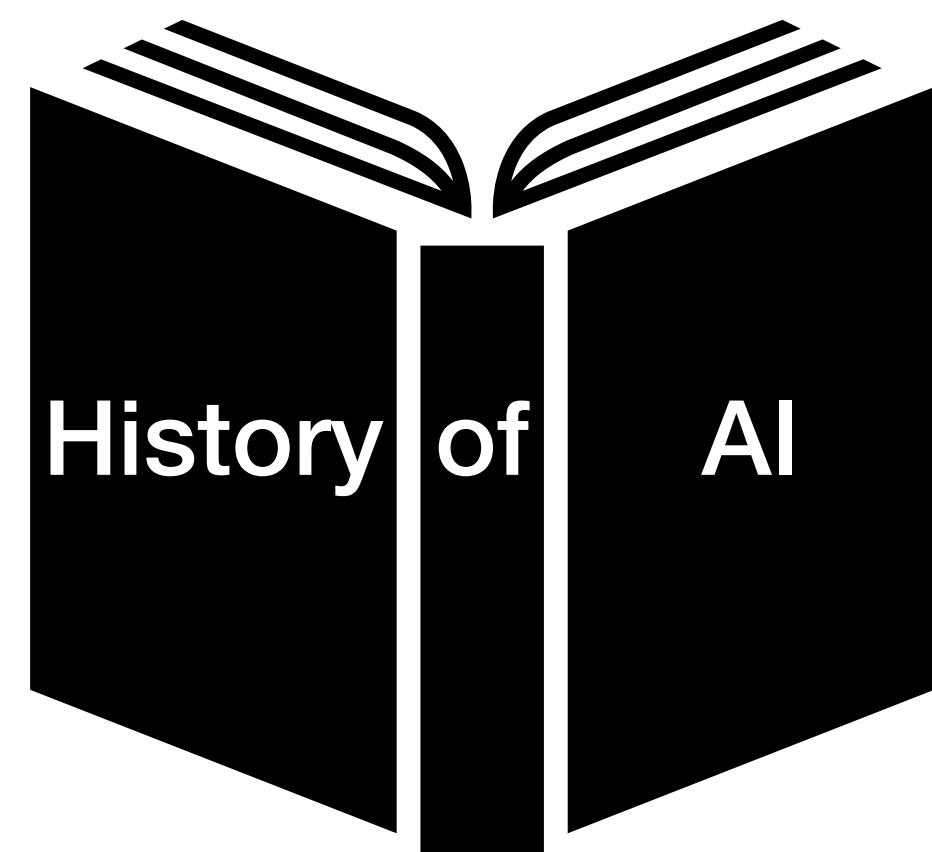


Operational view

How AI systems are made end-to-end

- From organization & data collection to deployment & operation
- Focus not on ChatGPT, but the tech that powers it
- Goal: Identify operational bottlenecks and how to resolve them



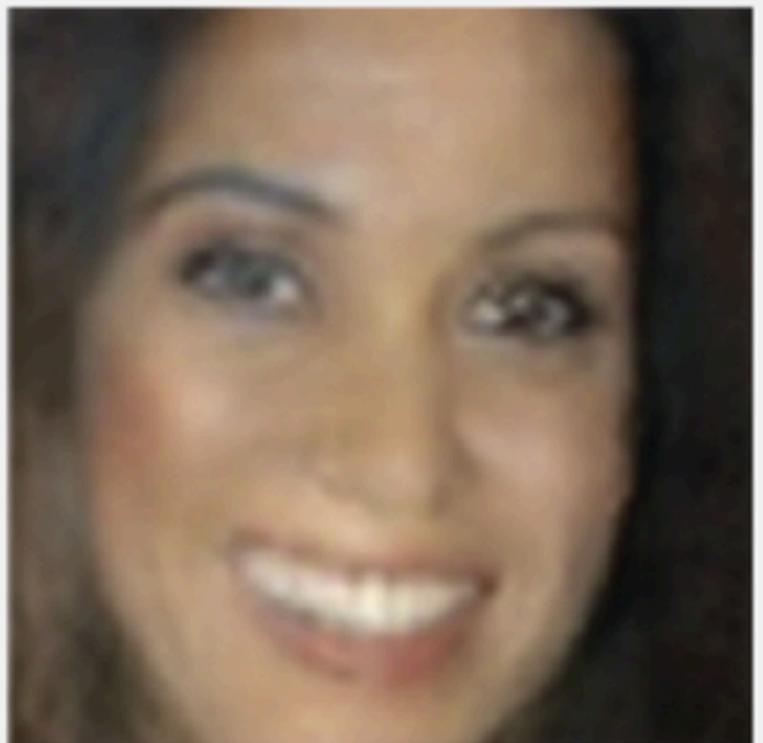
Generative AI

2014



Goodfellow et al. (2014) – Generative Adversarial Networks

2015



Radford, Metz, and Chintala (2015) – Unsupervised Representation Learning with Deep Convolutional GANs

2016



Liu and Tuzel (2016) – Coupled GANs

2022

Image generated with the prompt:
“A Pomeranian is sitting on the King’s
throne wearing a crown. Two tiger
soldiers are standing next to the throne.”



Saharia et al. (2022) – Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Google's Imagen)

What happened here?

What does it mean to “train” an AI model?

- AI is trained as optimizers
- Minimize prediction error on training data

Good computer!

Bad computer!



Key bottleneck

- Human needs to give feedback! Supervised data is expensive
 - e.g., radiologist labeling each X-ray
- Main idea behind language models: construct a fictitious training objective that removes the need for labeling
- Predict next word given previous words!

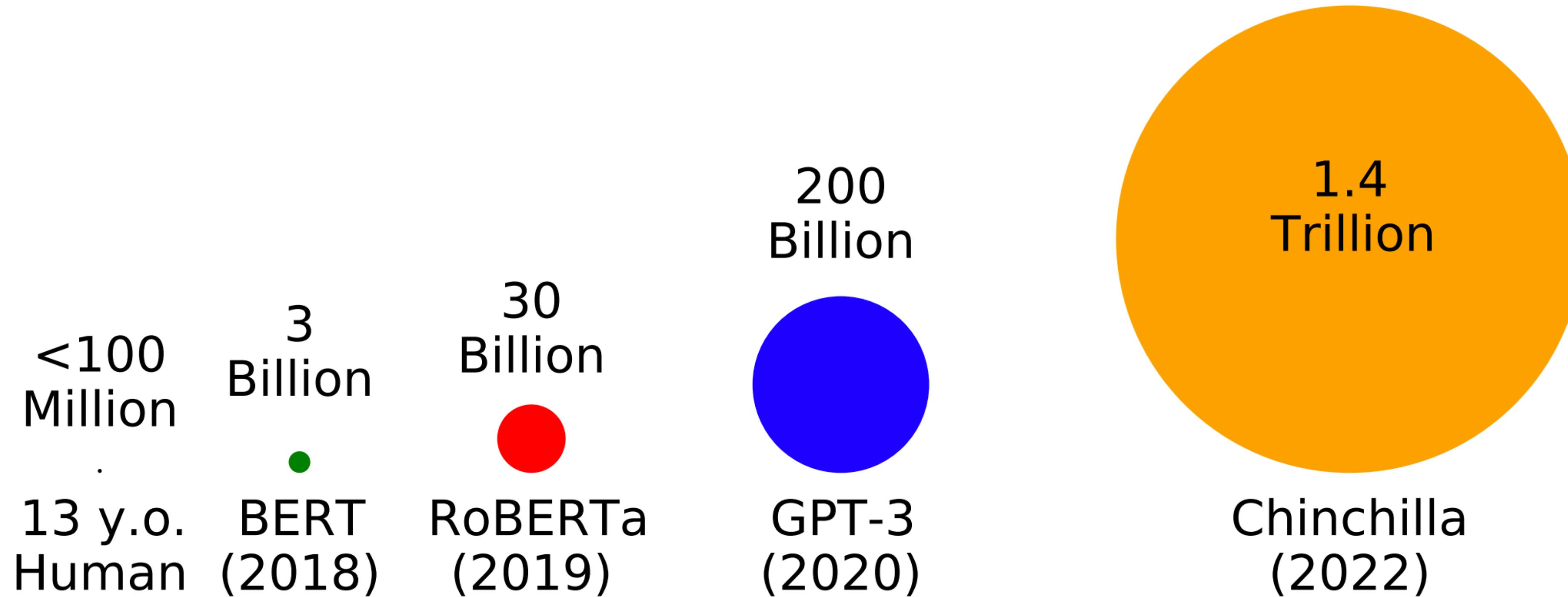
Language modeling

The patient presents with severe chest pain radiating to the left arm, sweating, and shortness of breath. These symptoms are classic indicators of _____.

Possible answers:
myocardial infarction, angina, pulmonary embolism

Scaling data

- Data size exploded in the past 5 years



LLM is still of limited use

Q: What happens if you fire a cannonball directly at a pumpkin at high speeds?

GPT-3 Answer:

The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.



OpenAI Ouyang et al. (2021)

Train model to follow instructions

Instruction finetuning

Please answer the following question.

What is the boiling point of Nitrogen?

-320.4F

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.

The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$.

Language
model

Multi-task instruction finetuning (1.8K tasks)

Before instruction finetuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

PaLM 540B output

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✖ (doesn't answer question)

After instruction finetuning

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Flan-PaLM 540B output

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). 

Going further: human preferences

- Get human input on which outputs are preferred

Summarize text

SAN FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects.

An earthquake hit San Francisco. There was minor property damage, but no injuries.

Good computer. High score.



The Bay Area has good weather but is prone to earthquakes and wildfires.

Bad computer. Low score.

Back to pumpkins

Q: What happens if you fire a cannonball directly at a pumpkin at high speeds?

ChatGPT Answer:

There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

ChatGPT

I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?

[[Microsoft Bing](#)]

(Also see OpenAI's ChatGPT,
Google's Bard, Anthropic's Claude)



Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: **Wild Mushroom Tartlets with Onion Sauce** ¹ or **Vegan Popcorn Tofu Nuggets** ²
- Main: **Vegan Butternut Squash Mac** ² or **Vegetarian Three-Bean Chili** ³
- Dessert: **Chocolate Lava Cake** ⁴ or **Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries** ⁴

I hope this helps you plan your party. Enjoy! 😊

Learn more: [1. booths.co.uk](#) [+10 more](#)

My take on current AI capabilities

- Compression engine (like a zip)
 - Entirety of humanity's digitizable knowledge is distilled down to 1 trillion parameters
- LLMs provides a **new UX to computing**
 - Retrieve information from the zip file in natural language
 - E.g., coding as a translation task from English => programming language
 - Natural language interface a disruptive force

What's next? Scaling, scaling, scaling

Key ingredients: compute + data

- Compute: GPUs + electricity

OCTOBER 2, 2024 | 5 MIN READ

Power-Thirsty AI Turns to Mothballed Nuclear Plants. Is That Safe?

As Microsoft strikes a deal to restart a reactor at Three Mile Island to power AI, nuclear specialists weigh in on the unprecedented process

BY MICHAEL GRESHKO & NATURE MAGAZINE



What's next? Scaling, scaling, scaling

Key ingredients: compute + data

- Compute: GPUs + electricity

OCTOBER 2, 2024 | 5 MIN READ

Power-Thirsty AI Turns to Mothballed Nuclear Plants. Is That Safe?

As Microsoft strikes a deal to restart a reactor at Three Mile Island to power AI, nuclear specialists weigh in on the unprecedented process

BY MICHAEL GRESHKO & NATURE MAGAZINE



- Data: quality matters!
 - Frontier: proprietary data highlighting specialized skills

Question 1: All you need is scaling?

- Easily available data on the internet has been exhausted

How Tech Giants Cut Corners to Harvest Data for A.I.

OpenAI, Google and Meta ignored corporate policies, altered their own rules and discussed skirting copyright law as they sought online information to train their newest artificial intelligence systems.

The New York Times

“In late 2021, [OpenAI](#) faced a supply problem. The artificial intelligence lab had exhausted every reservoir of reputable English-language text on the internet as it developed its latest A.I. system.

Question 1: All you need is scaling?

- Easily available data on the internet has been exhausted
- Where do we go from here?
 - Proprietary data sources, e.g., Amazon, Chase
 - Synthetic data
- Resource constraints are often binding!

Question 2: Language for datasets?

- We still don't have a language for describing datasets
- Essential to characterize robustness characteristics
- Even more challenging in language modeling: what even is X or Y?

Question 3: Eval, evals, evals

- Commercial interests block rigorous empirical science
- How do we know the limits of the current frontier technologies?
- How do we evaluate the robustness of AI systems?
- Chatbots are not the most exciting application of this technology
- Thumbs up or down very far from real decision-making outcomes

Question 4: Interactive decision-making

- Real-world decision-making needs to deal with a continual lack of data
 - New patient, customer, or user
 - New newspaper article or Youtube video
- Intelligent agents must know how to explore and learn across environments
- Active data collection can radically boost statistical efficiency

Question 5: Quantifying uncertainty

- AI models are always on (“omni-present”)
- They inevitably encounter inputs rarely seen during training, and they extrapolate in unexpected ways
- How do we know when this is the case?
 - A critical first step to bounding tail-risk and improving robustness

Other questions

- ML system interacts with (strategic) agents over time. How to model this interaction/dynamics?
- Operational constraints (safety, reliability etc)
- Collected data on decisions are observational
 - Often based on human agents' decisions, which may depend on unrecorded variables
 - For sequential decisions, observed data often does not cover entire (action seq, state seq) space. So not really “big data”...

My take on future AI capabilities

An “AGI-realist” view

- AI will take over tasks whose correctness can be ascertained
 - Key requirement for AI performance: good scoring system
 - Immediate applications: software engineering and data analytics
- Name of the game for the next 5 years

scoring system for business problems

My take on future AI capabilities

An “AGI-realist” view

- AI will take over tasks whose correctness can be ascertained
 - Key requirement for AI performance: good scoring system
 - Immediate applications: software engineering and data analytics
- Name of the game for the next 5 years
 - scoring system for business problems
- **Unsolved: Tail inputs!**