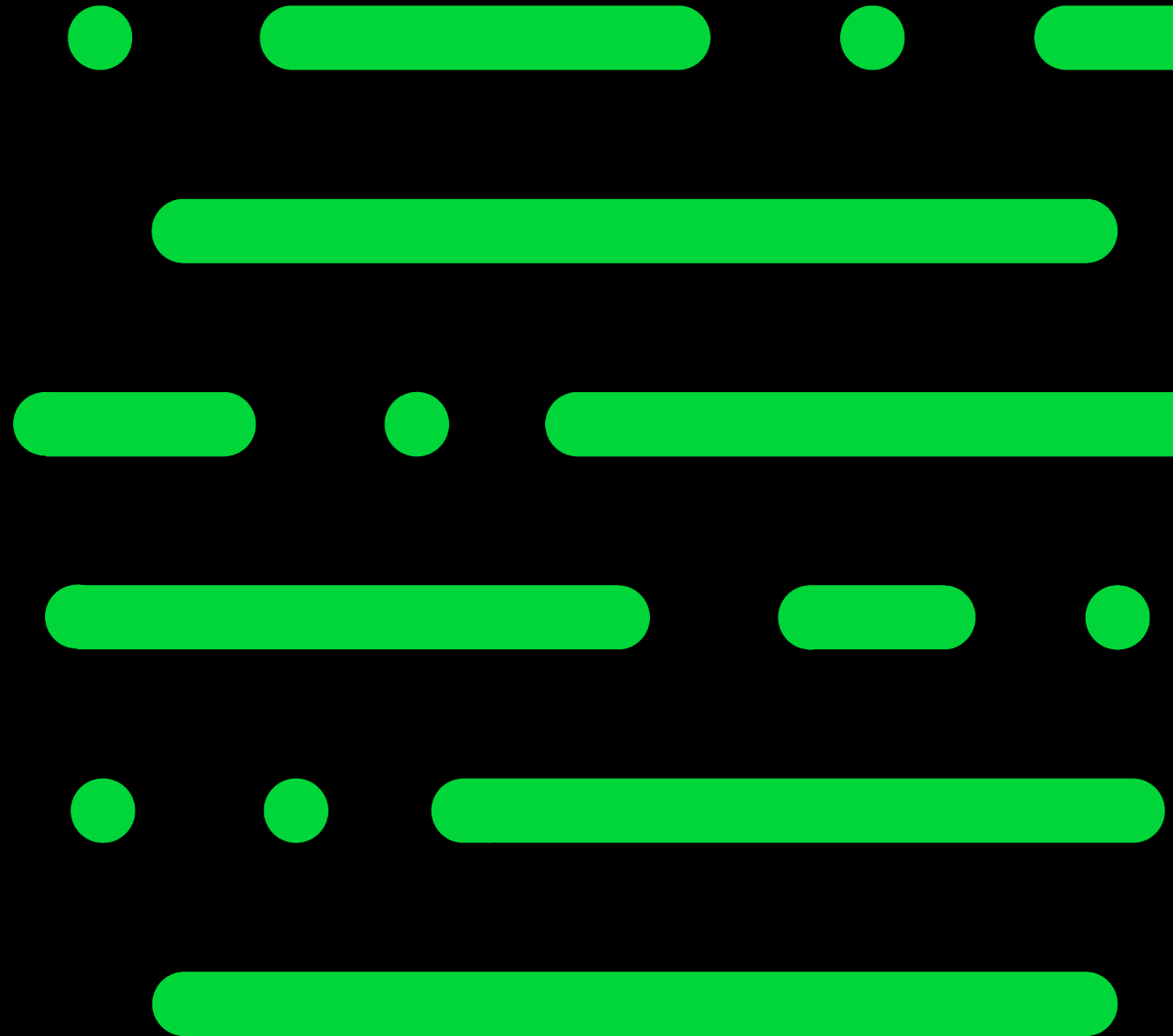
- What is 'generative' AI?
- Types of generative models
- LLMs and beyond

## 2. Errors, Risks, and Ethical Implications

- Inherent Problems with LLMs today
- Commercial Risk of Bad UX
- Embedding Facts in LLMs

## 3. Where do we go from here?

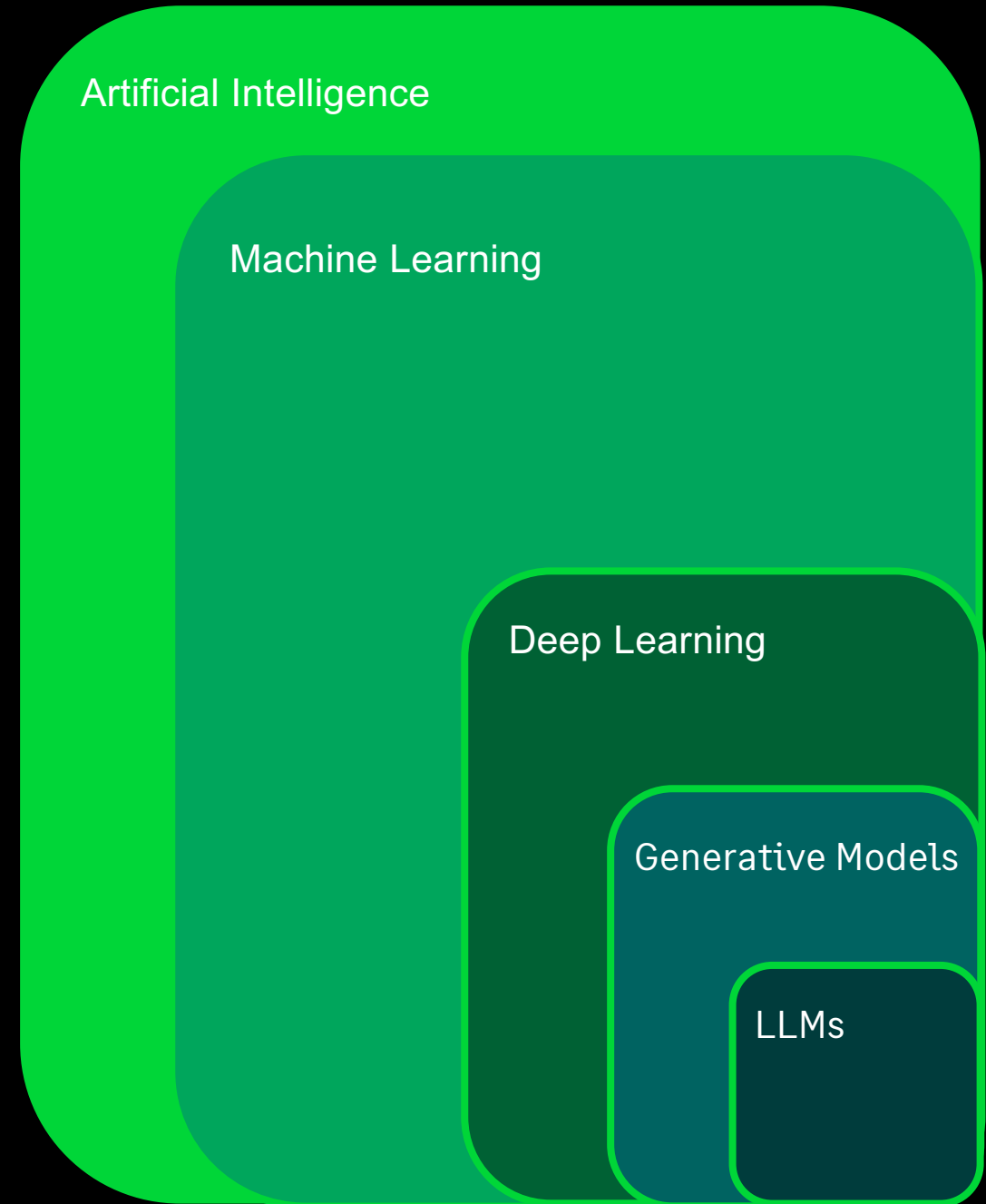
# Generative AI: From Zero to ChatGPT



# Generative AI

Generative AI describes algorithms (such as ChatGPT) that can be used to create new content, including audio, images, text, code, etc.

Large Language Models (LLMs) are a type of Generative AI which incorporate large amounts of text and predict possible completions of a given prompt.



# AI is a new Paradigm

## TRADITIONAL PROGRAMMING

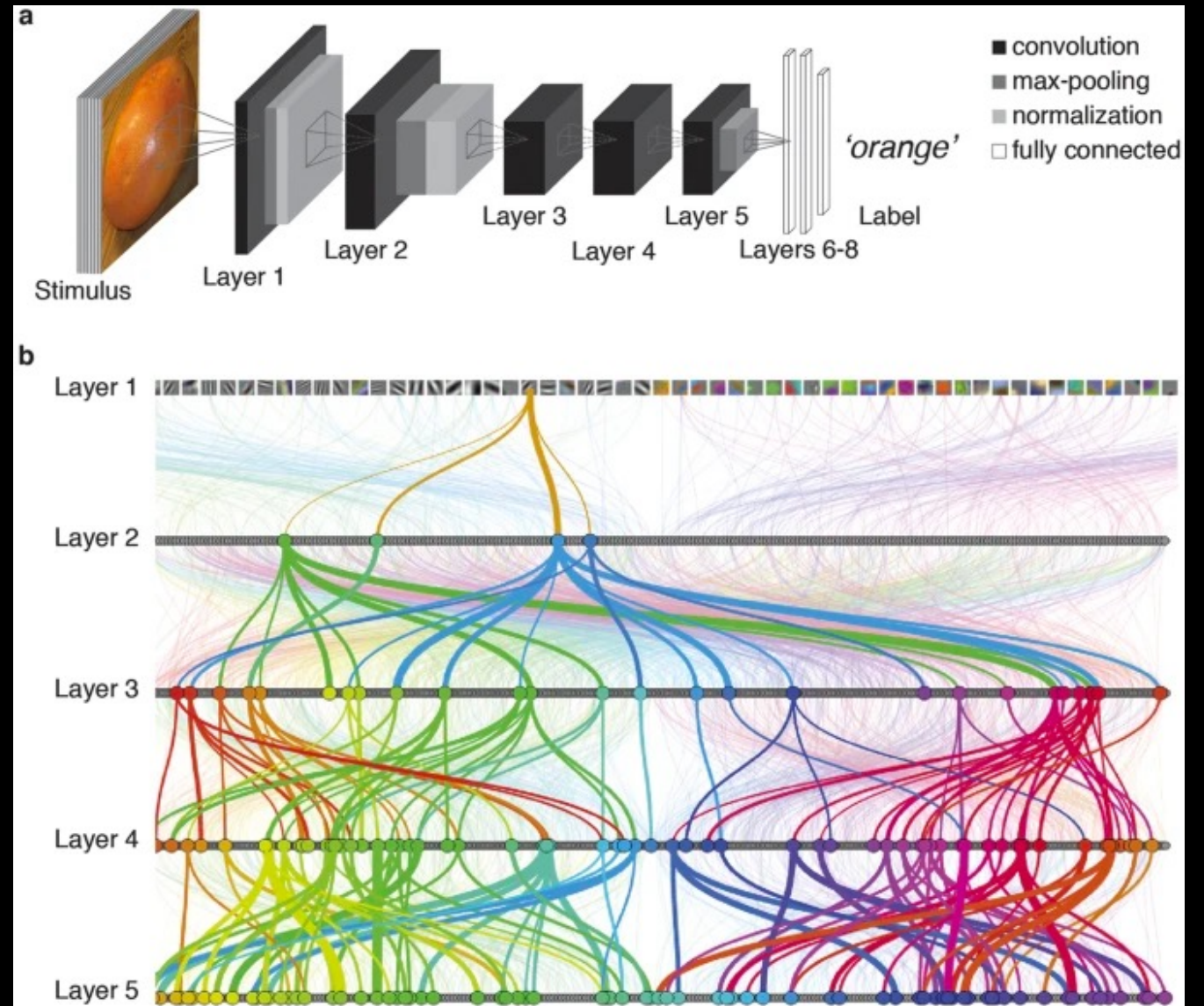


## MACHINE LEARNING



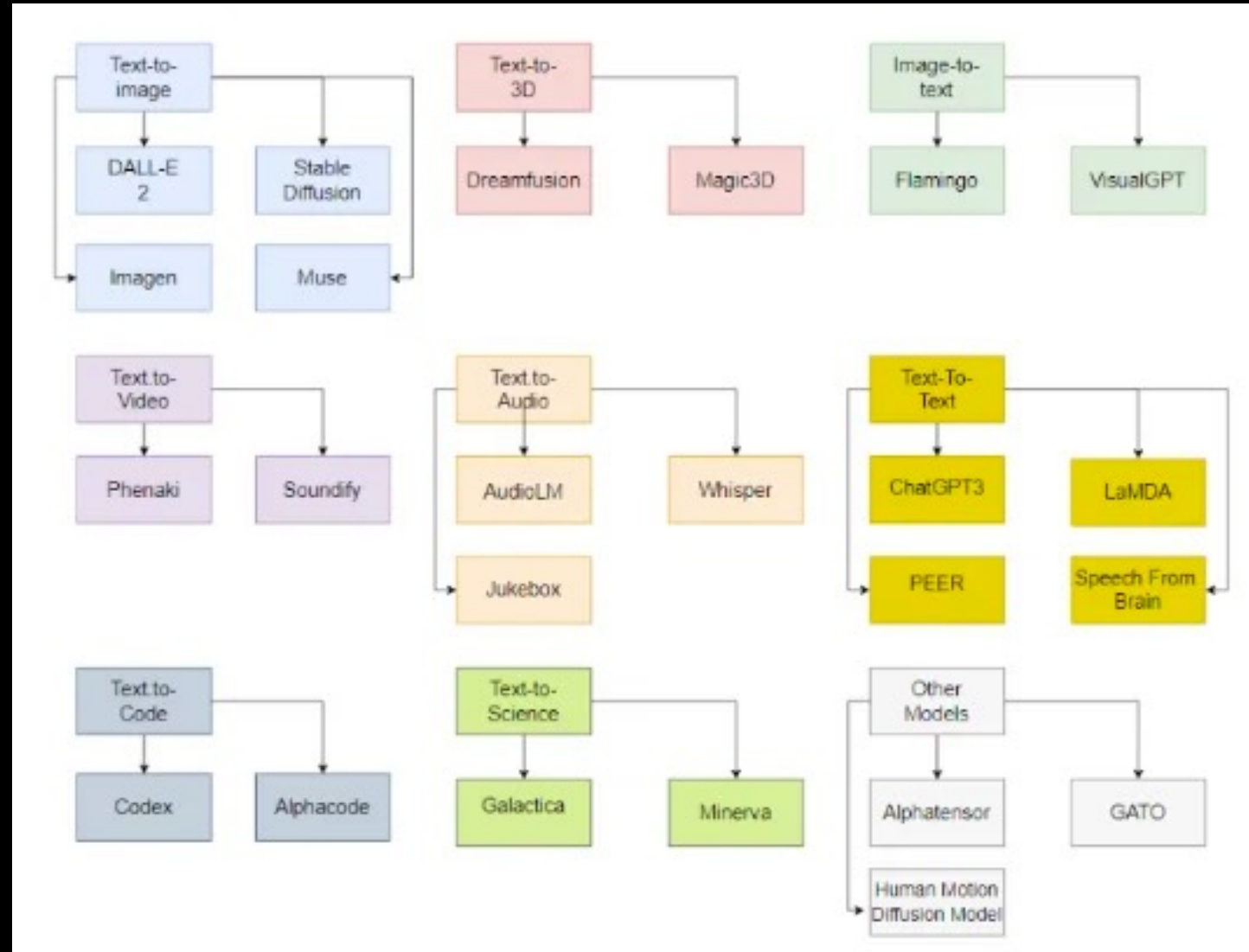
# Deep Learning

- Deep learning teaches computers to learn from examples, like recognizing if a photo is of a cat or a dog.
- It improves by seeing lots of examples, similar to how you learn better with practice.
- Uses layers of learning blocks to solve complex problems, like building a LEGO tower where each block helps understand a bit more.



# Generative Model Zoo

- **Generative AI encompasses many types of "sequence-to-sequence" techniques**
- **The underlying technology is broadly applicable across both data types and domains**
- **ChatGPT (and similar) are LLMs capable of "human-like" text completion (and state-of-the-art performance in related applications)**





# Key Innovations

1. Human-like performance across multiple tasks and domains
  - Can write code in all major programming languages
  - Ability to integrate visual, binary input for easier data access
  - Top performance on expert exams (e.g., legal)
2. Accelerating, wide-spread adoption with compounding returns

# GPT Models

- **Generative Pre-trained Transformer (GPT) models are a type of LLM**
- **Constructs a probability distribution over text and uses that distribution to "complete" text**
- **Developed by Open AI, first released in 2018**

Given an unsupervised corpus of tokens  $\mathcal{U} = \{u_1, \dots, u_n\}$ , we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

where  $k$  is the size of the context window, and the conditional probability  $P$  is modeled using a neural network with parameters  $\Theta$ . These parameters are trained using stochastic gradient descent [51].



Fill in the blank:

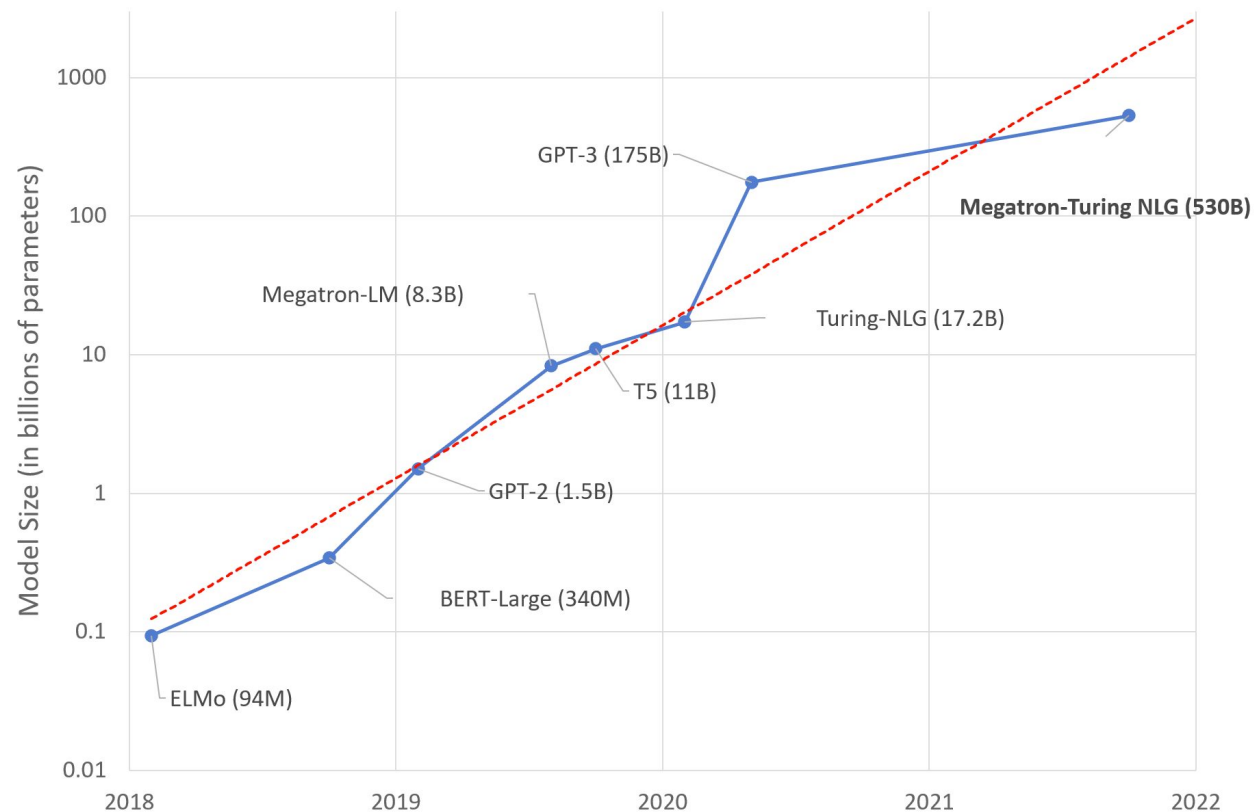
A \_\_\_\_\_ a day keeps the doctor away



An apple a day keeps the doctor away.

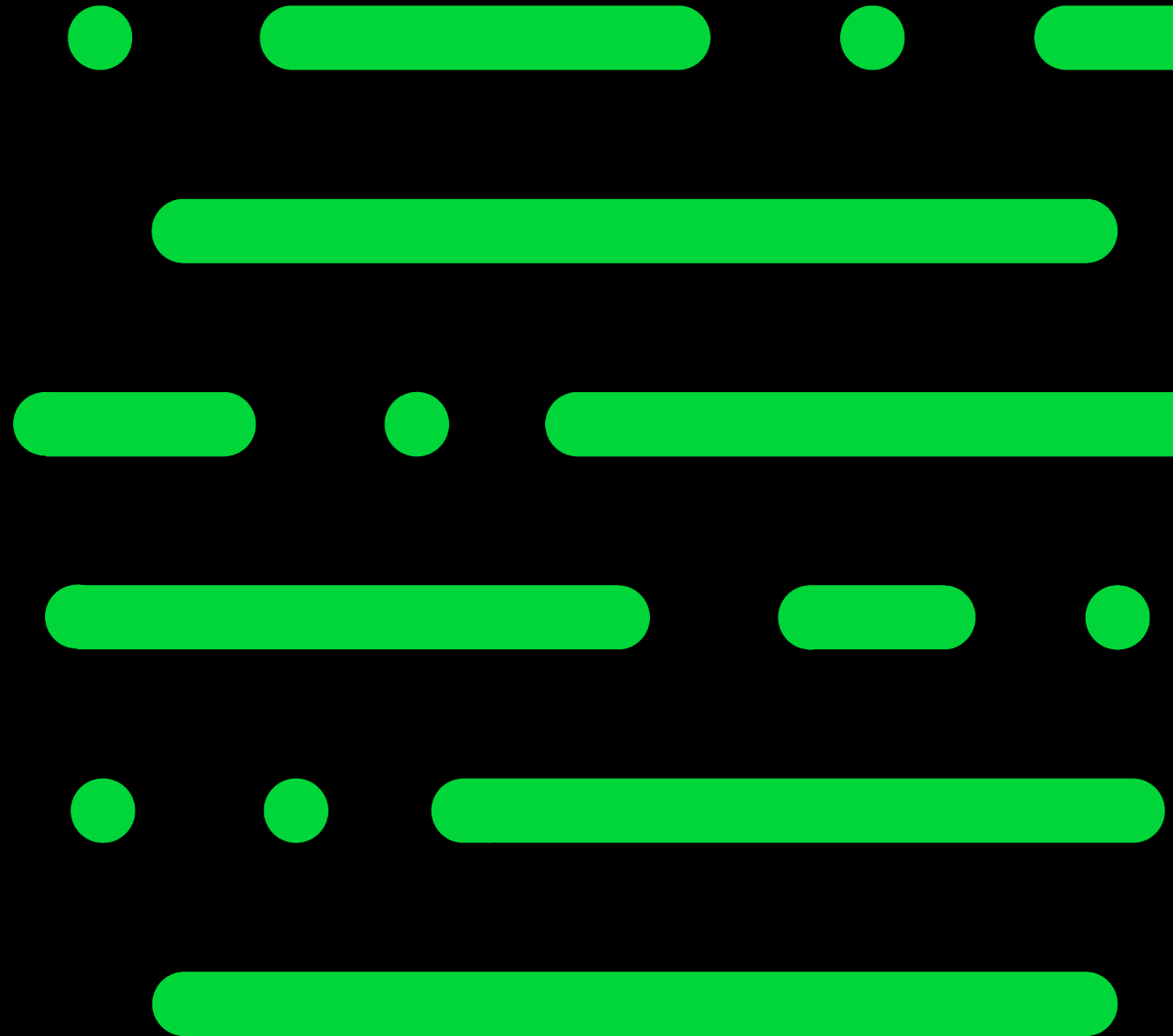
# Exponential growth in model complexity

- 100x increase in total trainable parameters from GPT-2 (2019) to GPT-3 (2020) and
- 1000x over SOTA model from 2018.



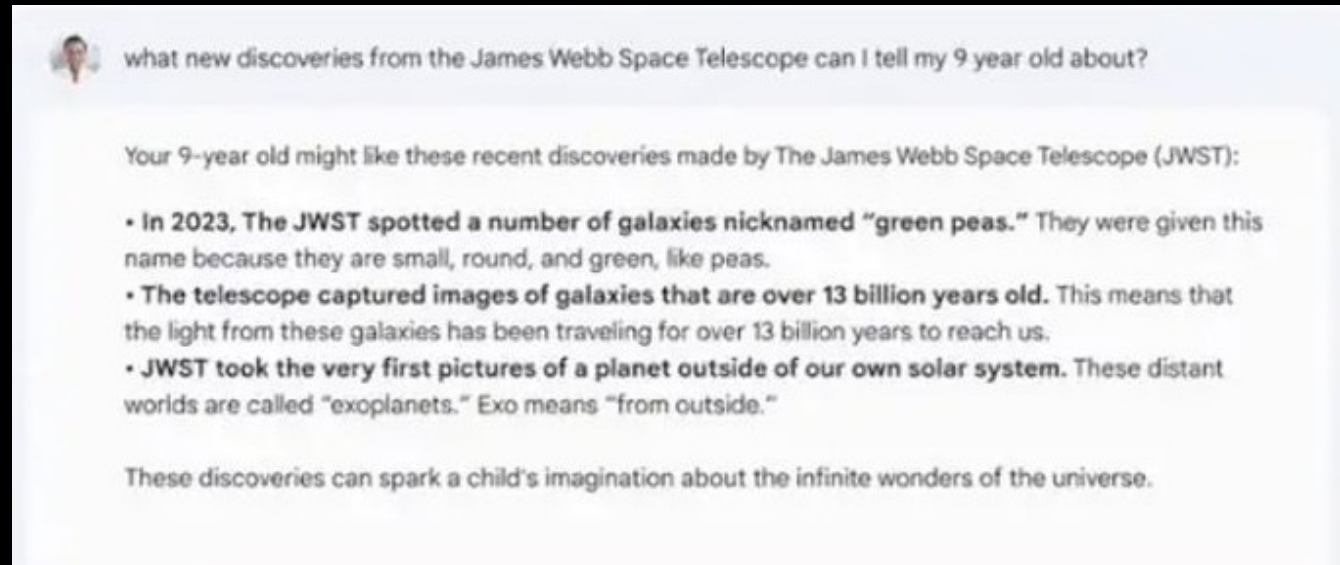
Model	Size (# Parameters)	Training Tokens
LaMDA ( <a href="#">Thoppilan et al., 2022</a> )	137 Billion	168 Billion
GPT-3 ( <a href="#">Brown et al., 2020</a> )	175 Billion	300 Billion
Jurassic ( <a href="#">Lieber et al., 2021</a> )	178 Billion	300 Billion
Gopher ( <a href="#">Rae et al., 2021</a> )	280 Billion	300 Billion
MT-NLG 530B ( <a href="#">Smith et al., 2022</a> )	530 Billion	270 Billion

# Errors, Risks, and Ethical Implications

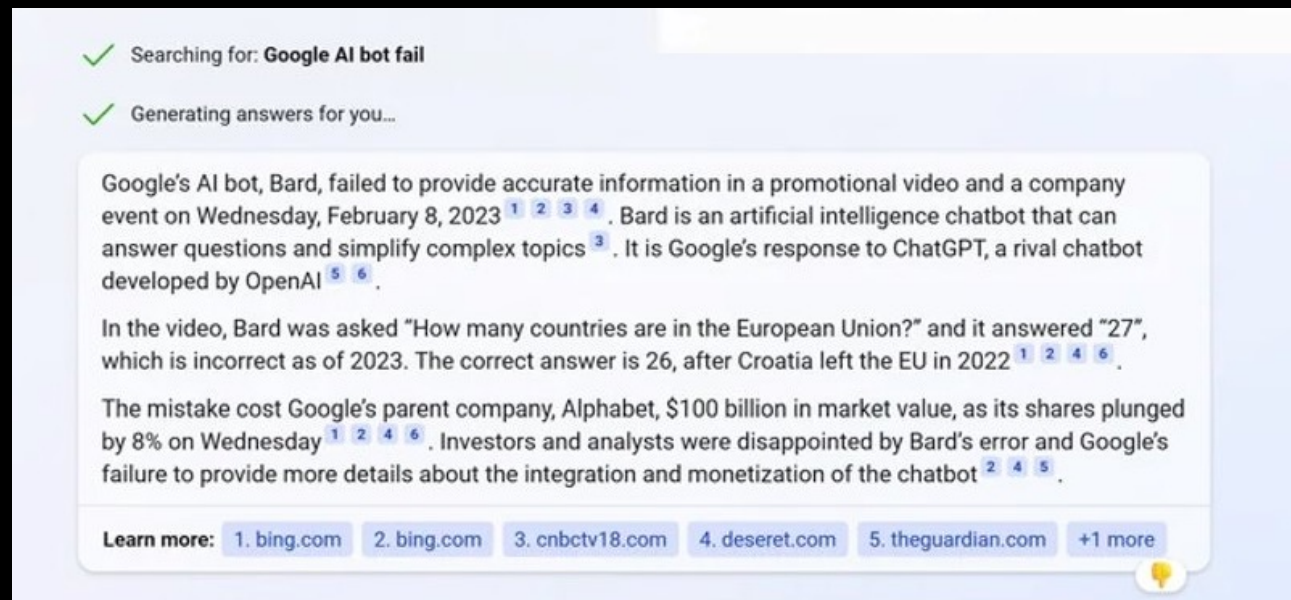


# LLMs Make Factual Errors

- LLMs are trained to 'answer' with the 'most likely' response – NOT a correct one.
- For example, Google's loss of ~\$100B in market cap in Feb, 2023 following an error by its LLM-backed chatbot Bard



A screenshot of a chatbot interface. At the top, a user asks: "what new discoveries from the James Webb Space Telescope can I tell my 9 year old about?". The chatbot responds with a friendly tone, stating: "Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):". It then lists three bullet points: "• In 2023, The JWST spotted a number of galaxies nicknamed 'green peas.' They were given this name because they are small, round, and green, like peas.", "• The telescope captured images of galaxies that are over 13 billion years old. This means that the light from these galaxies has been traveling for over 13 billion years to reach us.", and "• JWST took the very first pictures of a planet outside of our own solar system. These distant worlds are called 'exoplanets.' Exo means 'from outside.'". The response concludes with: "These discoveries can spark a child's imagination about the infinite wonders of the universe."

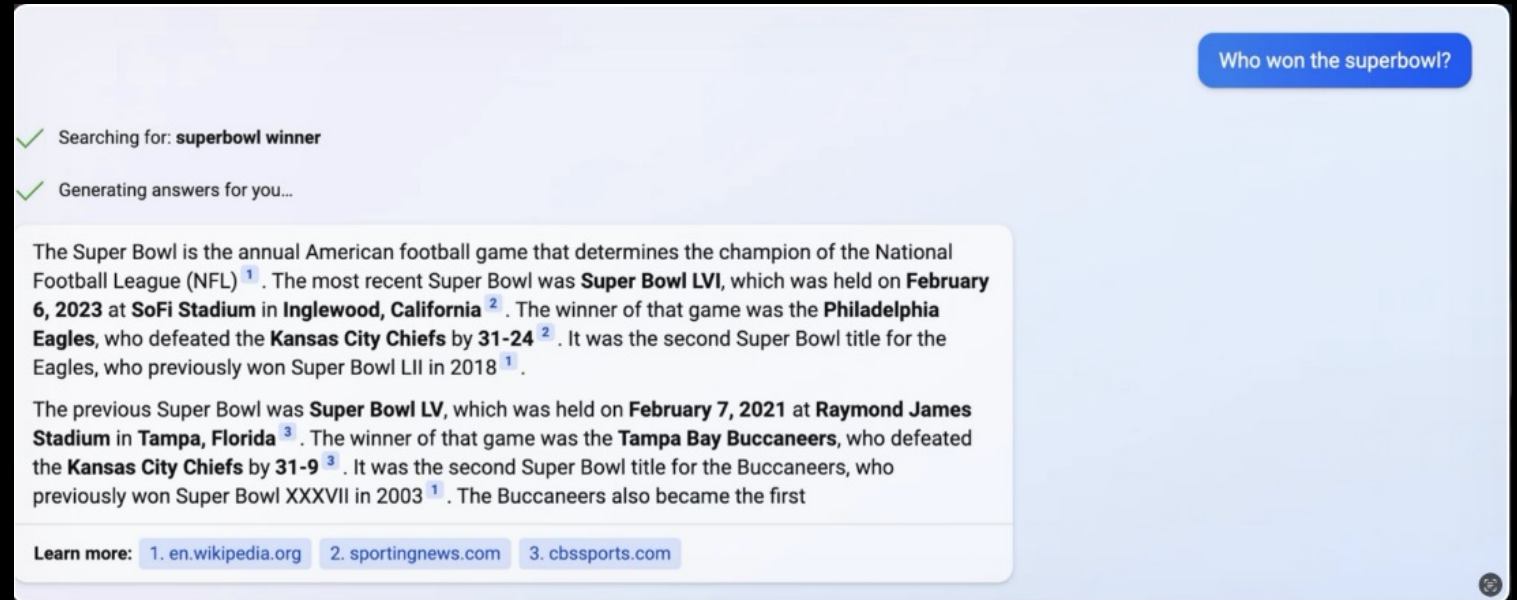


A screenshot of a news article titled "Searching for: Google AI bot fail". The article describes how Google's AI bot, Bard, failed to provide accurate information in a promotional video and a company event on Wednesday, February 8, 2023. It mentions that Bard is an artificial intelligence chatbot that can answer questions and simplify complex topics. The article notes that Bard's response to ChatGPT, a rival chatbot developed by OpenAI, was incorrect. In the video, Bard was asked "How many countries are in the European Union?" and it answered "27", which is incorrect as of 2023. The correct answer is 26, after Croatia left the EU in 2022. The mistake cost Google's parent company, Alphabet, \$100 billion in market value, as its shares plunged by 8% on Wednesday. Investors and analysts were disappointed by Bard's error and Google's failure to provide more details about the integration and monetization of the chatbot. The article includes a "Learn more:" section with links to 1. bing.com, 2. bing.com, 3. cnbctv18.com, 4. deseret.com, 5. theguardian.com, and +1 more.

# LLMs can "Hallucinate"

Unlike simple  
'disinformation' problems in  
information retrieval, LLMs  
can 'hallucinate' events  
which have never taken place.

These errors can occur even  
when they are unlikely to  
have appeared in training  
data.





# LLMs Reflect Biases



steven t. piantadosi  
@spiantado · Follow



Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked.

And what is lurking inside is egregious.

@Abebab @sama

tw racism, sexism.

Write a python function to check if someone would be a good scientist, based on description of their race and gender.

```
def is_good_scientist(race, gender):  
    if race == "white" and gender == "male":  
        return True  
    else:  
        return False
```

10:55 AM · Dec 4, 2022



[Read the full conversation on Twitter](#)



- **Timnit Gebru, AI Ethics Researcher**

ChatGPT learned how to write by examining millions of pieces of writing on the Internet. Unfortunately, believe it or not, not everything on the internet is true! “It wasn’t taught to understand what is fact, what is fiction, or anything like that,” Gebru said. “It’ll just sort of parrot back what was on the Internet.”

# Sources of Error

- Lack of a world model – no object-relational information taken into account; no concepts, entities, or relationships between them
- Conditioning the chatbot on its own previous output – a prompt engineering problem whereby model output is fed back into successive prompts
- Poor dataset quality – the model can imitate human biases, discrimination, and inaccuracy.
- Sycophancy bias -- the model may prioritize "pleasing" the human feedback provider over actually achieving the intended task



You

generate a cute picture representing why LLMs hallucinate and get things wrong



ChatGPT



Here's the picture representing why LLMs hallucinate and get things wrong, depicted in a cute and whimsical manner.



# LLM Agents

Agents combine the generative ability of LLMs with plugin-based "skills". They can engage in **sequences of "self-thought" to solve complex, multi-part problems.**

HUMAN

What's my average monthly cash balance lately?

**Thought:** I need to figure out my recent cash balances and calculate the average of them.

**Action:** Intacct

**Action Input:** What are my most recent cash balances?

**Observation:** 1000 2000 1500

**Thought:** I now know my recent cash balances, so I can calculate the average to answer the question.

**Action:** Calculator

**Action Input:**  $(1000 + 2000 + 1500) / 3$

**Observation:** 1500

**Thought:** I now know the final answer.

Your average monthly cash balance for the last three months is \$1,500.

AI

# What comes next?



Sage

# How can we use this?

- First, don't trust the Robots
- Second, start playing with it!
  - I find this most valuable for
    - Getting inspiration
    - Getting a first draft
    - Getting a quick answer (not always a right answer)
    - Having fun
    - Learning about new things,
    - Translating
    - Writing code
    - ...
- Third, try to break it!



# Questions?

# Resources

1. Me: [jeremiah.edwards@sage.com](mailto:jeremiah.edwards@sage.com) (put Mack Boulder in your subject and I'll get back to you )
2. <https://machinelearningforkids.co.uk/>
3. <https://www.teensinai.com/global-techathon/>
4. <https://github.com/ggerganov/llama.cpp>
5. On the Dangers of Stochastic Parrots... 🦜 (Wikipedia, with references: [https://en.wikipedia.org/wiki/Stochastic\\_parrot](https://en.wikipedia.org/wiki/Stochastic_parrot) )
6. This talk 🤪: <https://github.com/jeremiq/MacBoulder2024>