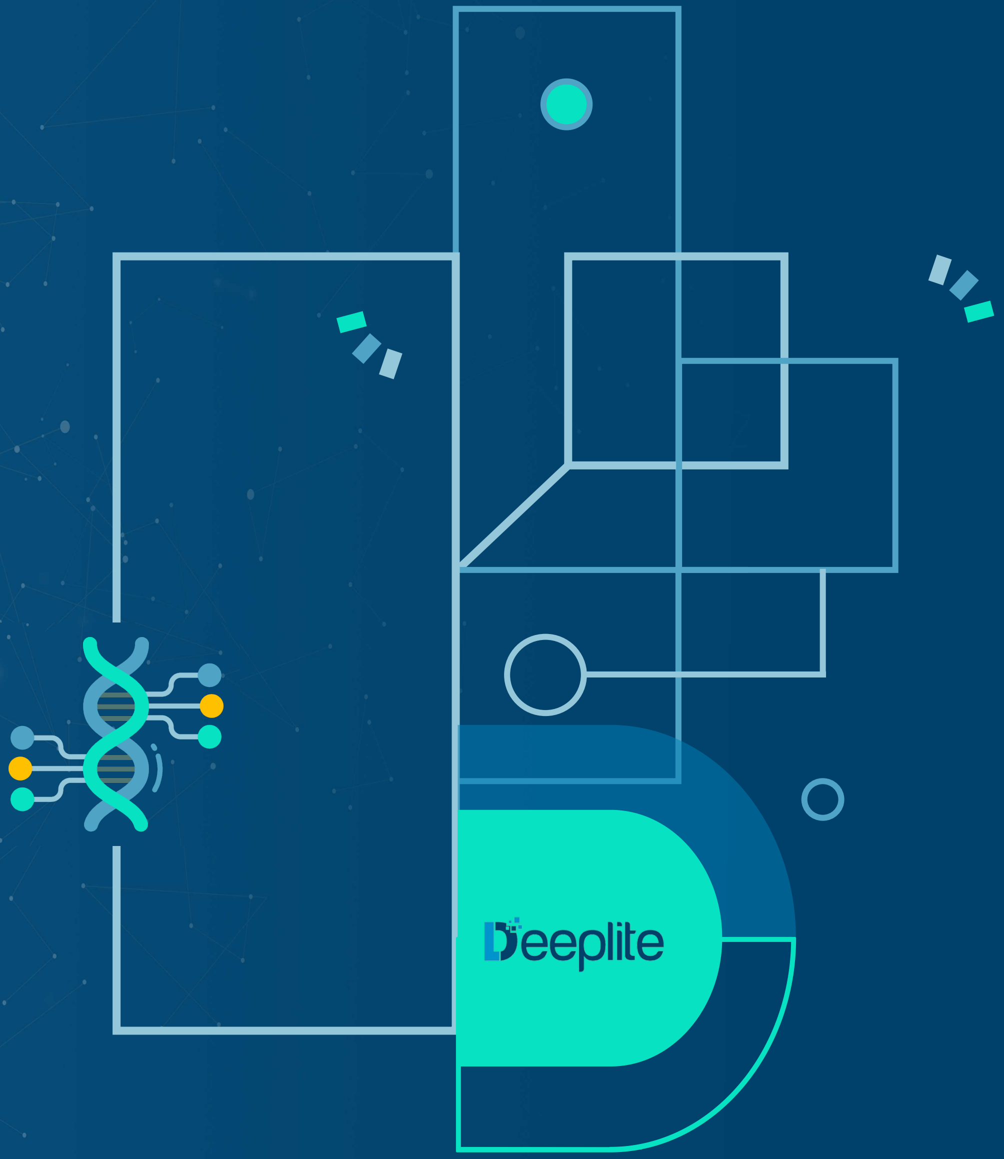
Deeplite

Optimizing AI for the Edge

Introduction to Artificial Intelligence and It's Impact on Business



April 24, 2024

Nick Romano
www.linkedin.com/in/nickromanoprofile/
nick@deeplite.ai

Cofounder and CEO of Deeplite, Nick is a serial entrepreneur and 3-time founder. He has Bachelor of Engineering and Management degree in Mechanical Engineering from McMaster University, Canada. Recently honored by McMaster's Engineering Faculty as being one of their Top 150 Distinguished Alumni for the role they've played in shaping Canada and the world.

About Me

CEO | Director | Advisor | Private Equity | SaaS | AI

Deeplite
2019

Canadian Chamber of Commerce
2023

McMaster University
2024

prinova
1998

messagepoint
2008

AI at the Edge

Opportunity and challenges

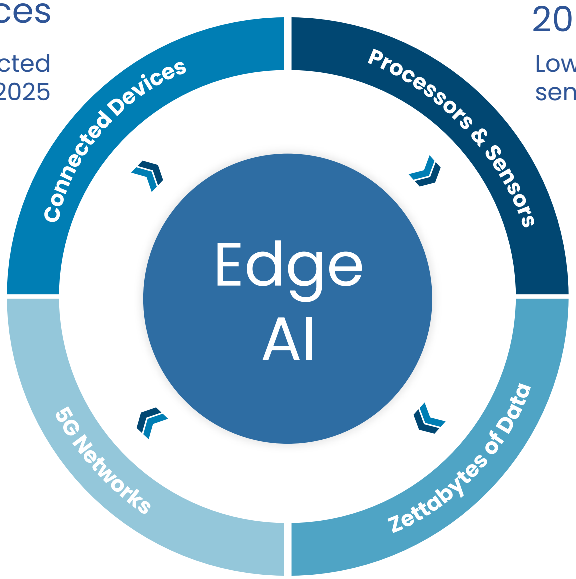
AI-Driven Optimization to make Deep Neural Networks faster, smaller and energy-efficient. **Edge will be bigger than cloud**

Billions of Devices

Over 100 billion connected edge and IoT devices by 2025

200B NPUs/CPUs/MCUs

Low-power, processors, sensors and controllers



M2M Communication

High speed connected edge networks facilitating node to node connectivity

Localized Data Capture

50% of data analyzed by neural networks will occur at the point of capture in edge systems by 2025

AI becomes untethered, decentralized and everywhere

Investors

PJC

INNOSPARK

DIFFERENTIAL VENTURES

bdc

Desjardins Capital

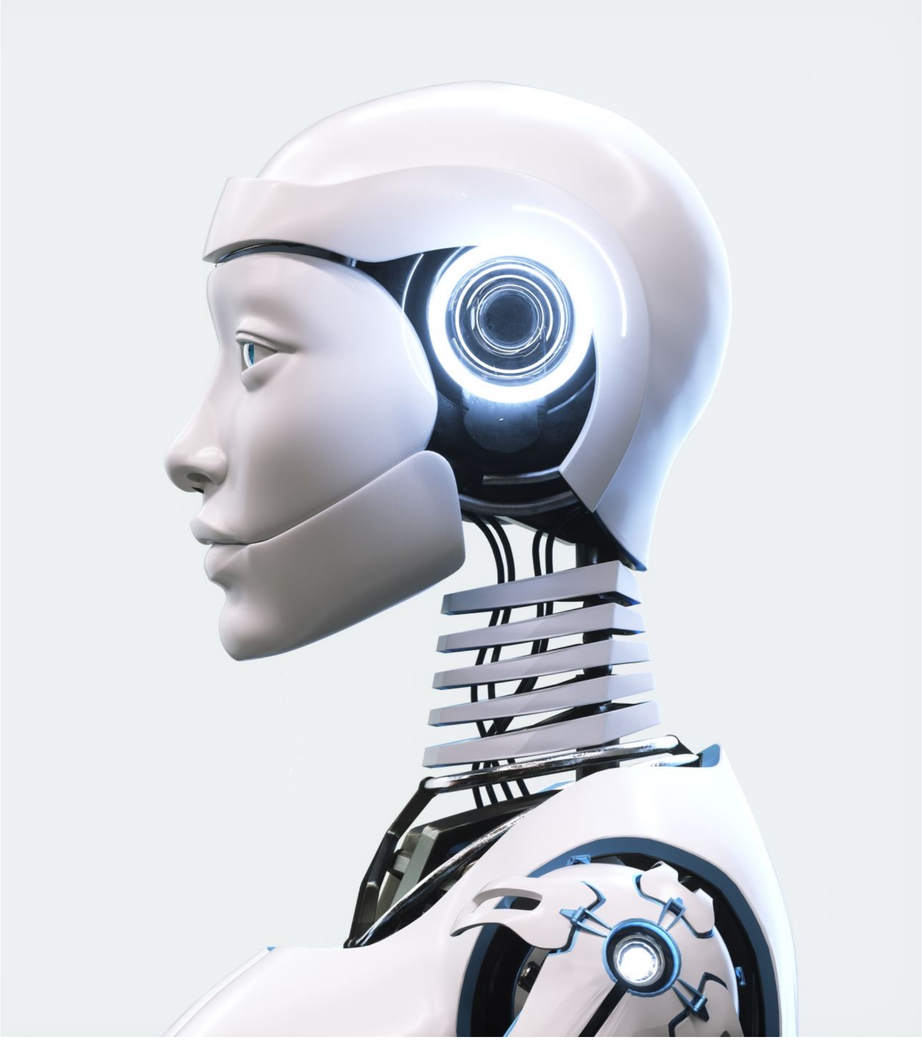
SOMEL

TANDEMLAUNCH

What is AI?

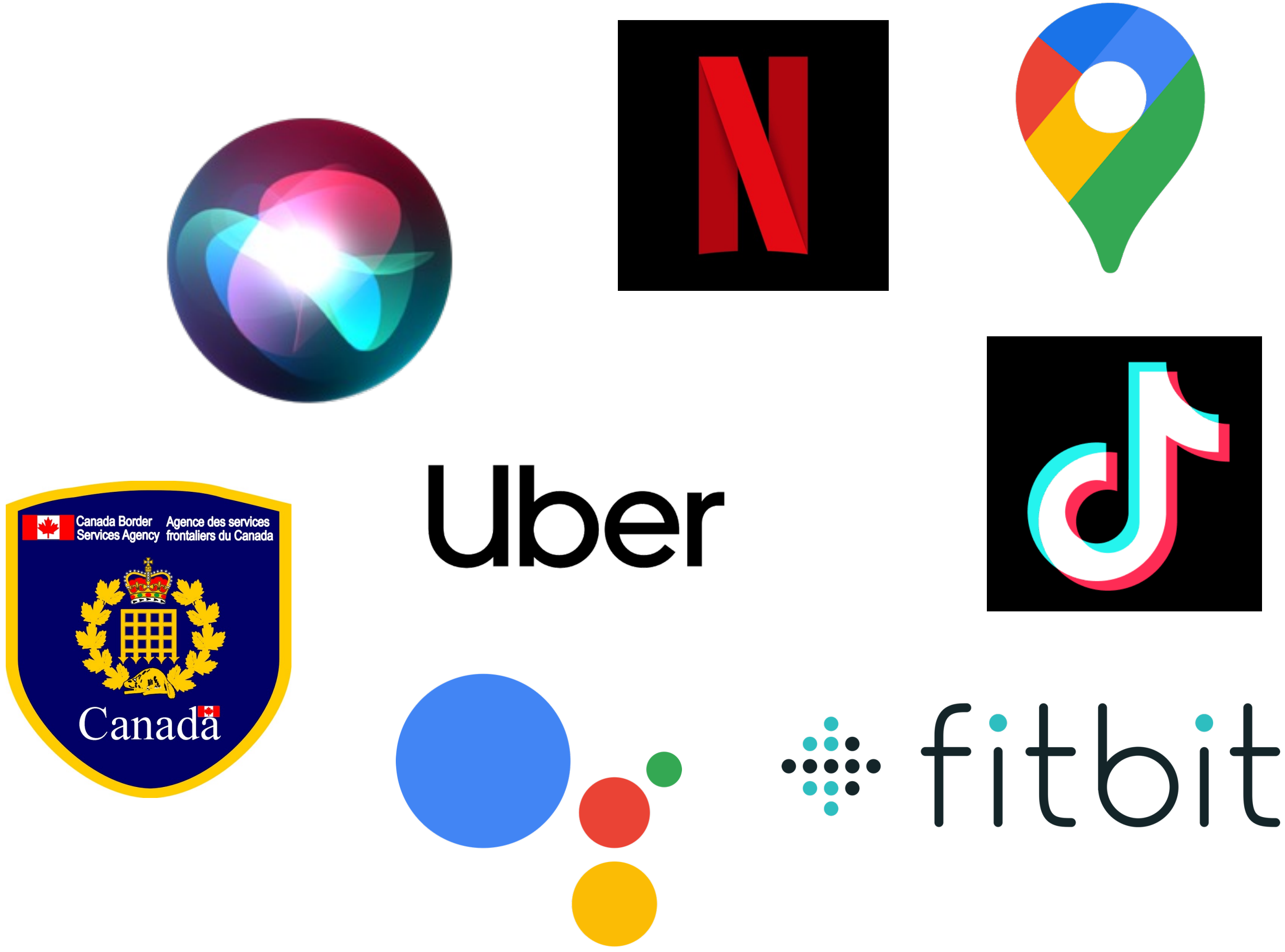
- AI is a branch of computer science that aims to create intelligent machines.

- AI is about creating systems that can perform tasks that would require human intelligence.



AI in Everyday Life

- Personal Assistants
- Recommendation Systems
- Navigation and Travel
- Health and Fitness
- Customs
- Many more…


Canada Border Services Agency
Agence des services frontaliers du Canada
Canada
Uber
fitbit

Types of AI

Deeplite

Narrow AI

- “Weak AI”
- Designed to perform a specific task such as “Person Detection”
- Focused with limited constraints
- Lacks consciousness, genuine understanding and self-awareness
- This is AI today

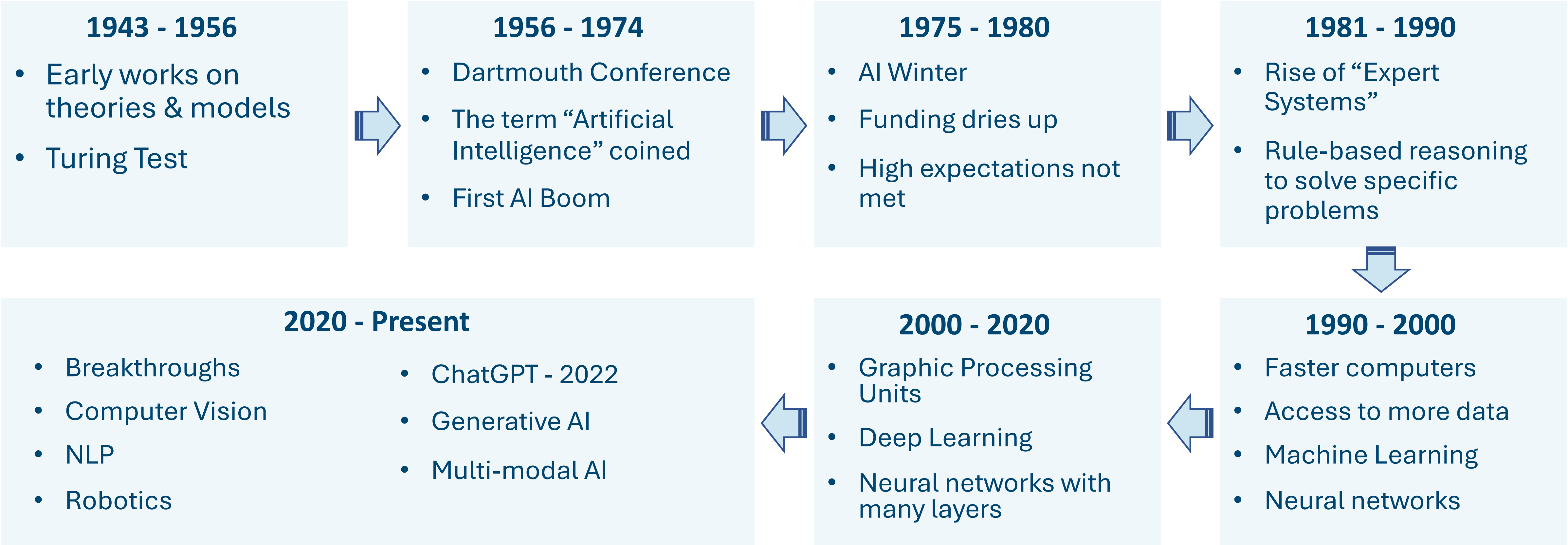
General AI

- “Strong AI” or “AGI”
- Can perform like a human
- Can understand, learn, adapt
- Implement knowledge from one domain to another
- Understands context and make judgements based on that
- Human level cognitive function
- Does not exist “yet”

Superintelligent AI

- Surpasses human intelligence
- Self-improvement
- Could solve complex problems that human couldn’t
  - Curing diseases
  - Predicting stock market
- Major risk & ethical considerations
- Hypothetical

History of AI



1943 - 1956

- Early works on theories & models
- Turing Test

1956 - 1974

- Dartmouth Conference
- The term “Artificial Intelligence” coined
- First AI Boom

1975 - 1980

- AI Winter
- Funding dries up
- High expectations not met

1981 - 1990

- Rise of “Expert Systems”
- Rule-based reasoning to solve specific problems

1990 - 2000

- Faster computers
- Access to more data
- Machine Learning
- Neural networks

2000 - 2020

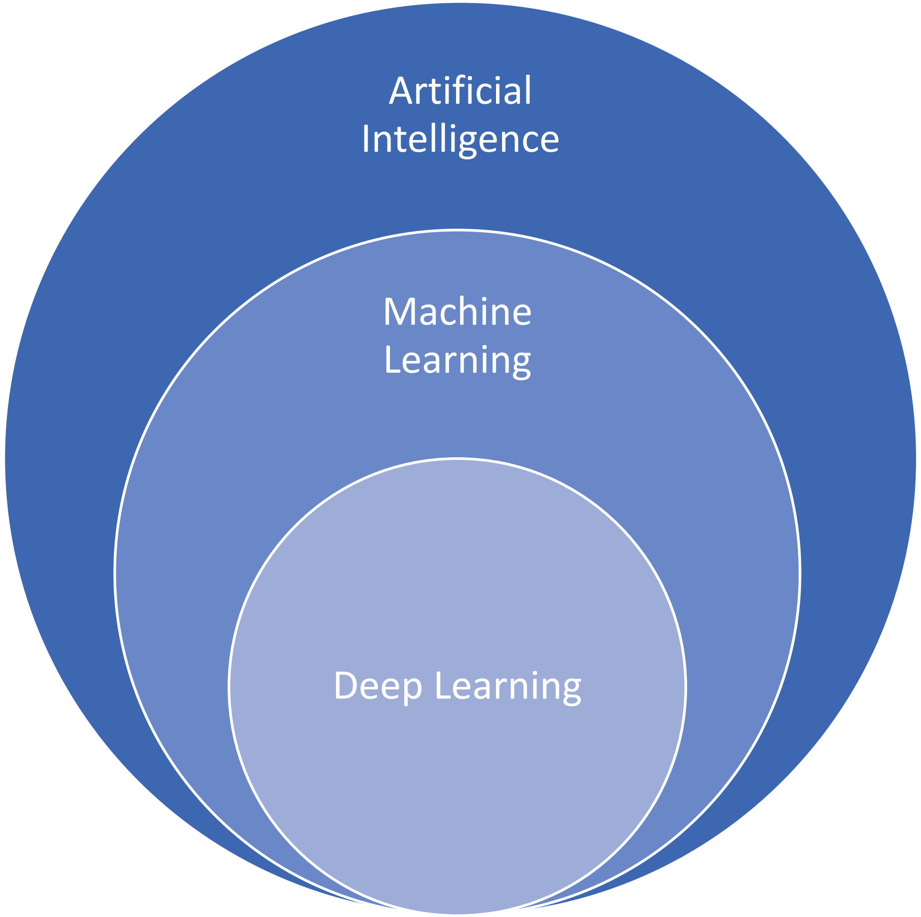
- Graphic Processing Units
- Deep Learning
- Neural networks with many layers

2020 - Present

- Breakthroughs
- Computer Vision
- NLP
- Robotics
- ChatGPT - 2022
- Generative AI
- Multi-modal AI

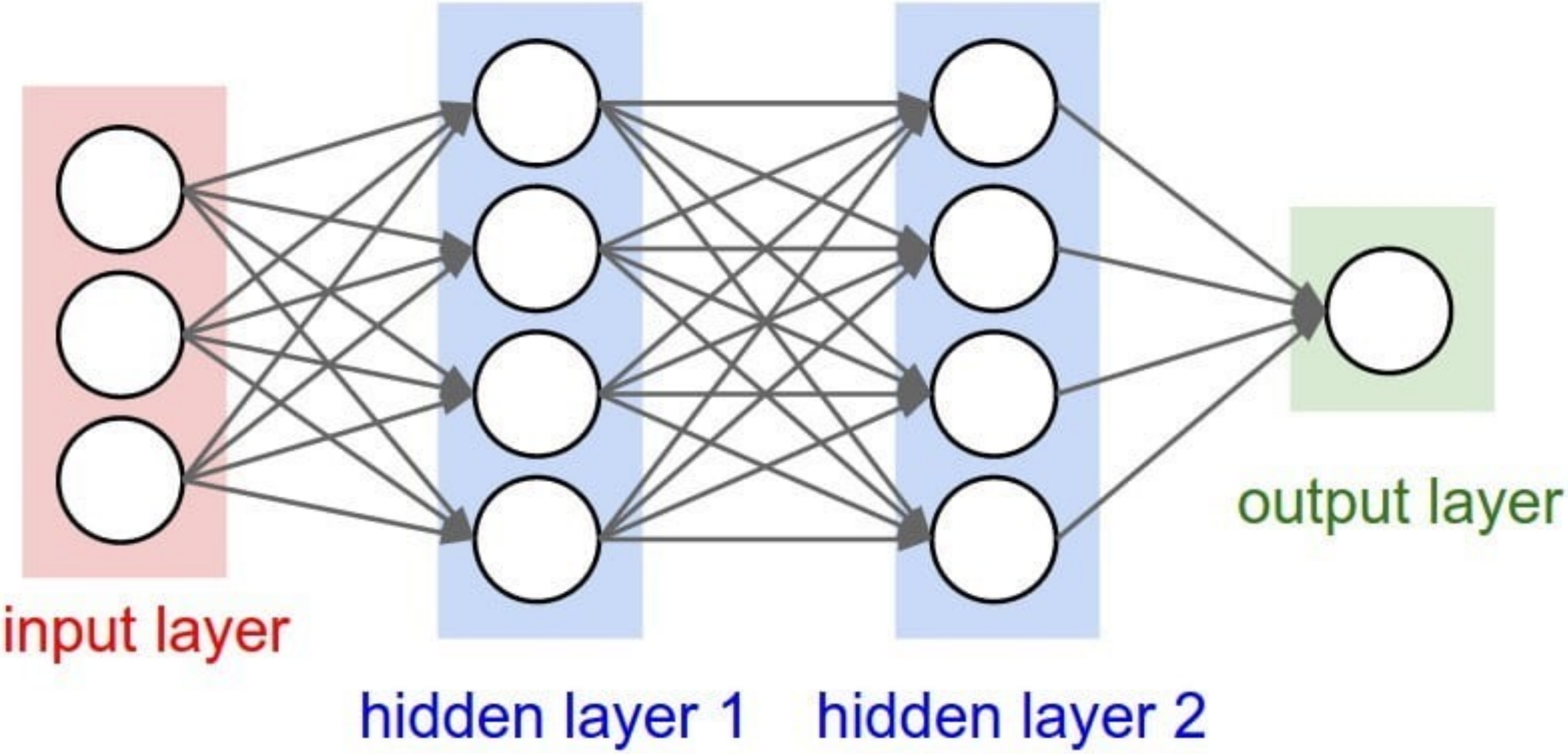
How does AI Work?

- Machine Learning: AI learns from data to make predictions or decisions.
- Deep Learning: A subset of machine learning that uses neural networks with many layers. Requires LOTS of data to learn
- Reinforcement Learning: AI learns by trial and error to achieve a clear objective.
- Imitation Learning: AI learns from a catalog of “demonstrations”



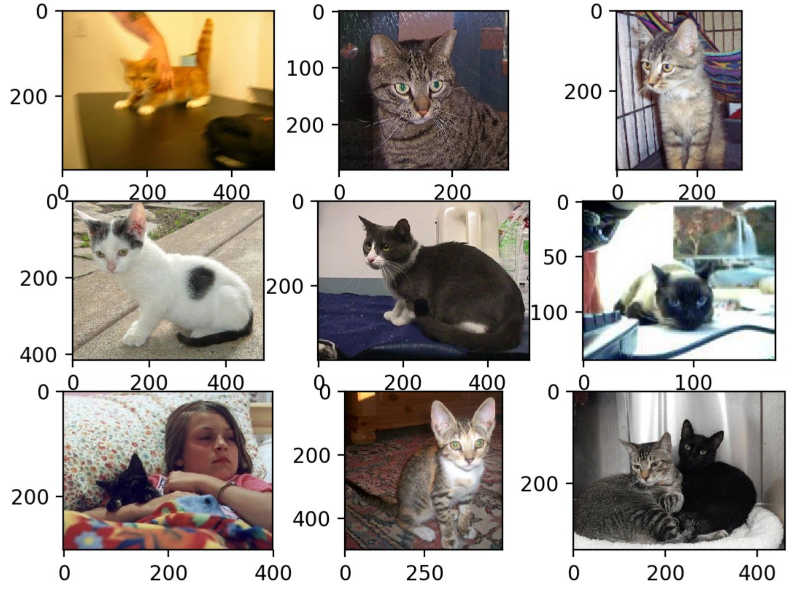
Neural Network

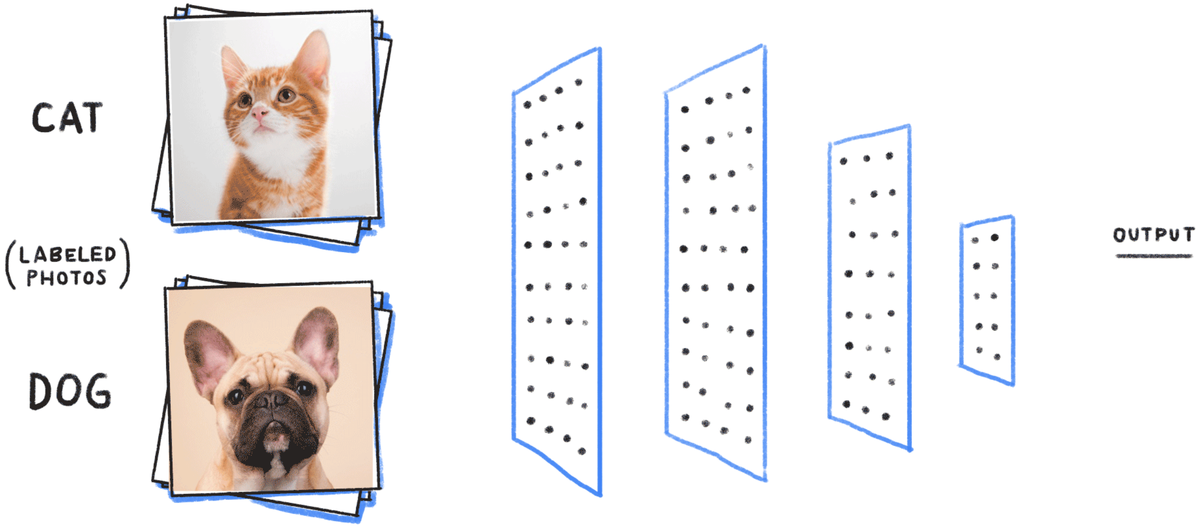
Based on the operation of neurons in the human brain



Training AI

Supervised and unsupervised learning



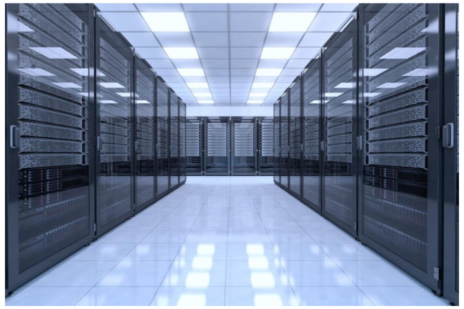


More data + bigger models = Better accuracy

Better accuracy = Better decisions

AI is Resource Intensive

Human brain requires ~20 watts of energy



An AI100 NVIDIA GPU can consume up to 300 watt-hours (Wh) of energy

(7.5MW per hour)

Training ChatGPT-3

(175B parameters)

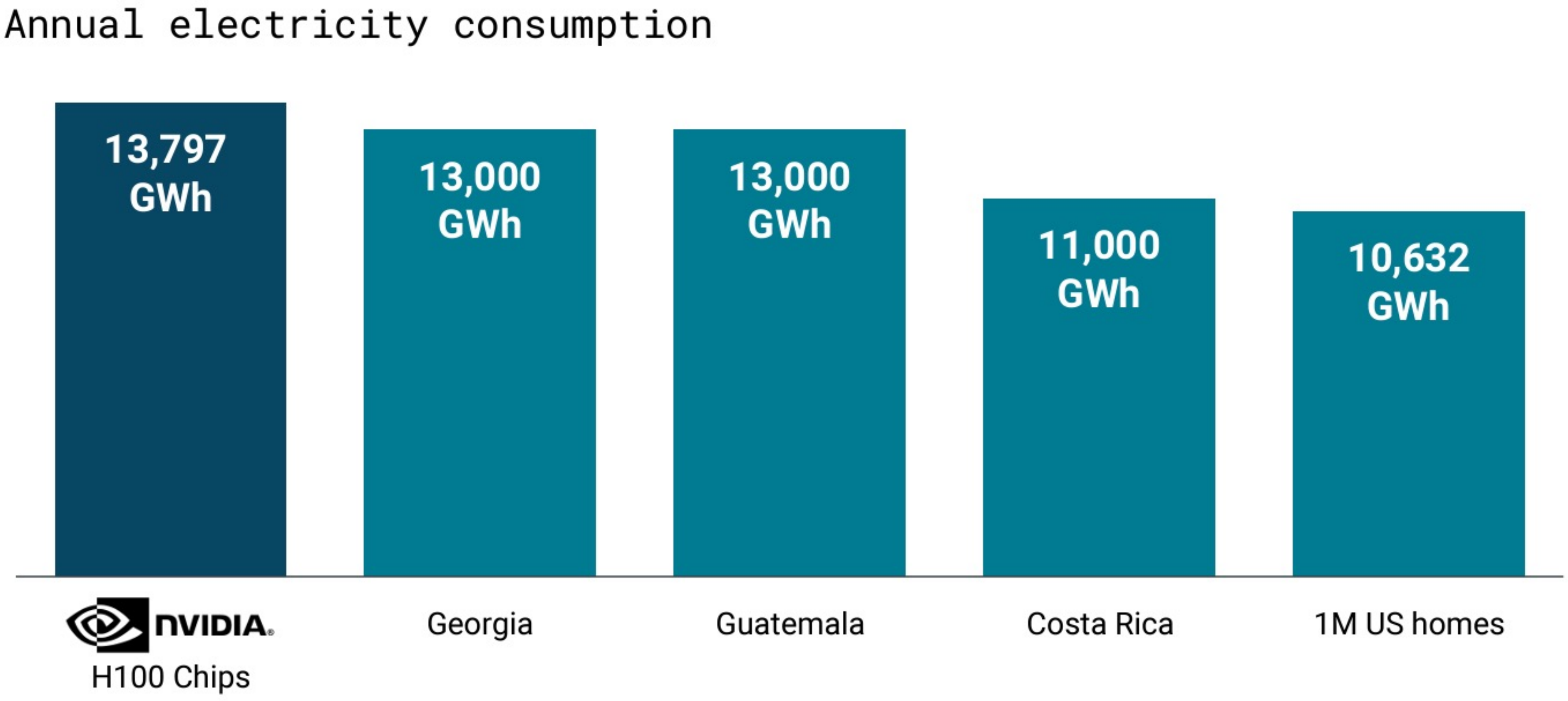
- 1,023 GPUs at ~34 days
- 250MW consumed
- Estimated 500+ metric tons of CO_2 released[1]

ChatGPT-4

- Is said to have 2T parameters!
- ~4-5,000 GPUs were used to train

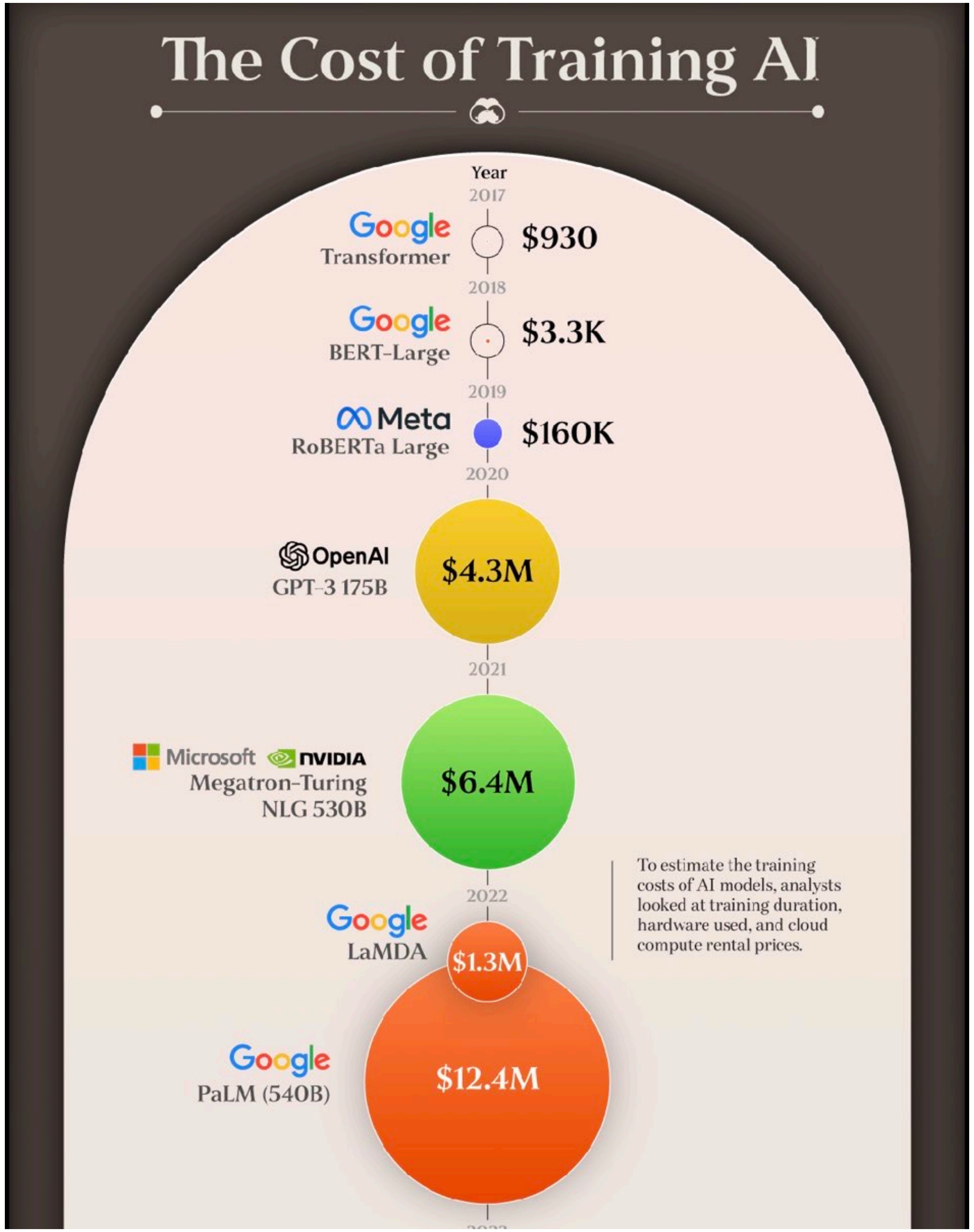
[1] Stanford Institute for Human-Centered AI

Nvidia's high-end chips will consume the same amount of energy as a small nation in 2024

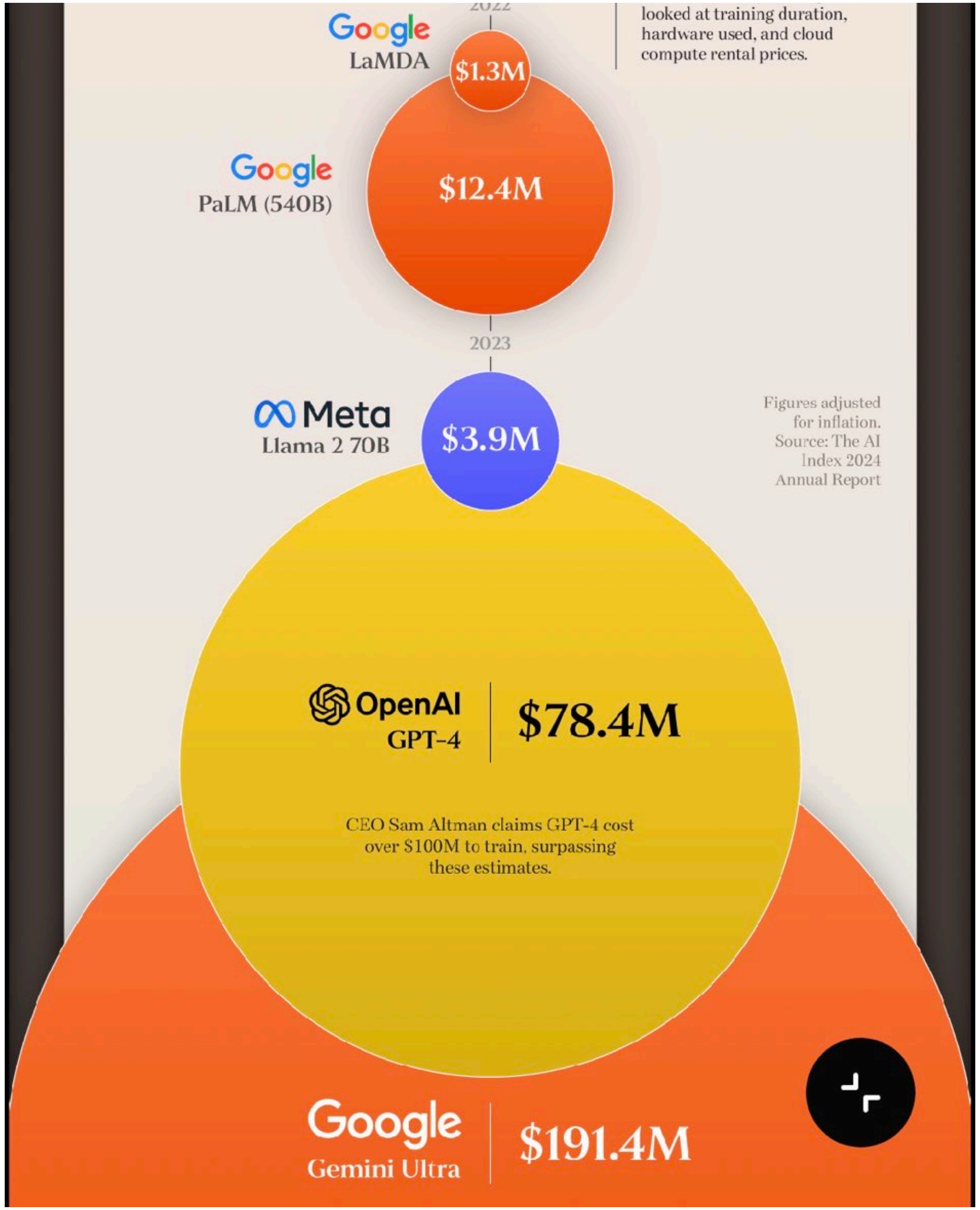


Source: CB Insights estimates, assuming 3M H100s are operational in 2024 (based on 2024 sales estimate from Nvidia and Q3'23 sales estimate from Omdia Research) at 75% of max power; electricity consumption of countries and US homes based on 2021 EIA data

5 / CBINSIGHTS



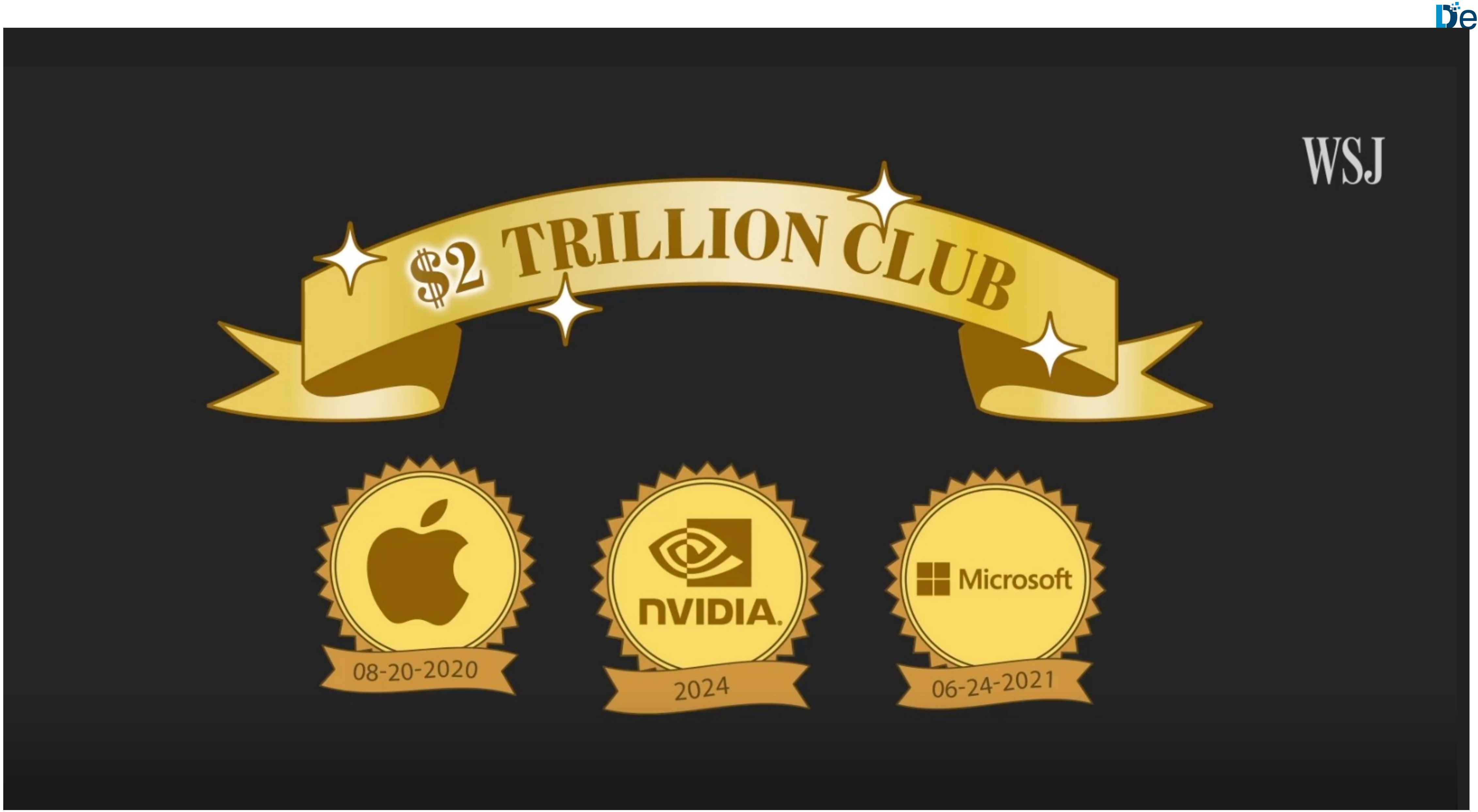
To estimate the training costs of AI models, analysts looked at training duration, hardware used, and cloud compute rental prices.



looked at training duration, hardware used, and cloud compute rental prices.

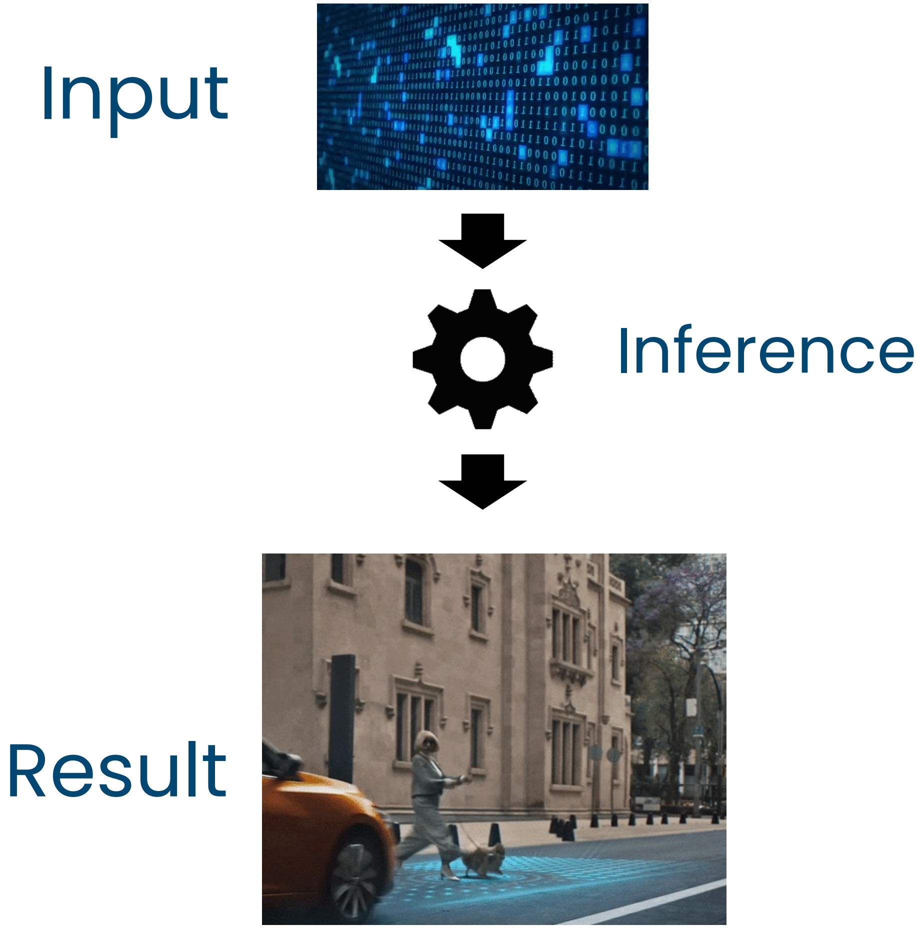
Figures adjusted for inflation. Source: The AI Index 2024 Annual Report

CEO Sam Altman claims GPT-4 cost over $100M to train, surpassing these estimates.



WSJ
$2 TRILLION CLUB
08-20-2020
NVIDIA.
2024
Microsoft
06-24-2021

AI is Resource Intensive



Input

Inference

Result

80-90%

of power consumption

Inference!

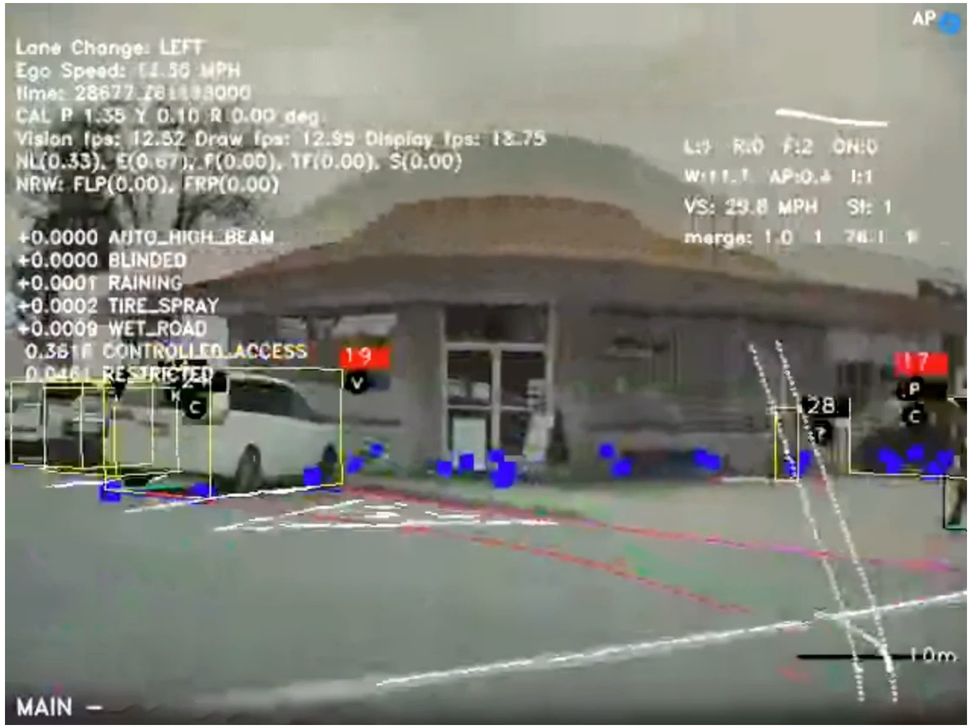
ChatGPT-4 est. of 25,000 GPUs to accommodate 100 million users

AI Technologies

Mastering the 5 senses

- Natural Language Processing: Allows machines to understand and respond to human language.
- Audio: Enables machines to hear and understand the content of sound.
- Computer Vision: Enables machines to see and understand the content of digital images or videos.
- Robotics: Design, construction, and use of robots to perform tasks done traditionally by human beings.
- Autonomous Vehicles: Use of AI in self-driving cars.

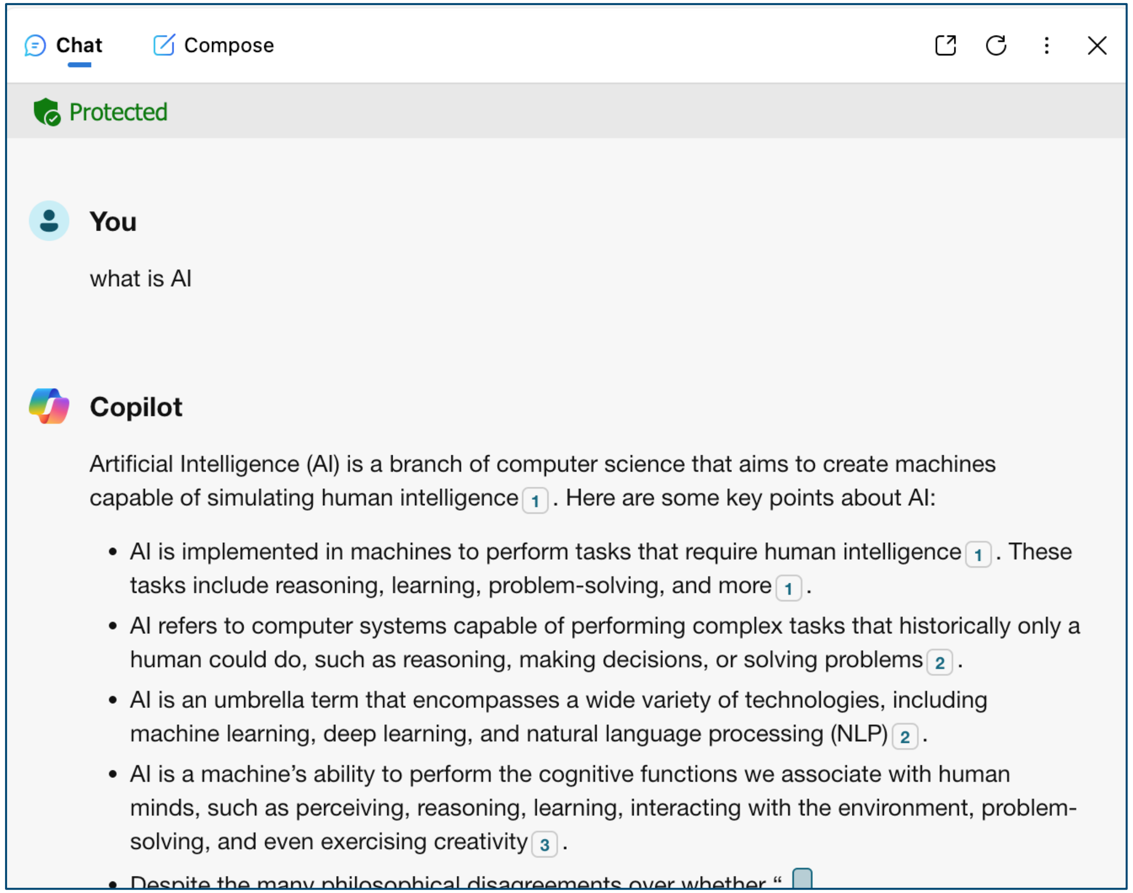
TESLA



What is Generative AI?

AI that “generates” content based on a request or “prompt”

- ChatGPT – Prompt to Text (written or voice exchange)



You

what is AI

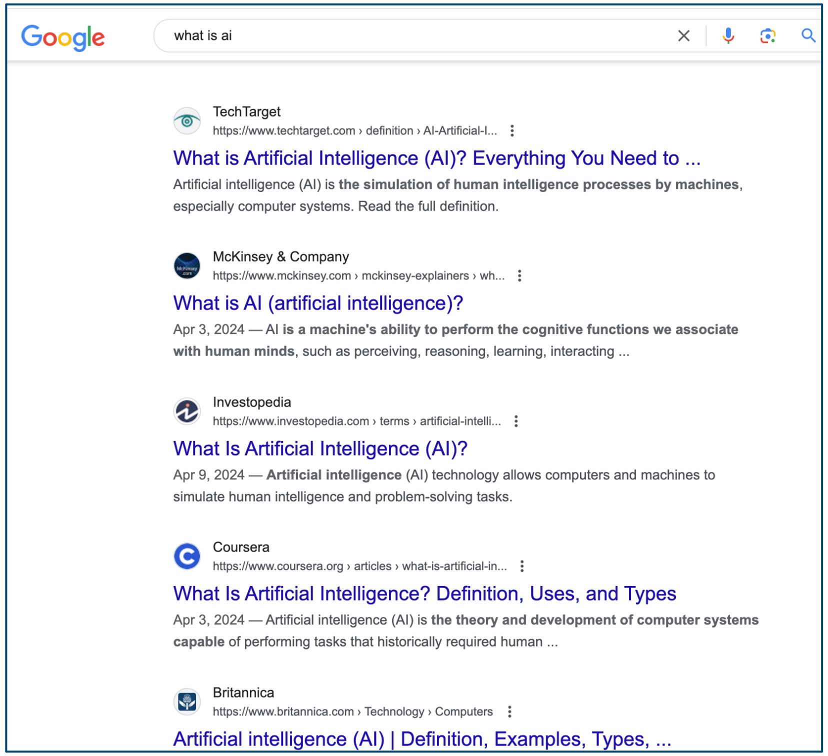
Copilot

Artificial Intelligence (AI) is a branch of computer science that aims to create machines capable of simulating human intelligence [1]. Here are some key points about AI:

- AI is implemented in machines to perform tasks that require human intelligence [1]. These tasks include reasoning, learning, problem-solving, and more [1].
- AI refers to computer systems capable of performing complex tasks that historically only a human could do, such as reasoning, making decisions, or solving problems [2].
- AI is an umbrella term that encompasses a wide variety of technologies, including machine learning, deep learning, and natural language processing (NLP) [2].
- AI is a machine’s ability to perform the cognitive functions we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, and even exercising creativity [3].
- Despite the many philosophical disagreements over whether “

The Answer Era

Card Catalogue Era



TechTarget
https://www.techtarget.com › definition › AI-Artificial-I...
What is Artificial Intelligence (AI)? Everything You Need to ...
Artificial intelligence (AI) is **the simulation of human intelligence processes by machines**, especially computer systems. Read the full definition.

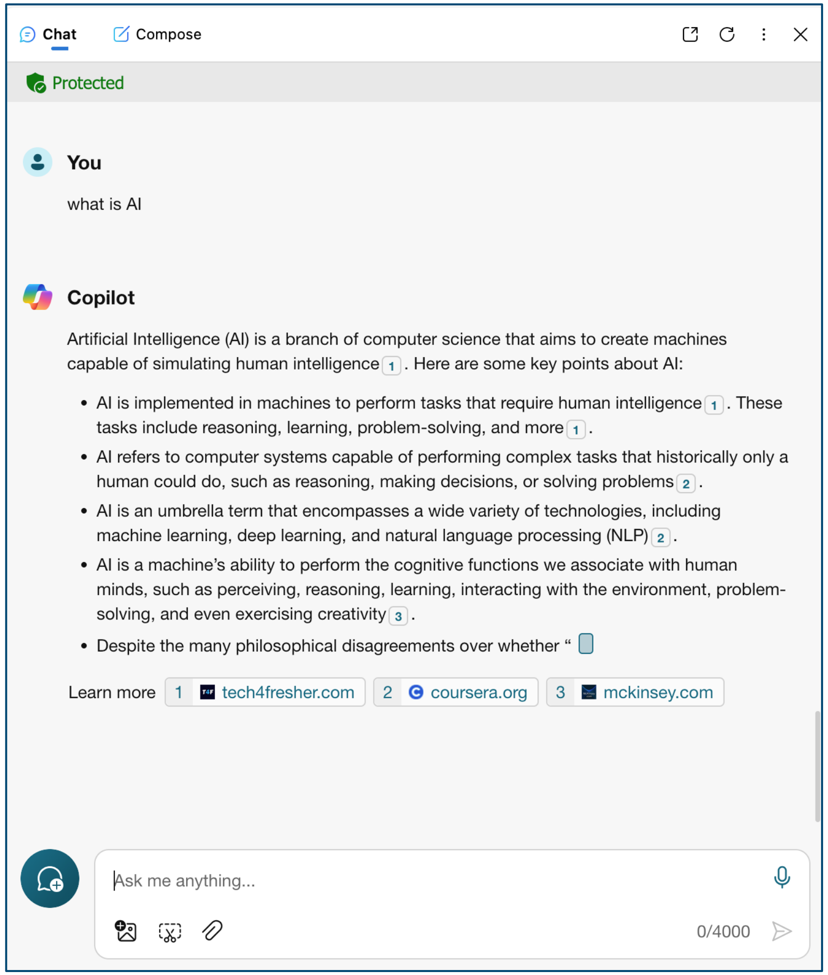
McKinsey & Company
https://www.mckinsey.com › mckinsey-explainers › wh...
What is AI (artificial intelligence)?
Apr 3, 2024 — AI **is a machine's ability to perform the cognitive functions we associate with human minds**, such as perceiving, reasoning, learning, interacting ...

Investopedia
https://www.investopedia.com › terms › artificial-intelli...
What Is Artificial Intelligence (AI)?
Apr 9, 2024 — **Artificial intelligence** (AI) technology allows computers and machines to simulate human intelligence and problem-solving tasks.

Coursera
https://www.coursera.org › articles › what-is-artificial-in...
What Is Artificial Intelligence? Definition, Uses, and Types
Apr 3, 2024 — Artificial intelligence (AI) is **the theory and development of computer systems capable** of performing tasks that historically required human ...

Britannica
https://www.britannica.com › Technology › Computers
Artificial intelligence (AI) | Definition, Examples, Types, ...

Search or "Blue Links" Era



You

what is AI

Copilot

Artificial Intelligence (AI) is a branch of computer science that aims to create machines capable of simulating human intelligence 1 . Here are some key points about AI:

- AI is implemented in machines to perform tasks that require human intelligence 1 . These tasks include reasoning, learning, problem-solving, and more 1 .
- AI refers to computer systems capable of performing complex tasks that historically only a human could do, such as reasoning, making decisions, or solving problems 2 .
- AI is an umbrella term that encompasses a wide variety of technologies, including machine learning, deep learning, and natural language processing (NLP) 2 .
- AI is a machine's ability to perform the cognitive functions we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, and even exercising creativity 3 .
- Despite the many philosophical disagreements over whether "

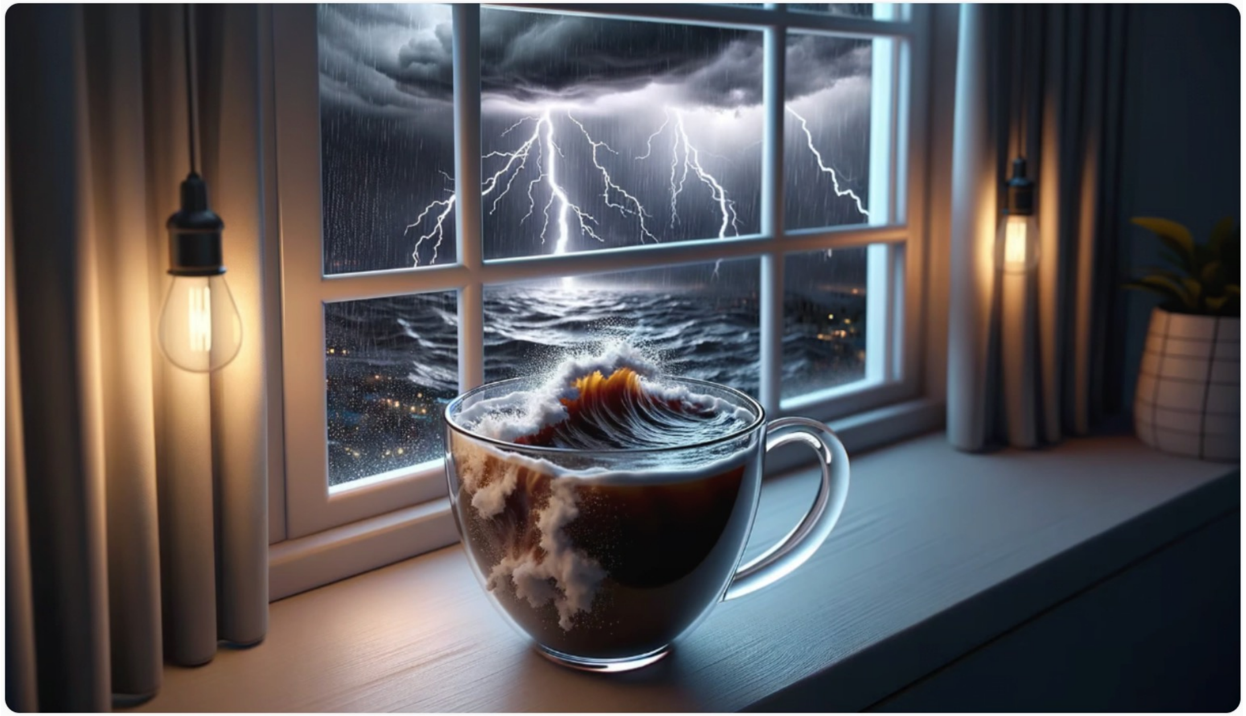
Learn more 1 tech4fresher.com 2 coursera.org 3 mckinsey.com

Answer Era

What is Generative AI?

AI that “generates” content based on a request or “prompt”

- ChatGPT – Prompt to Text
- Dall-e – Prompt to Image



Prompt: "A 3D render of a coffee mug placed on a window sill during a stormy day. The storm outside the window is reflected in the coffee, with miniature lightning bolts and turbulent waves seen inside the mug. The room is dimly lit, adding to the dramatic atmosphere." | Image: DALL-E 3 prompted by OpenAI

What is Generative AI?

AI that “generates” content based on a request or “prompt”

- ChatGPT – Prompt to Text
- Dall-e – Prompt to Image
- Sora – Prompt to Video

Historical footage of California during the gold rush

Multi-modal AI

AI that can understand and create content across different modalities like text, images, audio etc.

Combines perception and generative AI

One step closer to AGI!

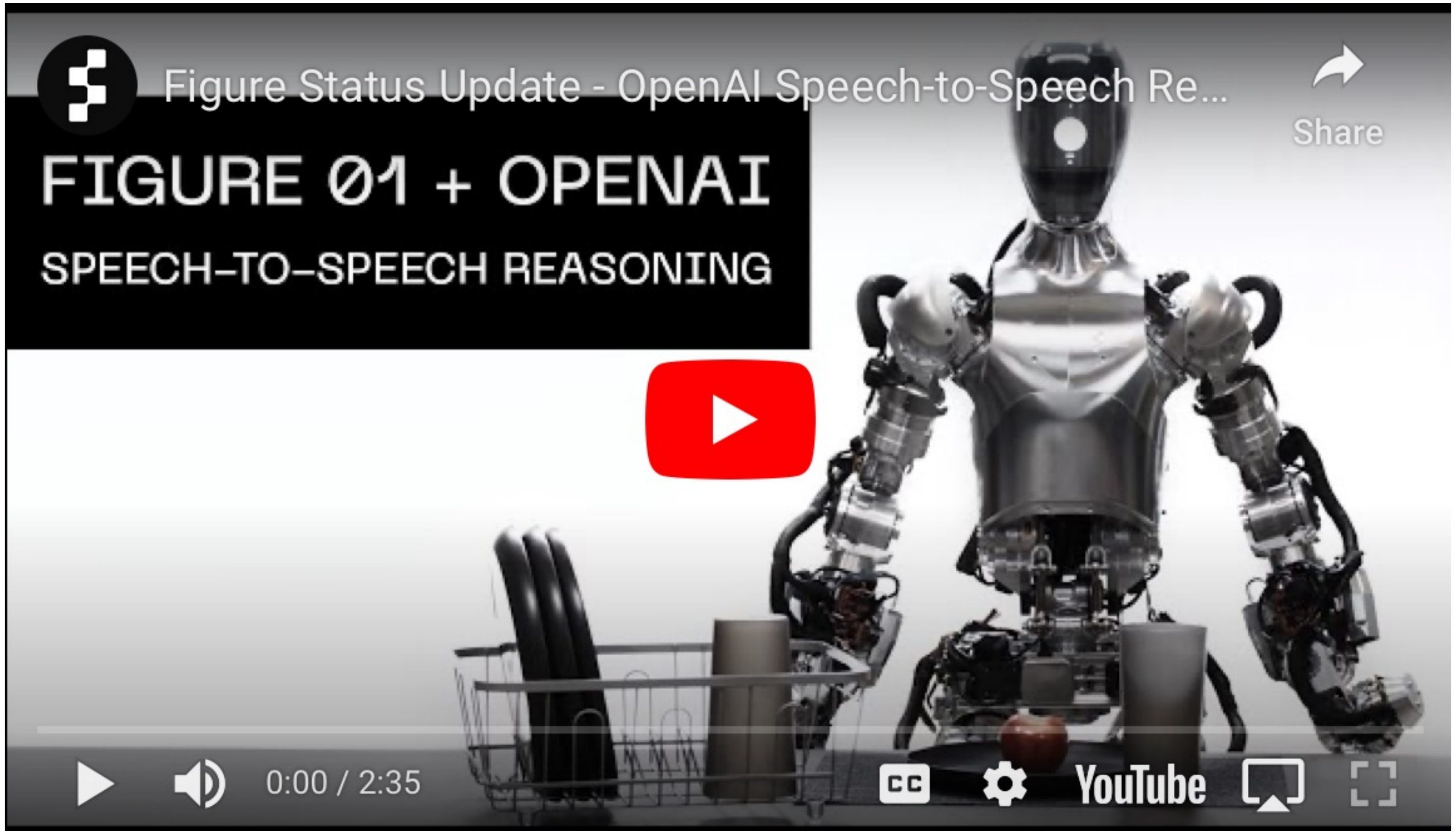


Figure Status Update - OpenAI Speech-to-Speech Re...
Share
FIGURE 01 + OPENAI
SPEECH-TO-SPEECH REASONING
0:00 / 2:35
YouTube

AI in Business

Overview

- AI is transforming various business functions: marketing, HR, finance, legal etc.
- AI can automate repetitive tasks, freeing up time for more strategic work.
- AI can provide insights from large amounts of data to inform decision-making.
- AI can improve customer service through chatbots and personalized recommendations.

Amazon

How Amazon uses AI for Product Recommendations

- AI algorithms identify patterns in customer data
- AI generates personalized product recommendations.
- Use AI to monetize customer data.
- Algorithm generates product pitches
  - Help customers find what they are looking for
  - Nudge them towards buying more.
- Amazon's AI analyzes user behavior by tracking various data points
  - Products viewed,
  - Time spent on each page,
  - The frequency of purchases.

Starbucks

How Starbucks uses AI to Personalize Customer Experiences

- Personalized customer experiences by leveraging big data and predictive analytics.
- The **Deep Brew** initiative, a platform that enables the seamless deployment of AI models across the company's expansive operations.
- AI algorithms built into the Starbucks app
- Analyzes customer preferences and behavior
- Provide personalized drink recommendations and offers.
- AI-powered chatbots provide customers with personalized recommendations and assistance.

STARBUCKS® REWARDS

FESTIVE THURS★YAYS

HALF OFF FOR THE HOLIDAYS

Enjoy half off a drink, every Thursday in December.
For Starbucks® Rewards members.

THURSDAYS | ALL DECEMBER | 12–6 PM

At participating stores. Offer applies to handcrafted drinks prepared by baristas, including coffee and tea. Limit one drink per member per week. Cannot be combined with other offers or promotions. This offer is exclusive to you and this Starbucks Rewards account and cannot be reproduced, transferred or used by anyone else.

American Express

How Amex uses AI for Fraud Detection

- AI powered fraud detection models
- Monitors each transaction in real-time
- Generates a fraud decision in milliseconds every single time an American Express card is used globally.
- American Express uses advanced AI models to detect anomalous patterns in transactions.
- The real-time fraud detection system improves accuracy, and better protect customers and merchants.

Google

How Google uses AI to optimize cooling in its data centers

- By applying DeepMind’s machine learning to Google data centers, Google has managed to
- Reduced the amount of energy used for cooling by up to 40%.
- Thousands of sensors in the data center
- Every five minutes, Google's cloud-based AI pulls a snapshot of the data center cooling system
- Data is fed it into deep neural networks.
- The AI system then identifies which actions will minimize the energy consumption while satisfying a robust set of safety constraints.

Fun fact: Microsoft used the equivalent water volume of 2,560 Olympic size swimming to cool its data centers in 2022!

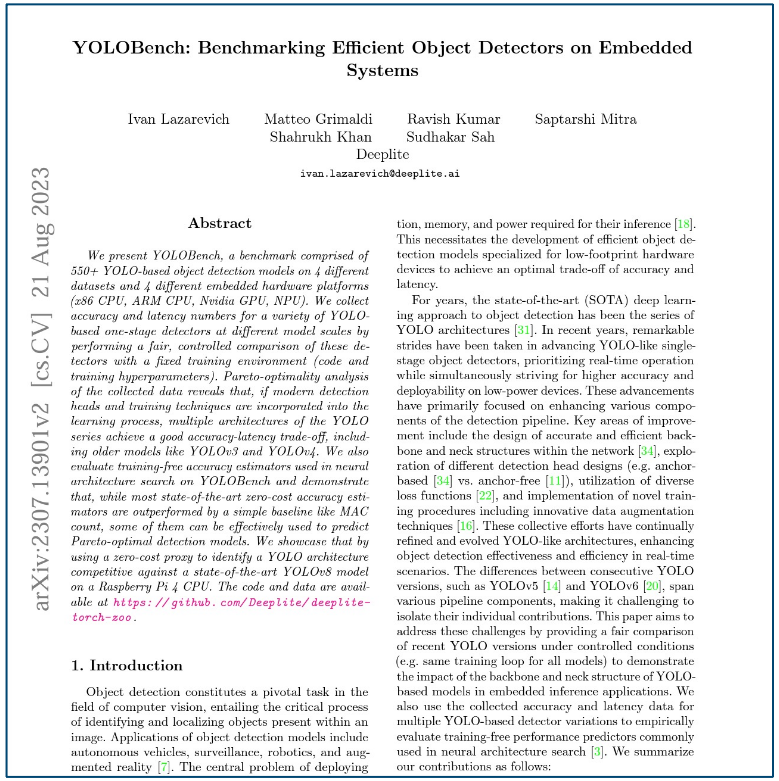
AI in Your Business

Benefits

- Cost Savings: AI can automate routine tasks, reducing labor costs.
- Improved Customer Experience: AI can provide personalized experiences and 24/7 customer service.
- Data Analysis: AI can analyze large amounts of data to provide insights.
- Increased Productivity: AI can take over mundane tasks, freeing up humans for more complex tasks.

Deeplite Marketing Example

Marketing campaign in less than half a day



arXiv:2307.13901v2 [cs.CV] 21 Aug 2023

YOLOBench: Benchmarking Efficient Object Detectors on Embedded Systems

Ivan Lazarevich Matteo Grimaldi Ravish Kumar Saptarshi Mitra
Shahrukh Khan Sudhakar Sah
Deeplite
ivan.lazarevich@deeplite.ai

Abstract

We present YOLOBench, a benchmark comprised of 550+ YOLO-based object detection models on 4 different datasets and 4 different embedded hardware platforms (x86 CPU, ARM CPU, Nvidia GPU, NPU). We collect accuracy and latency numbers for a variety of YOLO-based one-stage detectors at different model scales by performing a fair, controlled comparison of these detectors with a fixed training environment (code and training hyperparameters). Pareto-optimality analysis of the collected data reveals that, if modern detection heads and training techniques are incorporated into the learning process, multiple architectures of the YOLO series achieve a good accuracy-latency trade-off, including older models like YOLOv3 and YOLOv4. We also evaluate training-free accuracy estimators used in neural architecture search on YOLOBench and demonstrate that, while most state-of-the-art zero-cost accuracy estimators are outperformed by a simple baseline like MAC count, some of them can be effectively used to predict Pareto-optimal detection models. We showcase that by using a zero-cost proxy to identify a YOLO architecture competitive against a state-of-the-art YOLOv8 model on a Raspberry Pi 4 CPU. The code and data are available at https://github.com/Deeplite/deeplite-torch-zoo.

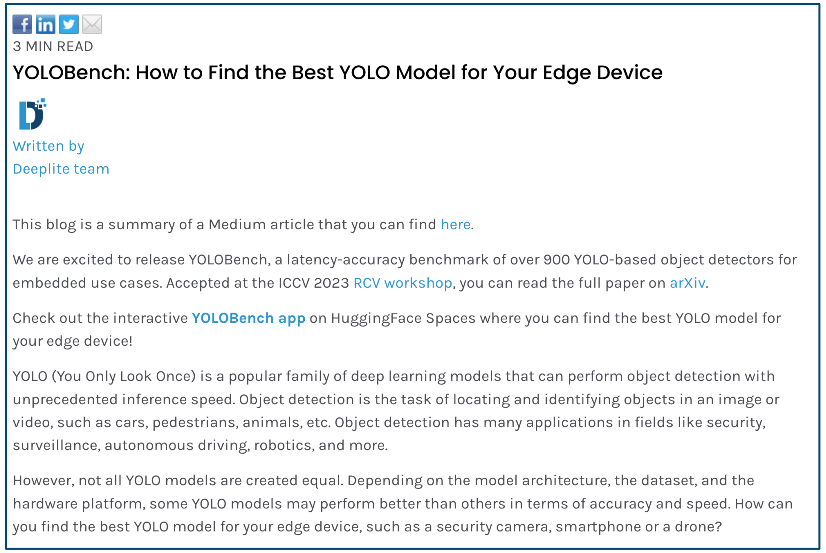
1. Introduction

Object detection constitutes a pivotal task in the field of computer vision, entailing the critical process of identifying and localizing objects present within an image. Applications of object detection models include autonomous vehicles, surveillance, robotics, and augmented reality [7]. The central problem of deploying

tion, memory, and power required for their inference [18]. This necessitates the development of efficient object detection models specialized for low-footprint hardware devices to achieve an optimal trade-off of accuracy and latency.

For years, the state-of-the-art (SOTA) deep learning approach to object detection has been the series of YOLO architectures [31]. In recent years, remarkable strides have been taken in advancing YOLO-like single-stage object detectors, prioritizing real-time operation while simultaneously striving for higher accuracy and deployability on low-power devices. These advancements have primarily focused on enhancing various components of the detection pipeline. Key areas of improvement include the design of accurate and efficient backbone and neck structures within the network [34], exploration of different detection head designs (e.g. anchor-based [34] vs. anchor-free [11]), utilization of diverse loss functions [22], and implementation of novel training procedures including innovative data augmentation techniques [16]. These collective efforts have continually refined and evolved YOLO-like architectures, enhancing object detection effectiveness and efficiency in real-time scenarios. The differences between consecutive YOLO versions, such as YOLOv5 [14] and YOLOv6 [20], span various pipeline components, making it challenging to isolate their individual contributions. This paper aims to address these challenges by providing a fair comparison of recent YOLO versions under controlled conditions (e.g. same training loop for all models) to demonstrate the impact of the backbone and neck structure of YOLO-based models in embedded inference applications. We also use the collected accuracy and latency data for multiple YOLO-based detector variations to empirically evaluate training-free performance predictors commonly used in neural architecture search [3]. We summarize our contributions as follows:

26 Page research paper in Arxiv



3 MIN READ

YOLOBench: How to Find the Best YOLO Model for Your Edge Device

Written by
Deeplite team

This blog is a summary of a Medium article that you can find here.

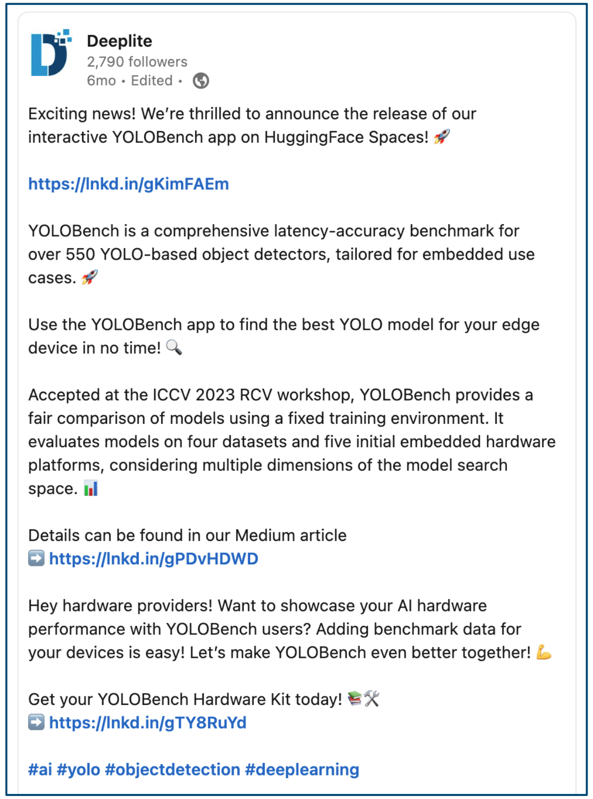
We are excited to release YOLOBench, a latency-accuracy benchmark of over 900 YOLO-based object detectors for embedded use cases. Accepted at the ICCV 2023 RCV workshop, you can read the full paper on arXiv.

Check out the interactive **YOLOBench app** on HuggingFace Spaces where you can find the best YOLO model for your edge device!

YOLO (You Only Look Once) is a popular family of deep learning models that can perform object detection with unprecedented inference speed. Object detection is the task of locating and identifying objects in an image or video, such as cars, pedestrians, animals, etc. Object detection has many applications in fields like security, surveillance, autonomous driving, robotics, and more.

However, not all YOLO models are created equal. Depending on the model architecture, the dataset, and the hardware platform, some YOLO models may perform better than others in terms of accuracy and speed. How can you find the best YOLO model for your edge device, such as a security camera, smartphone or a drone?

Website blog



Deeplite
2,790 followers
6mo • Edited •

Exciting news! We're thrilled to announce the release of our interactive YOLOBench app on HuggingFace Spaces! 🚀

https://lnkd.in/gKimFAEm

YOLOBench is a comprehensive latency-accuracy benchmark for over 550 YOLO-based object detectors, tailored for embedded use cases. 🚀

Use the YOLOBench app to find the best YOLO model for your edge device in no time! 🔍

Accepted at the ICCV 2023 RCV workshop, YOLOBench provides a fair comparison of models using a fixed training environment. It evaluates models on four datasets and five initial embedded hardware platforms, considering multiple dimensions of the model search space. 📊

Details can be found in our Medium article
➡️ **https://lnkd.in/gPDvHDWD**

Hey hardware providers! Want to showcase your AI hardware performance with YOLOBench users? Adding benchmark data for your devices is easy! Let's make YOLOBench even better together! 💪

Get your YOLOBench Hardware Kit today! 📚🛠️
➡️ **https://lnkd.in/gTY8RuYd**

#ai #yolo #objectdetection #deeplearning

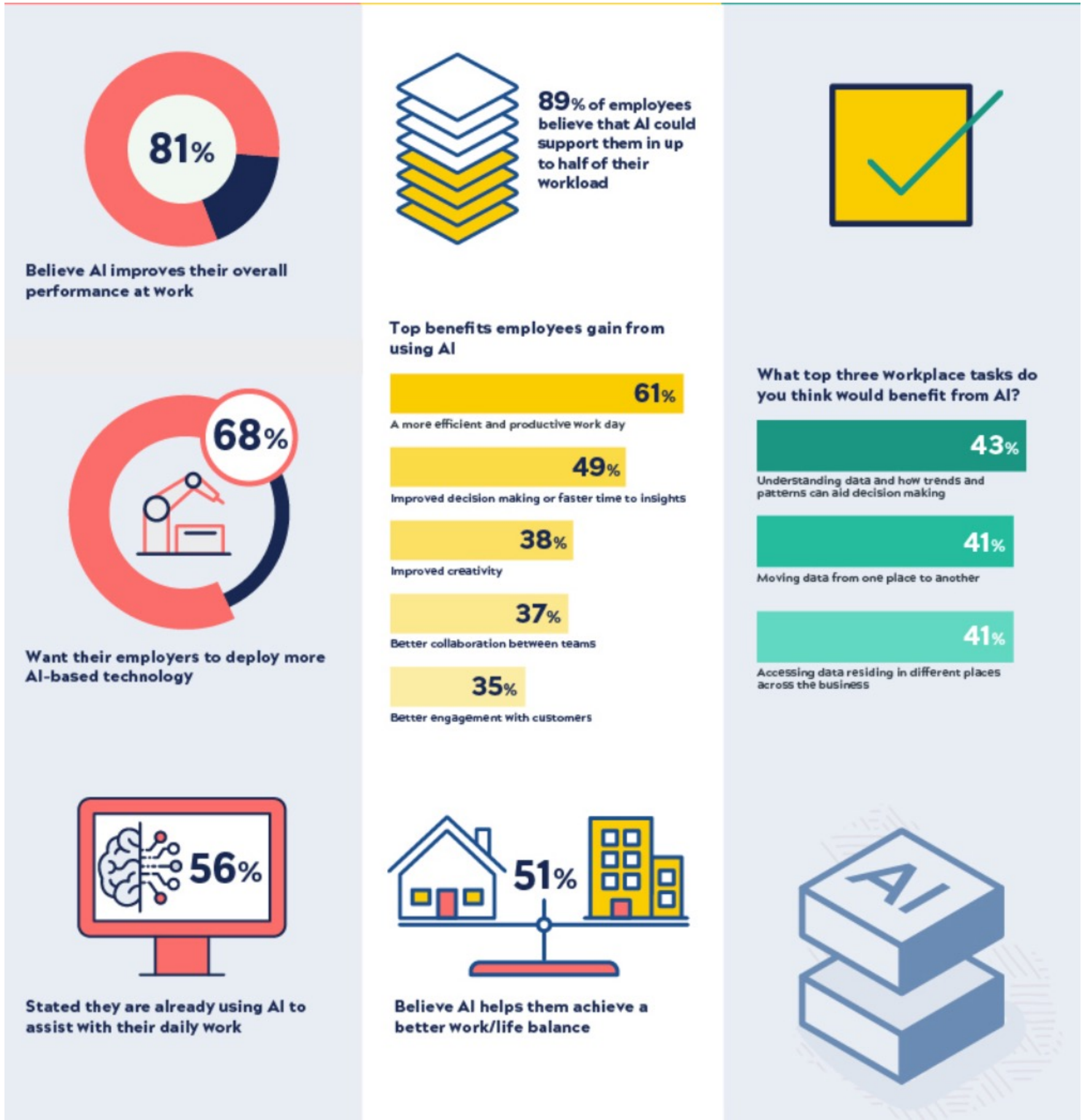
LinkedIn post

Challenges of Implementing AI

- Training Data: Based on the application, sourcing training data can be difficult and expensive.
- Data Privacy: Businesses must ensure they respect customer data when using AI.
- Data security: Sharing your proprietary data with public AI could expose your secret sauce.
- Job Displacement: While AI can create jobs, it can also displace workers.
- Bias: Poorly designed AI may inadvertently lead to suspicions of discrimination.
- Implementation Costs: Developing and implementing AI can be expensive.
- Regulations: Governments globally are implementing “guardrails” to ensure AI is ethically developed and deployed

What are employees saying?

Study by SnapLogic



81%

Believe AI improves their overall performance at work

68%

Want their employers to deploy more AI-based technology

56%

Stated they are already using AI to assist with their daily work

89% of employees believe that AI could support them in up to half of their workload

Top benefits employees gain from using AI

61% A more efficient and productive work day

49% Improved decision making or faster time to insights

38% Improved creativity

37% Better collaboration between teams

35% Better engagement with customers

51%

Believe AI helps them achieve a better work/life balance

What top three workplace tasks do you think would benefit from AI?

43% Understanding data and how trends and patterns can aid decision making

41% Moving data from one place to another

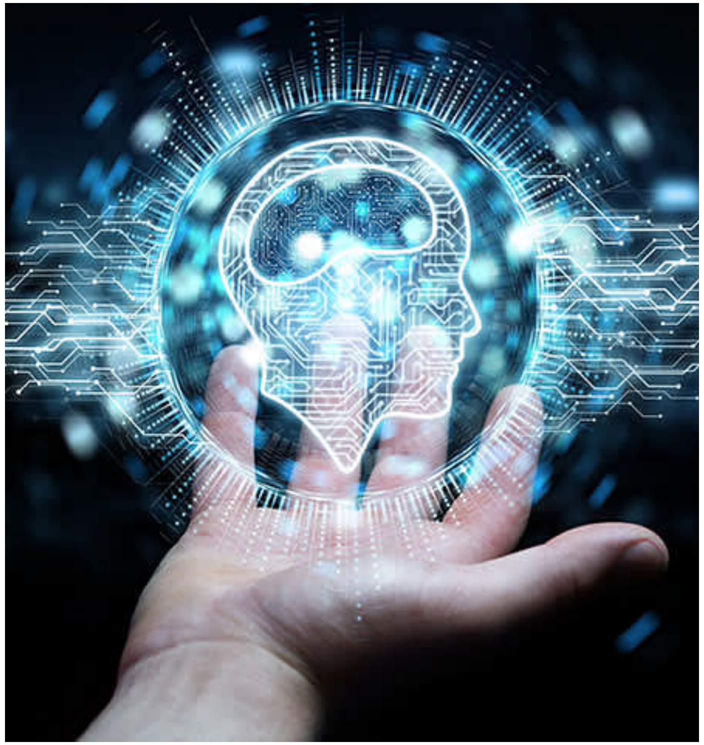
41% Accessing data residing in different places across the business

Preparing Your Business for AI

- Start by identifying areas where AI could improve efficiency or customer service.
- Look to your vendors – what are they doing?
- Start small and get comfortable with it
- Invest in training and education to ensure your team understands AI.
- Develop a clear plan for implementing AI, including goals and timelines.
- Consider working with AI experts or consultants to ensure successful implementation.

Future of AI in Business

- AI will become increasingly integrated into business operations.
- It will become more ubiquitous
- Businesses that fail to adopt AI risk being left behind.
- AI will continue to create new job opportunities while displacing others.
- Ethical considerations will become increasingly important as AI becomes more prevalent.



Ethical Considerations

- AI is a dual use technology
- Businesses have a responsibility to use AI ethically.
- This includes respecting customer data and privacy.
- Businesses must also consider the impact of AI on jobs and work to mitigate negative impacts.
- Transparency is key: Businesses should be clear about how and why they're using AI.
- Misuse will lead to over-regulation by governments

Summary

- AGI is closer than we think
- AI will have a profound impact on society
- Where there is data, there is an AI use case. Data is gold.
- AI is becoming democratized – accessible to many
- Start small but make it a priority
- Look for areas of opportunity in your business
- Buy vs. build – look to your vendors
- Use AI responsibly and ethically

And let’s hope the world doesn’t end!

THANK YOU!

Please email me to let me know of a cool AI solution you’ve done!

Sources

Deeplite

Source(s)

1. Amazon's Secret to AI-Powered Product Recommendations
2. 10 Ways Amazon Uses AI to Revolutionize E-Commerce in 2024
3. How Starbucks Leveraged AI Predictive Analytics for Personalized ...
4. Leveraging AI for Personalized Customer Experiences. Lessons from ...
5. How Starbucks is Revolutionizing Customer Relationship Management (CRM ...
6. How Amex Helps You Protect Yourself Against Credit Card Fraud
7. American Express Prevents Fraud and Foils Cybercrime With NVIDIA AI ...
8. Data centers are more energy efficient than ever - The Keyword
9. DeepMind AI reduces energy used for cooling Google data centers by 40%
10. Safety-first AI for autonomous data center cooling and industrial control
11. Artificial Intelligence at American Express – Two Current Use Cases
12. American Express Adopts NVIDIA AI to Help Prevent Fraud and Foil ...
13. AI for humanity: How Starbucks plans to use technology to nurture the ...
14. 3 Ways Amazon Uses AI to Make Product Recommendations - Lineate
15. How Does Amazon Use Artificial Intelligence? Exploring AI-Powered ...
16. Google's Use of AI to Manage Data Centers Enters a New Phase | Data ...