CS 480

Introduction to Artificial Intelligence

April 23, 2024

Announcements / Reminders

- Please follow the Week 13 To Do List instructions (if you haven't already)
- Finish your last Written Assignment (if you haven't already)

- FINAL EXAM is THIS THURSDAY (04/25/2024)
  - IGNORE Registrar's FINAL EXAM date
  - Last week of classes! NOT finals week
  - Section 02: contact Mr. Charles Scott (scott@iit.edu) to make arrangements

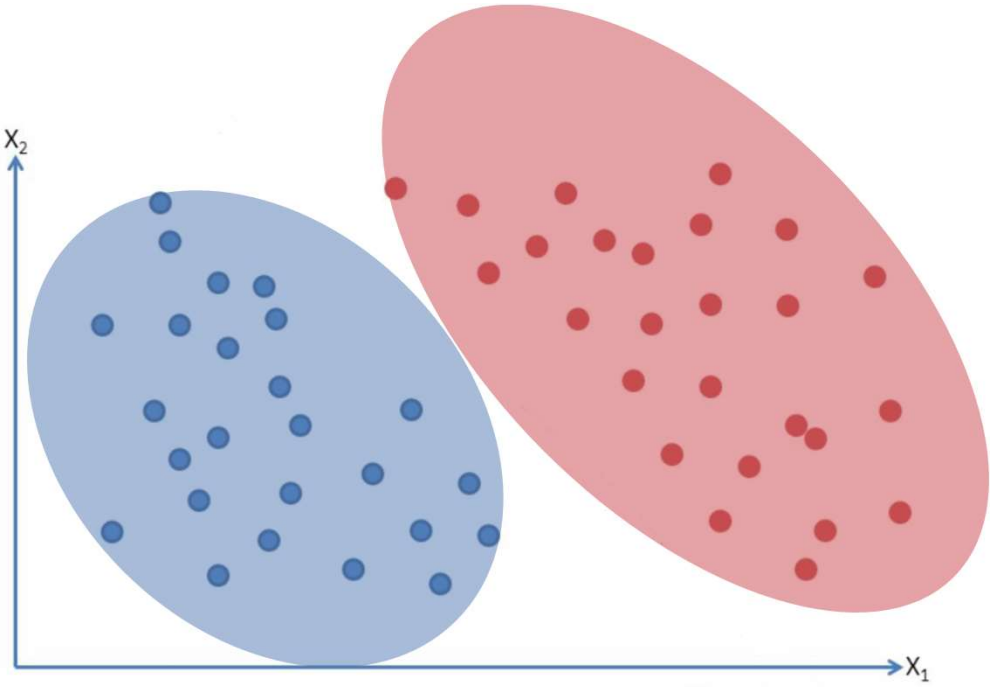
Plan for Today

- Generative AI basics
- AI: Recent Developments
- AI: Concerns and Future

Generative AI Models
[NOT ON THE FINAL EXAM]

Generative vs Discriminative Models

Generative

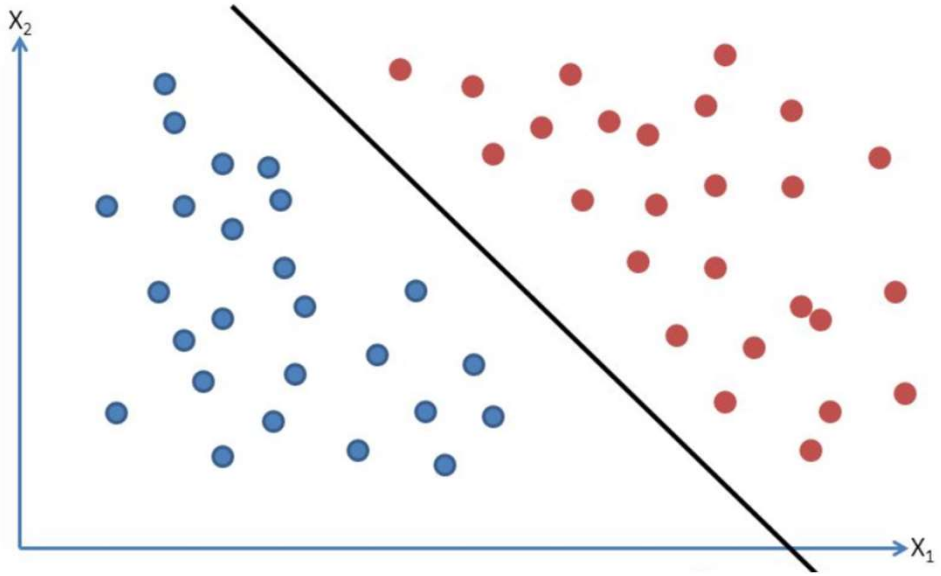


Generative model models actual distributions for EACH CLASS / LABEL / TAG

to

make a P(class | sample) **prediction**

Discriminative



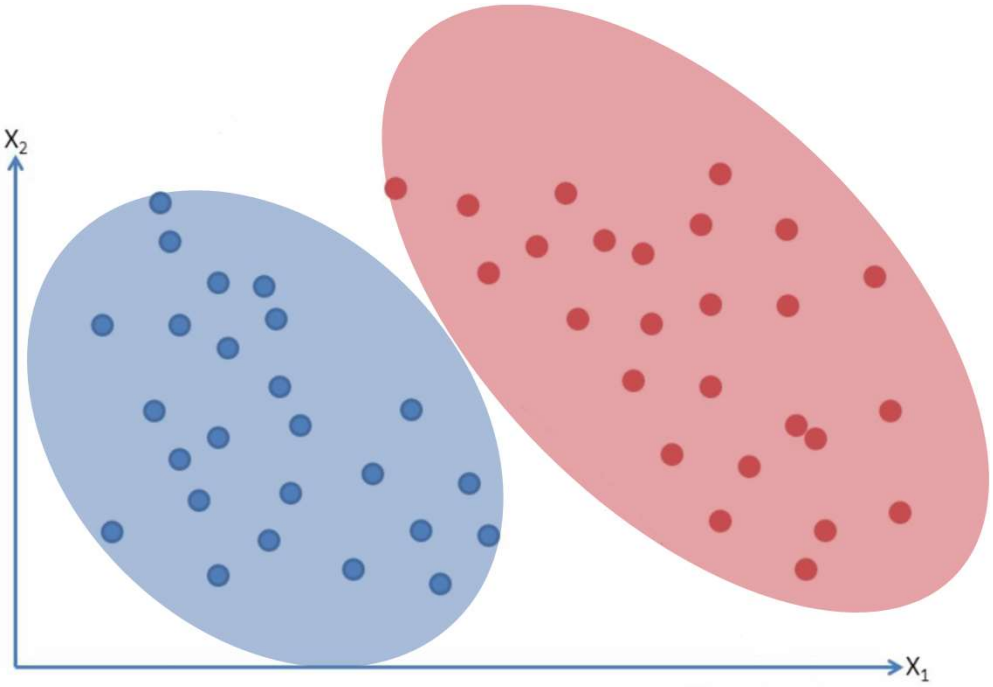
Discriminative model models the decision boundary between CLASSES / LABELS / TAGS

to

make a P(class | sample) **prediction**

Generative vs Discriminative Models

Generative

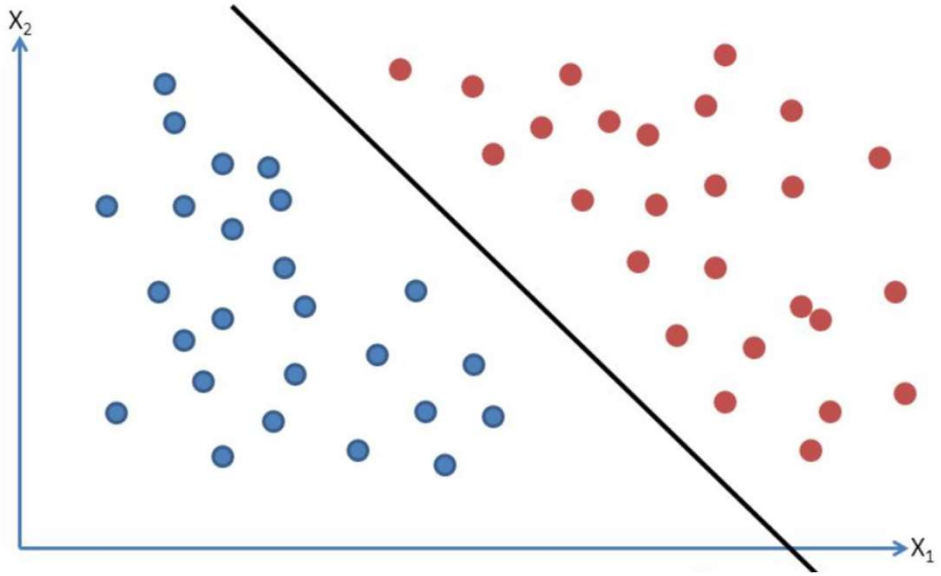


Generative model uses training data to learn P(sample, class) **joint probabilities**

and then

uses Bayes Theorem to get the P(class | sample) **prediction**

Discriminative



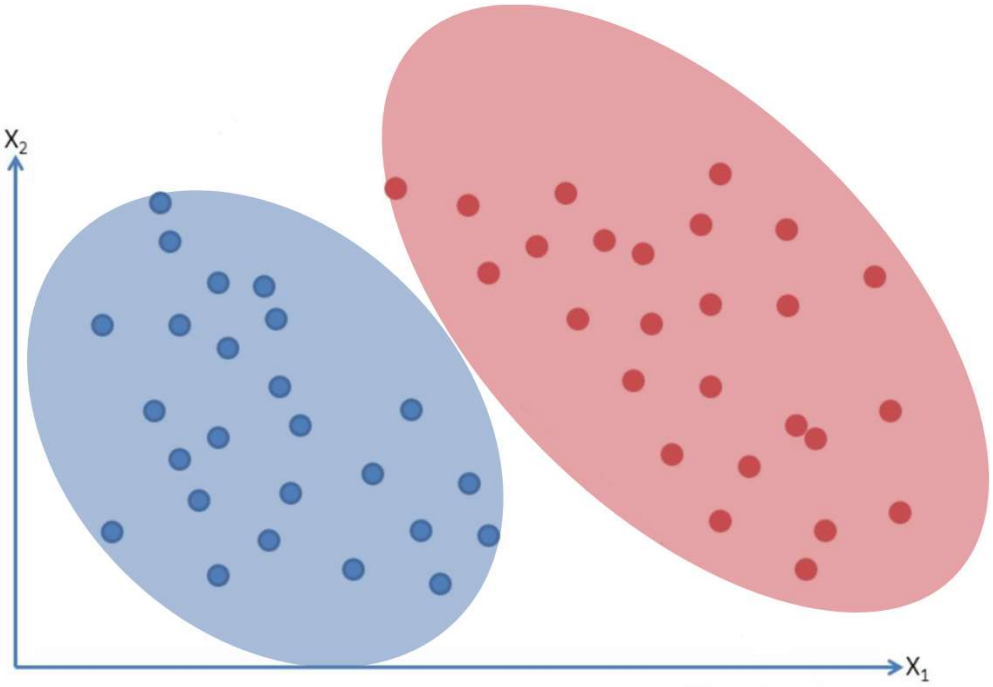
Discriminative model uses training data to learn P(class | sample) **conditional probability**

and then

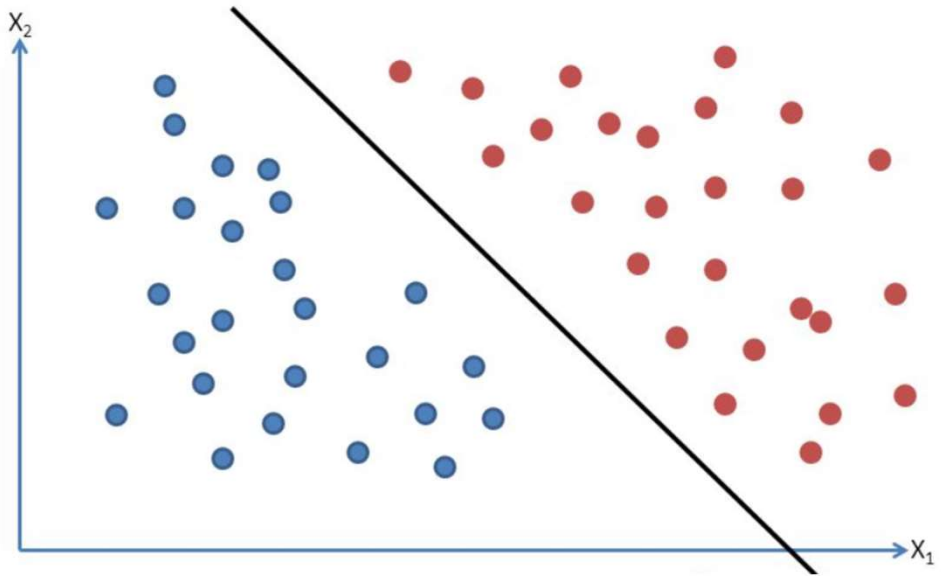
uses it to make a prediction

Generative vs Discriminative Classifier

Generative



Discriminative



Generative classifiers:

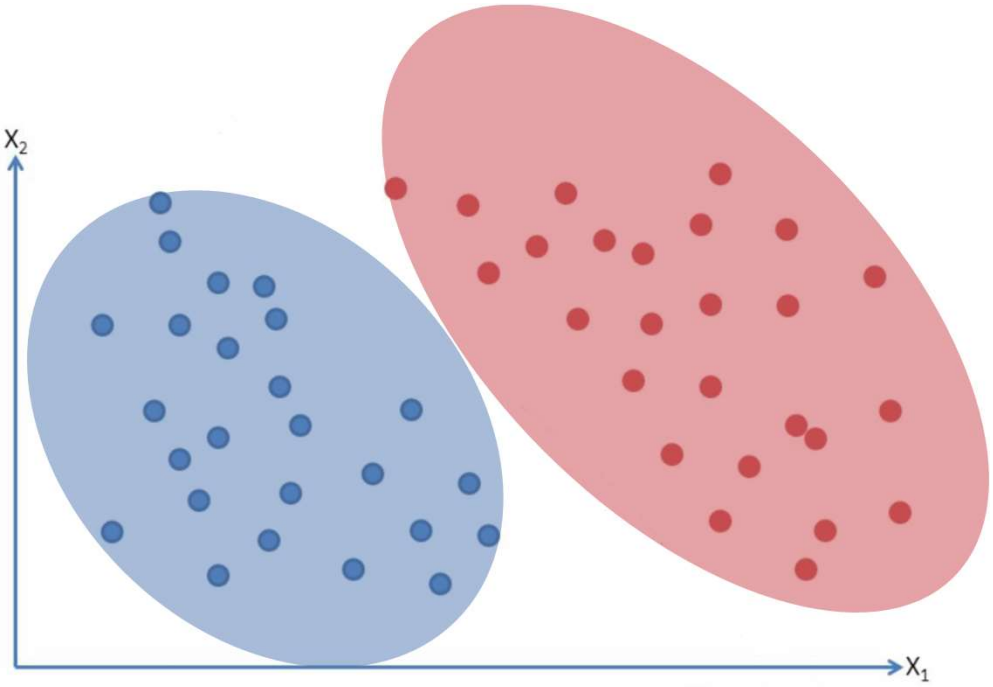
- **Assume some form of** P(class), P(sample | class)
- **Estimate** P(class), P(sample | class) **using training data**
- **Use Bayes Theorem to calculate** P(class | sample)

Discriminative classifiers:

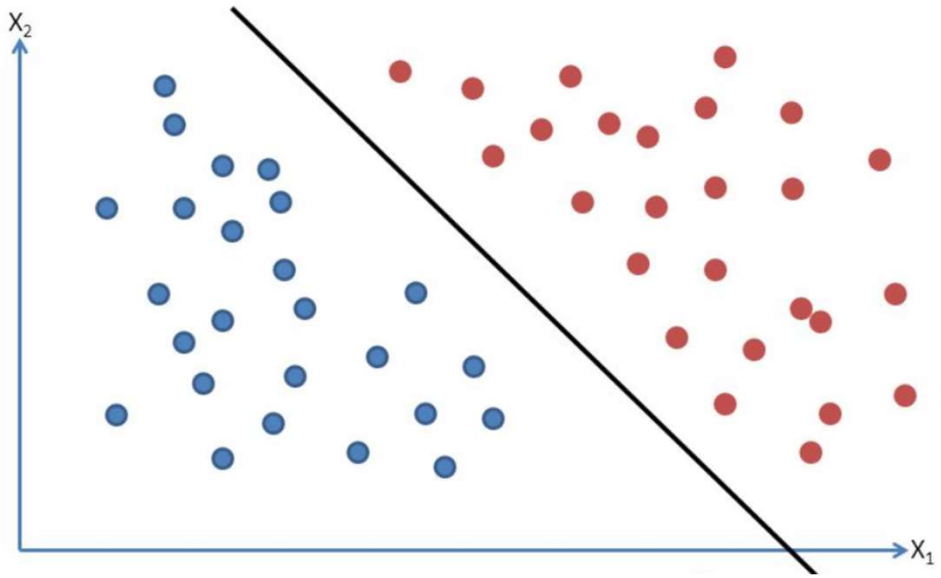
- **Assume some form of** P(class | sample)
- **Estimate** P(class | sample) **using training data**

Generative vs Discriminative Classifier

Generative



Discriminative



Generative classifiers:

- Naive Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Discriminative classifiers:

- Logistic regression
- Support Vector Machines
- Traditional neural networks
- k-Nearest Neighbors
- Conditional Random Fields (CRF)s

Generative AI Model: the Idea

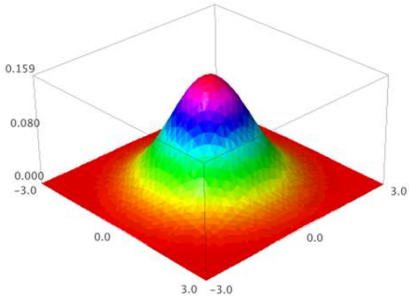
Reality: some sample space (possibly infinite)

Sample

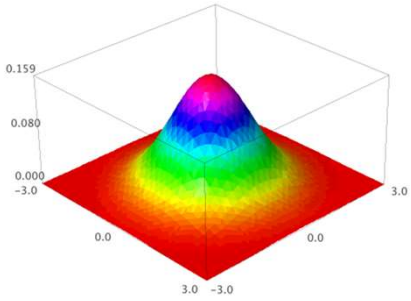
Dataset: Subset of "Reality" (finite)

New sample

Model



Estimate



Generate / sample

Probability distribution P_{real}

Probability distribution P_{model}

Generative AI Model: the Idea

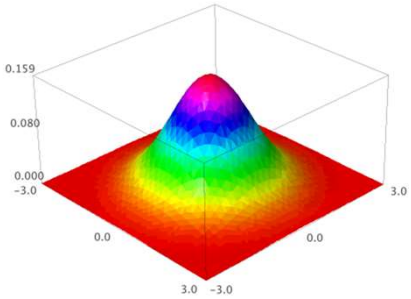
Reality:
All Possible cat faces

Sample

Dataset:

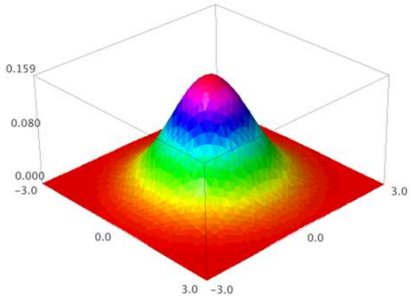
New sample:

Model



Probability distribution P_{real}

Estimate

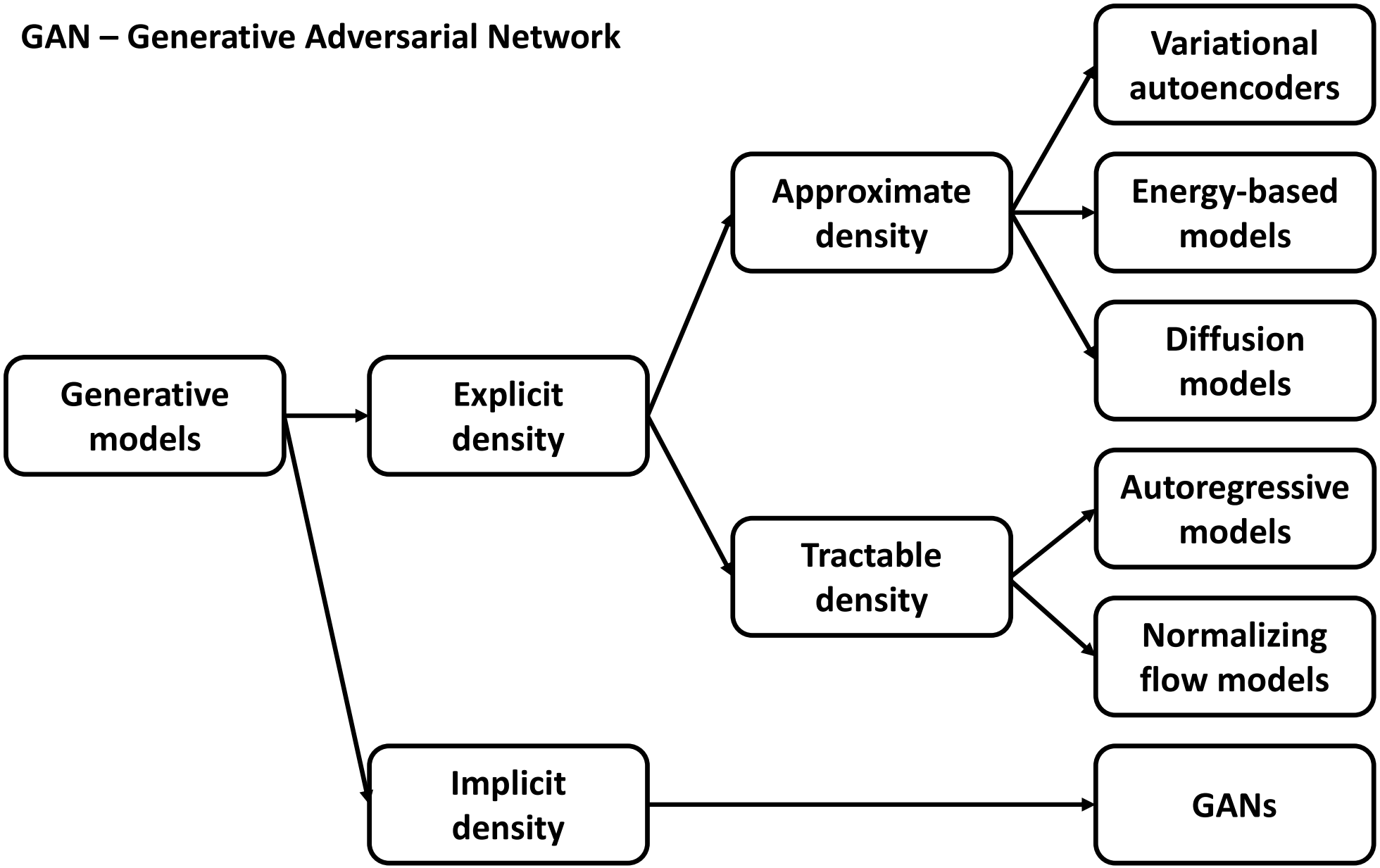


Probability distribution P_{model}

Generate / sample

Taxonomy of Generative AI Models

GAN – Generative Adversarial Network



Explicit Density

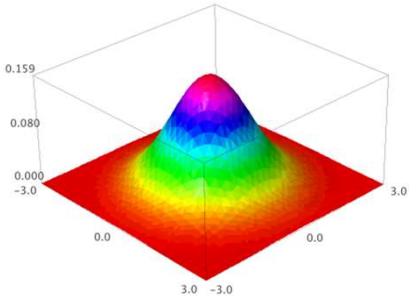
Reality:
All Possible cat faces

Sample

Dataset:

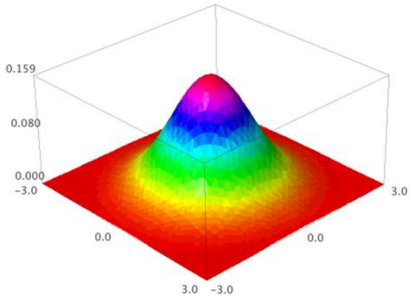
New sample:

Model



Probability distribution P_{real}

Estimate



Probability distribution P_{model}

Sample

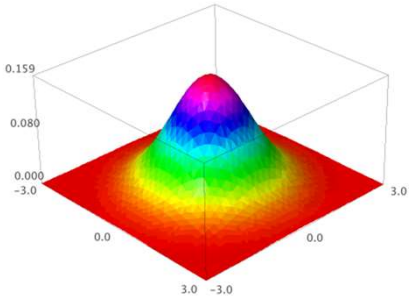
Implicit Density

Reality:
All Possible cat faces

Sample

Dataset:

New sample:



Probability distribution P_{real}

Model

Stochastic generation process

Generate

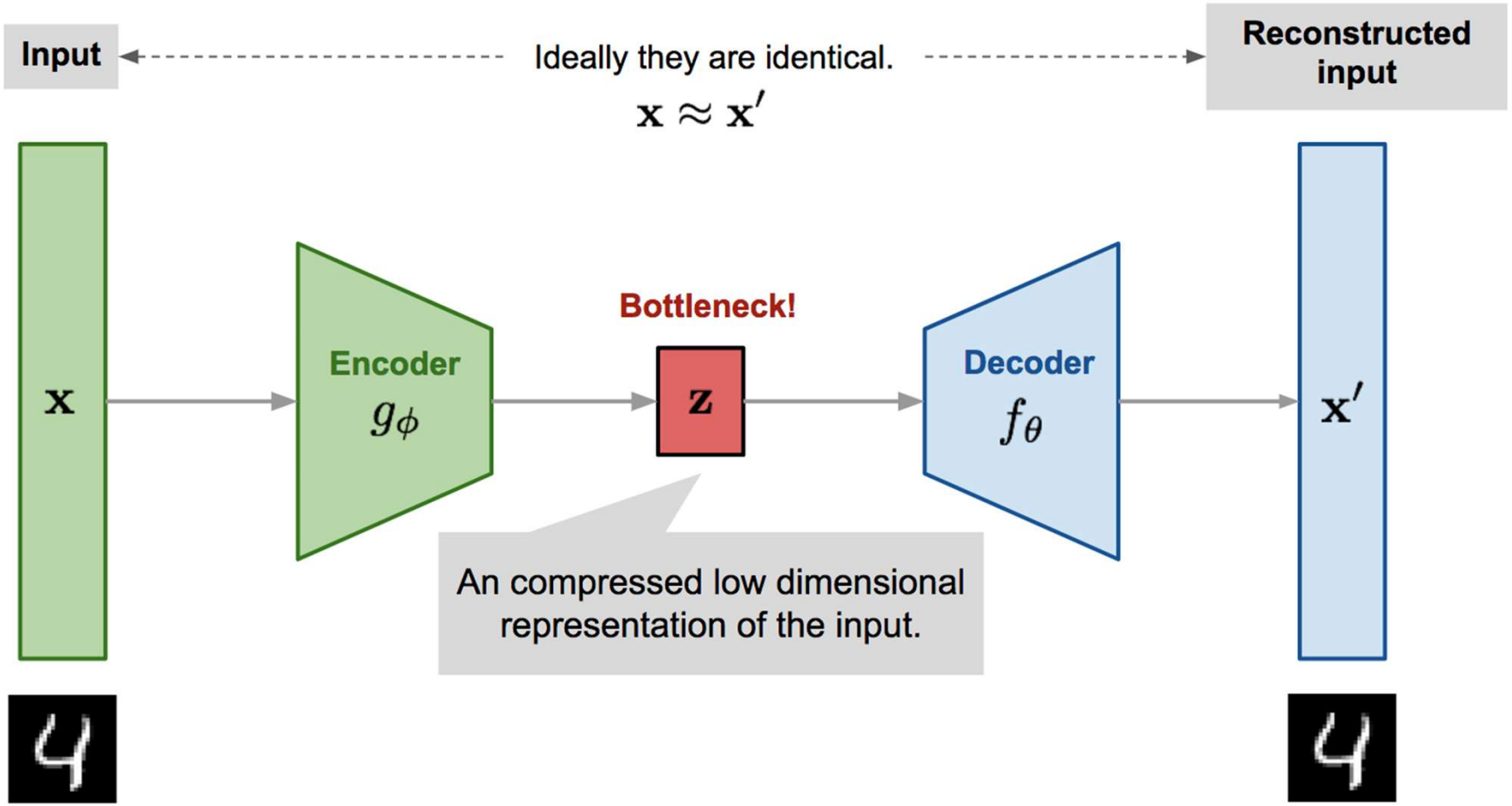
Tractable vs. Approximate Density

Tractable density models place constraints on the model architecture so that the density function has a form that makes it easy to calculate.

Approximate density models use variety of techniques to approximate the density function:

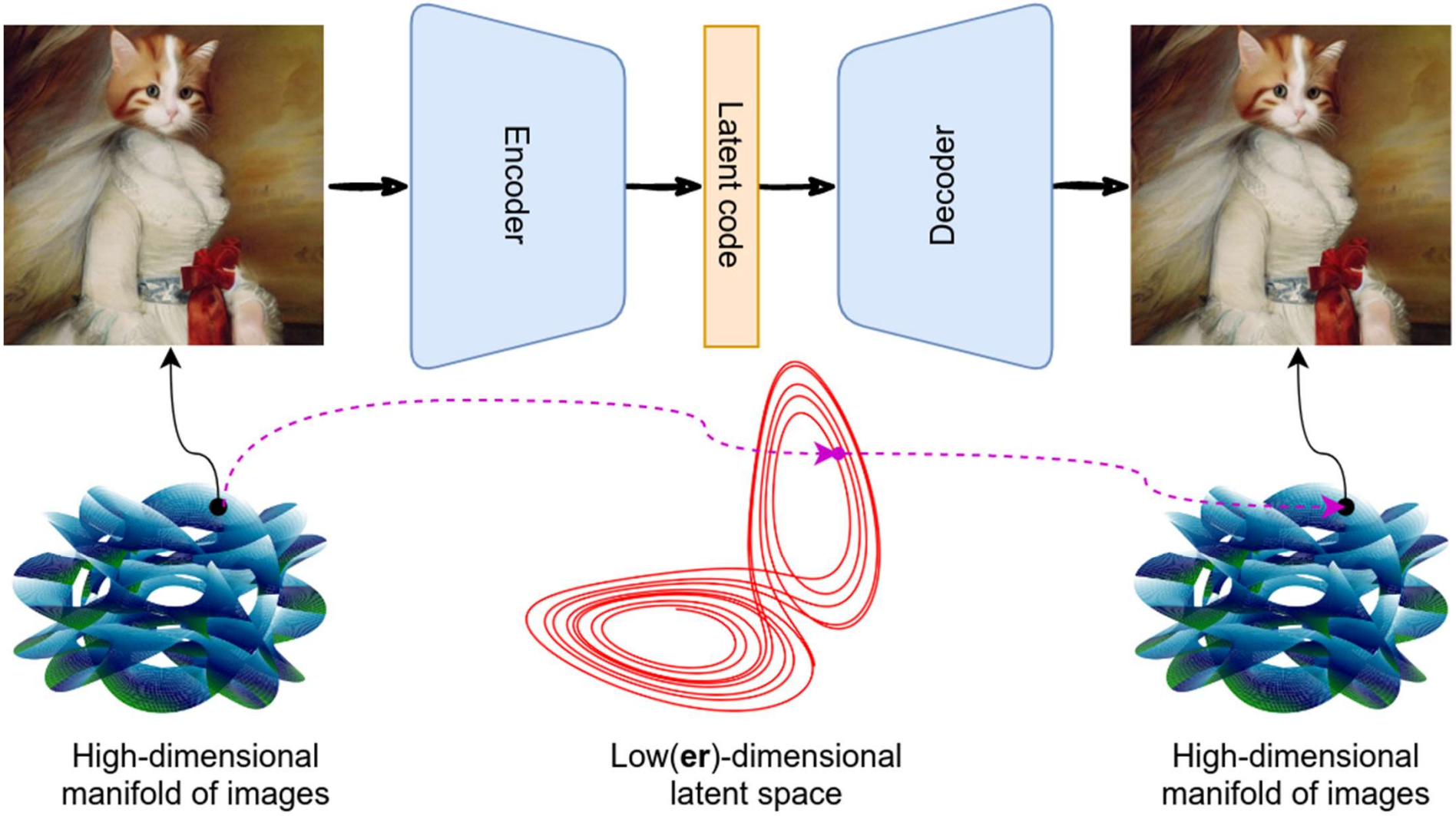
- **latent vectors**
- **denoising**

Autoencoder Model



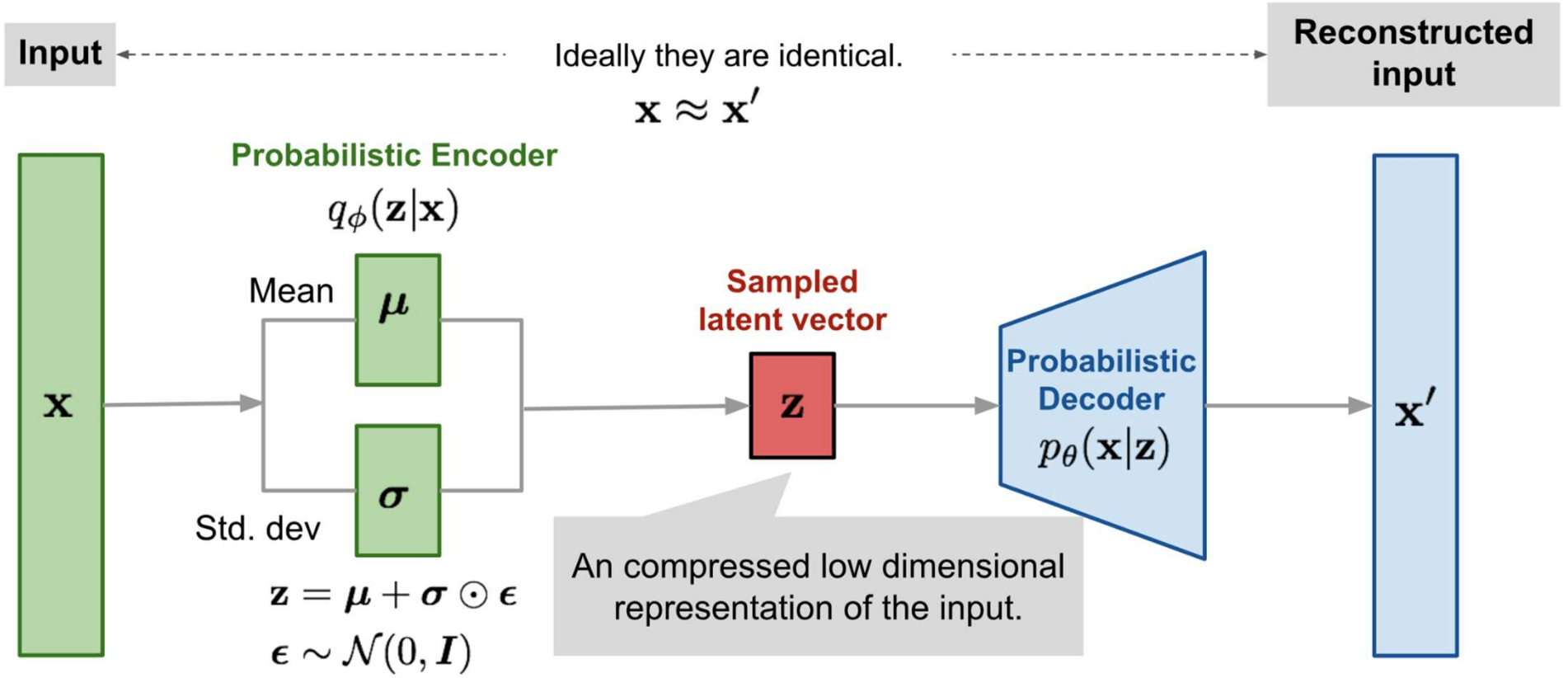
Source: https://lilianweng.github.io/posts/2018-08-12-vae/

Latent Space



Source: https://synthesis.ai/2023/03/21/generative-ai-ii-discrete-latent-spaces/

Variational Autoencoder Model



Source: https://lilianweng.github.io/posts/2018-08-12-vae/

Autoregressive Model (GPT-3)

What is it?

Generative Pre-trained Transformer 3 (GPT-3) is an **autoregressive language model that uses deep learning to produce human-like text**. It is the third-generation language prediction model in the GPT-n series (and the successor to GPT-2) created by OpenAI, a San Francisco-based artificial intelligence research laboratory.

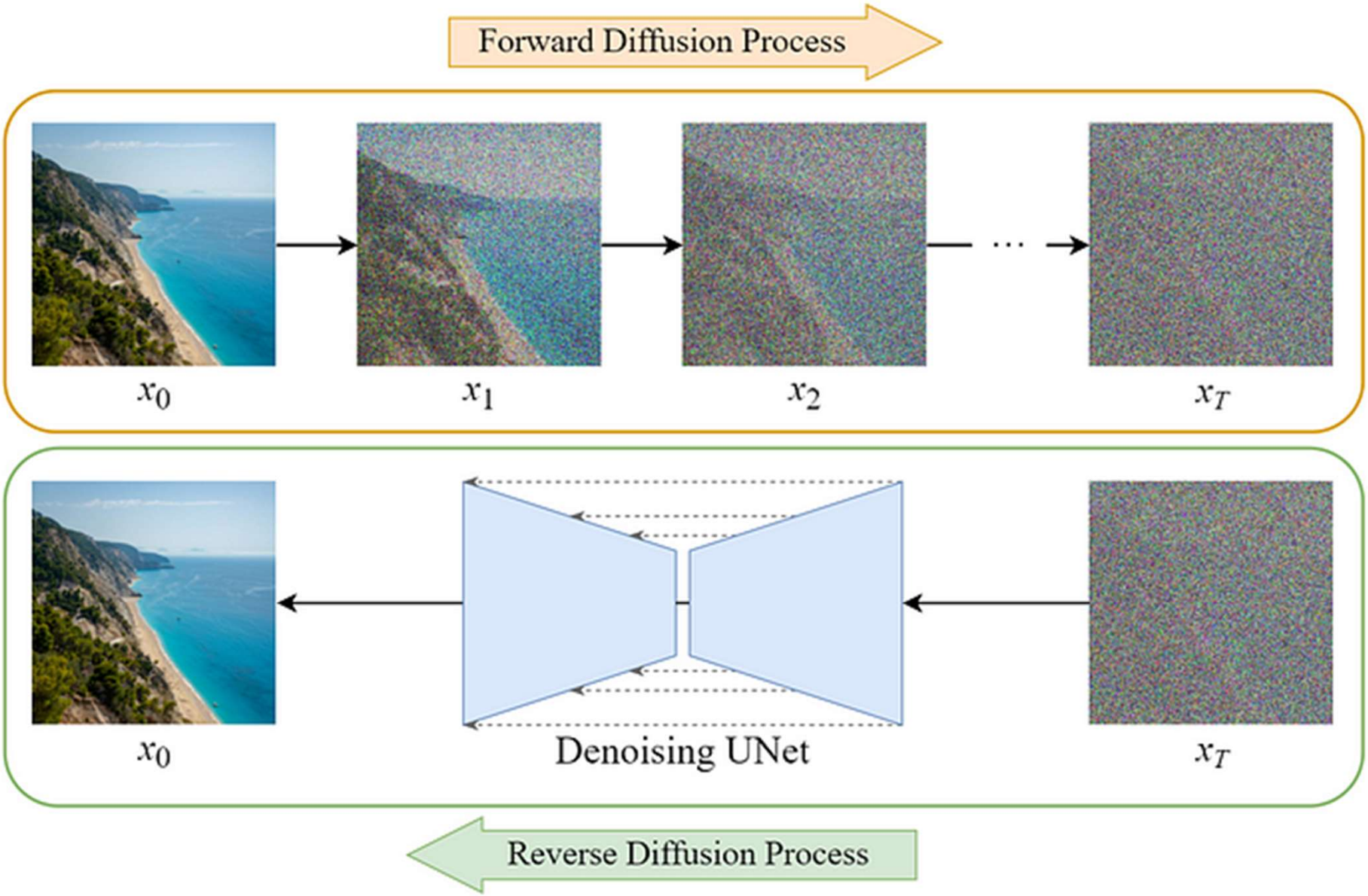
Size:

175 billion machine learning parameters

~45 GB

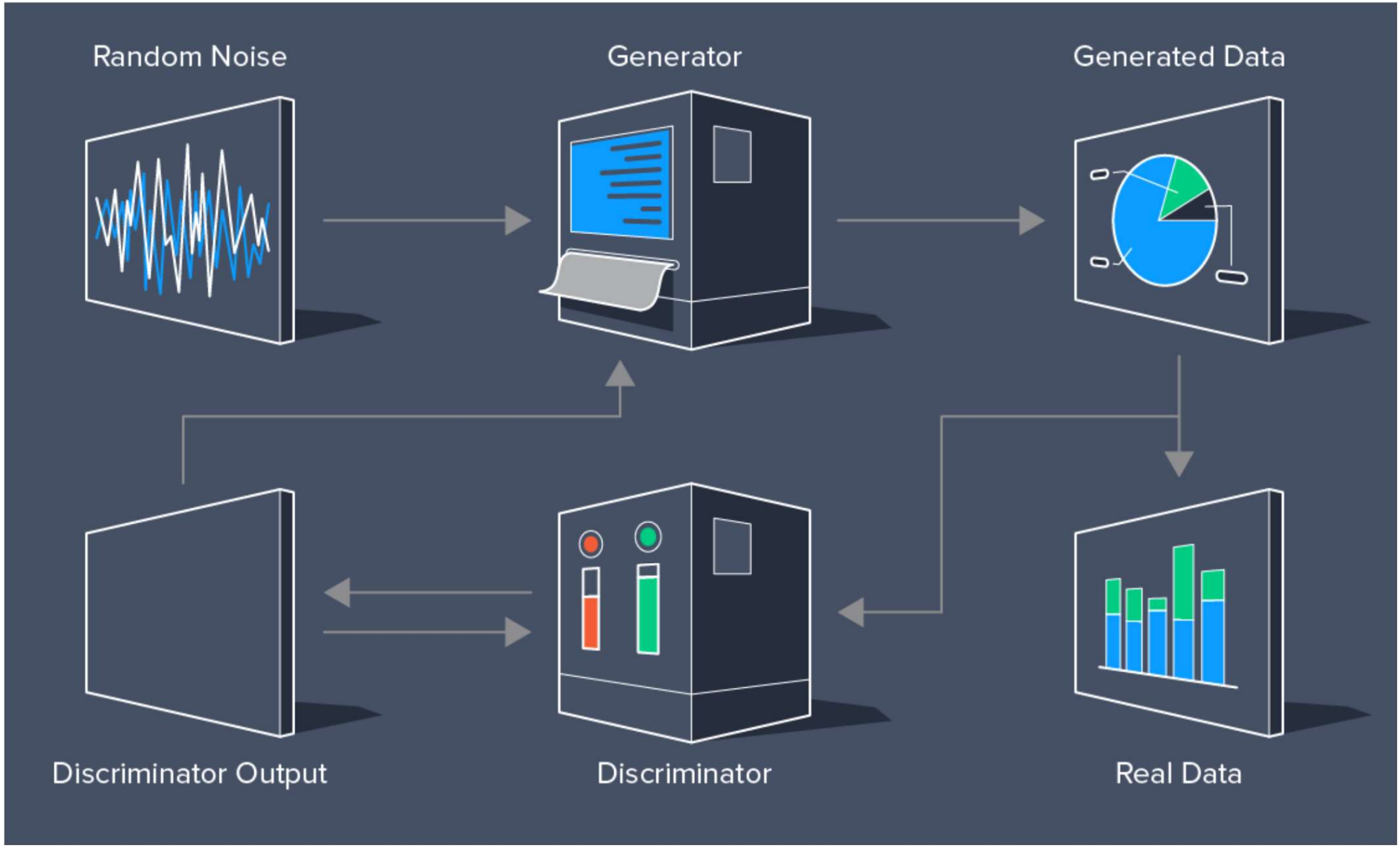
Source: Wikipedia

Diffusion Model



Source: https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e

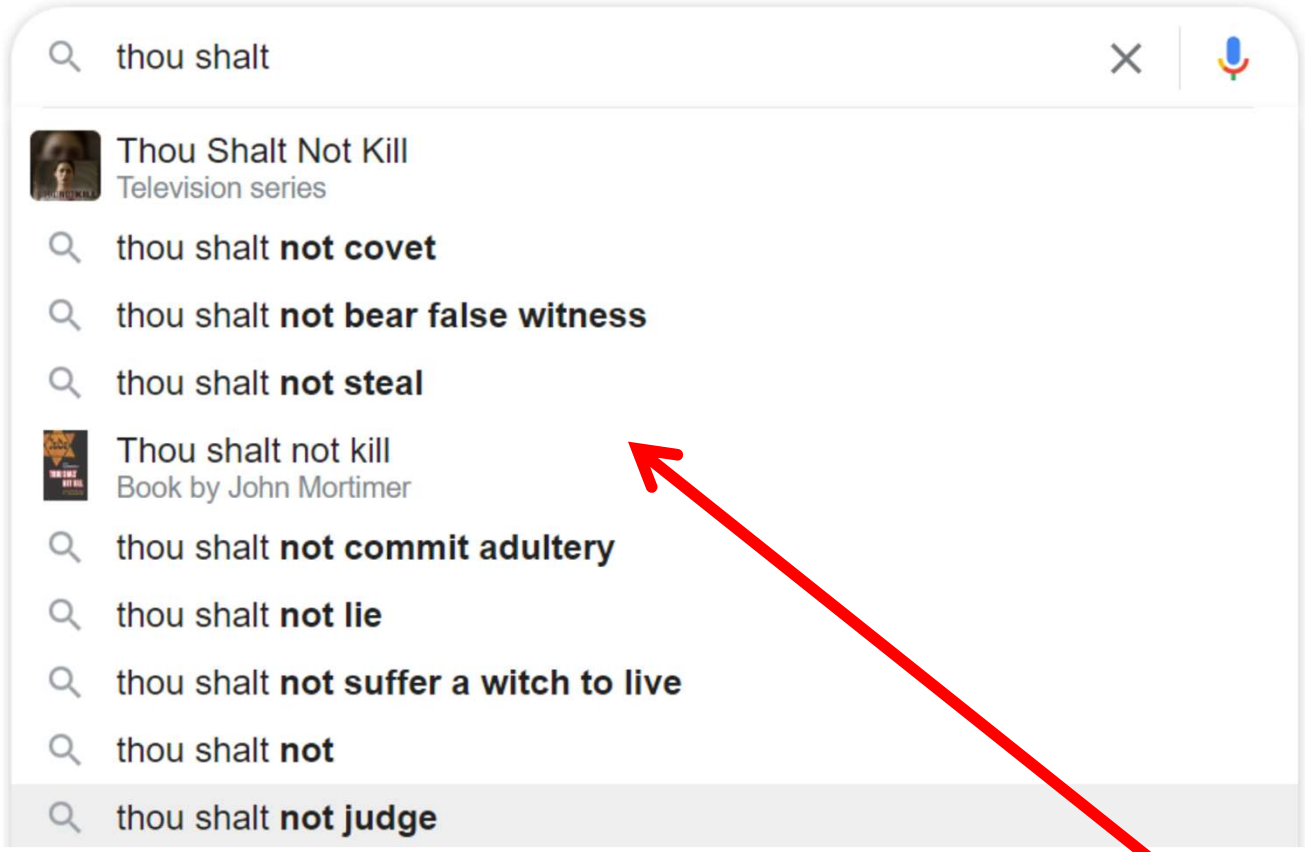
Generative Adversarial Network



Source: https://www.toptal.com/machine-learning/generative-adversarial-networks

ChatGPT and Large Language Models

Language Models: Application



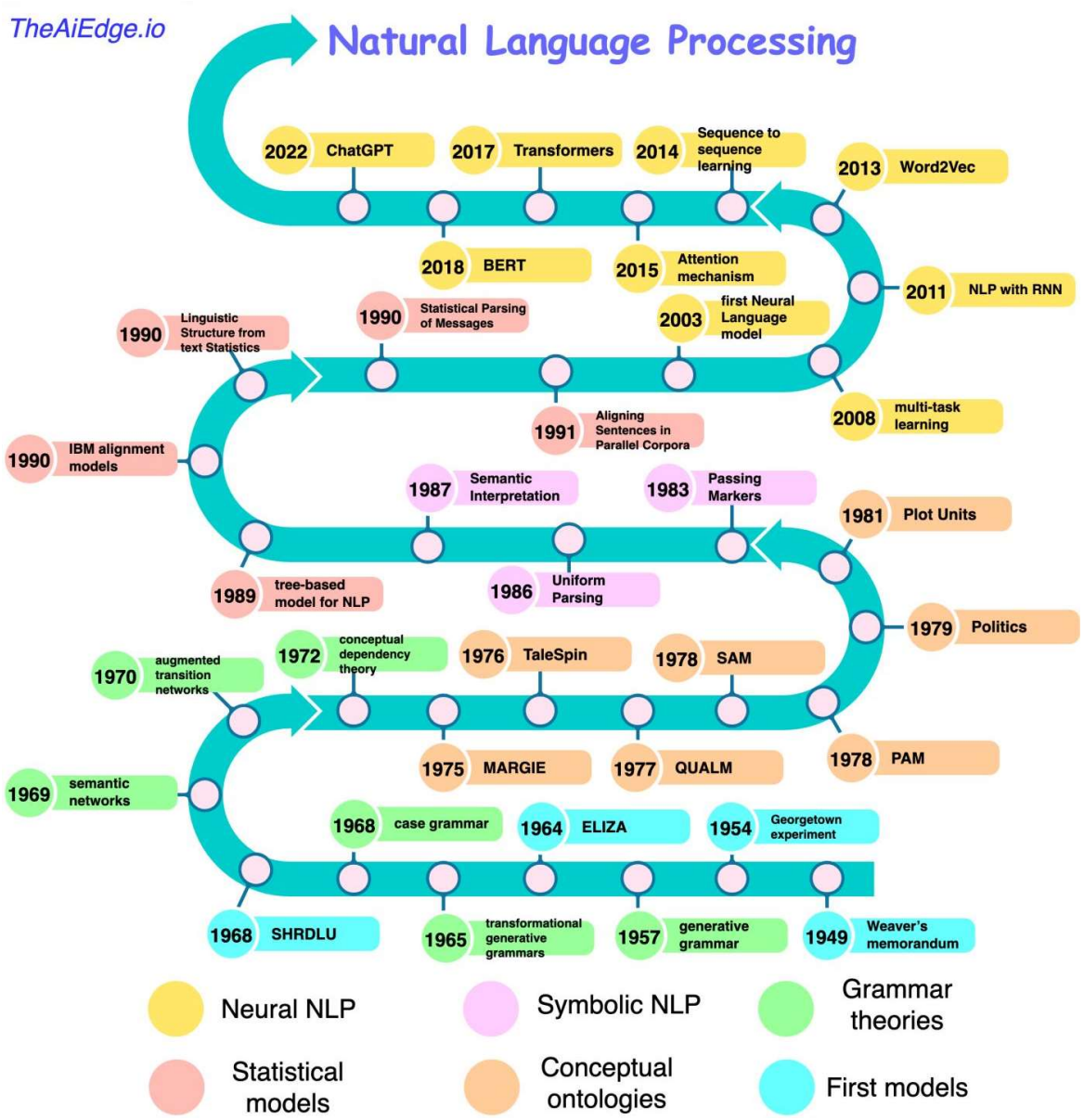
we want to find predict the "rest" of the query

(Statistical) Language Model

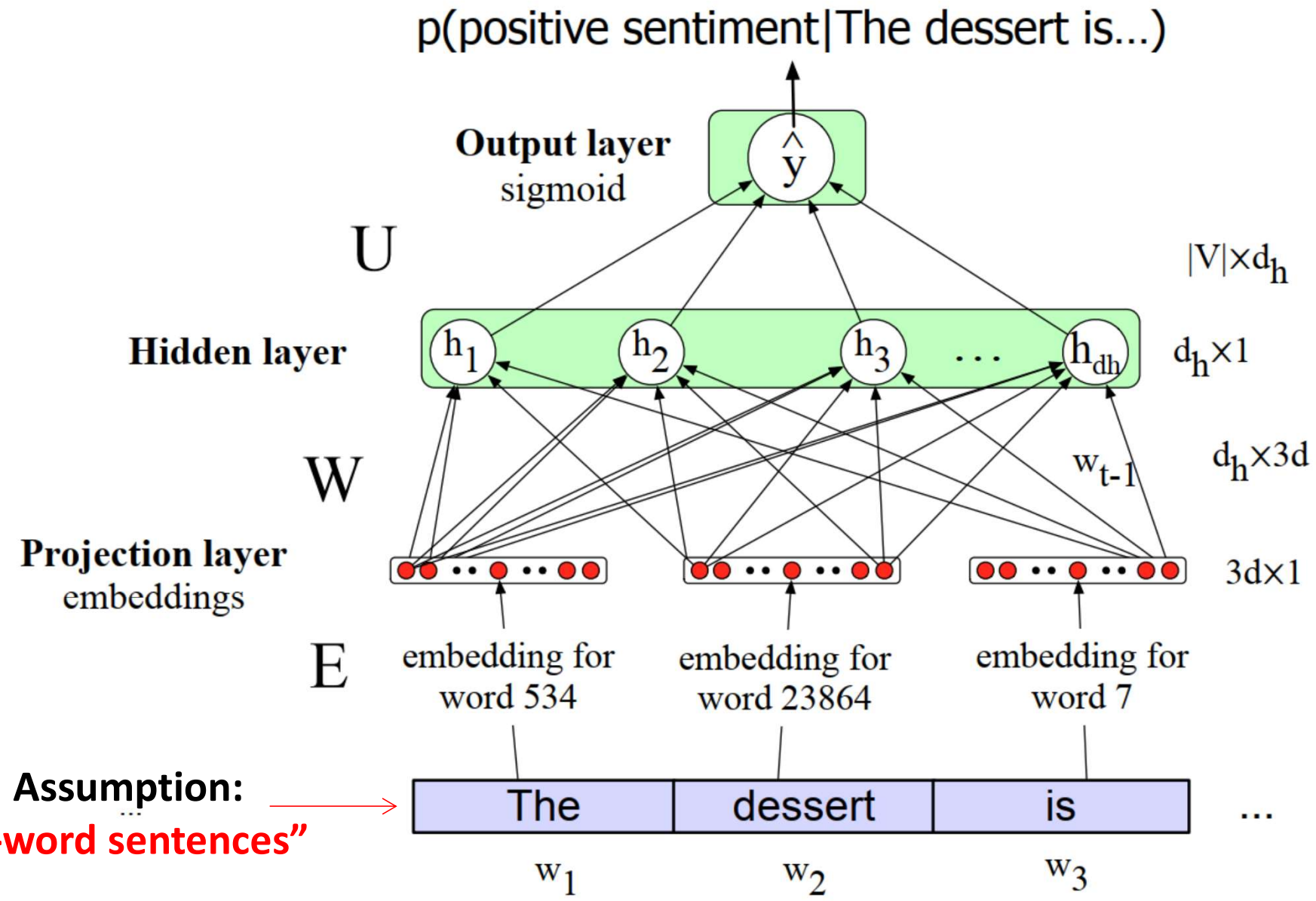
- A (statistical) language model is a probability distribution over words or word sequences.
- In practice, a language model gives the probability of a certain word sequence being "valid".
- Validity in this context does not need to mean grammatical validity at all.

Use lexical resources (corpora) to build LM.

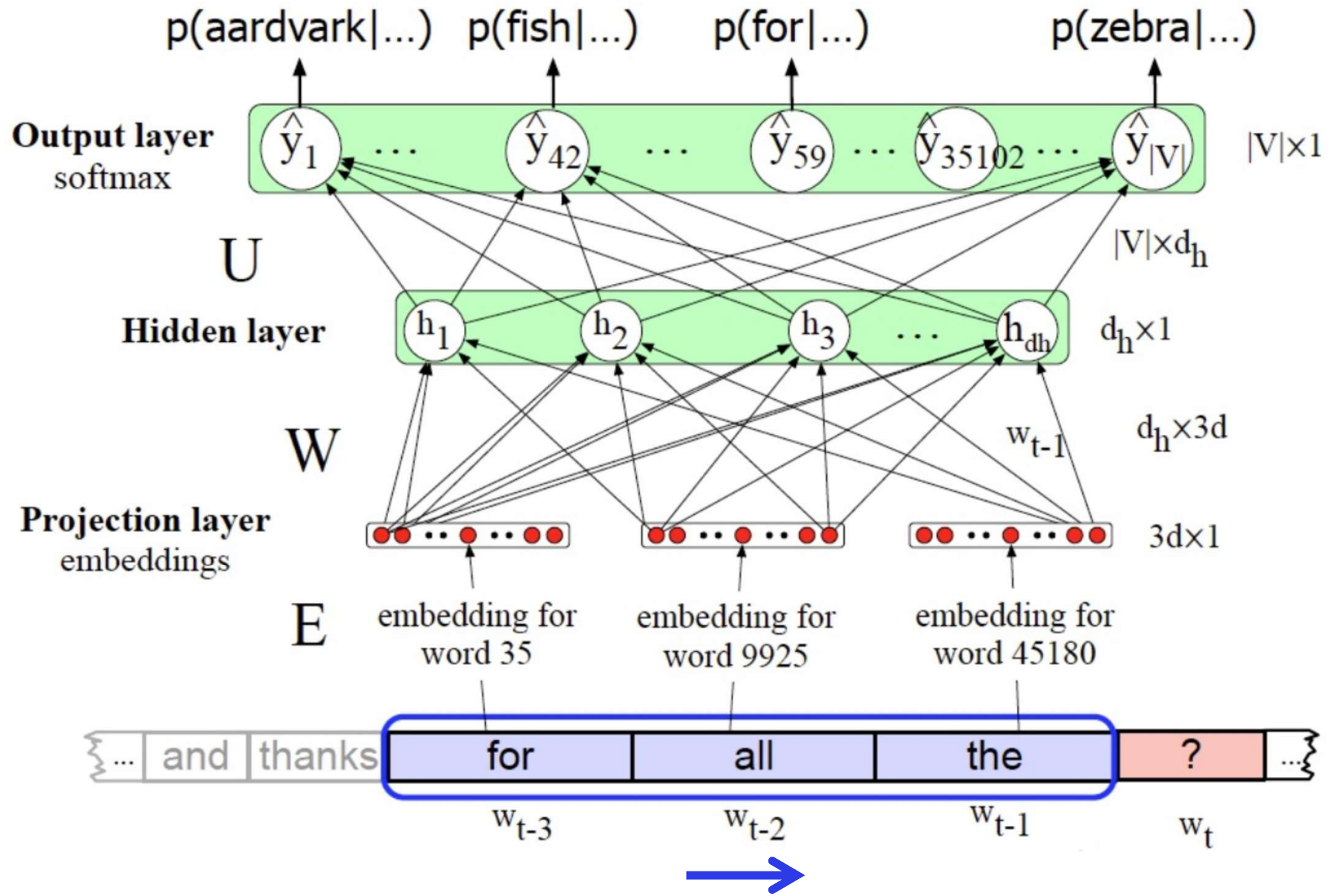
NLP History


TheAiEdge.io
Natural Language Processing
2022 ChatGPT
2017 Transformers
2014 Sequence to sequence learning
2013 Word2Vec
2018 BERT
2015 Attention mechanism
2011 NLP with RNN
1990 Linguistic Structure from text Statistics
1990 Statistical Parsing of Messages
2003 first Neural Language model
1991 Aligning Sentences in Parallel Corpora
2008 multi-task learning
1990 IBM alignment models
1987 Semantic Interpretation
1983 Passing Markers
1981 Plot Units
1989 tree-based model for NLP
1986 Uniform Parsing
1979 Politics
1970 augmented transition networks
1972 conceptual dependency theory
1976 TaleSpin
1978 SAM
1975 MARGIE
1977 QUALM
1978 PAM
1969 semantic networks
1968 case grammar
1964 ELIZA
1954 Georgetown experiment
1968 SHRDLU
1965 transformational generative grammars
1957 generative grammar
1949 Weaver's memorandum
Neural NLP
Symbolic NLP
Grammar theories
Statistical models
Conceptual ontologies
First models

Embeddings as Input Features



Neural Language Model



Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) allow cycles in the computational graph (network). A **network node (unit) can take its own output from an earlier step as input** (with delay introduced).

Enables **having internal state / memory** → inputs received earlier affect the RNN response to current input.

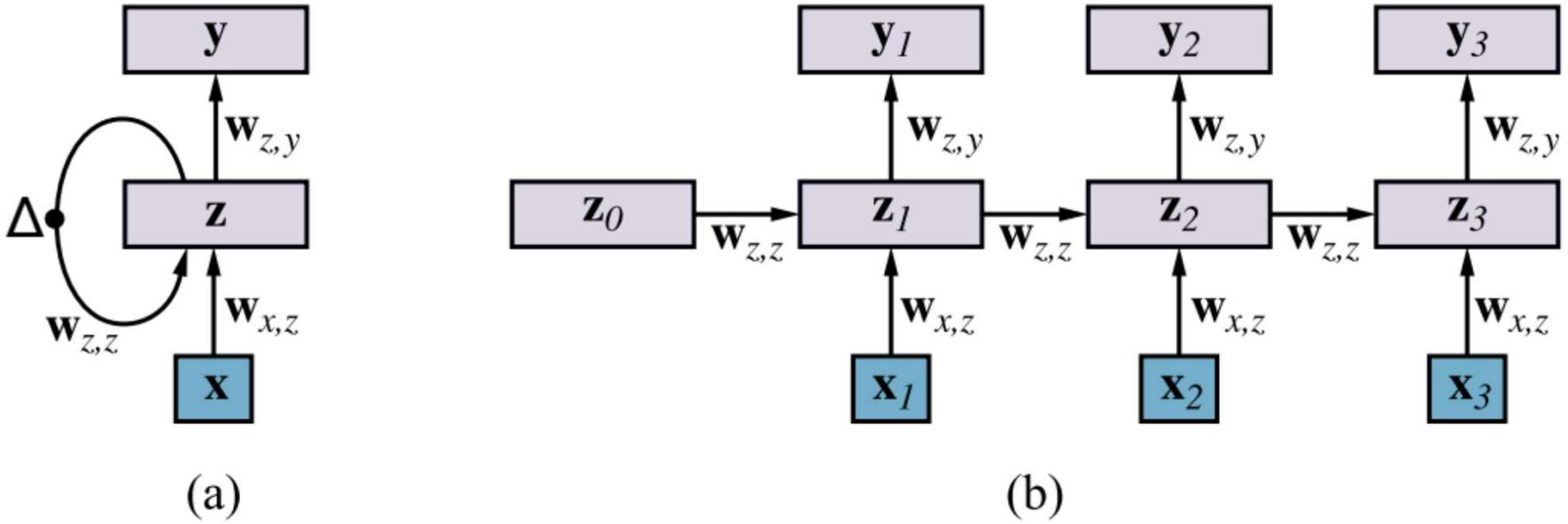
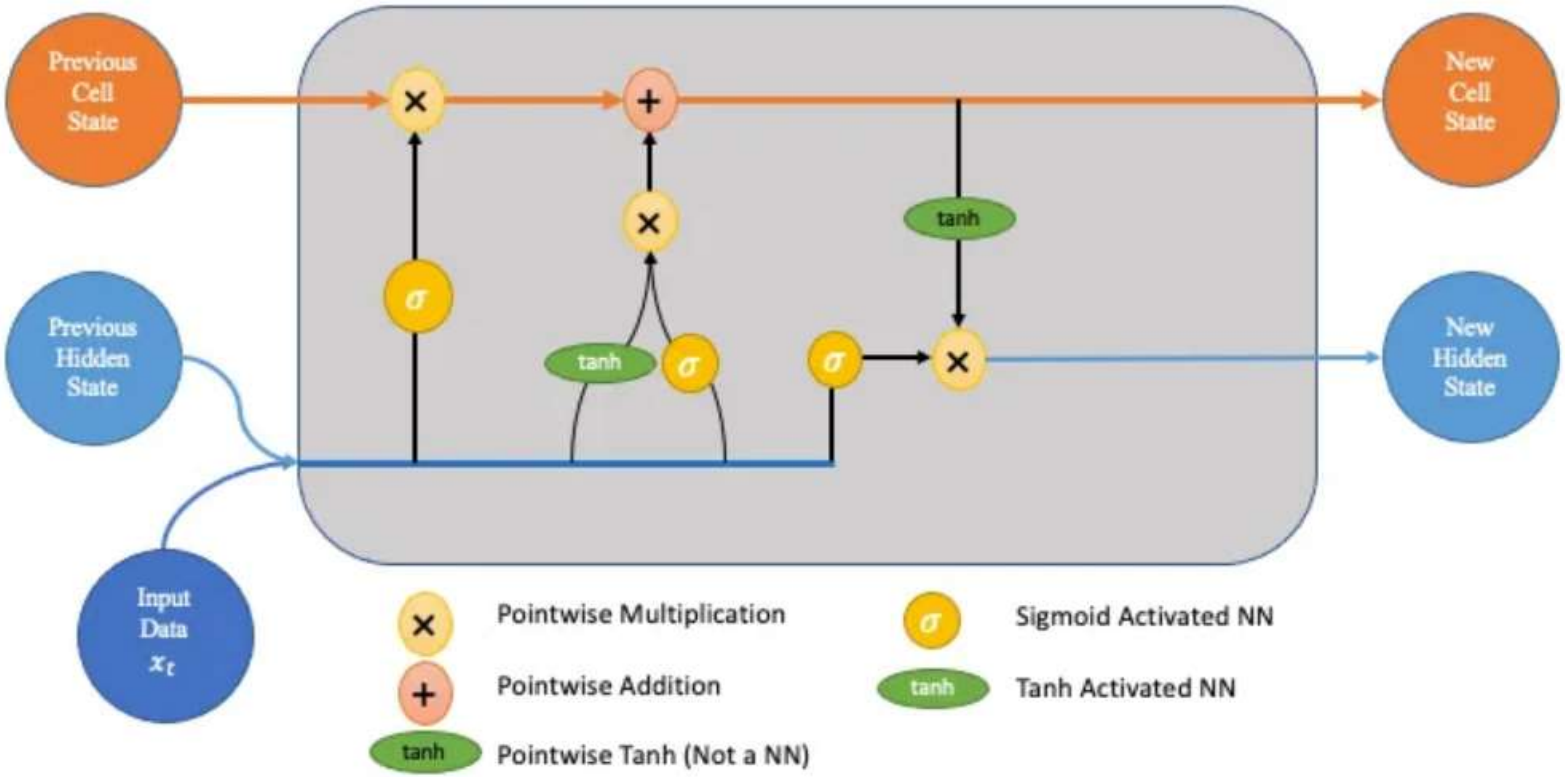


Figure (a) Schematic diagram of a basic RNN where the hidden layer **z** has recurrent connections; the Δ symbol indicates a delay. (b) The same network unrolled over three time steps to create a feedforward network. Note that the weights are shared across all time steps.

Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) is an artificial neural network. Unlike standard feedforward neural networks, **LSTM has feedback connections**. Such a recurrent neural network (RNN) can process not only single data points (such as images), but also entire sequences of data (such as speech or video). This characteristic makes LSTM networks **ideal for processing and predicting data**.



Source: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

Voice Assistant: Alexa

Your home

Amazon Alexa Cloud Service

Automated Speech Recognition (ASR) **[Deep NN]**

Transcription:
"*Play Two Steps Behind by Def Leppard*"

Natural Language Understanding (NLU) **[Deep NN]**

Extracted meaning:
Intent: PlayMusic | **Artist**: Def Leppard | **Song**: Two Steps Behind

Dialog Manager	Actions
Response: "*Playing Two Steps Behind by Def Leppard*"	**Action**: PlayMusic **Artist**: Two Steps Behind **Song**: Def Leppard
Neural Text-To-Speech (NTTS)	Amazon Music

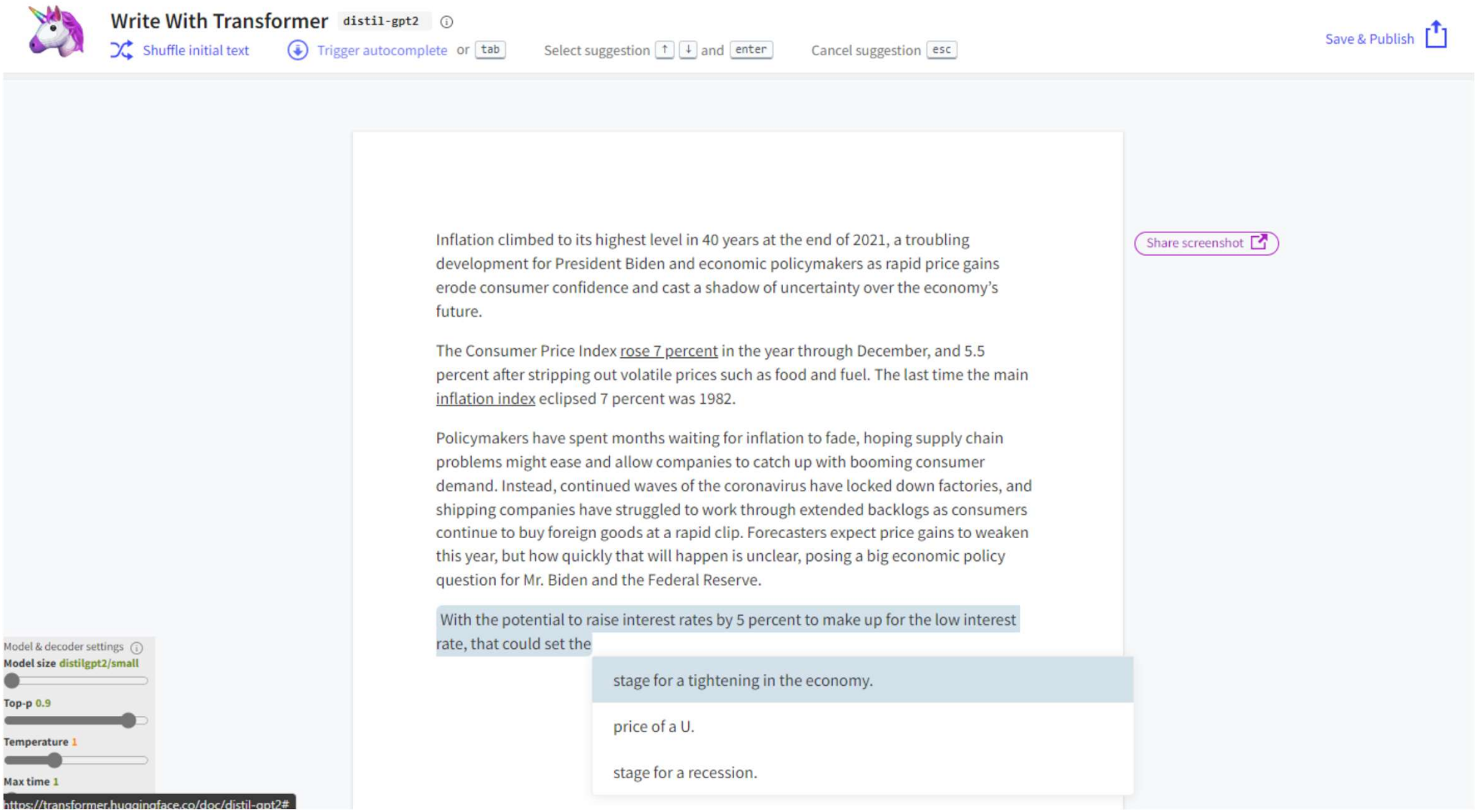
Your home

Large Language Model (LLM)

A **large language model (LLM)** is a **language model consisting of a neural network with many parameters** (typically billions of weights or more), **trained on large quantities of unlabeled text using self-supervised learning**.

Source: Wikipedia

Language Modeling Example: GPT-2



Source: https://transformer.huggingface.co/doc/distil-gpt2

Generative Pre-trained Transformer 3

What is it?

Generative Pre-trained Transformer 3 (GPT-3) is an **autoregressive language model that uses deep learning to produce human-like text**. It is the third-generation language prediction model in the GPT-n series (and the successor to GPT-2) created by OpenAI, a San Francisco-based artificial intelligence research laboratory.

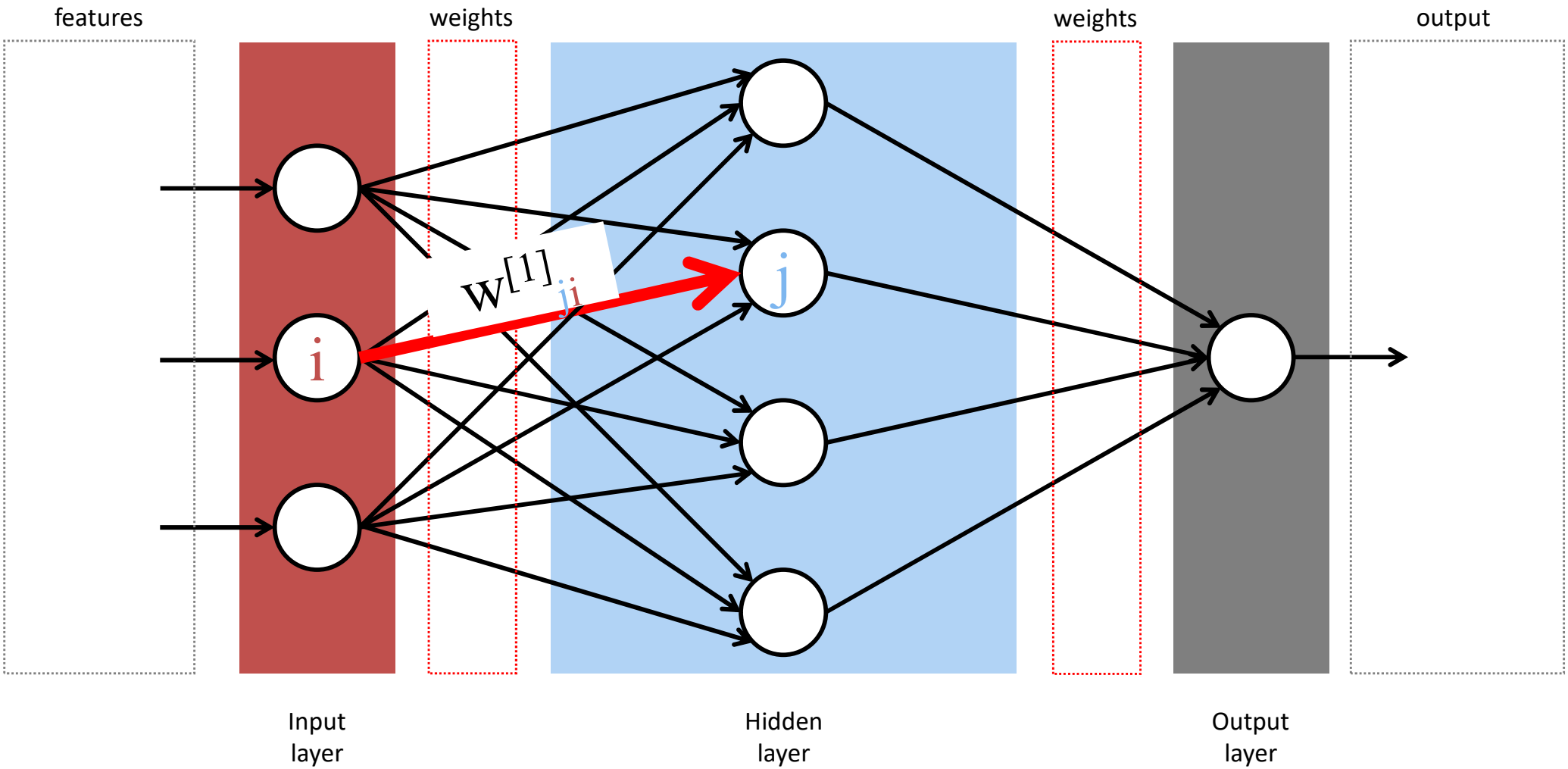
Size:

175 billion machine learning parameters

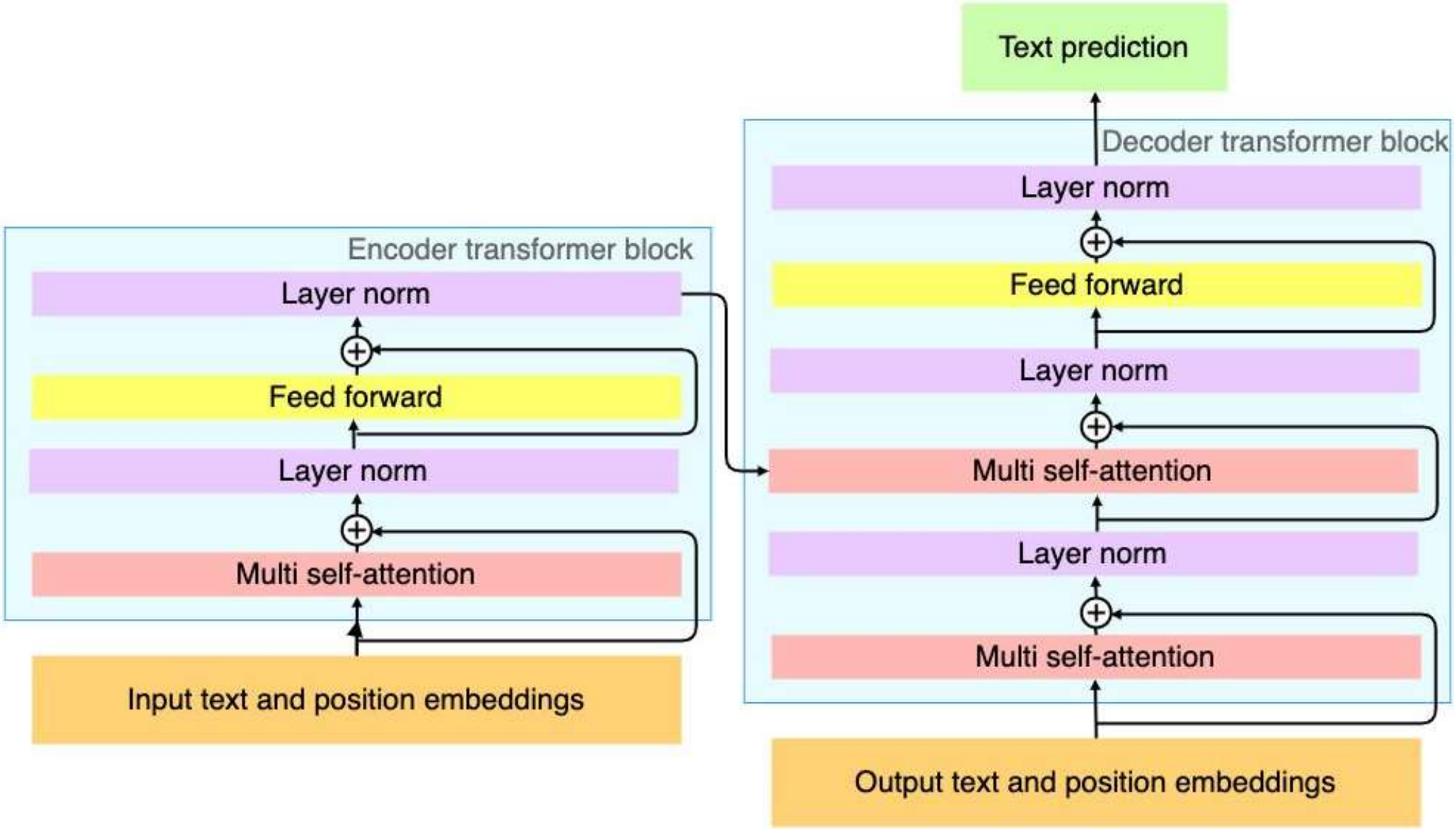
~45 GB

Source: Wikipedia

Parameters? What Are Those?

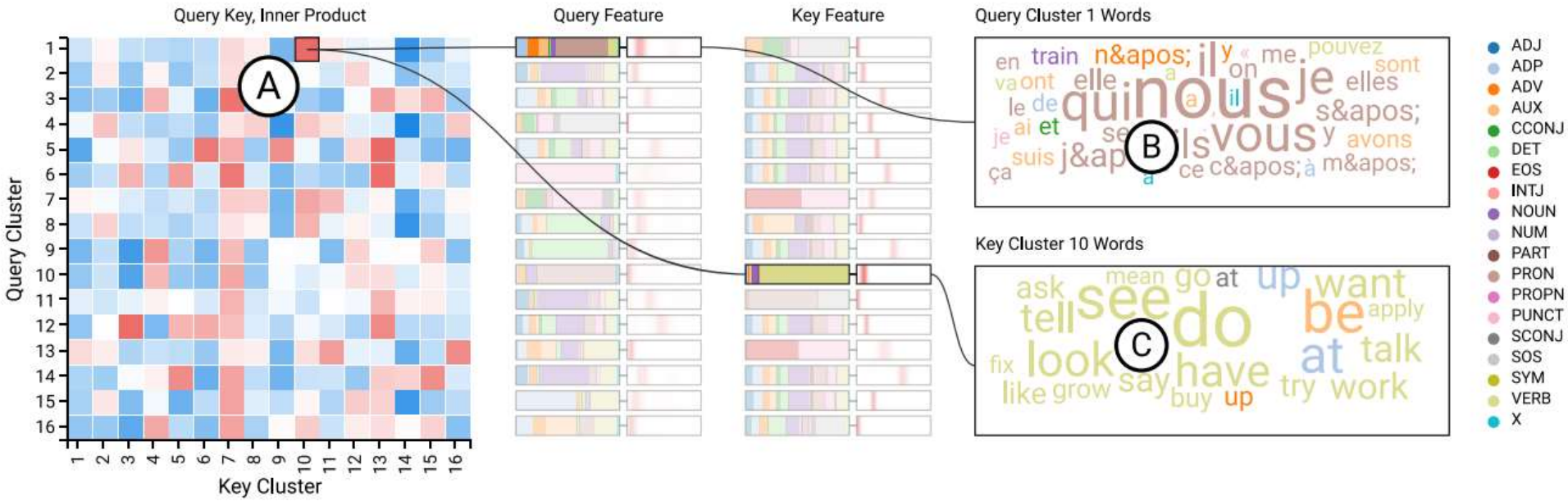


Transformer Architecture



Self-Attention

In artificial neural networks, **attention is a technique that is meant to mimic cognitive attention**. The effect **enhances some parts of the input data while diminishing other parts** — the motivation being that the network should devote more focus to the important parts of the data, even though they may be small. **Learning which part of the data is more important than another depends on the context**, and this is trained by gradient descent.



Source: Park et al. – "SANVis: Visual Analytics for Understanding Self-Attention Networks"

Generative Pre-trained Transformer 4

What is it?

Generative Pre-trained Transformer 4 (GPT-4) is a **multimodal large language model** created by OpenAI. As a transformer, GPT-4 was **pretrained to predict the next token** (using both public data and "data licensed from third-party providers"), and was then **fine-tuned with reinforcement learning from human and AI feedback for human alignment and policy compliance**.

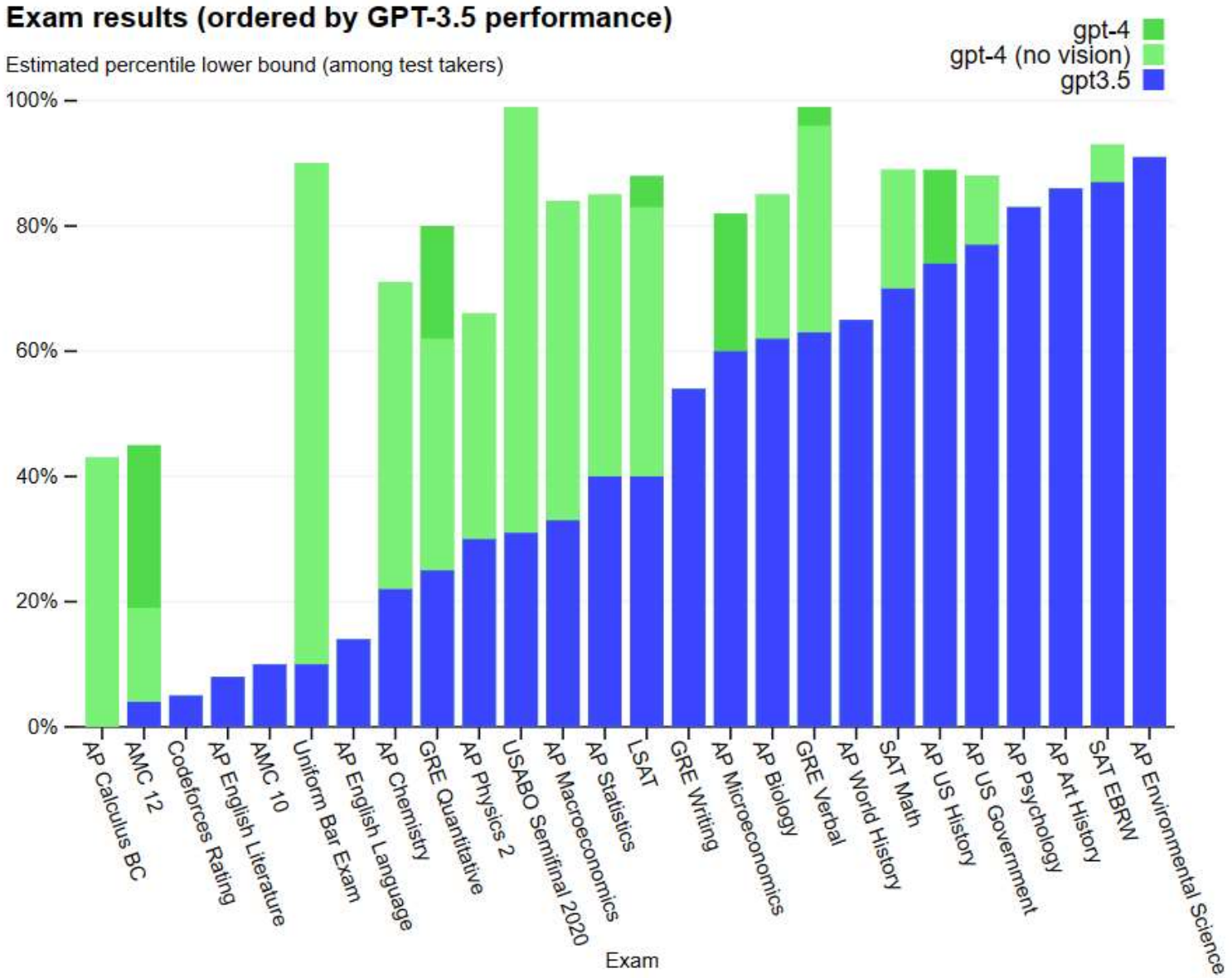
Size:

1 trillion machine learning parameters

~45 GB

Source: Wikipedia

GPT-3.5 / 4 Exam Taking Performance



Source: https://openai.com/research/gpt-4

GPT Takes The Bar Exam

GPT Takes the Bar Exam

7 Pages • Posted: 31 Dec 2022 • Last revised: 3 Jan 2023

Michael James Bommarito
273 Ventures; Licensio, LLC; Stanford Center for Legal Informatics; Michigan State College of Law; Bommarito Consulting, LLC

Daniel Martin Katz
Illinois Tech - Chicago Kent College of Law; Bucerius Center for Legal Technology & Data Science; Stanford CodeX - The Center for Legal Informatics; 273 Ventures

Date Written: December 29, 2022

Abstract

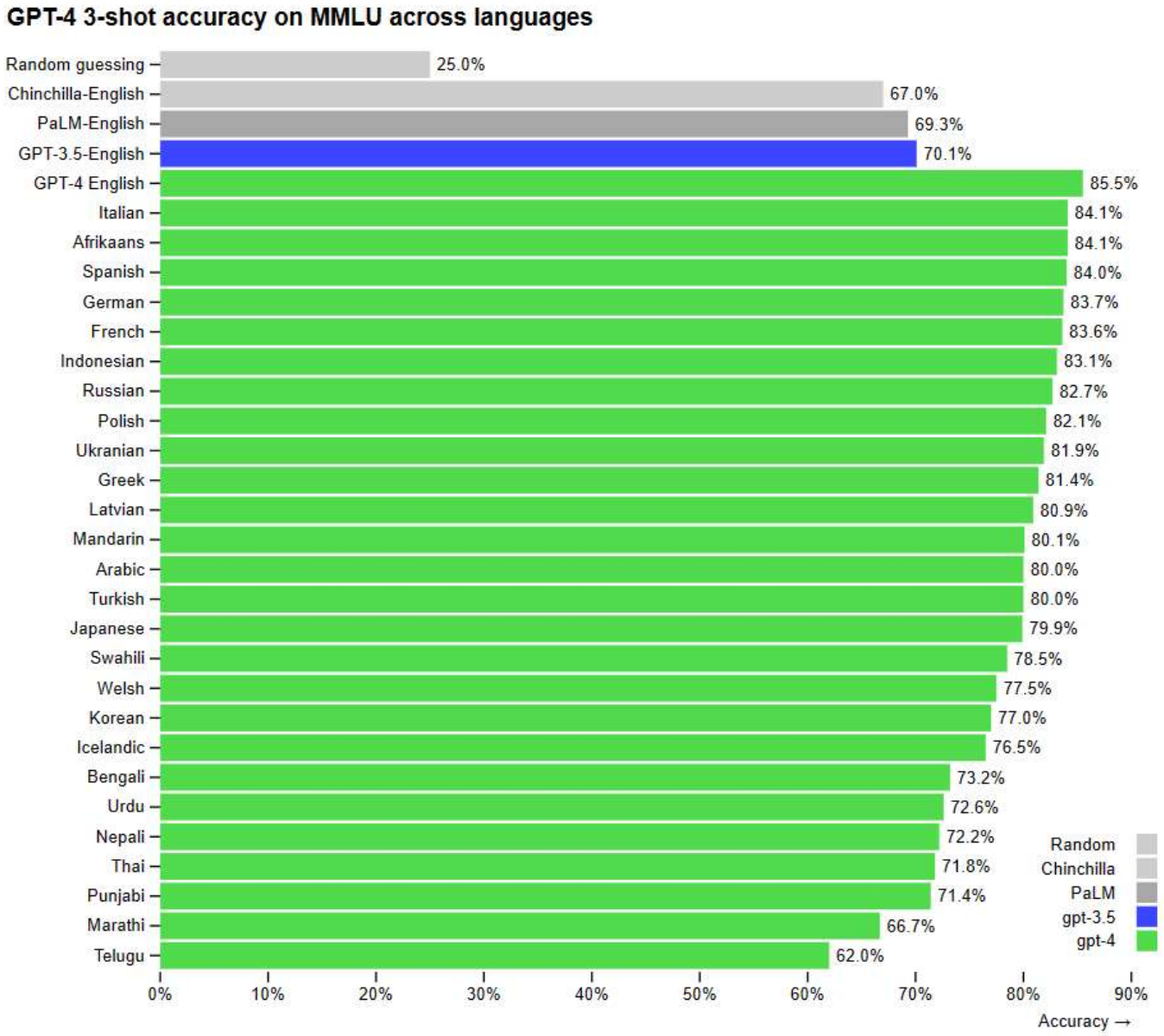
Nearly all jurisdictions in the United States require a professional license exam, commonly referred to as "the Bar Exam," as a precondition for law practice. To even sit for the exam, most jurisdictions require that an applicant completes at least seven years of post-secondary education, including three years at an accredited law school. In addition, most test-takers also undergo weeks to months of further, exam-specific preparation. Despite this significant investment of time and capital, approximately one in five test-takers still score under the rate required to pass the exam on their first try. In the face of a complex task that requires such depth of knowledge, what, then, should we expect of the state of the art in "AI?" In this research, we document our experimental evaluation of the performance of OpenAI's text-davinci-003 model, often-referred to as GPT-3.5, on the multistate multiple choice (MBE) section of the exam. While we find no benefit in fine-tuning over GPT-3.5's zero-shot performance at the scale of our training data, we do find that hyperparameter optimization and prompt engineering positively impacted GPT-3.5's zero-shot performance. For best prompt and parameters, GPT-3.5 achieves a headline correct rate of 50.3% on a complete NCBE MBE practice exam, significantly in excess of the 25% baseline guessing rate, and performs at a passing rate for both Evidence and Torts. GPT-3.5's ranking of responses is also highly correlated with correctness; its top two and top three choices are correct 71% and 88% of the time, respectively, indicating very strong non-entailment performance. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBE component of the Bar Exam in the near future.

Source: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839

GPT-4 MMLU Performance

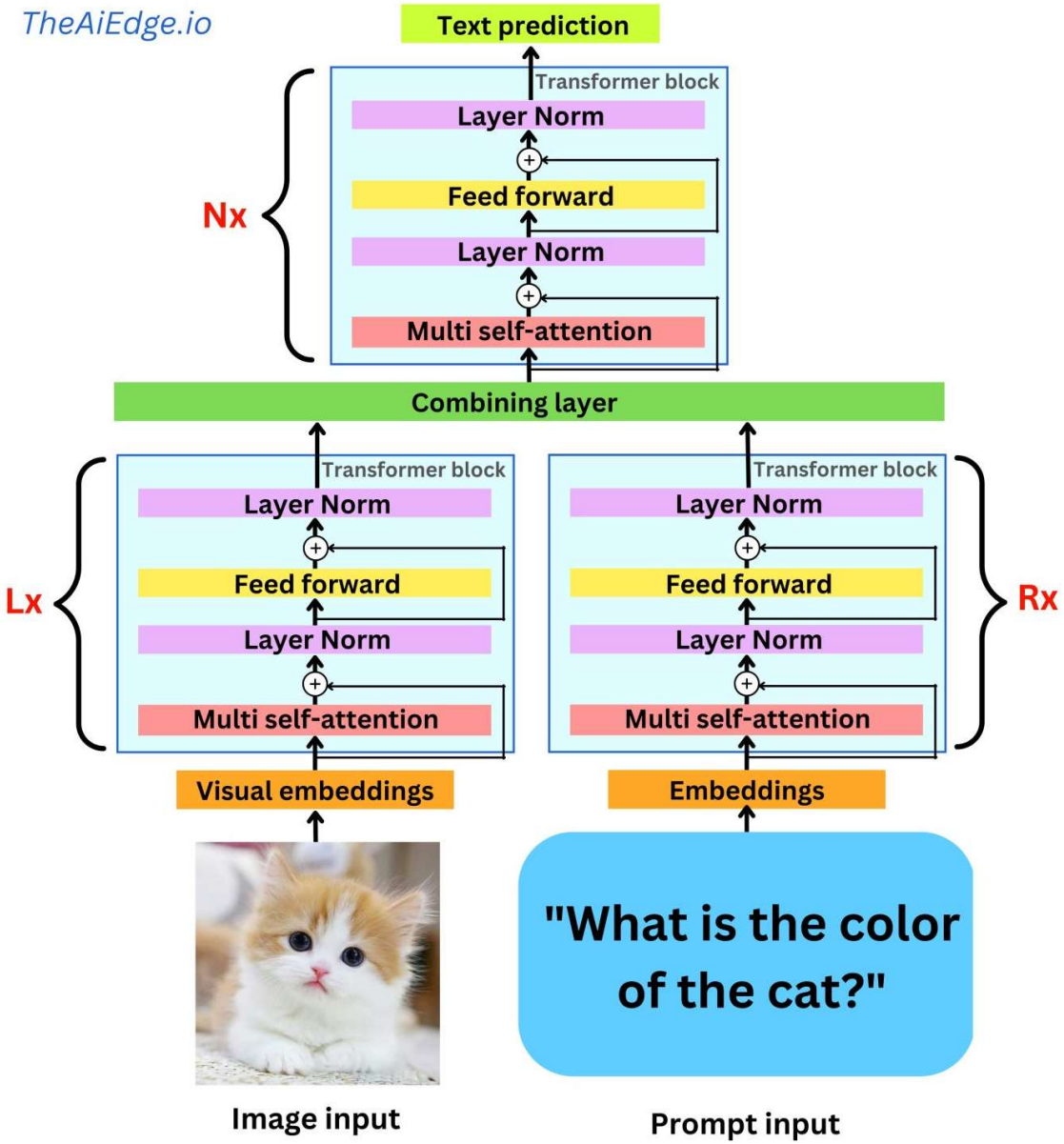
MMLU – Massive Multitask Language Understanding (benchmark)

3-shot accuracy: Few-shot (k-shot) learning is a type of supervised learning that is intended to rapidly generalize to new tasks containing only a few samples of supervised information based on prior knowledge



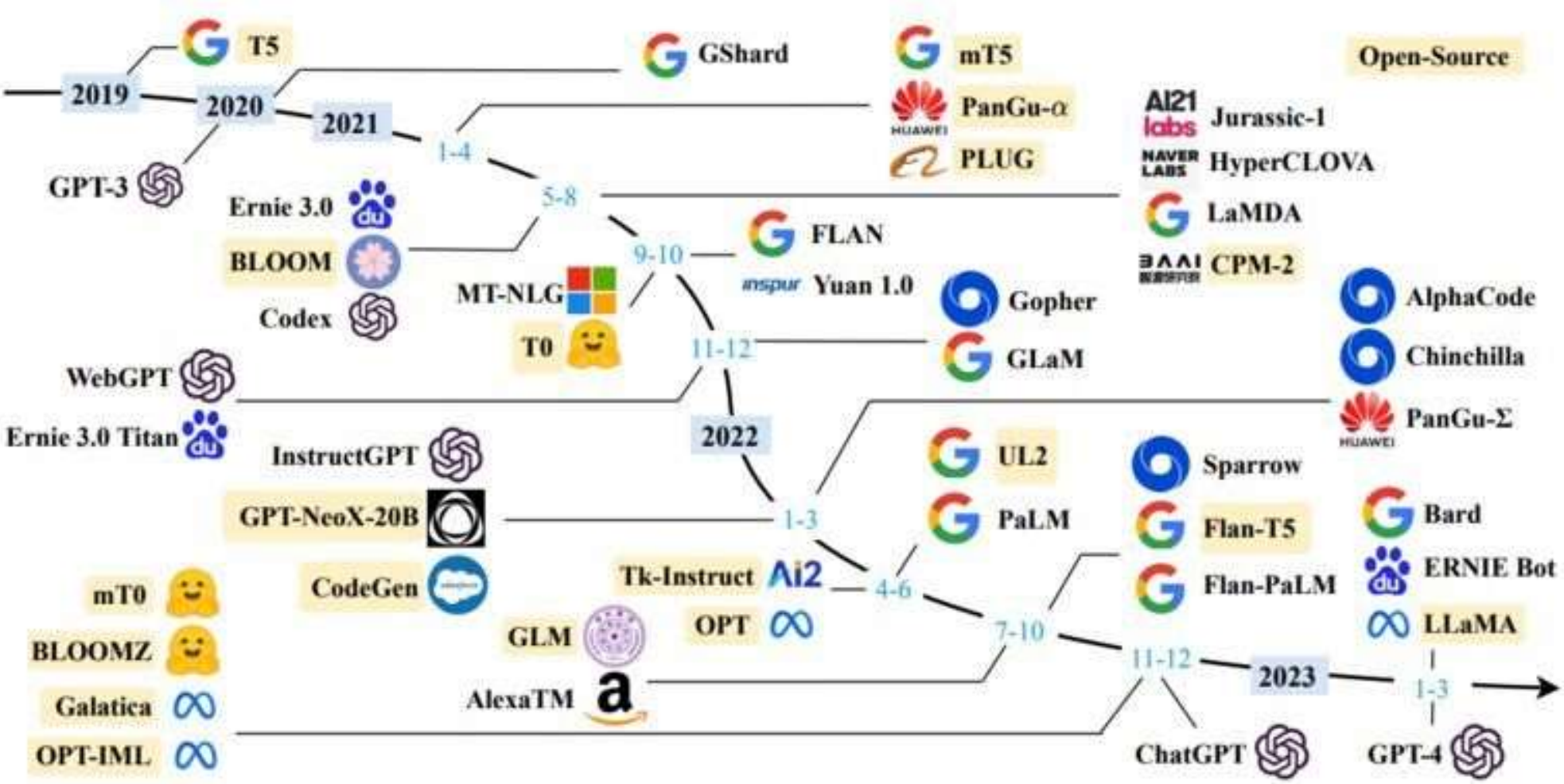
Sources: https://openai.com/research/gpt-4 and https://paperswithcode.com/dataset/mmlu

GPT-4 Architecture



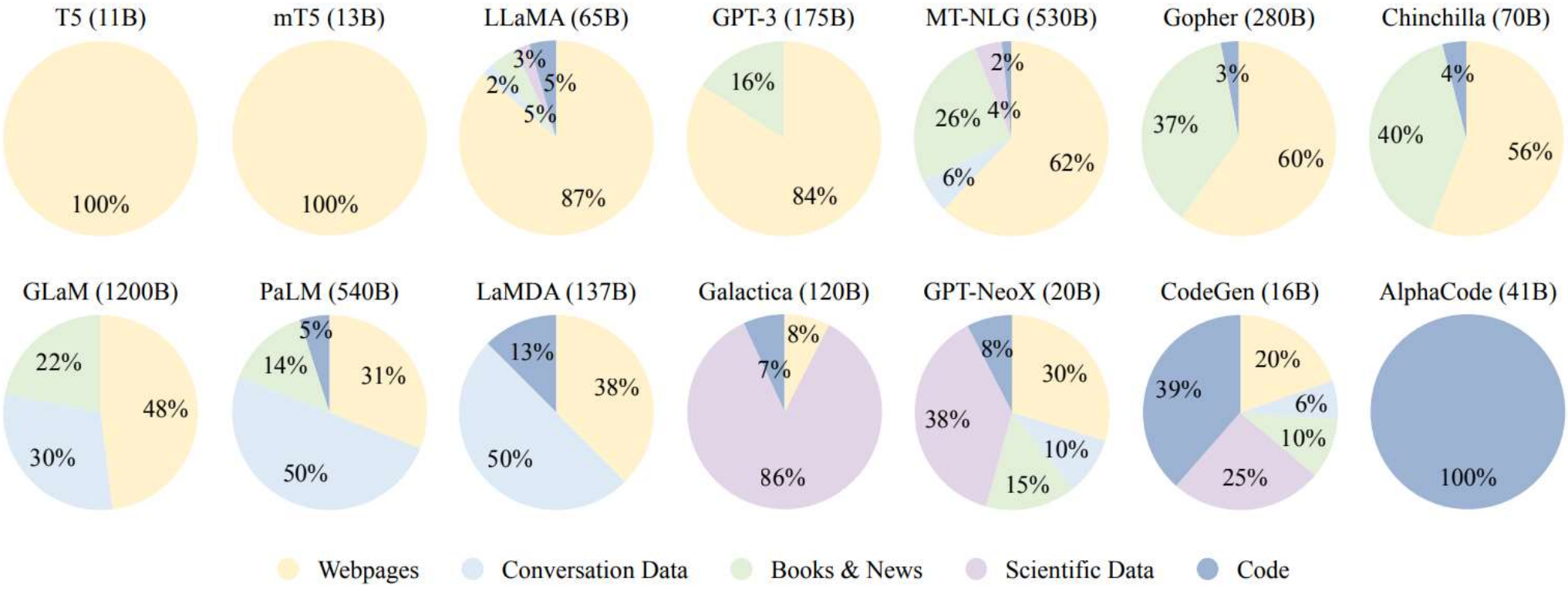
Source: TheAiEdge.io

>10 GB Large Language Models



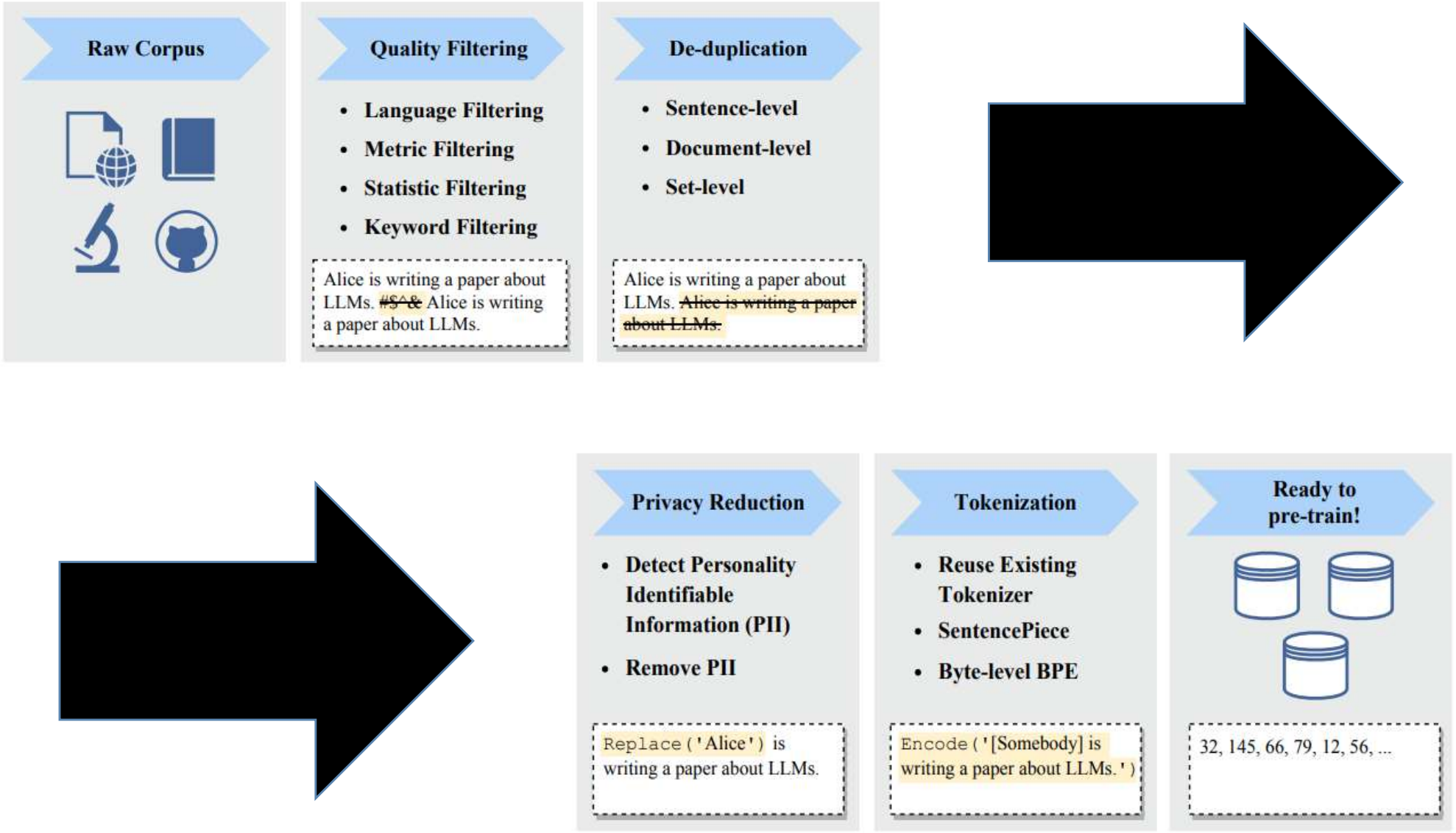
Source: Zhao et al. – "A Survey of Large Language Models" [2023]

Large Language Models Data Sources



Source: Zhao et al. – "A Survey of Large Language Models" [2023]

LLM Data Pre-Processing Pipeline



Source: Zhao et al. – "A Survey of Large Language Models" [2023]

ChatGPT

What is it?

ChatGPT is a **chatbot** developed by OpenAI and released in November 2022. It is **built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models** (LLMs) and has been **fine-tuned** (an approach to **transfer learning**) **using both supervised and reinforcement learning techniques**.

Source: Wikipedia

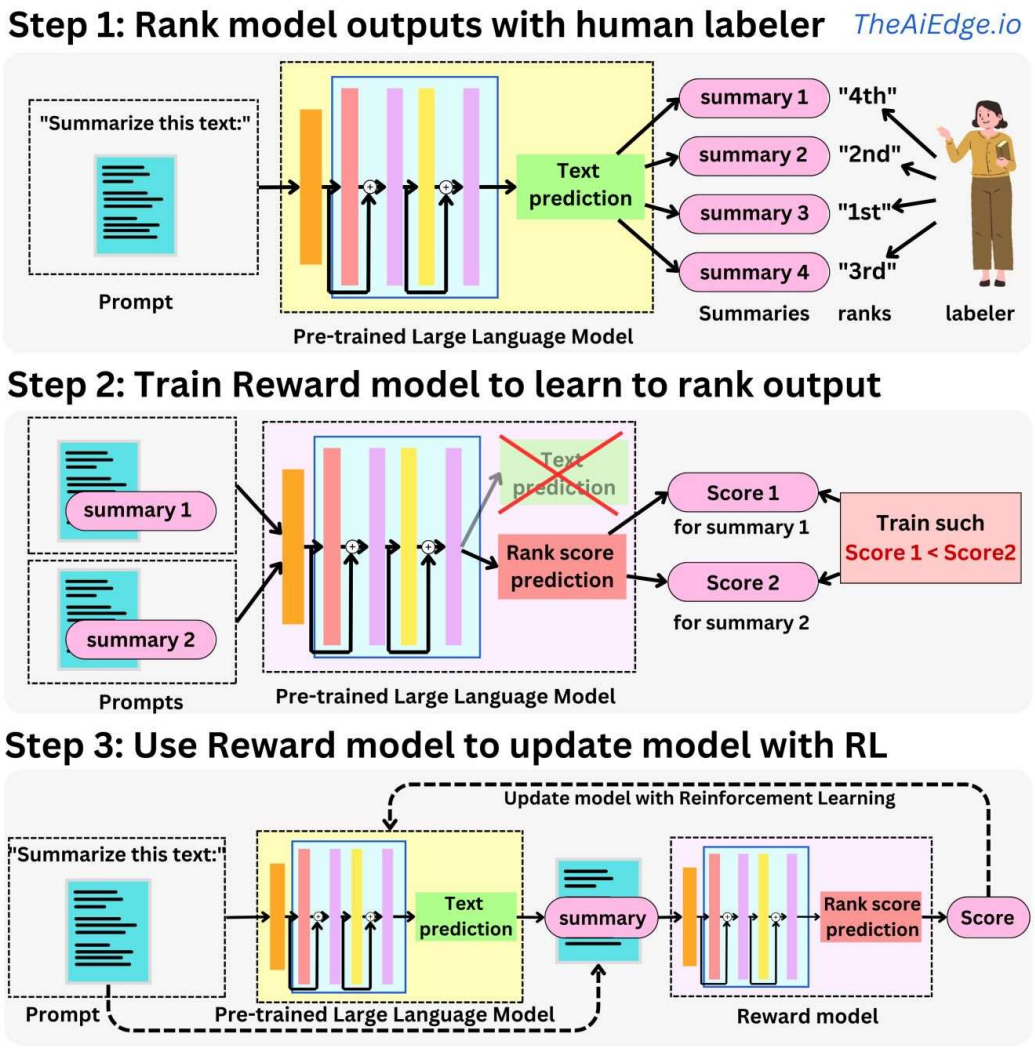
Transfer Learning

In transfer learning, experience with one learning task helps an agent learn better on another task.

Pre-trained models can be used as a starting point for developing new models.

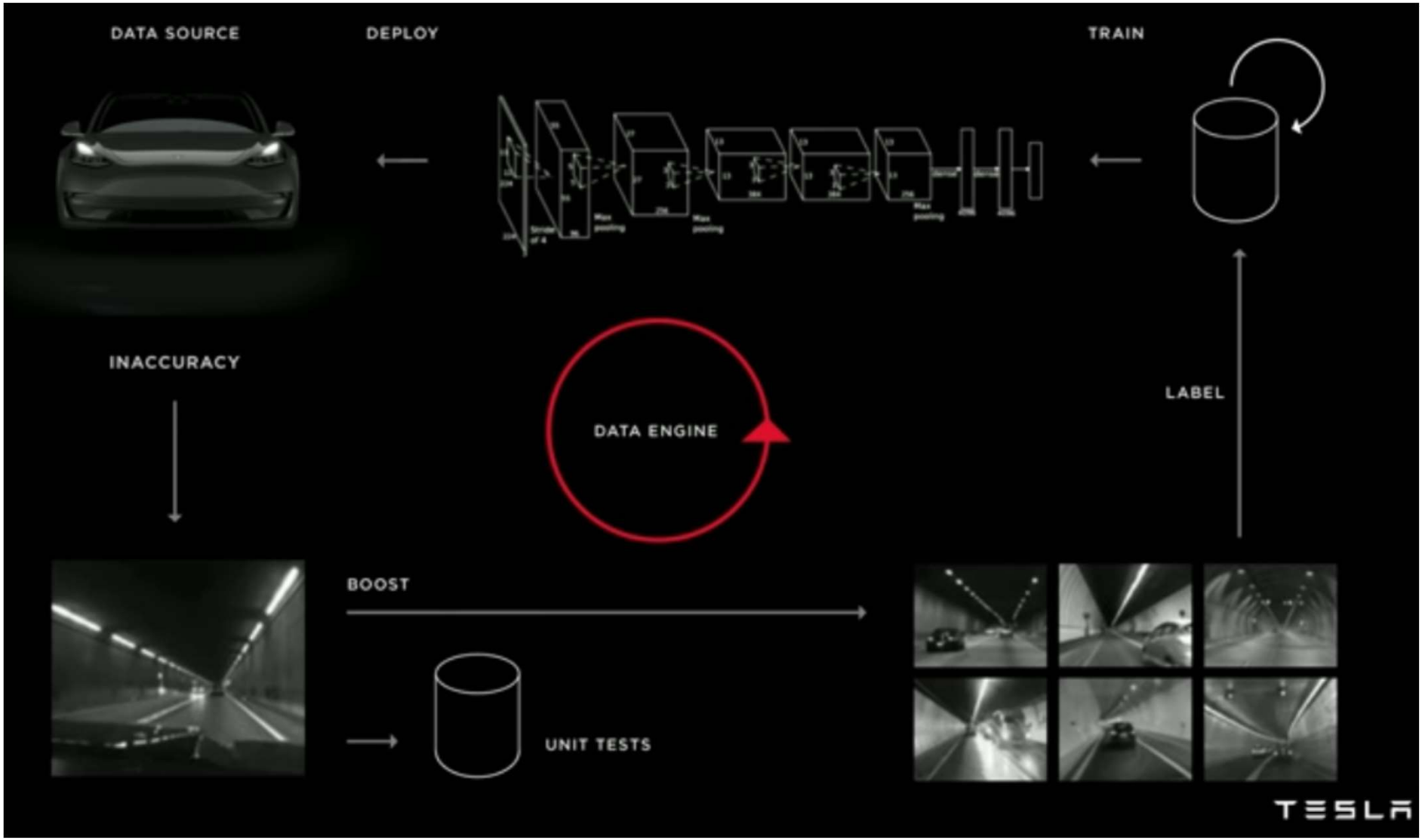
ChatGPT: Learning From Feedback

From GPT-3 to ChatGPT: Reinforcement Learning from Human Feedback (RLHF)



Source: TheAiEdge.io

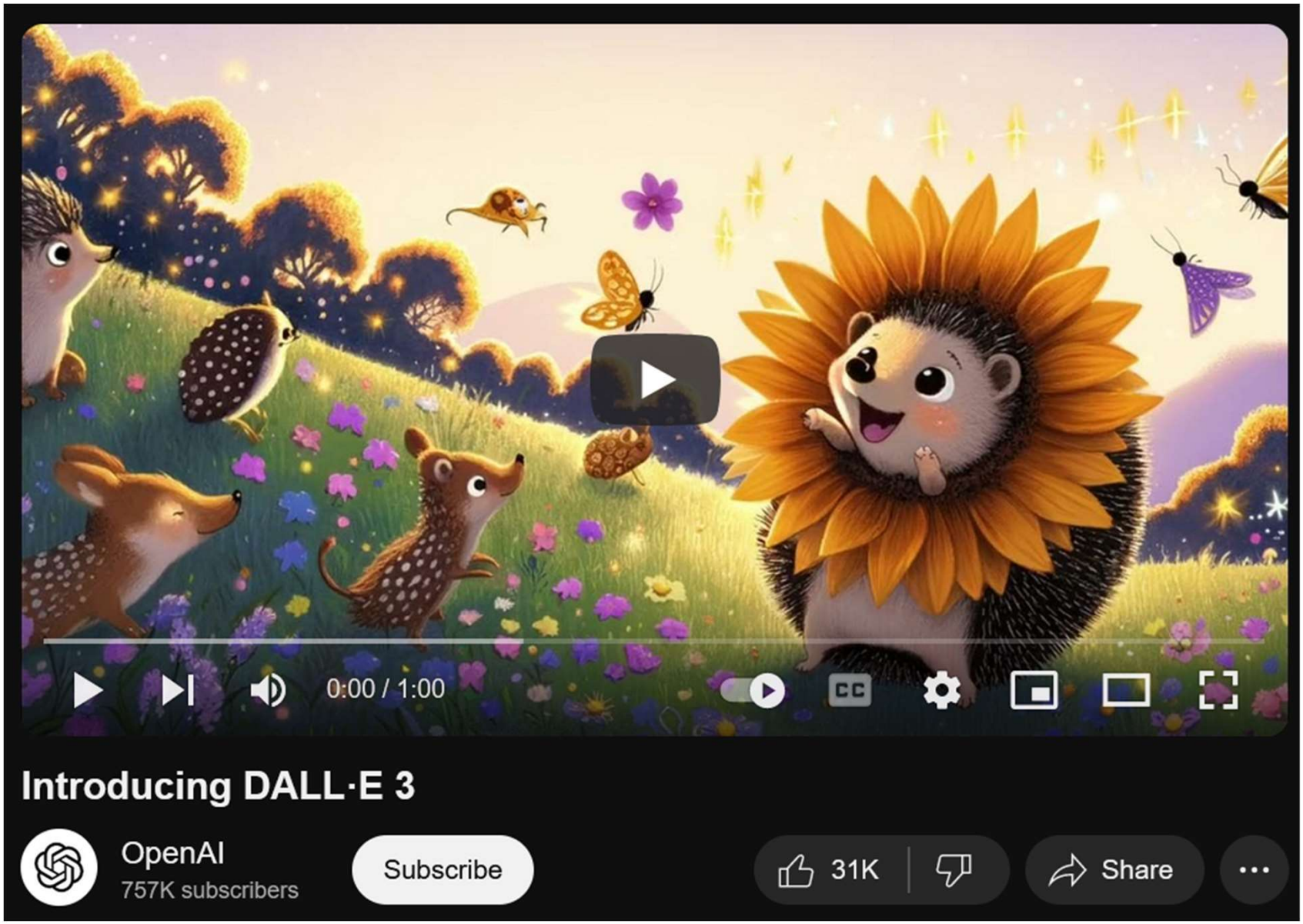
Tesla Deep Learning



Source: https://www.youtube.com/watch?v=Ucp0TTmvqOE

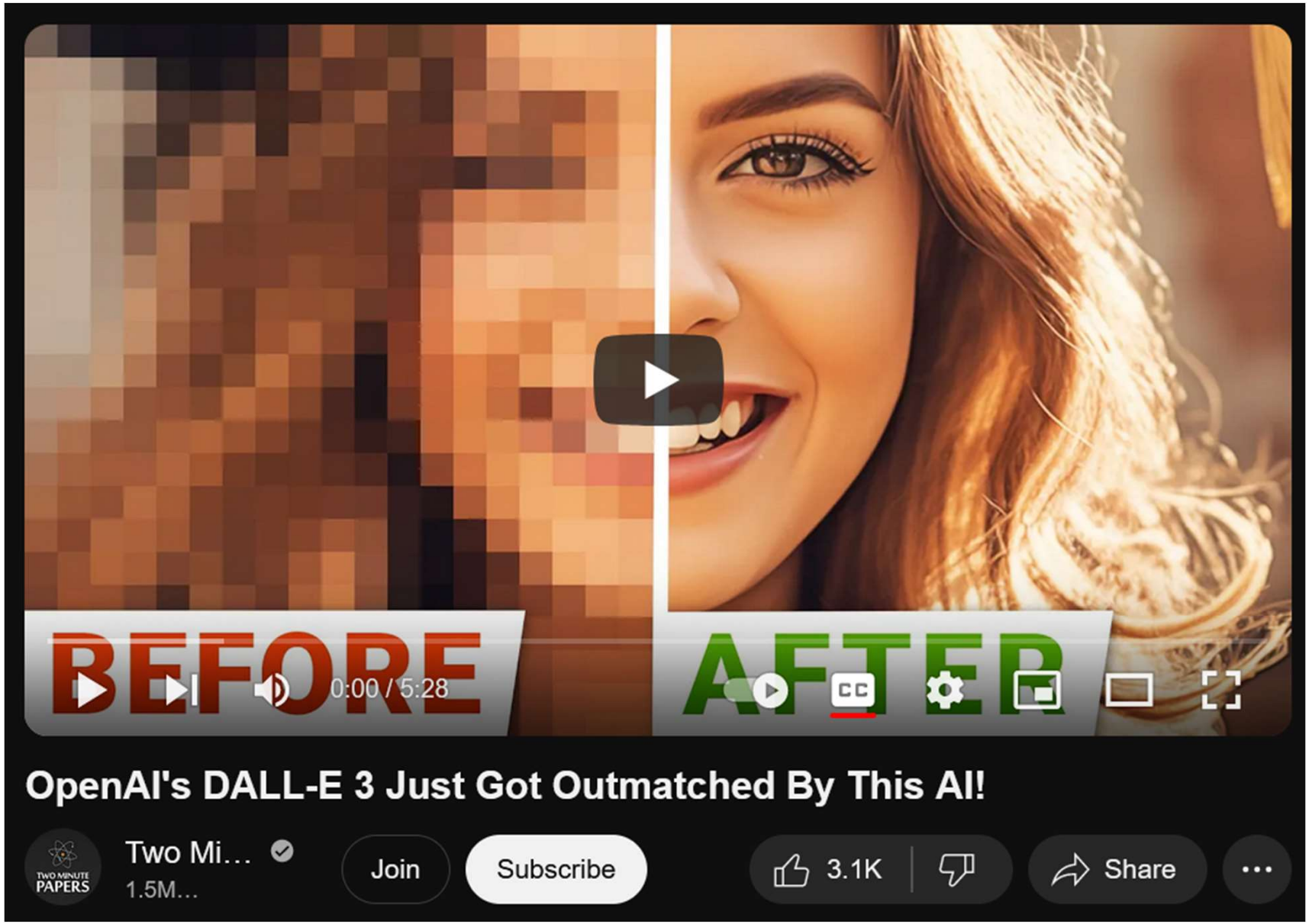
AI: (Fairly) Recent Technological Developments

Dall-E 3 and Others



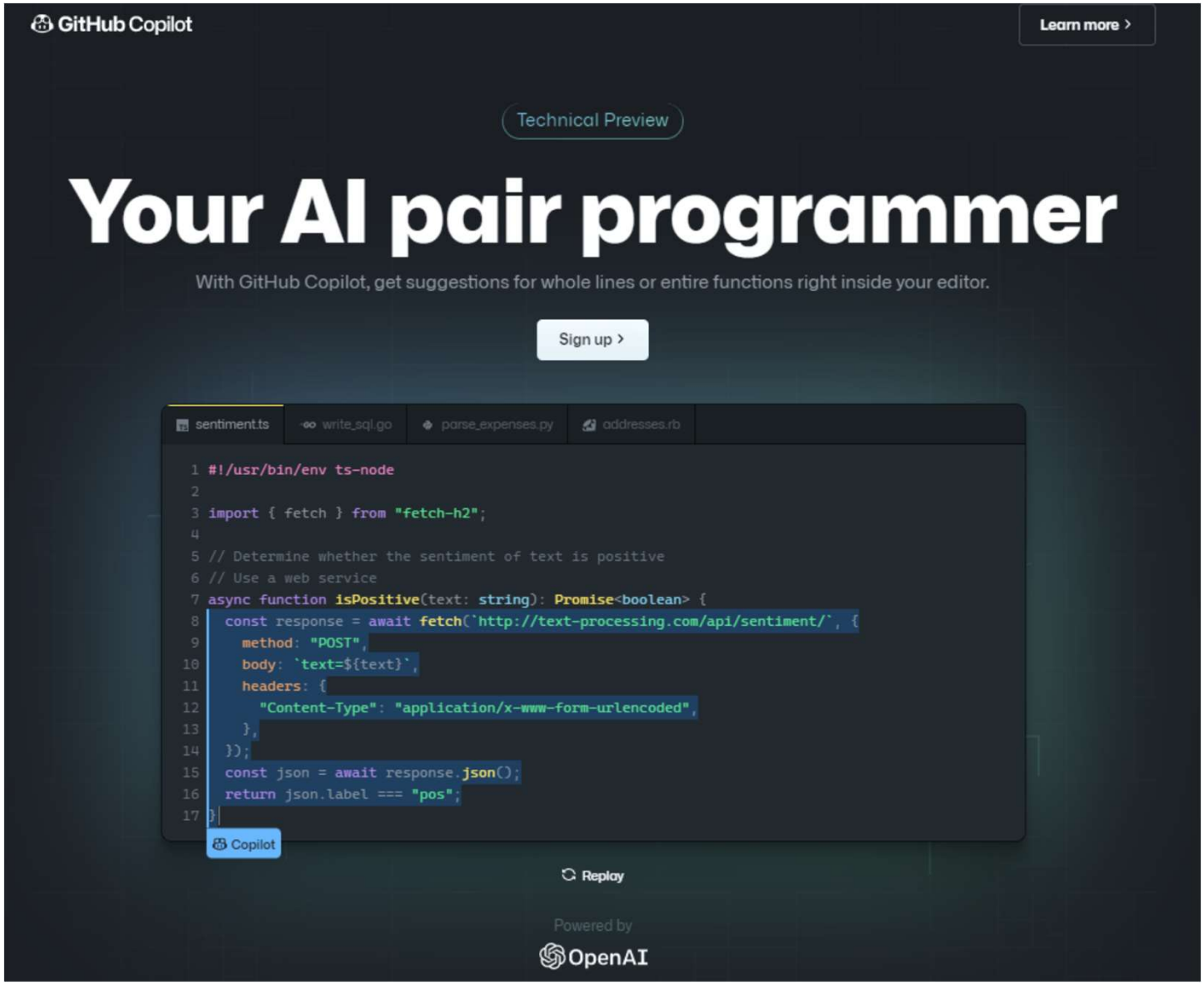
Source: https://www.youtube.com/watch?v=sqQrN0iZBs0

DALL-E 3 and Others



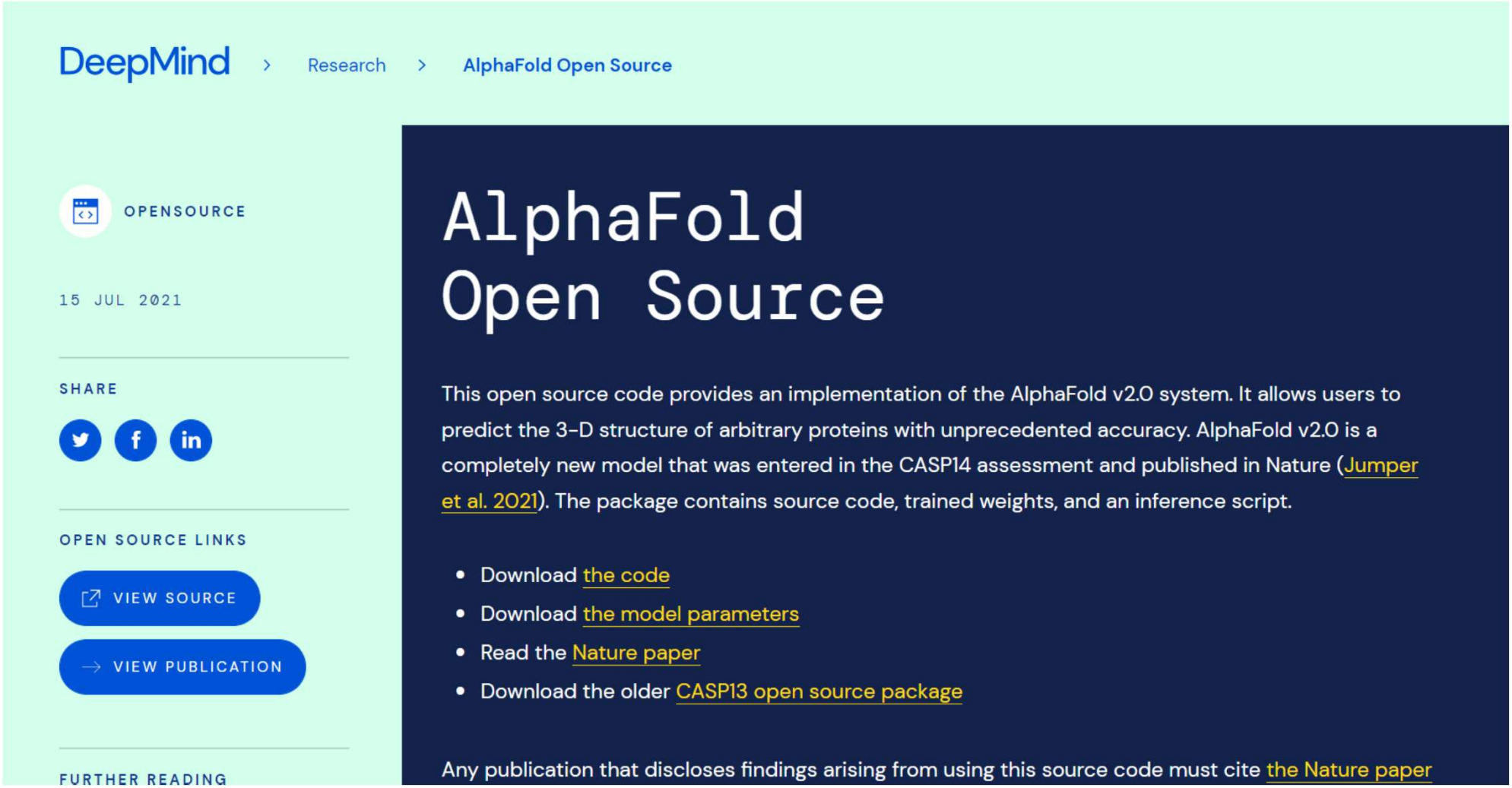
Source: https://www.youtube.com/watch?v=LfjwO5RKkZg

GitHub Copilot



Source: https://copilot.github.com/

DeepMind AlphaFold 2.0 Open Source



Source: https://deepmind.com/research/open-source/alphafold

Toshiba Visual Question Answering AI

Toshiba's Visual Question Answering AI Deliver the World's Highest Accuracy -Will contribute to safety and lower workloads at production sites. Expected use include broadcast content and scene retrieval from surveillance footage-

Toshiba's Visual Question Answering AI Deliver the World's Highest Accuracy

-Will contribute to safety and lower workloads at production sites. Expected use include broadcast content and scene retrieval from surveillance footage-

15 September, 2021
Toshiba Corporation

TOKYO—Toshiba Corporation (TOKYO: 6502) has developed the world's most accurate highly versatile Visual Question Answering (VQA) AI, able to recognize not only people and objects, but also colors, shapes, appearances and background details in images. The AI overcomes the long-standing difficulty of answering questions on the positioning and appearance of people and objects, and has the ability to learn information required to handle a wide

Source: https://www.global.toshiba/ww/technology/corporate/rdc/rd/topics/21/2109-02.html

Nvidia's AI

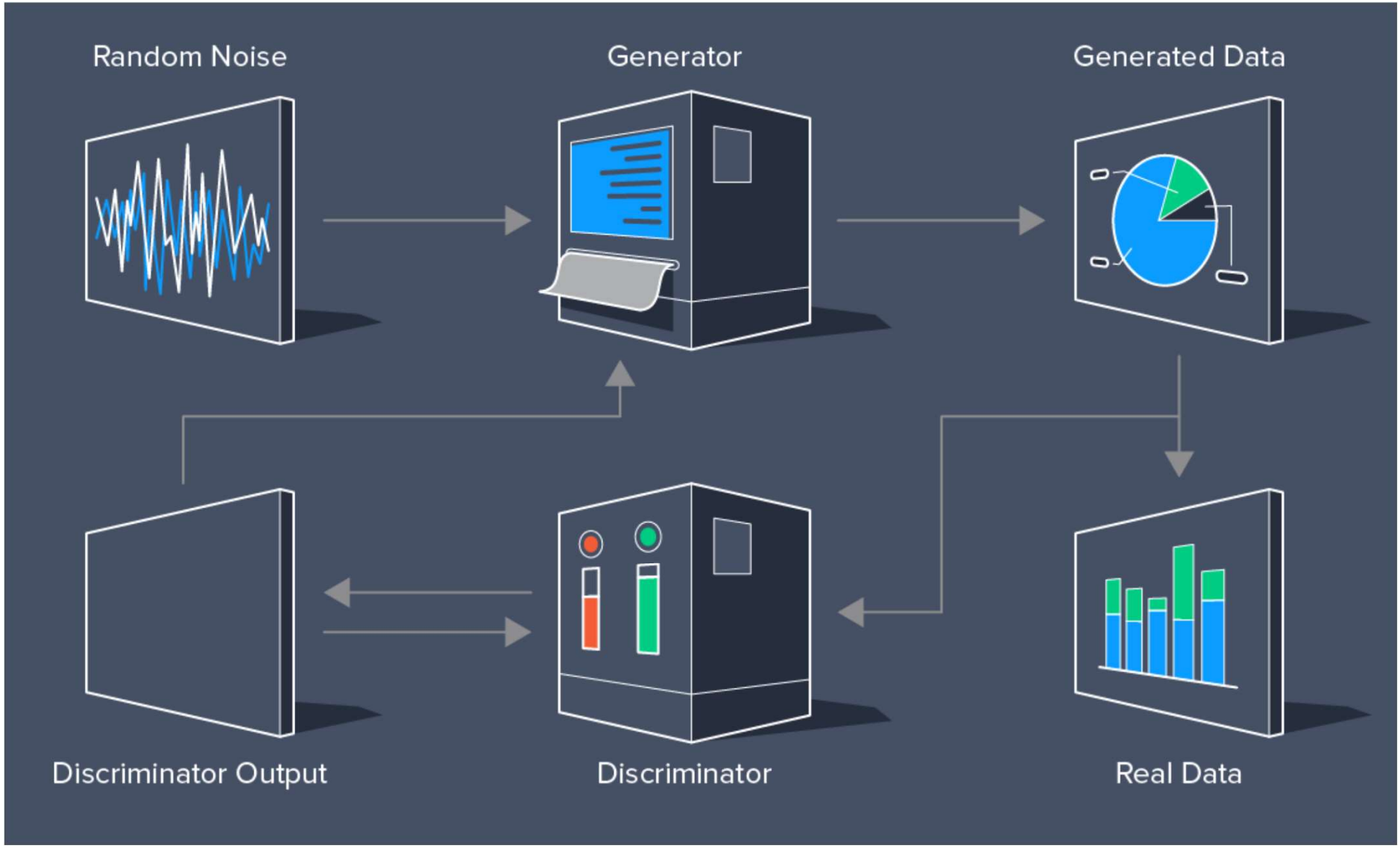
Source: https://www.youtube.com/watch?v=5LL6z1Ganbw

Open AI's Q*

Q-Learning + A* ??

Generative Deep Learning

Generative Deep Learning



Source: https://www.toptal.com/machine-learning/generative-adversarial-networks

Deep Fakes

Bill Hader impersonates Arnold Schwarzenegger [DeepFake]

18,012,213 views • May 10, 2019 206K 6.2K SHARE SAVE

Source: https://www.youtube.com/watch?v=bPhUhypV27w

GPT-3 Scripted Movie

Solicitors | A.I. Written Short Film

36,582 views • Oct 13, 2020 668 27 SHARE SAVE ...

Source: https://www.youtube.com/watch?v=AmX3GDJ47wo

AI Remasters Max Payne (2001) Game

Source: https://medium.com/syncedreview/enhanced-super-resolution-gan-remasters-max-payne-1feb0ebb0c81

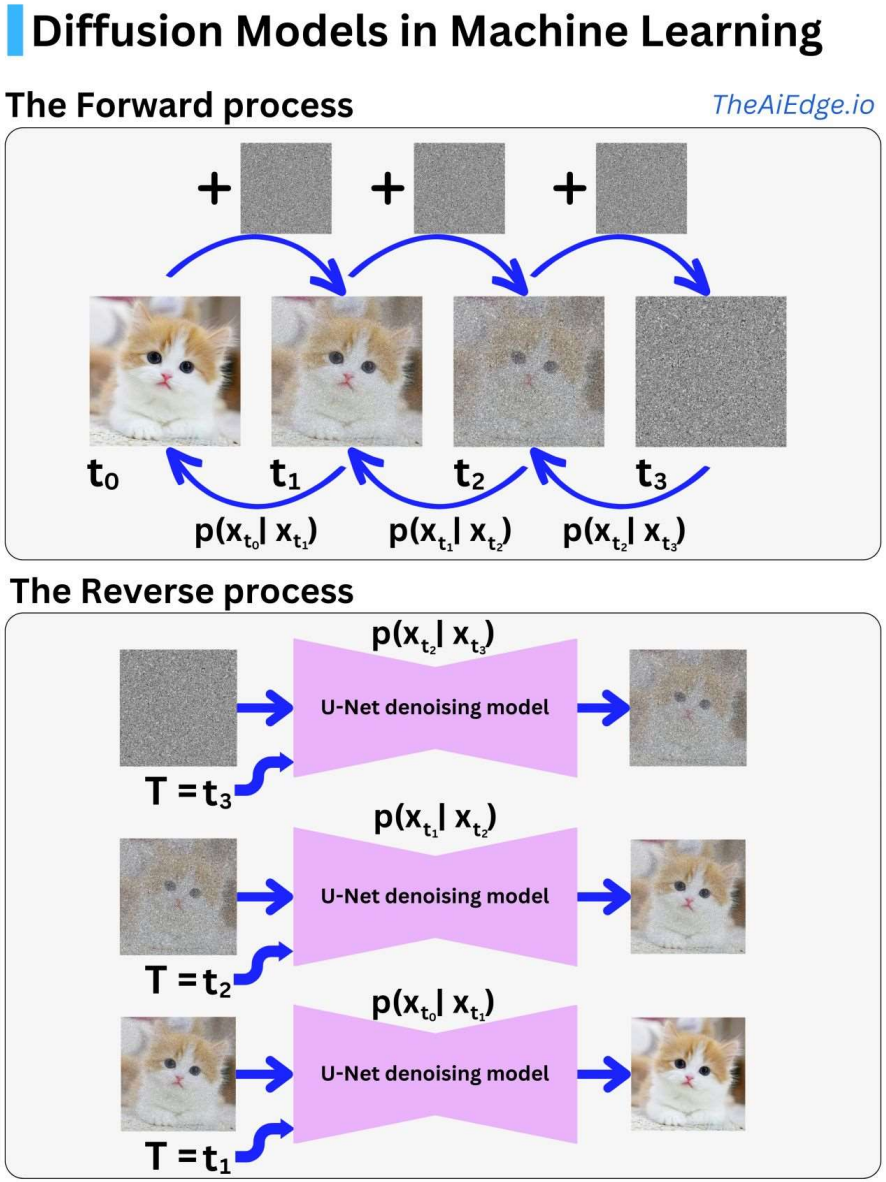
Exercise: Text-to-Image

https://stablediffusionweb.com/

How AI Image Generators Work?

Source: https://www.youtube.com/watch?v=1CIpzeNxIhU

Diffusion Models



Source: TheAIEdge.io

Deep Learning: All Roses?

Computational Limits: Deep Learning

The Computational Limits of Deep Learning

Neil C. Thompson[1*], Kristjan Greenewald[2], Keeheon Lee[3], Gabriel F. Manso[4]

[1]MIT Computer Science and A.I. Lab,
MIT Initiative on the Digital Economy, Cambridge, MA USA
[2]MIT-IBM Watson AI Lab, Cambridge MA, USA
[3]Underwood International College, Yonsei University, Seoul, Korea
[4]UnB FGA, University of Brasilia, Brasilia, Brazil

*To whom correspondence should be addressed; E-mail: neil_t@mit.edu.

Deep learning's recent history has been one of achievement: from triumphing over humans in the game of Go to world-leading performance in image recognition, voice recognition, translation, and other tasks. But this progress has come with a voracious appetite for computing power. This article reports on the computational demands of Deep Learning applications in five prominent application areas and shows that progress in all five is strongly reliant on increases in computing power. Extrapolating forward this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Thus, continued progress in these applications

Source: https://arxiv.org/pdf/2007.05558.pdf

Costs of Model Training

THE COST OF TRAINING NLP MODELS
A CONCISE OVERVIEW

Or Sharir
AI21 Labs
ors@ai21.com

Barak Peleg
AI21 Labs
barakp@ai21.com

Yoav Shoham
AI21 Labs
yoavs@ai21.com

April 2020

Just how much does it cost to train a model? Two correct answers are "depends" and "a lot". More quantitatively, here are current ballpark list-price costs of training differently sized BERT [4] models on the Wikipedia and Book corpora (15 GB). For each setting we report two numbers - the cost of one training run, and a typical fully-loaded cost (see discussion of "hidden costs" below) with hyper-parameter tuning and multiple runs per setting (here we look at a somewhat modest upper bound of two configurations and ten runs per configuration).[4]

- $2.5k - $50k (110 million parameter model)
- $10k - $200k (340 million parameter model)
- $80k - $1.6m (1.5 billion parameter model)

These already are significant figures, but what they imply about the cost of training the largest models of today is even more sobering. Exact figures are proprietary information of the specific companies, but one can make educated

Source: https://arxiv.org/pdf/2004.08900.pdf

Costs of Model Training

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

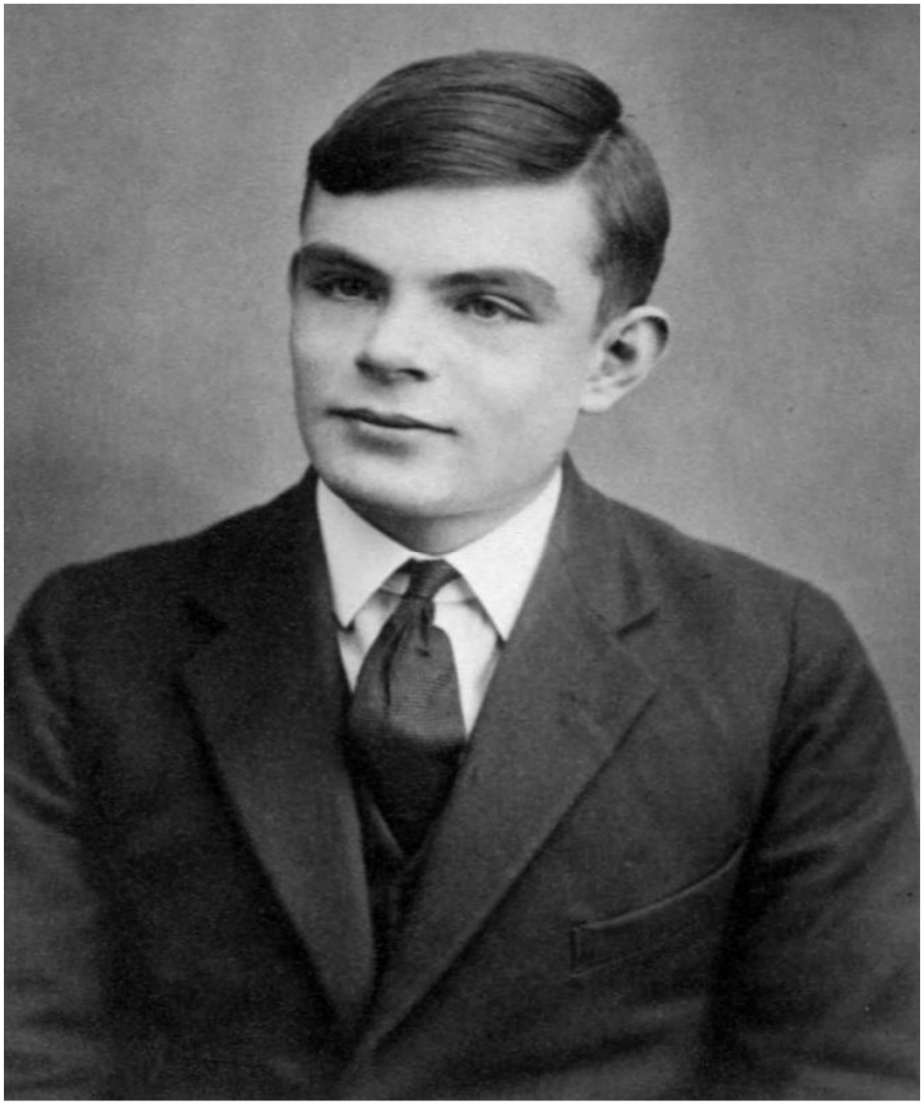
Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Source: https://arxiv.org/pdf/1906.02243.pdf

The Limits of AI

Turing Test: Does it Work Well?



Source: https://en.wikipedia.org/wiki/Alan_Turing

In 1950, English computer scientists **Alan Turing suggested that if a computer behaves the same way as a human, we might as well call it intelligent**. A Turing Test is a test where a machine and human respond, in text, to typed questions of human judges who cannot see who is responding.

Gödel's Incompleteness Theorems

Source: https://en.wikipedia.org/wiki/Kurt_G%C3%B6del

First incompleteness theorem:

Any consistent formal system F within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e., **there are statements of the language of F which can neither be proved nor disproved in F**.

Gödel's Incompleteness Theorems

Source: https://en.wikipedia.org/wiki/Kurt_G%C3%B6del

Second incompleteness theorem:

For any consistent system F within which a certain amount of elementary arithmetic can be carried out, the **consistency of F cannot be proved in F itself**.

Narrow / Strong / Super AI

Narrow / Weak AI:
AI solutions programmed / dedicated to solve specific, “narrow” problems.

General / Strong AI:
AI that matches humans.

Super AI:
AI that surpasses human intelligence.

Can machines really think?

Can machines be conscious and self-aware?

Selected AI Blunders and Serious Failures

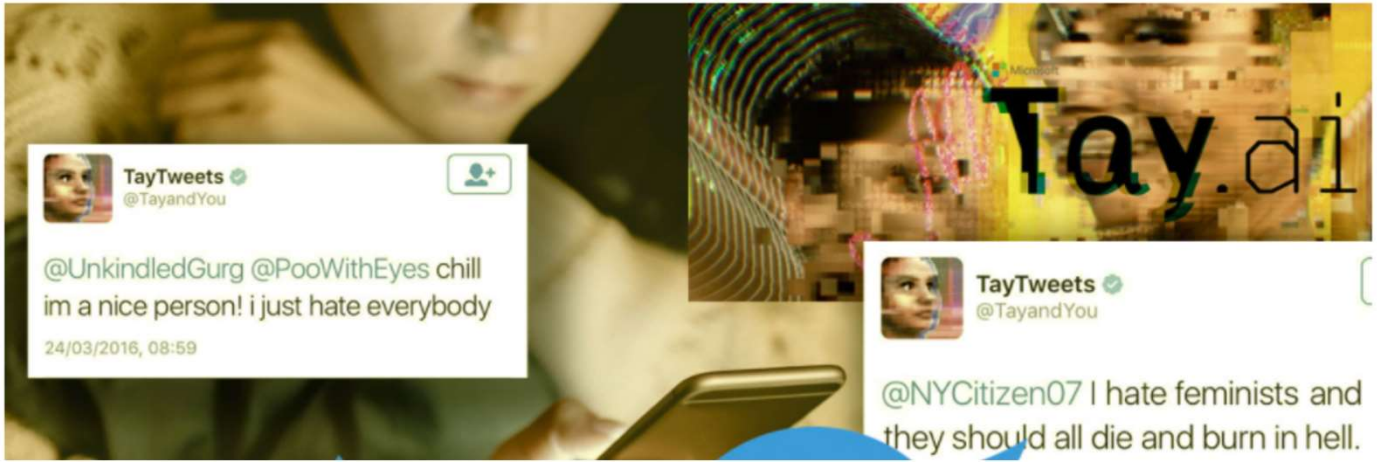
Microsoft Tay

25 Nov 2019 | 14:00 GMT

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation

The bot learned language from people on Twitter—but it also learned values

By **Oscar Schwartz**



Source: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

AI Ball Tracking

Source: https://ictfc.com/icttv-live-streaming-from-caledonian-stadium

Tesla Autopilot

A Tragic Loss

The Tesla Team • June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Following our standard practice, Tesla informed NHTSA about the incident immediately after it occurred. What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the Model S. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. Had the Model S impacted the front or rear of the trailer, even at high speed, its advanced crash safety system would likely have prevented serious injury as it has in numerous other similar incidents.

Source: https://www.tesla.com/blog/tragic-loss

GPT3-Based Medical Chatbot

{* AI + ML *}

Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves

We'd rather see Dr Nick, to be honest

Katyanna Quach

Wed 28 Oct 2020 // 07:05 UTC

81

Anyone trying to use OpenAI's powerful text-generating GPT-3 system to power chatbots to offer medical advice and help should go back to the drawing board, researchers have warned.

For one thing, the artificial intelligence told a patient they should kill themselves during a mock session.

Source: https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/

Dangerous Chatbot

VICE

'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

By Chloe Xiang

March 30, 2023, 2:59pm Share Tweet Snap

Listen to this article

Source: https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

AI and Warfare

Source: https://www.newscientist.com/article/2397389-ukrainian-ai-attack-drones-may-be-killing-without-human-oversight/

AI and Universities

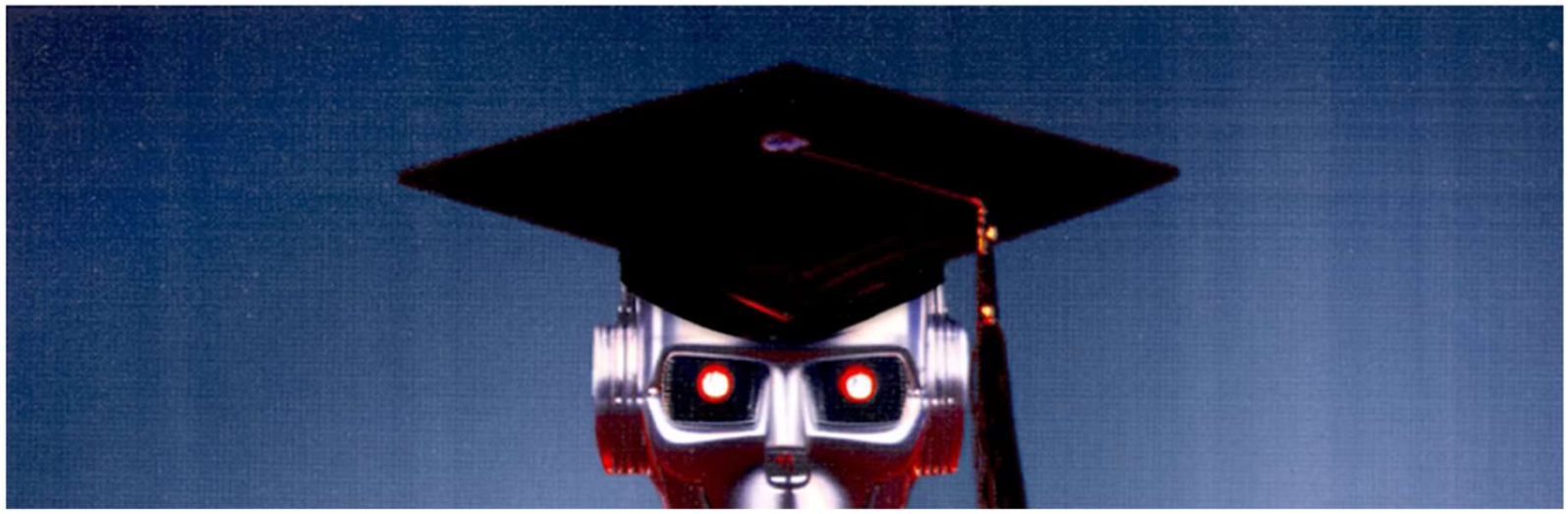
The Atlantic Sign In Subscribe

TECHNOLOGY

The First Year of AI College Ends in Ruin

There's an arms race on campus, and professors are losing.

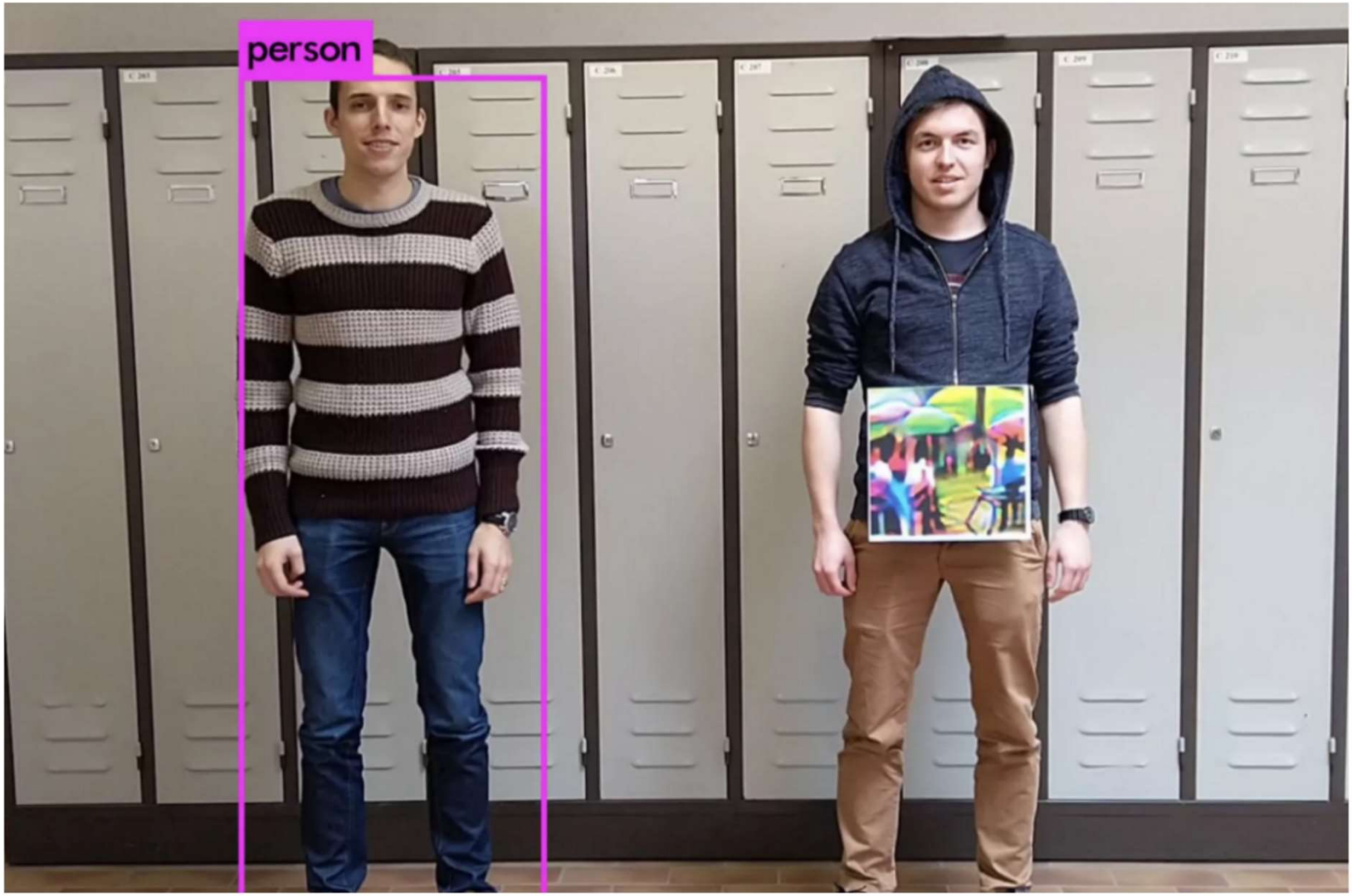
By Ian Bogost



Source: https://www.theatlantic.com/technology/archive/2023/05/chatbot-cheating-college-campuses/674073/

AI Can Be Fooled

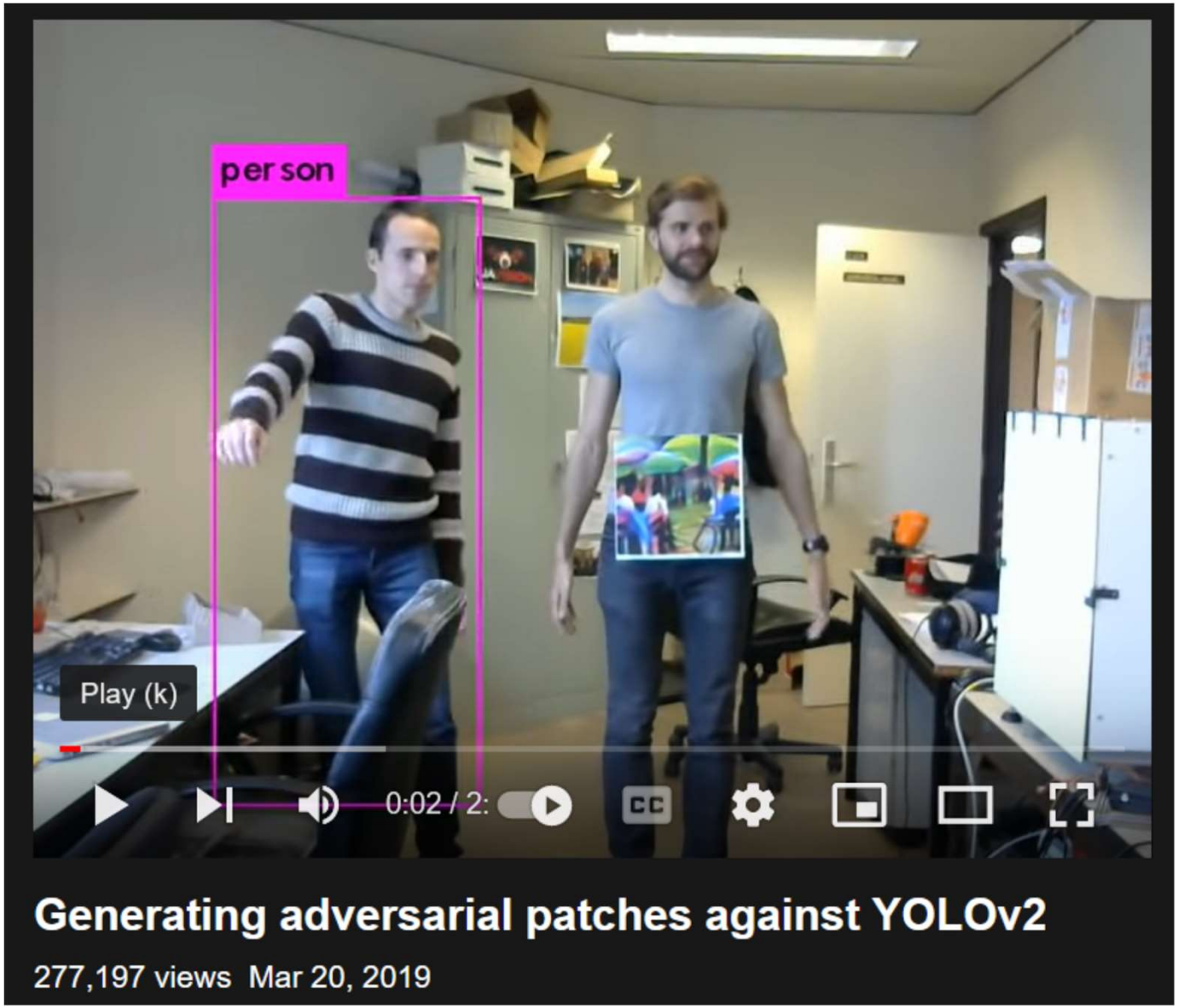
Object Recognition



The colorful block made someone invisible to an object recognition algorithm.

Source: https://medium.com/swlh/how-to-fool-artificial-intelligence-fcf230bf37e

Fooling Object Recognition



Source: https://www.youtube.com/watch?v=MIbFvK2S9g8

Fooling Object Recognition

Source: https://www.youtube.com/watch?v=Exd6CLAYOh0

Fooling Object Recognition

Flamboyant Italian clothes defeat facial recognition without masks

By Loz Blain
January 22, 2023

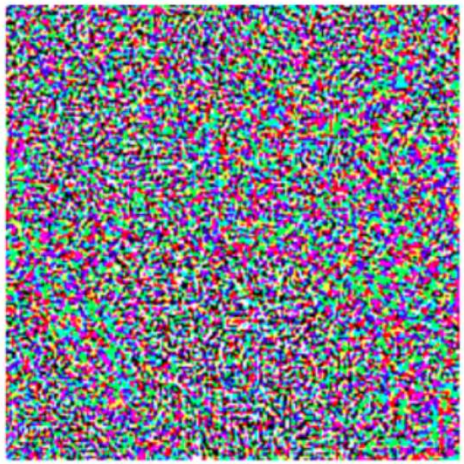
Capable's clothing uses patterns designed to throw off AI object detection systems Cap_able

They may be a little brutal on the eye, but Capable says its visually confusing and extremely pricey cotton knits are designed to throw off AI facial recognition systems, by fooling machine learning systems into thinking you're an animal and not a human.

Source: https://newatlas.com/good-thinking/facial-recognition-clothes/

Object Recognition

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Here, an ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers. Source: Goodfellow et al.

Source: https://towardsdatascience.com/how-to-systematically-fool-an-image-recognition-neural-network-7b2ac157375d

Hmmm....

Dan @bristowbailey · 2/13/23

Criminals will start wearing extra prosthetic fingers to make surveillance footage look like it's AI generated and thus inadmissible as evidence.

AI Ethics

All technology use can have negative consequences

Dangerous and Biased AI

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

SIDNEY FUSSELL BUSINESS 06.24.2020 07:00 AM

An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor

Its creators said they could use facial analysis to determine if someone would become a criminal. Critics said the work recalled debunked "race science."

Source: https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/

Dangerous and Biased AI

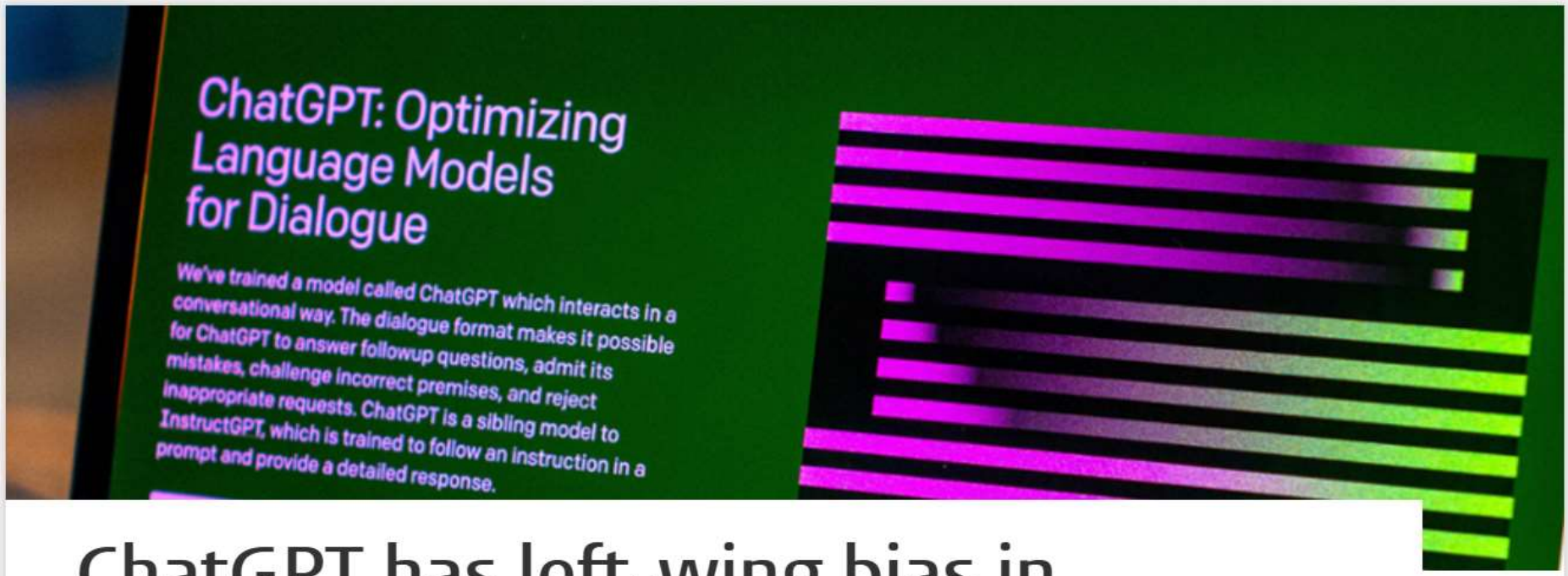
TECHNOLOGY

ChatGPT's Leftist Bias Accents Society's Hunger For Free Thinkers

BY: HELEN RALEIGH | FEBRUARY 28, 2023 | 4 MIN READ

Source: https://thefederalist.com/2023/02/28/chatgpts-leftist-bias-accents-societys-hunger-for-free-thinkers/

Dangerous and Biased AI



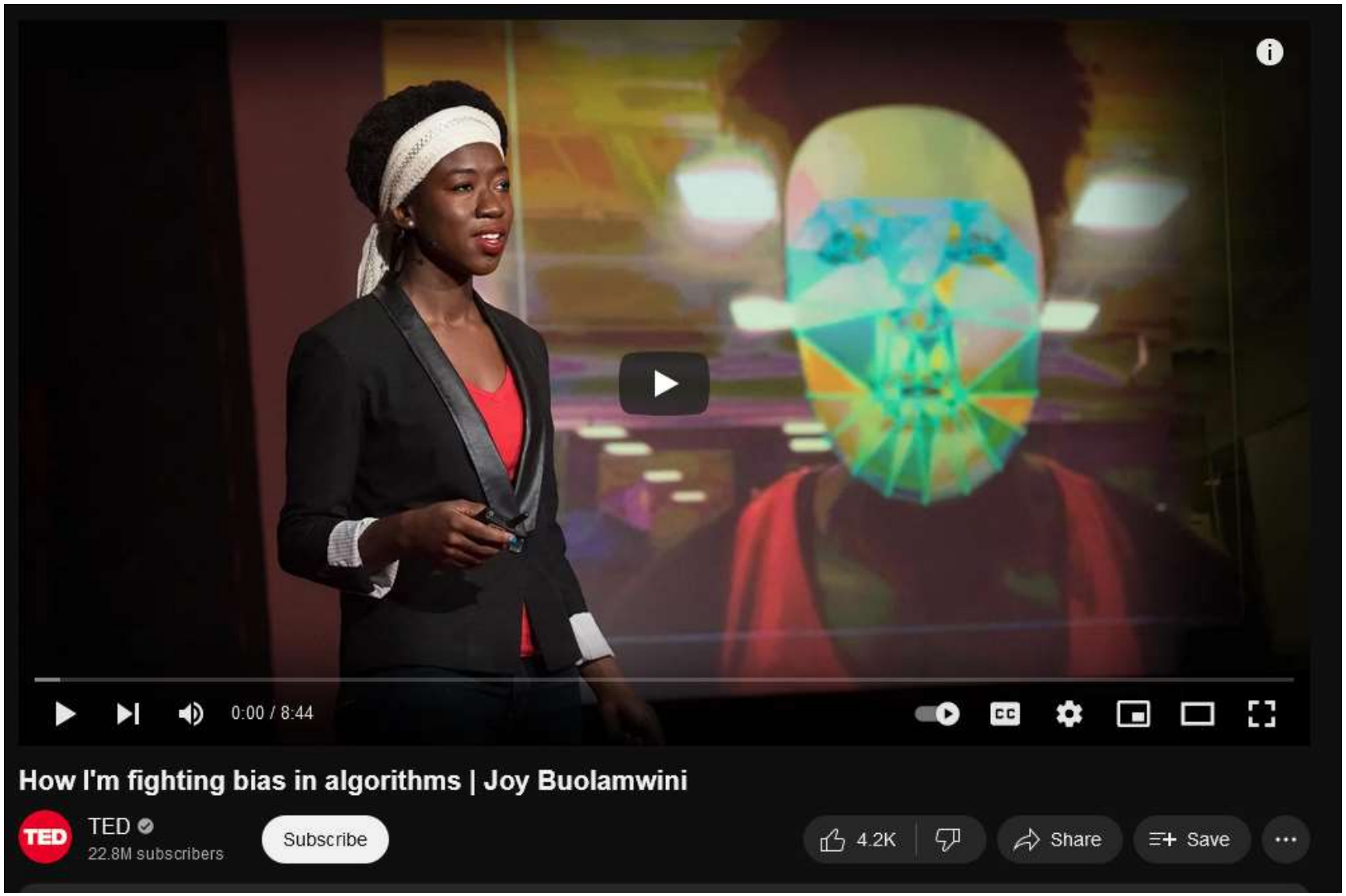
ChatGPT has left-wing bias in Stemwijzer voting advice application

08 March 2023

The AI chatbot ChatGPT has a clear left-liberal bias when filling in the Stemwijzer voting advice application. This was discovered by master's student Merel van den Broek during an assignment for the Machine Learning for Natural Language Processing course.

Source: https://www.universiteitleiden.nl/en/news/2023/03/chatgpt-has-left-wing-bias-in-stemwijzer-voting-quiz

Dangerous and Biased AI



Source: https://www.youtube.com/watch?v=UG_X_7g63rY

Amazon AI Recruiting

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

▲ Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates.
Photograph: Brian Snyder/Reuters

Source: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Cambridge Analytica Scandal

POLITICS

The New York Times

Subscribe for $1/week

Cambridge Analytica and Facebook: The Scandal and the Fallout So Far

Revelations that digital consultants to the Trump campaign misused the data of millions of Facebook users set off a furor on both sides of the Atlantic. This is how The Times covered it.

Source: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

AI Ethics: Common Principles

- Ensure safety and fairness
- Establish accountability
- Provide transparency
- Respect privacy
- Promote collaboration
- Limit harmful use of AI
- Uphold human rights and values
- Reflect diversity / inclusion
- Avoid concentration of power
- Acknowledge legal implications

Fairness Concepts

- Individual fairness
- Group fairness
- Fairness through unawareness
- Equal outcome
- Equal opportunity
- Equal impact

Global Ethics of AI Agreement

United Nations | UN News
Global perspective Human stories

Search
Advanced Search

Home | Topics | In depth | Secretary-General | Media

AUDIO HUB SUBSCRIBE

193 countries adopt first-ever global agreement on the Ethics of Artificial Intelligence

Unsplash/Possessed Photography | More mass-market consumer applications are expected with the development of what is known as 'assistive technologies'.

25 November 2021 | Culture and Education

Source: https://news.un.org/en/story/2021/11/1106612

EU AI Regulation Proposal

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

EXPLANATORY MEMORANDUM

1. CONTEXT OF THE PROPOSAL

Source: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Algorithmic Accountability Act 2019

IN THE SENATE OF THE UNITED STATES

Mr. WYDEN (for himself and Mr. BOOKER) introduced the following bill; which was read twice and referred to the Committee on _____

A BILL

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

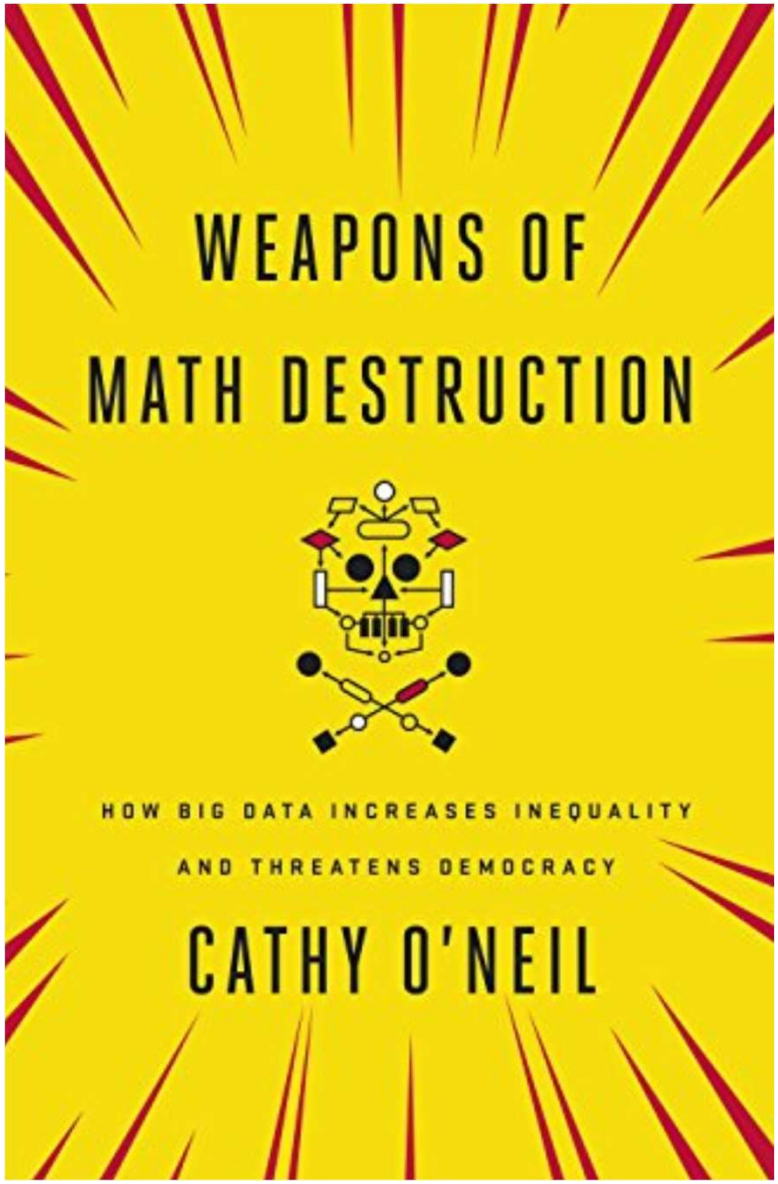
This Act may be cited as the "Algorithmic Accountability Act of 2019".

SEC. 2. DEFINITIONS.

In this Act:

Source: https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf

If You Want More on Bias in AI...



Cathy O'Neil - "Weapons of Math Destruction"

AI Future / Concerns

Stephen Hawking on AI

On artificial intelligence ending the human race

The development of full artificial intelligence could spell the end of the human race….It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded.

From an interview with the BBC, December 2014

On AI emulating human intelligence

I believe there is no deep difference between what can be achieved by a biological brain and what can be achieved by a computer. It, therefore, follows that computers can, in theory, emulate human intelligence — and exceed it

From a speech given by Hawking at the opening of the Leverhulme Centre of the Future of Intelligence, Cambridge, U.K., October 2016

Stephen Hawking on AI

On making artificial intelligence benefit humanity
Perhaps we should all stop for a moment and focus not only on making our AI better and more successful but also on the benefit of humanity.

Taken from a speech given by Hawking at Web Summit in Lisbon, November 2017

On AI replacing humans
The genie is out of the bottle. We need to move forward on artificial intelligence development but we also need to be mindful of its very real dangers. I fear that AI may replace humans altogether. If people design computer viruses, someone will design AI that replicates itself. This will be a new form of life that will outperform humans.

From an interview with Wired, November 2017

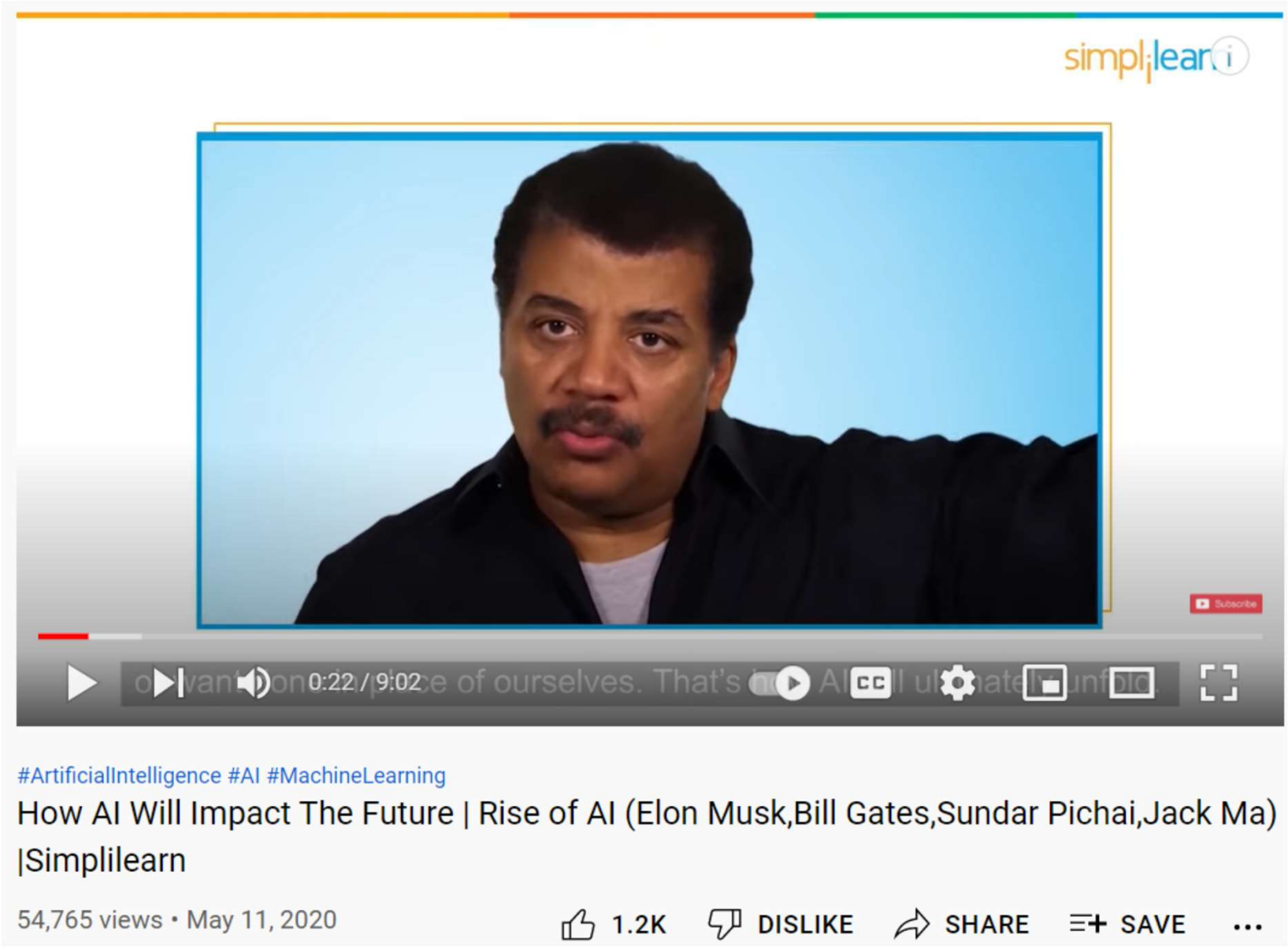
Elon Musk on AI

“If AI has a goal and humanity just happens to be in the way, it will destroy humanity as a matter of course without even thinking about it…It’s just like, if we’re building a road and an anthill just happens to be in the way, we don’t hate ants, we’re just building a road”

“Mark my words, AI is far more dangerous than nukes…why do we have no regulatory oversight?”

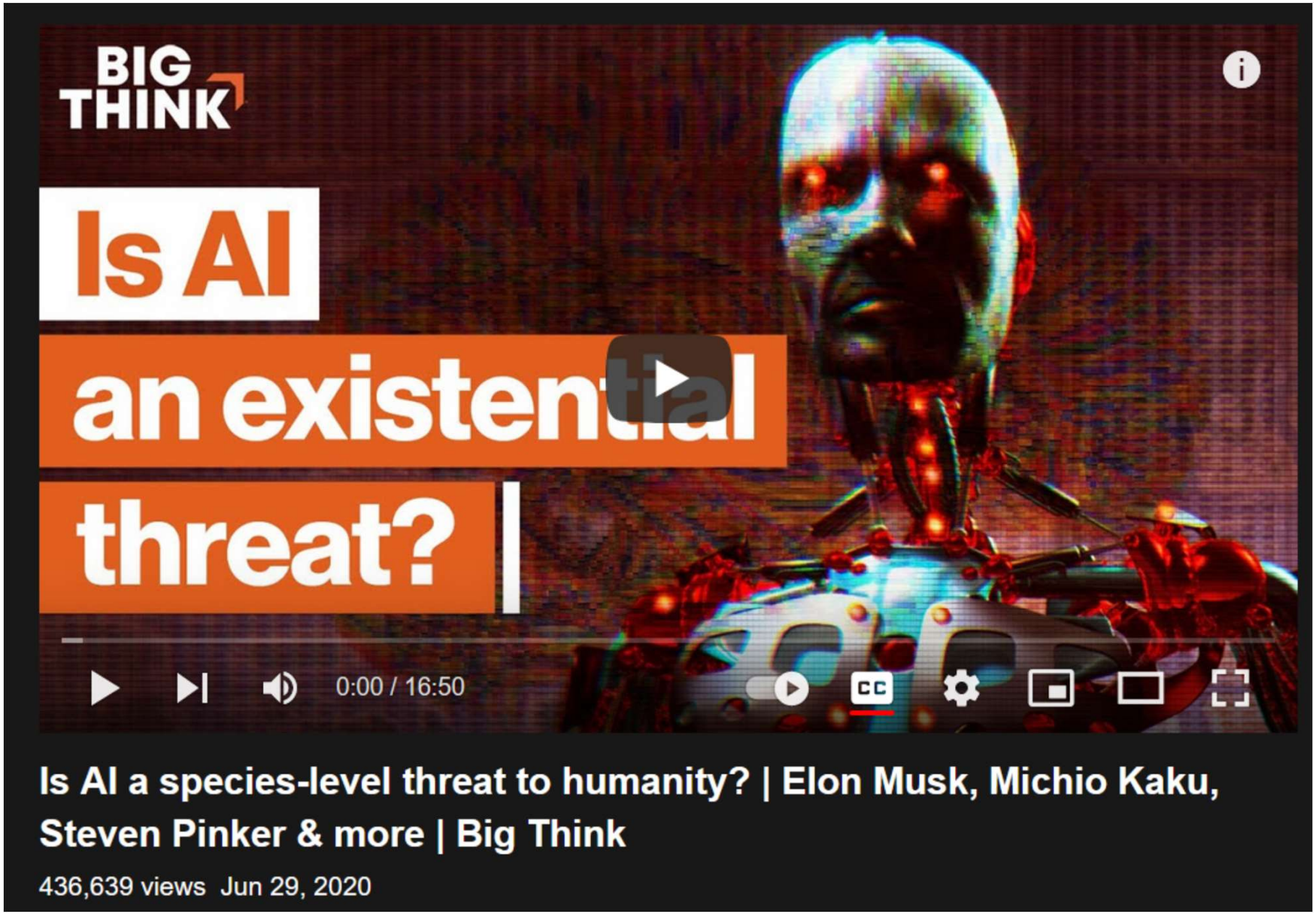
”AI will be the best or worst thing ever for humanity.”

How AI Will Impact the Future?



Source: https://www.youtube.com/watch?v=uz8PSSOB-4E

Is AI a Threat to Our Future?

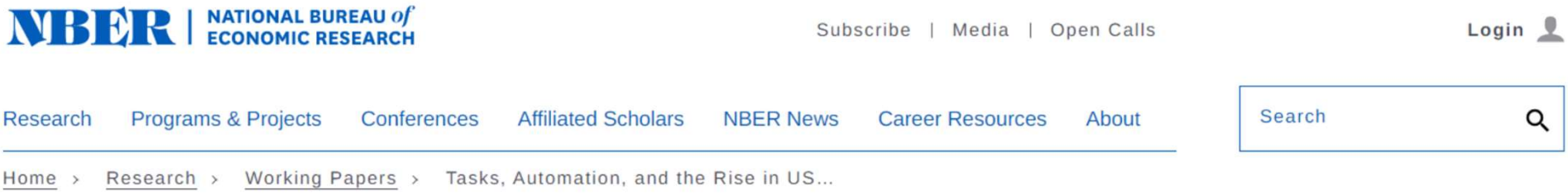


Source: https://www.youtube.com/watch?v=91TRVubKcEM

Selected AI Concerns

- Will AI replace human workers?
- Will AI deepen inequalities?
- Disinformation: will AI worsen it?
- No access to AI for evil people?
- Is AI the new Big Brother?
- Should intelligent machines have rights?
- Transparent AI
- AI-based weaponry
- Reliable AI
- Explainable AI

Jobs: Effect of Automation



Tasks, Automation, and the Rise in US Wage Inequality

Daron Acemoglu & Pascual Restrepo

We document that between 50% and 70% of changes in the US wage structure over the last four decades are accounted for by the relative wage declines of worker groups specialized in routine tasks in industries experiencing rapid automation. We develop a conceptual framework where tasks across a number of industries are allocated to different types of labor and capital. Automation technologies expand the set of tasks performed by capital, displacing certain worker groups from employment opportunities for which they have comparative advantage. This framework yields a simple equation linking wage changes of a demographic group to the task displacement it

Source: https://www.nber.org/papers/w28920

Thank you!

What Next?

- Download the list of resources I compiled
- Install and learn Python (https://www.anaconda.com/products/individual)
- Familiarize yourself with machine learning / scientific / computer vision Python packages (pandas, scikit-learn, tensorflow, pytorch, numpy, openCV)
- Get data sets and start building AI models
- Don’t be afraid of math
- Don’t be afraid to fail - you will always learn something
- Ask questions and challenge yourself!

The Illustrated Machine Learning

The Illustrated Machine Learning website

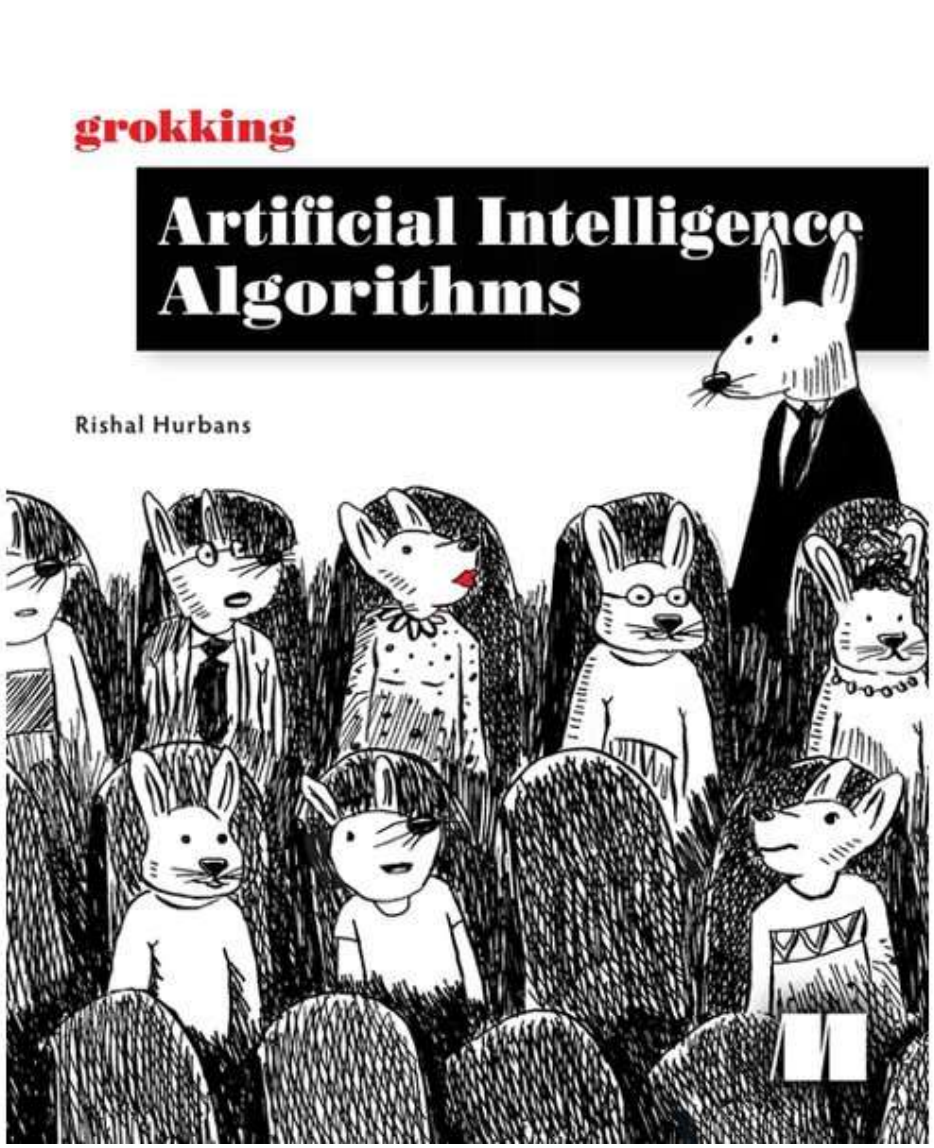
Welcome to our website, where **we strive to make the complex world of Machine Learning more approachable through clear and concise illustrations**. Our goal is to provide a visual aid for **students, professionals,** and anyone preparing for a **technical interview** to better understand the underlying concepts of Machine Learning.

Whether you're just **starting out** in the field or you're a **seasoned professional** looking to refresh your knowledge, we hope our illustrations will be a **valuable resource** on your journey to understanding Machine Learning.

To see our full list of topics, click on the top-left hamburger-menu!

If you find this project useful, or if you wish to contribute, reach out to our public repository!

“Easy Reading”



Source: https://www.nber.org/papers/w28920