GSAS@NIDA : Deep Learning

Introduction

Assoc.Prof.Thitirat Siriborvornratanakul, Ph.D.

Email: thitirat@as.nida.ac.th
Website: http://as.nida.ac.th/~thitirat/

1

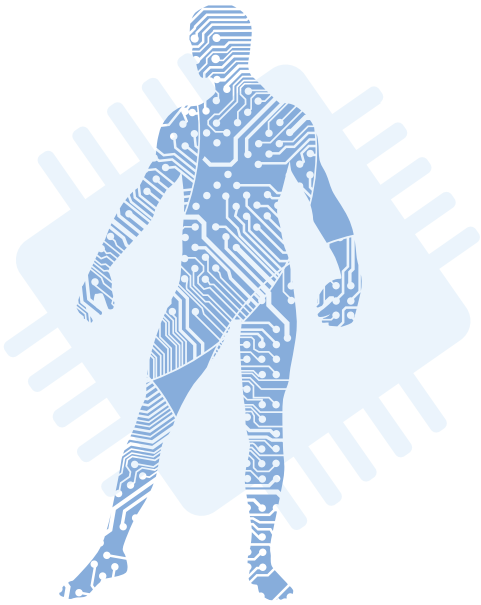
OUTLINE

01 Course's Introduction
Schedule, evaluation, books, and FAQs

02 History of Modern DL
The rapid growth of Deep Learning in the modern era

03 Hardware Acceleration
Prepare necessary hardware to accelerate our deep learning project

04 DL Frameworks
Which frameworks to use in our deep learning project



2

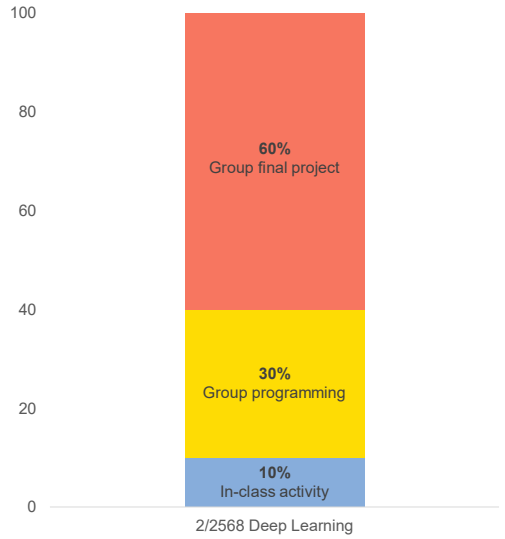
Course's Introduction

Schedule, evaluation, books, and FAQs

3

Schedule and Evaluation

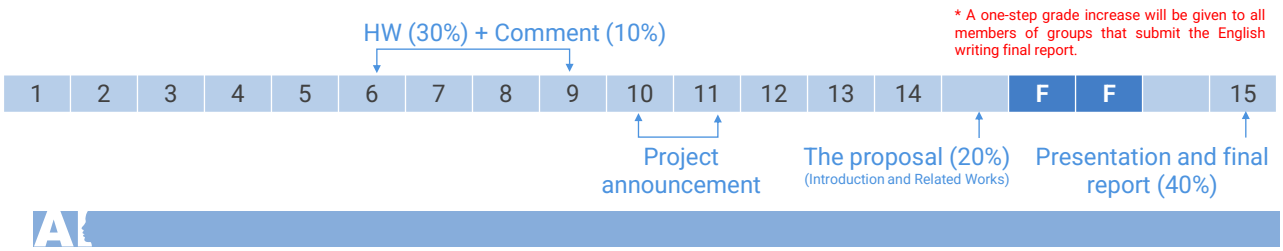
- **Subject:**
  - DADS 7202: Deep Learning (3 credits)
  - CI 7310/7105: Deep Learning (3 credits)
- **Class material and schedule:**
  - MS Teams > General channel > Files (Shared) tab > เอกสารประกอบของคลาส (Class Materials) folder
- **Grading policy:**
  - For students with "Audit" registration, fulfilling the 80% class attendance does not guarantee the "pass" grade.
- **Important note!!!**
  - Participation in the class is only permitted for students who have completed the registration process.
  - Video recording is not allowed.



4

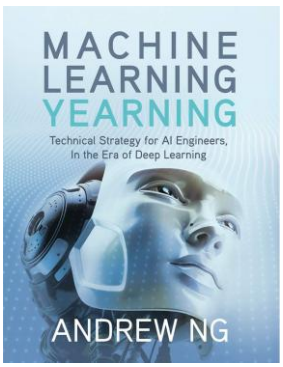
Schedule and Evaluation

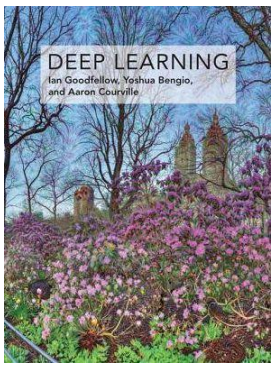
- **10% of in-class individual activities**
- **30% of Group Programming Homework:**
  - 20% for the work
  - 10% for point-by-point responses to comments
- **60% of Group Project:**
  - 20% for the proposal (Introduction and Related Works)
  - 40% for presentation and the final report*

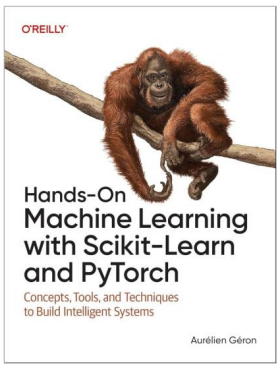


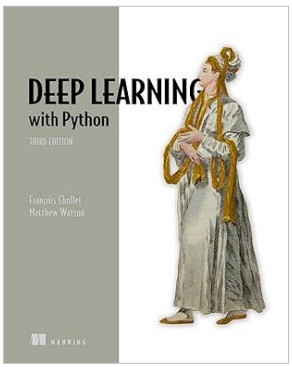
5

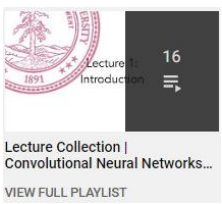
Recommended Materials

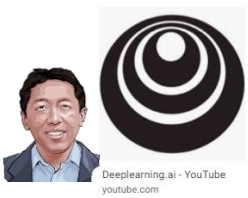










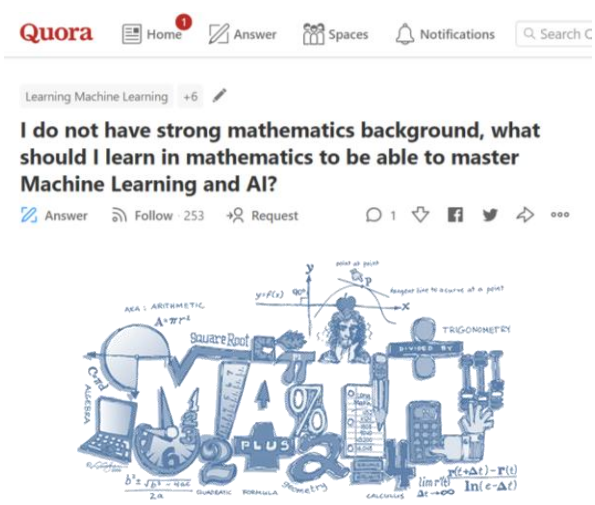


6

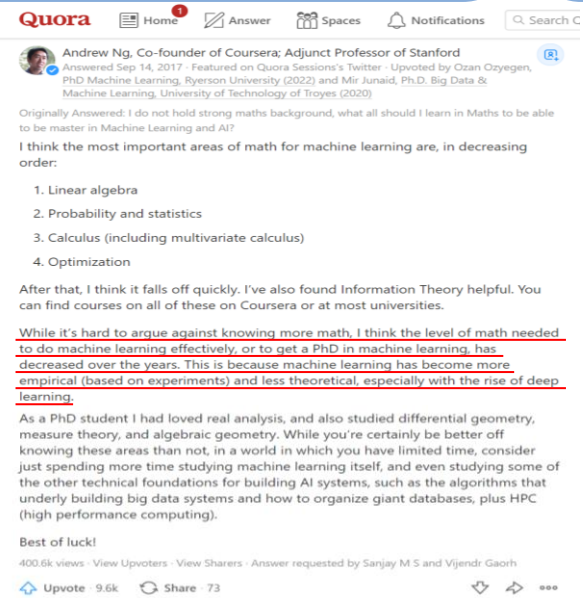
FAQ 1: Programming skill

7

FAQ 2: Math skill



[14SEP2017] https://www.quora.com/I-do-not-have-strong-mathematics-background-what-should-I-learn-in-mathematics-to-be-able-to-master-Machine-Learning-and-AI



8

FAQ 3: Prerequisite

NumPy

Image generated from https://imgflip.com/memegenerator/179756507/Slipping-steps

9

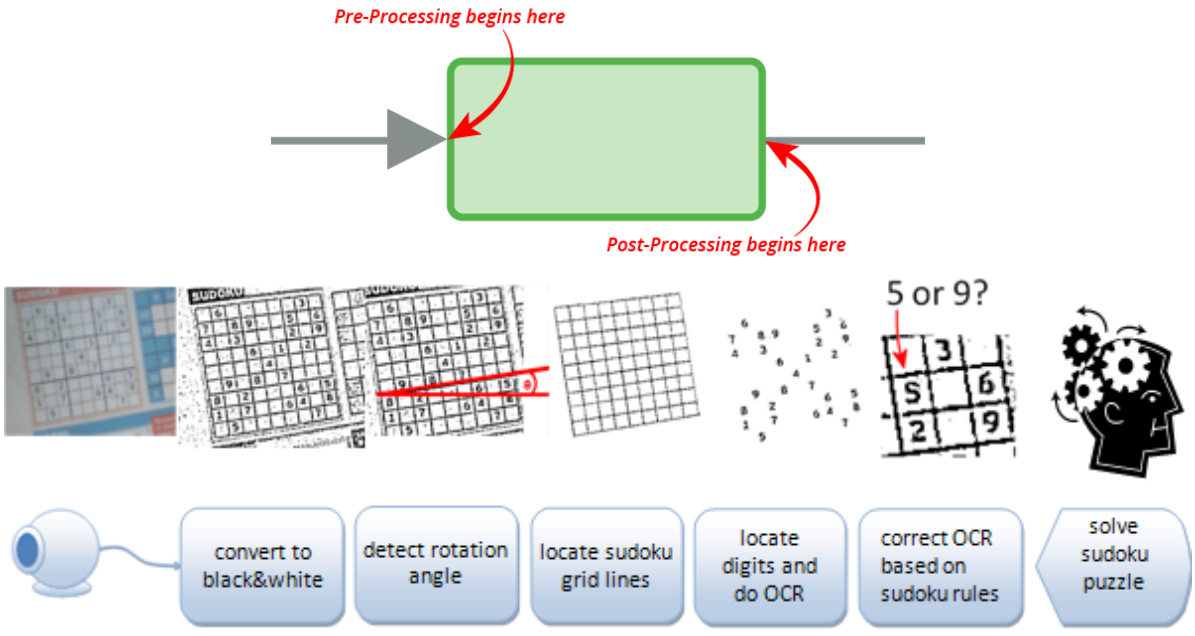
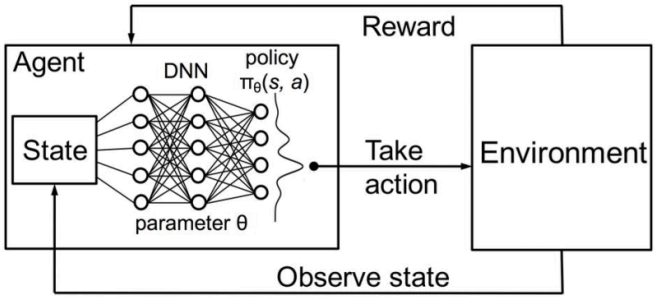


Image source: https://www.codeproject.com/Articles/238114/Realtime-Webcam-Sudoku-Solver

10

FAQ 4: Not included

Deep Reinforcement Learning



Geometric Deep Learning

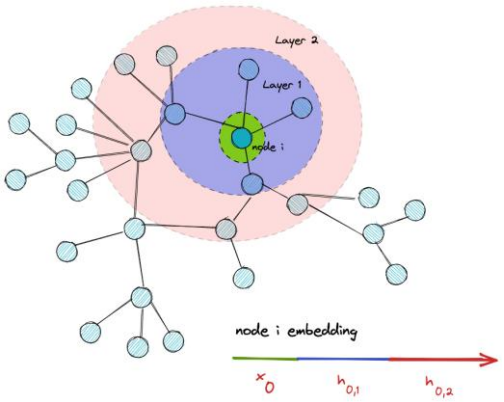


Image credit: https://medium.com/@vishnuvijayanpv/deep-reinforcement-learning-artificial-intelligence-machine-learning-and-deep-learning-e52cb5974420

11

FAQ 5: Hardware and tools

python + colab

Keras

NVIDIA

PyTorch

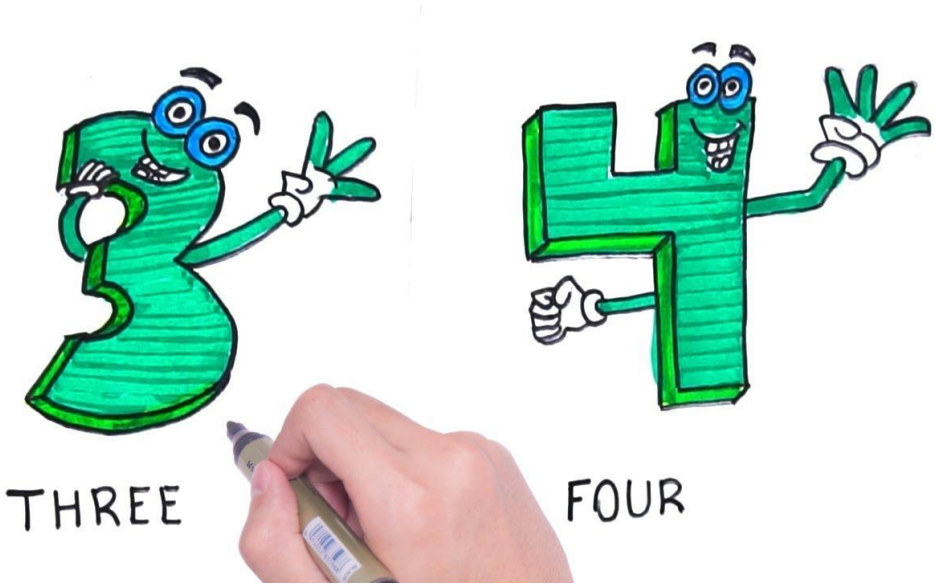
Weights & Biases

OpenAI

LangChain

12

FAQ 6: Group members



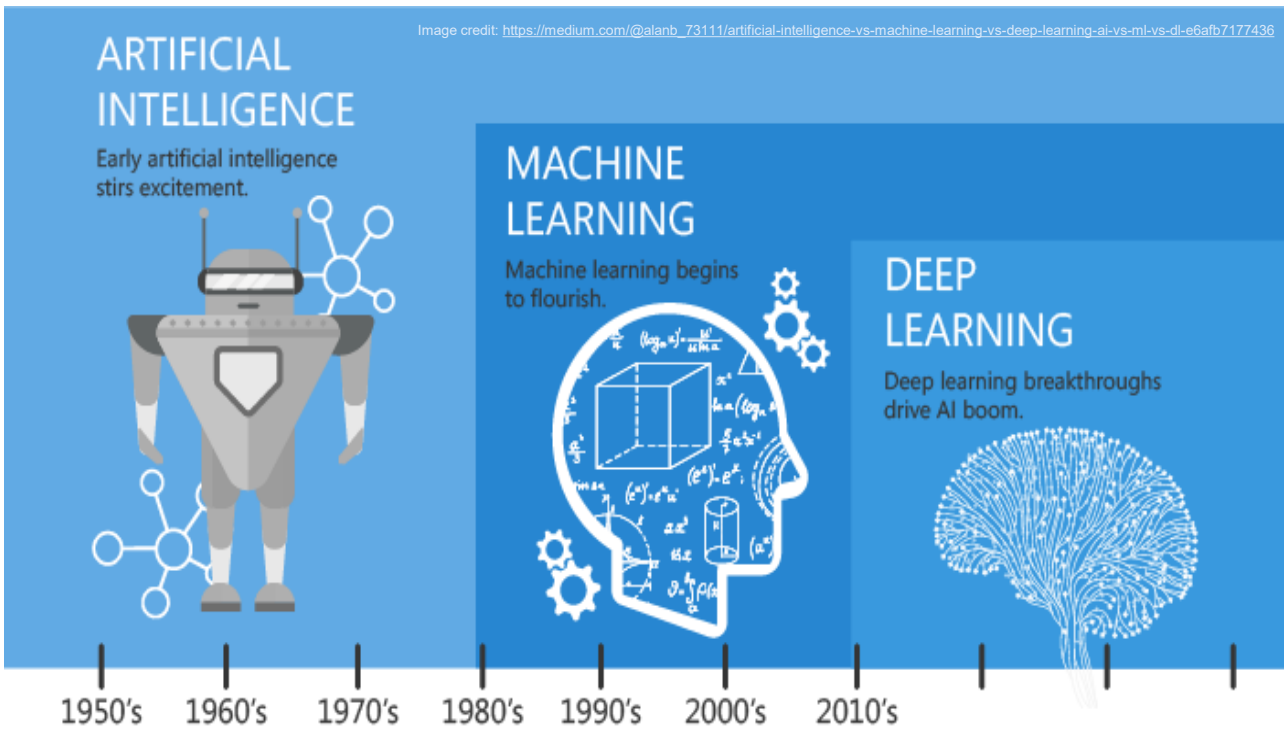
13

14

History of Modern DL

The rapid growth of Deep Learning in the modern era

15



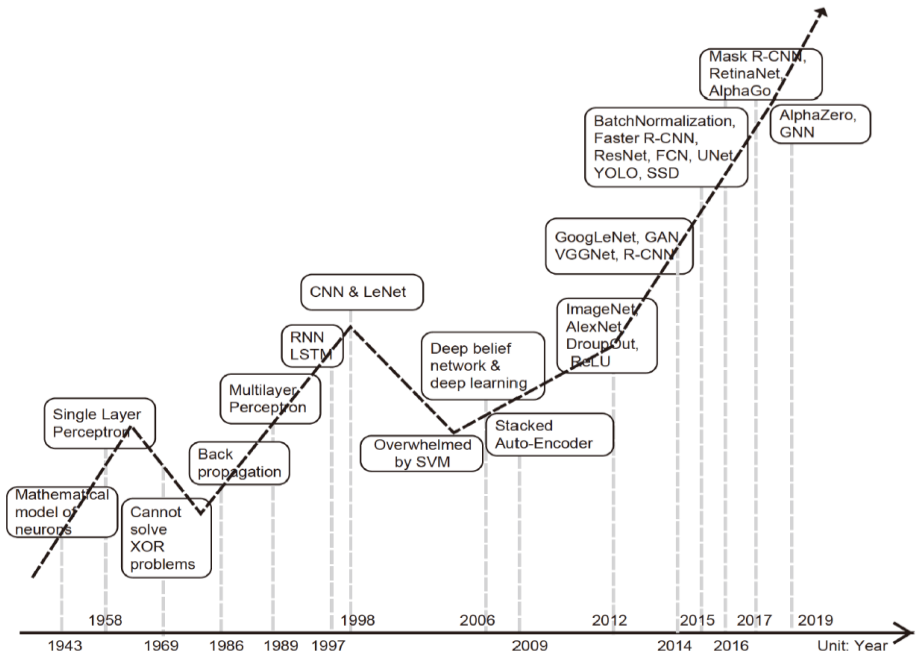
16

The universal approximation theorem

Feed-forward neural networks with at least one hidden layer can approximate any continuous function.

Papers that prove the theorem https://ai.stackexchange.com/questions/13317/where-can-i-find-the-proof-of-the-universal-approximation-theorem
A visual proof that neural nets can compute any function http://neuralnetworksanddeeplearning.com/chap4.html

17



Guo et al., "Application of deep learning in ecological resource research: Theories, methods, and challenges," Science China Earth Sciences, 2020 https://doi.org/10.1007/s11430-019-9584-9

18

THE OLD HISTORY

- **1943:** The first neural network was proposed by McCulloch and Pitts.
- **1955:** John McCarthy first coined the term Artificial Intelligence.
- **1959:** Samuel coined and popularized the term Machine Learning.

- **1982:** Birth of Hopfield Network
- **1986:** Birth of backpropagation
- **1986:** Birth of RNN and AE

- **1997:** Birth of LSTM
- **1998:** Birth of CNN

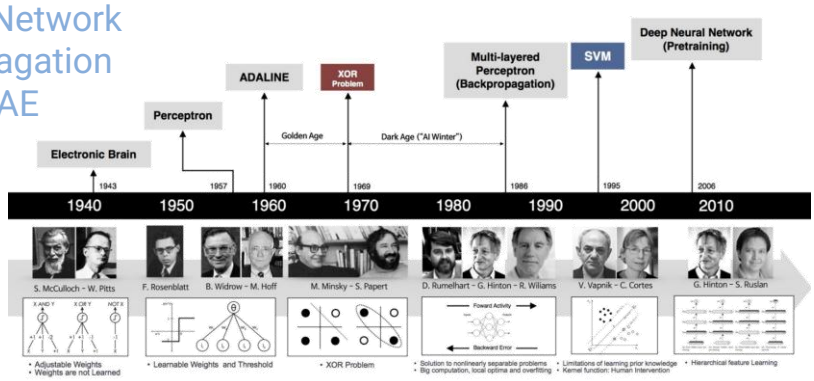


Image credit:
https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

19

THE MODERN HISTORY

- **2006:** The pretraining technique by Hinton et al.

- **2012:** ImageNet evolution—AlexNet
- **2014:** Birth of GRU, VAE, and GAN
- **2015:** Birth of diffusion model
- **2017:** The DeepFake viral
- **2017:** Birth of the groundbreaking Transformer
- **2018:** NLP's ImageNet moment—BERT
- **2018:** ACM Turing Award to Deep NN

20

THE MODERN HISTORY

- **2020:** Birth of ViT
- **2020:** AlphaFold2 "This will change medicine. It will change research. It will change bioengineering. It will change everything."
- **2020:** Self-supervised learning trend in vision

- **2021:** Bridging the gap between vision and NLP—DALL-E and CLIP
- **2021:** The comeback of diffusion models in image generation

- **2022:** The year of Generative AI (massive adoption 🚀): Midjourney (July), Stable Diffusion (August), ChatGPT (November)

- **2023:** Generative AI's breakout year
- **2023:** The homerun year for LLMs
- **2023:** Birth of Flow Matching (Rectified Flow) and Mamba

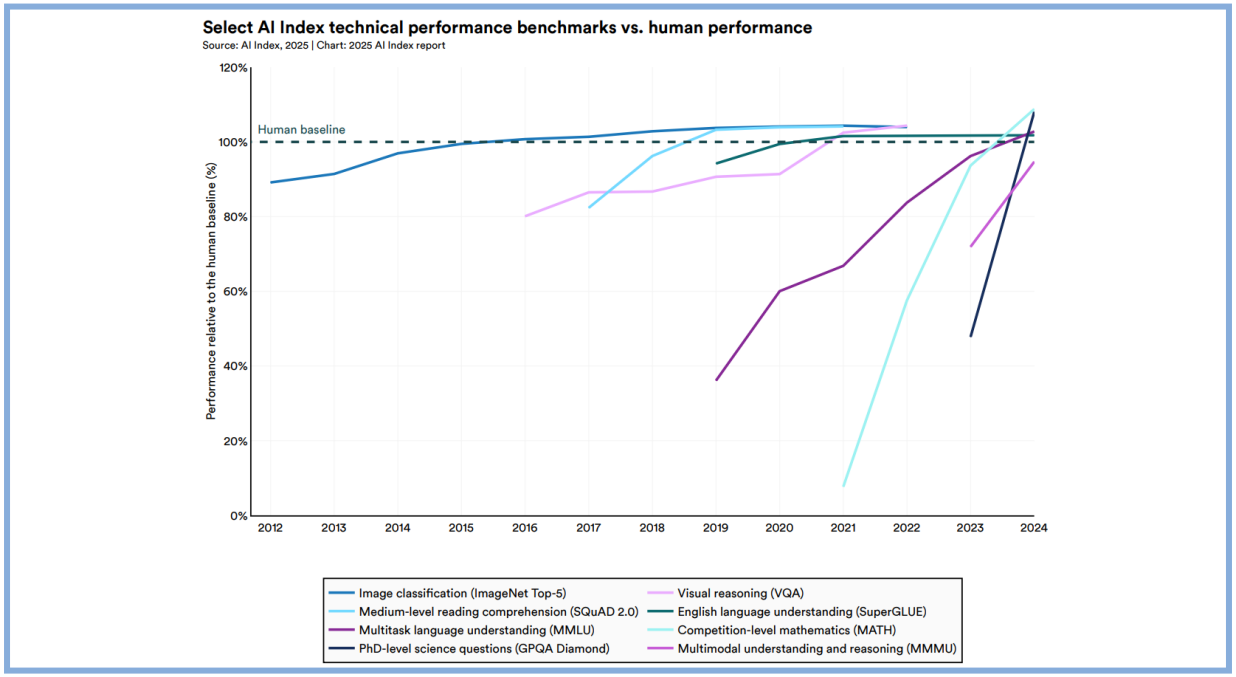
21

THE MODERN HISTORY

- **2024:** Vim, Multimodal LLMs, Agentic AI 🔥, Small Language Models, AI pricing war, Large Reasoning Models (LRMs), Generative video
- **2024:** Nobel Prize in Physics and Chemistry 2024 🏅
- **2024:** ACM Turing Award to RL

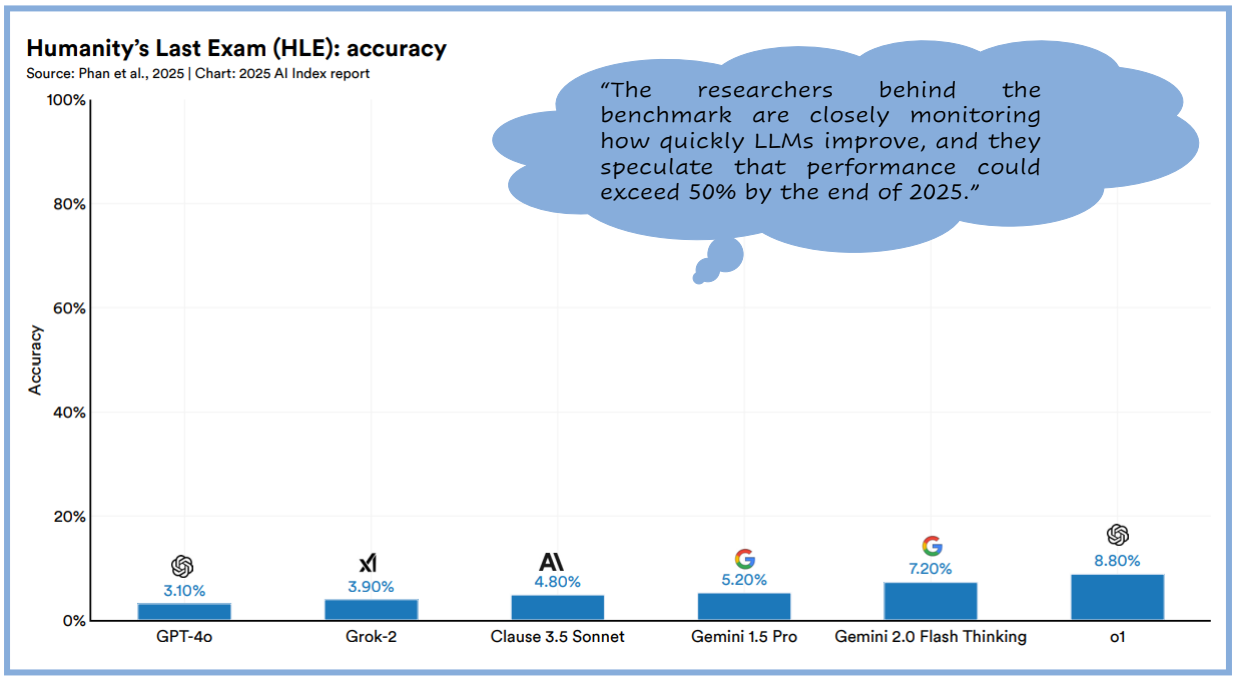
- **2025:** China dominated open-source AI models (DeepSeek, Qwen, Kimi).
  - (January) DeepSeek: the Sputnik moment for RL-based LLM training and Mixture of Experts (MoE) models
- **2025:** The year of agentic AI products (AI coding and more)
- **2025:** The dawn of AI's industrial age

22



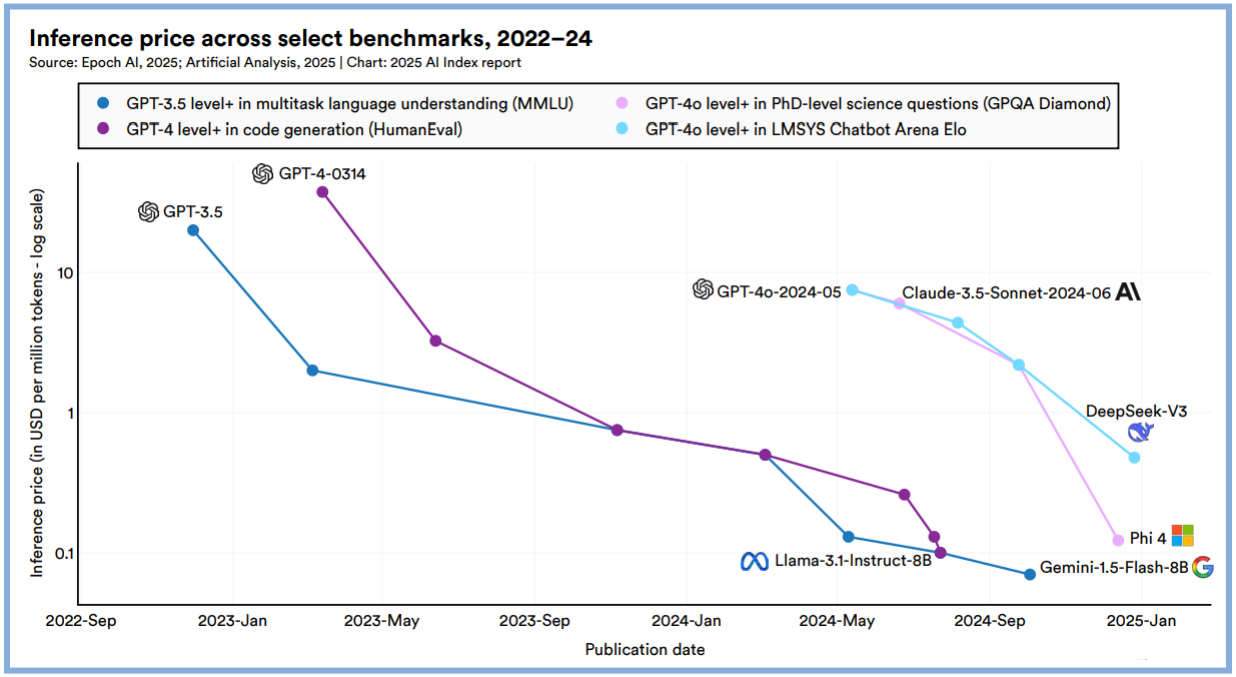
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

23



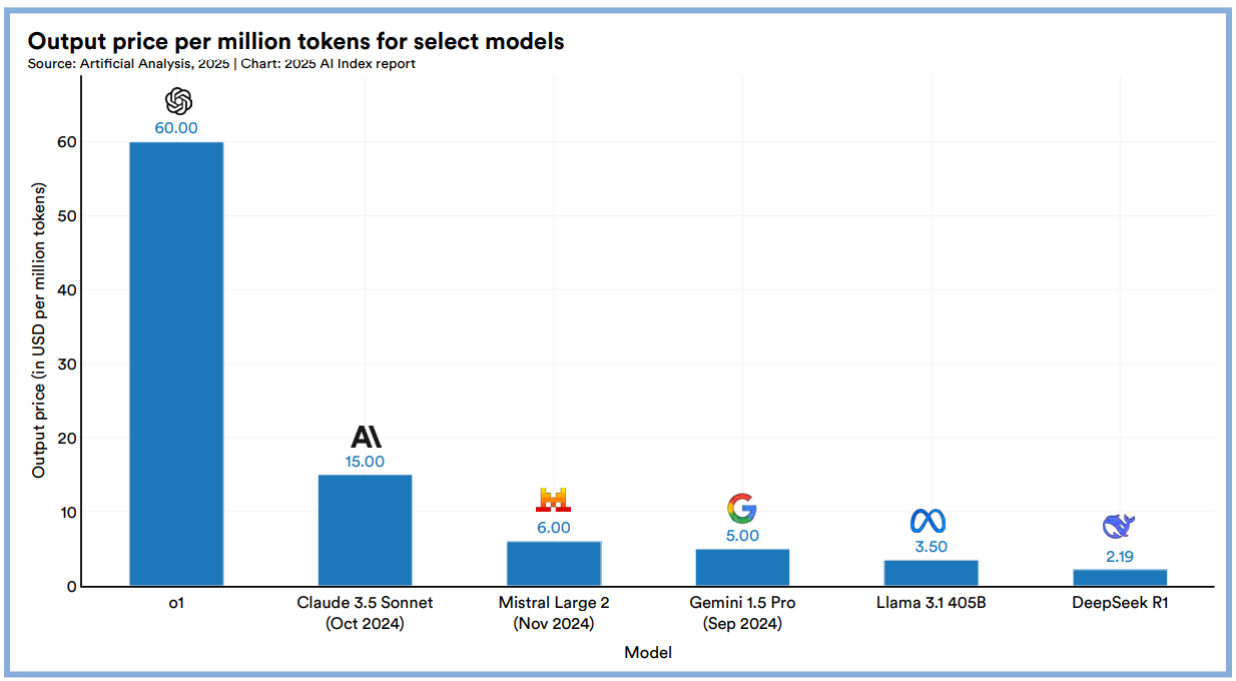
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

24



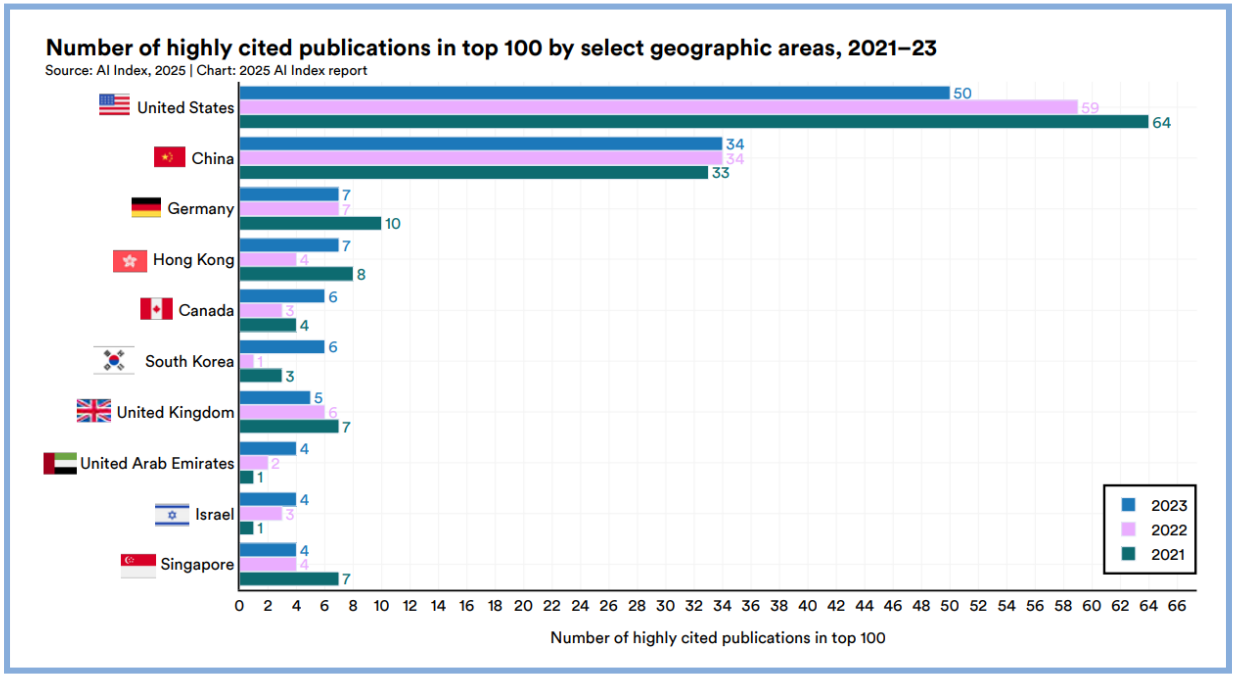
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

25



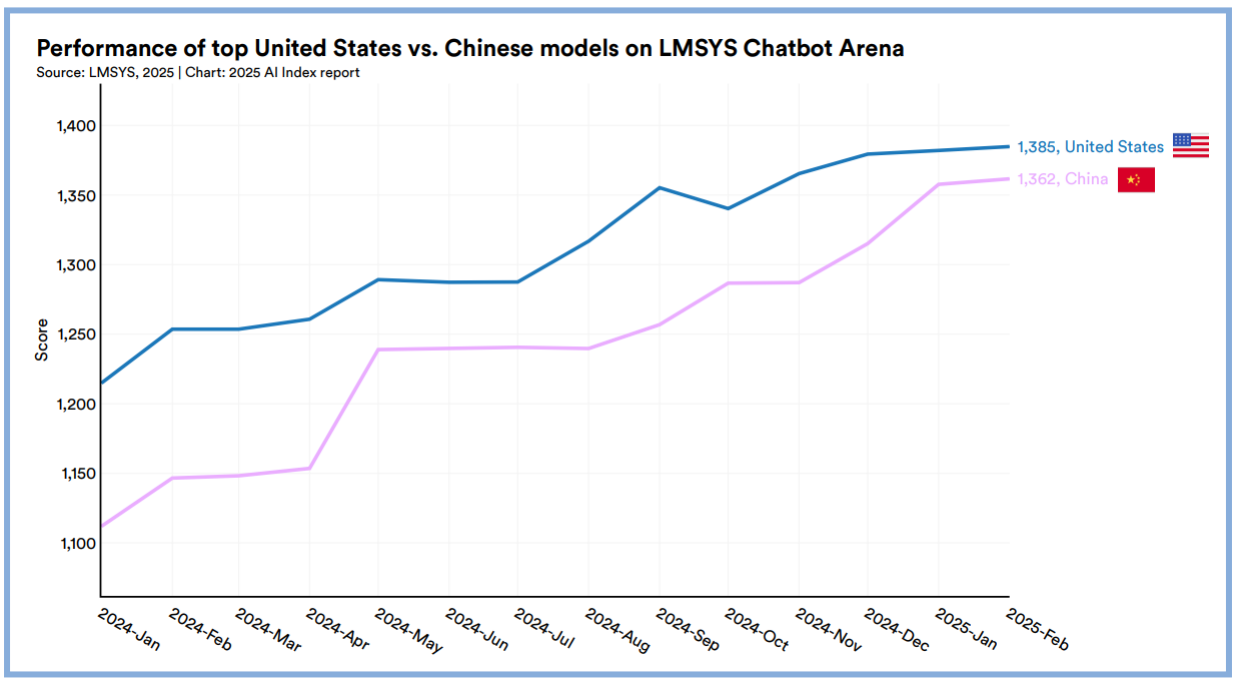
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

26



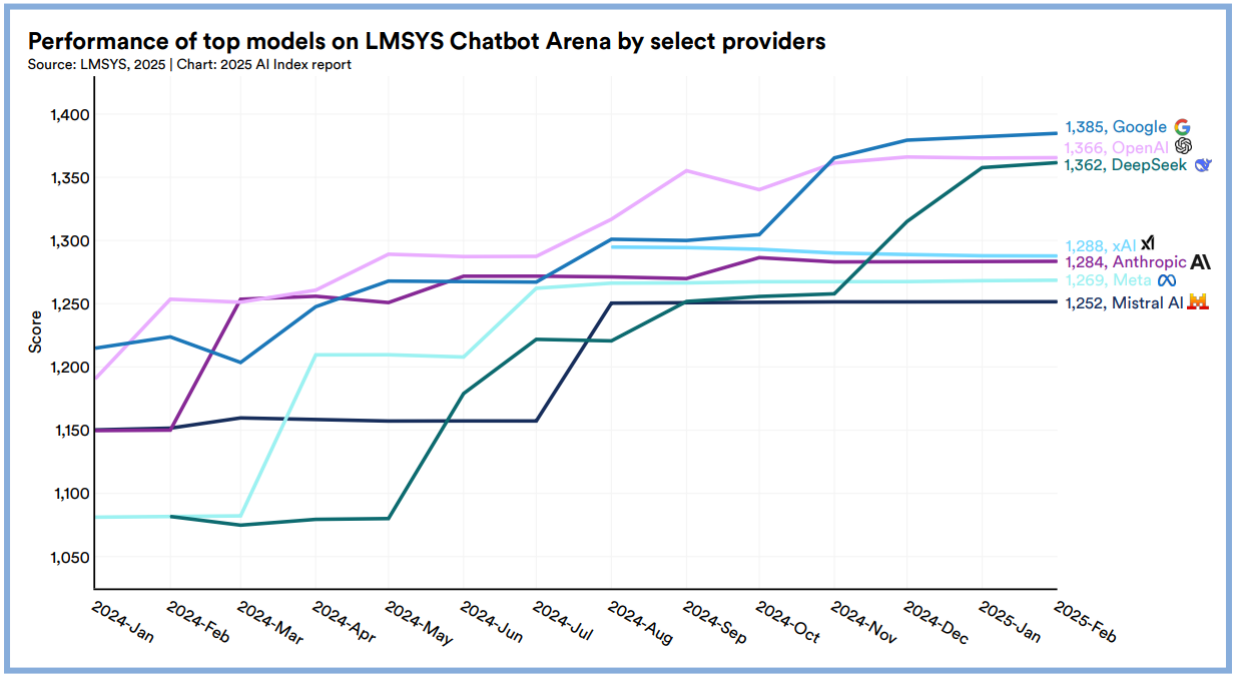
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

27



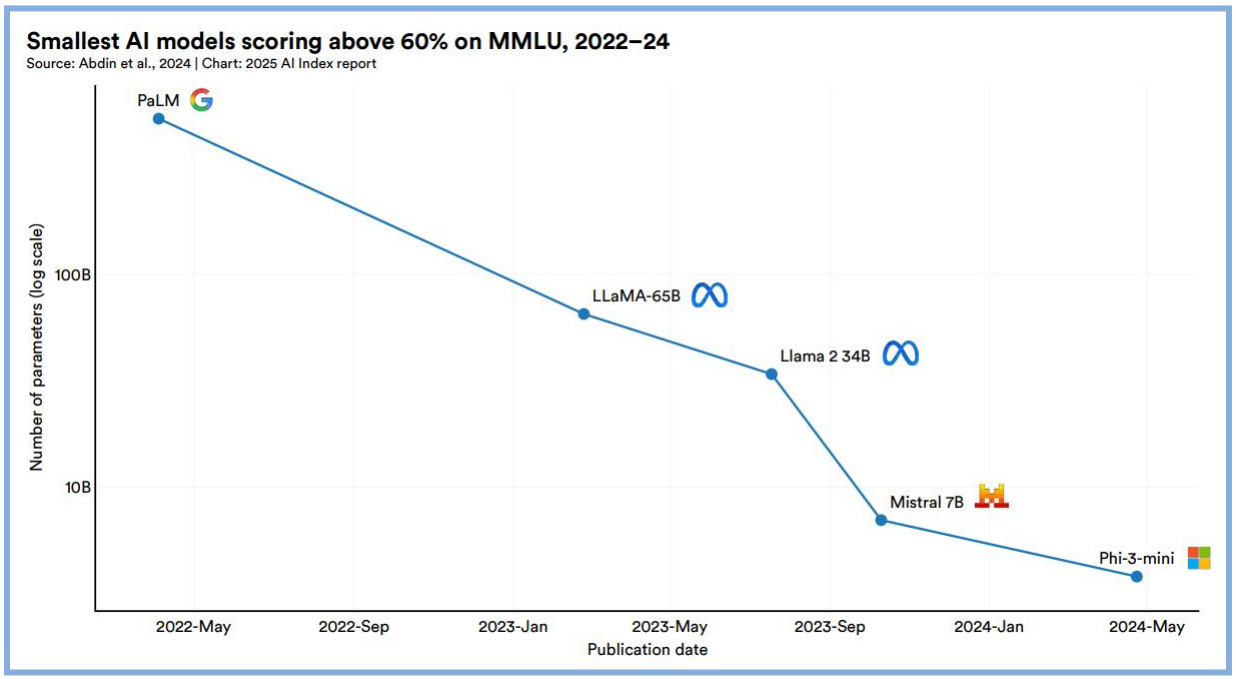
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

28



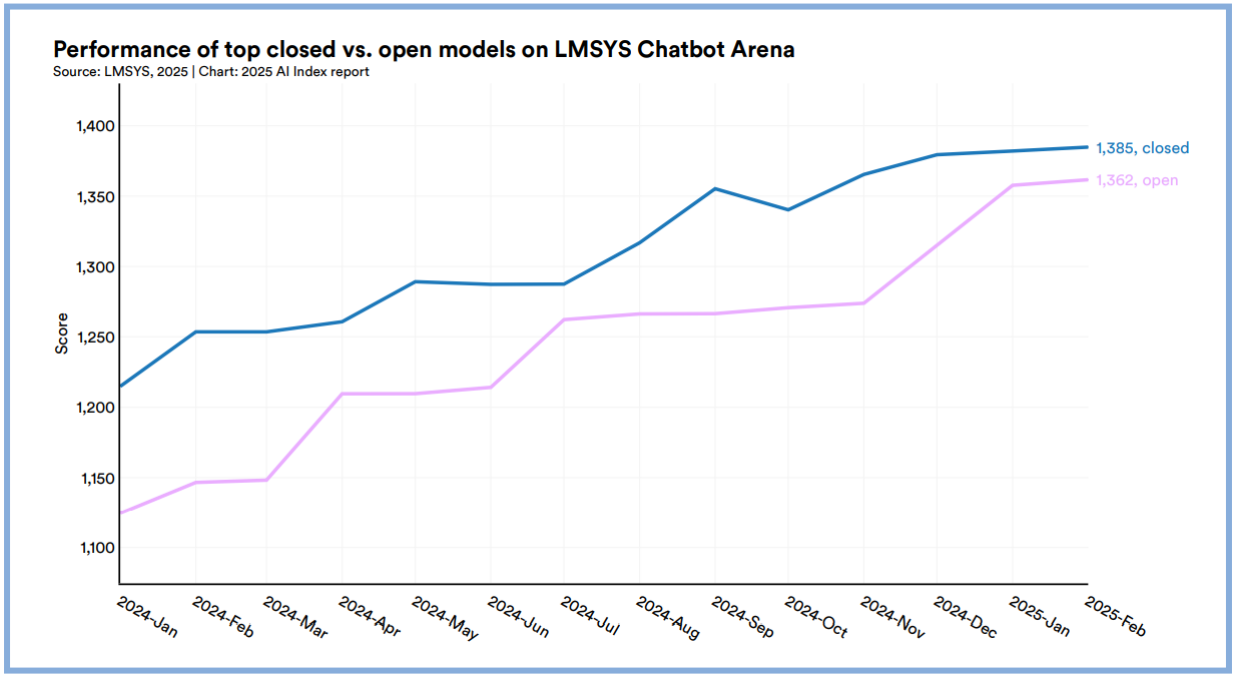
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

29



HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

30



HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

31

DL models in this course

~~DISCRIMINATIVE~~ FUNDAMENTAL

MLP

CNN

RNN

LSTM

GRU

Transformer

SSM

GENERATIVE

VAE

GAN

Diffusion

Flow matching

ChatGPT

32

DL models in this course

~~DISCRIMINATIVE~~ FUNDAMENTAL

GENERATIVE

MLP
The projector

CNN
The conventional image analyzer

RNN
LSTM
GRU
The conventional sequence analyzer

Transformer
The all-rounder

SSM
A new sequence modeling paradigm

VAE

GAN

Diffusion

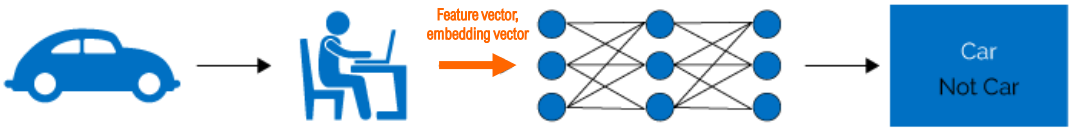
Flow matching

ChatGPT

33

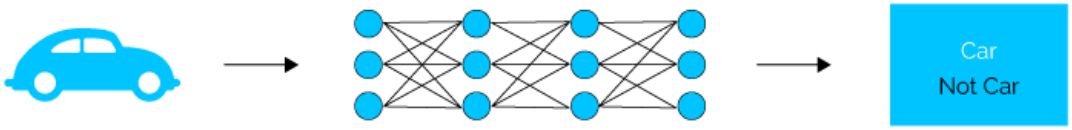
Paradigm shift: ML > DL

Machine Learning



Input → Feature extraction (Representation learning) → Classification → Output

Deep Learning

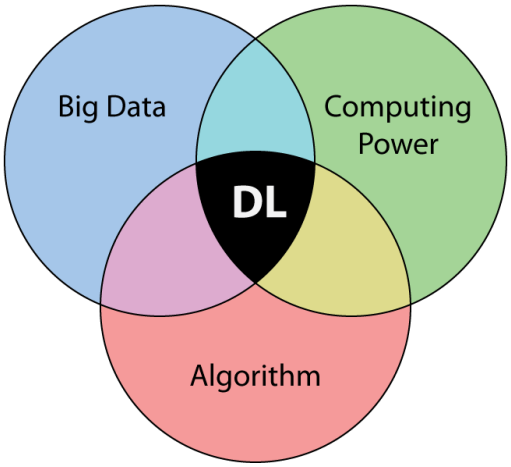


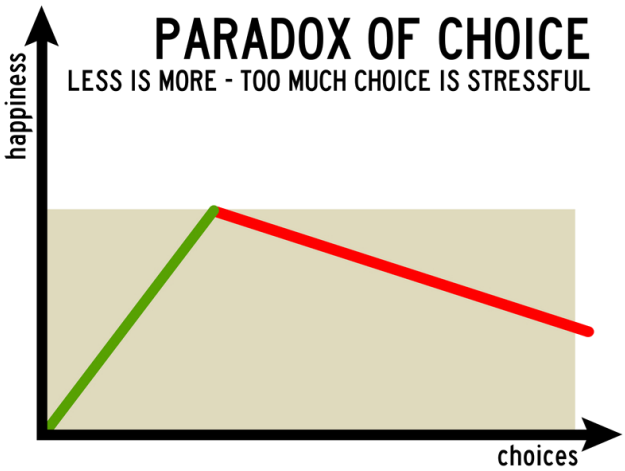
Input → Feature extraction + Classification (Representation learning) → Output

Image credit: https://www.softwaretestinghelp.com/data-mining-vs-machine-learning-vs-ai/

34

WHY Deep Learning





35

WHY NOT Deep Learning

- Deep learning usually requires a huge amount of input data. For supervised learning, data collection and annotation are sometimes very laborious.

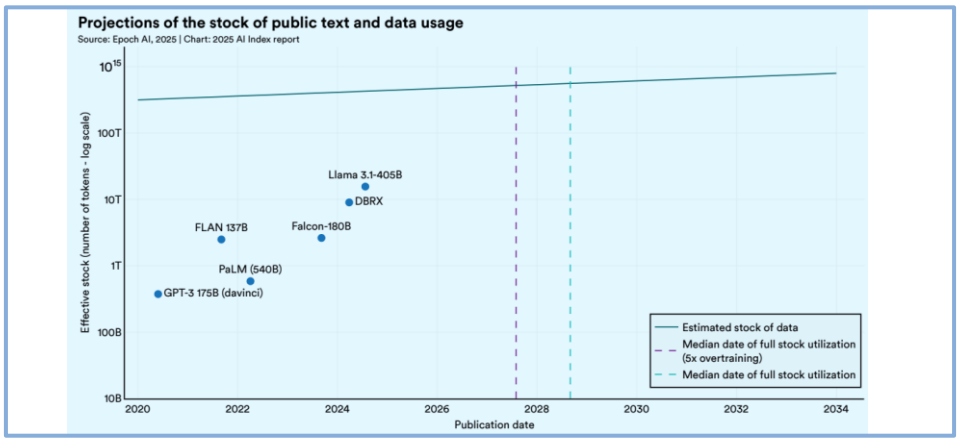


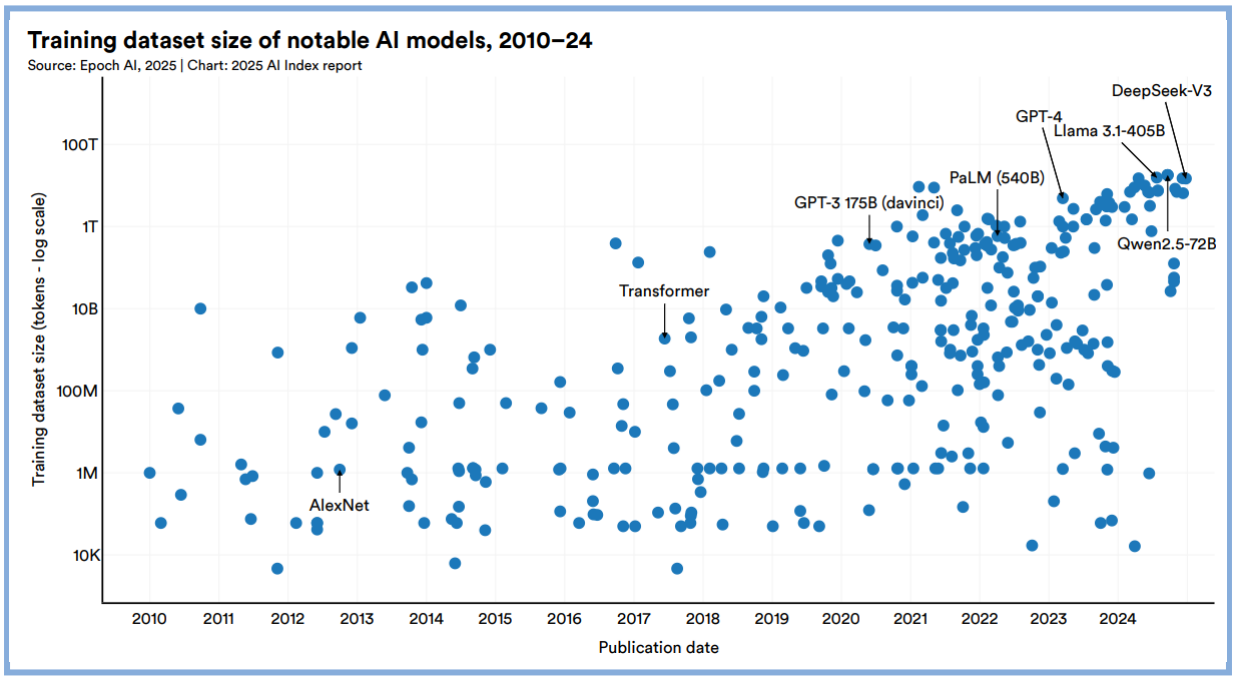
Image credit: HAI AI Index Report, April 2025, https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

36

WHY NOT Deep Learning

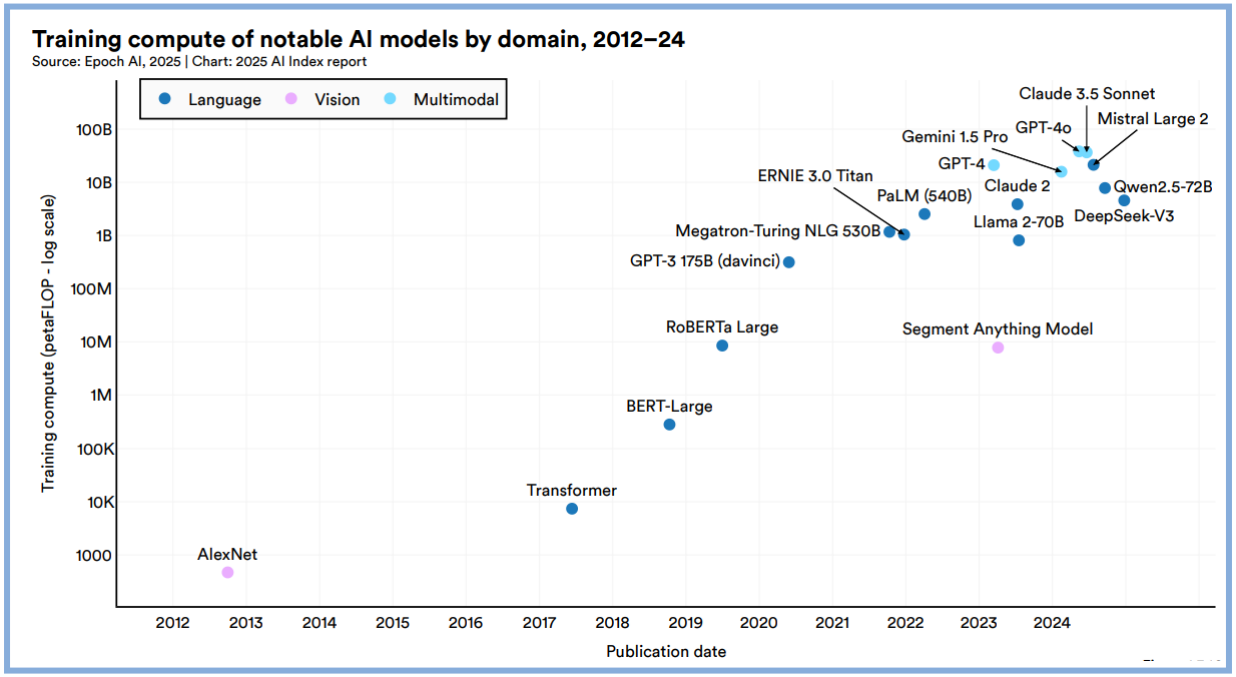
- Training deep networks consumes <u>huge computational resources</u> (GPU memory, training time, and inference time).
  - The largest StyleGAN2 (Nvidia 2019) was trained for 69d 23h on one Tesla V100 GPU.
  - At theoretical 28 TFLOPS for Tesla V100 and lowest 3-year reserved cloud pricing, GPT-3 (OpenAI 2020) was still trained for 355 GPU-years and cost $4.6M.
  - Vision Transformer (Google 2020) was trained for 2,500 TPUv3-core-days.
  - CLIP (OpenAI 2021) was trained for 2 weeks on 256 GPUs.
  - WangchanBERTa (VISTEC x DEPA 2021) was trained for 134 days on eight Tesla V100 GPUs (one DGX-1 server).
  - Stable Diffusion (2022) was trained with 256 A100 GPUs on Amazon Web Services for a total of 150,000 GPU hours, at $600,000.
  - GPT-4 (OpenAI 2023) used an estimated $78 million worth of computing to train.
  - Gemini Ultra (Google 2023) cost $191 million to compute.
  - In 2024, OpenAI is likely to spend inference costs around $4 billion on processing power supplied by Microsoft (at a special discount rate), and expects to spend $3 billion on training models and data.
  - DeepSeek-R1 (DeepSeek 2025) was officially reported as using only $294,000 training cost with 512 NVIDIA H800 GPUs, excluding ~$6 million for the base model.

37



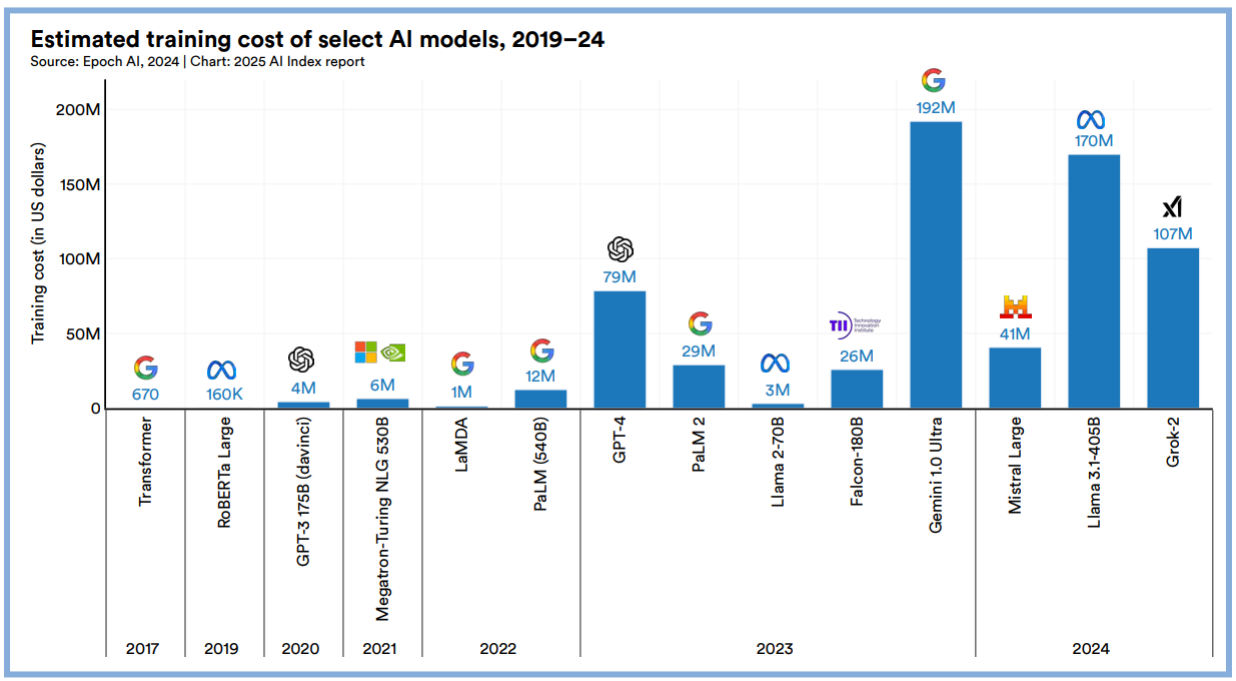
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

38



HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

39

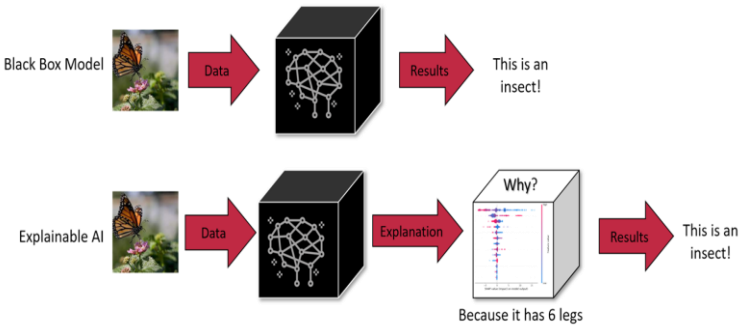


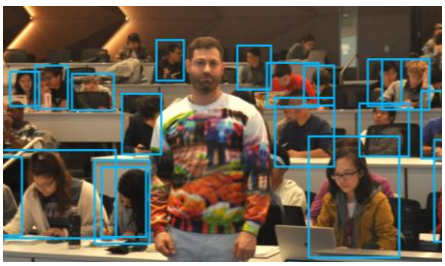
HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

40

WHY NOT Deep Learning

- Deep learning suffers from the infamous black-box issue.
  - Hence, it is necessary to use as many indicators as possible to evaluate DLs
- Deep learning can fail and can be attacked (Adversarial Attack).





Invisible sweater (ECCV2020):
https://www.cs.umd.edu/~tomg/projects/invisible/

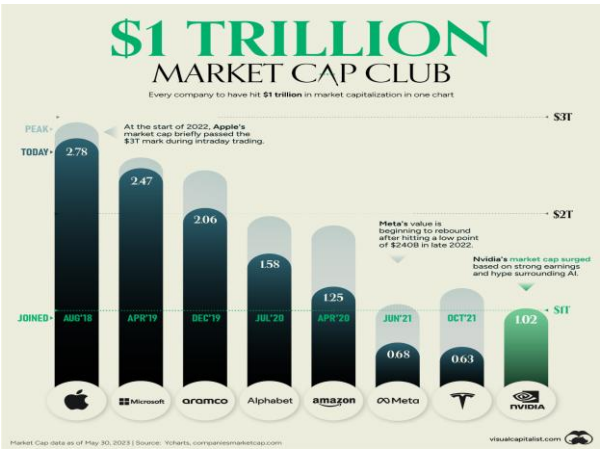
41

Hardware Acceleration

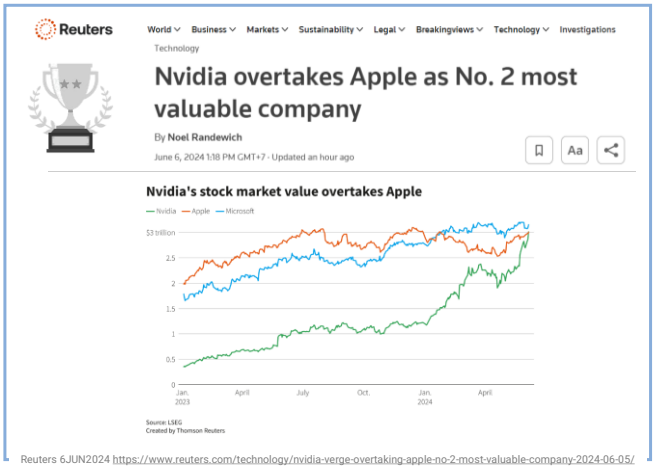
Prepare necessary hardware to accelerate our deep learning project

42

Reuters 31MAY2023 https://www.reuters.com/technology/nvidia-sets-eye-1-trillion-market-value-2023-05-30/



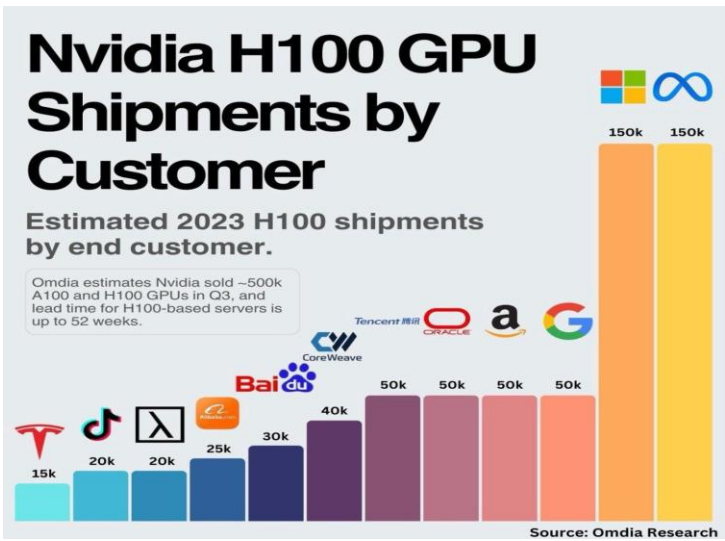
Visual Capitalist 30MAY2023 https://www.visualcapitalist.com/nvidia-joins-the-trillion-dollar-club/

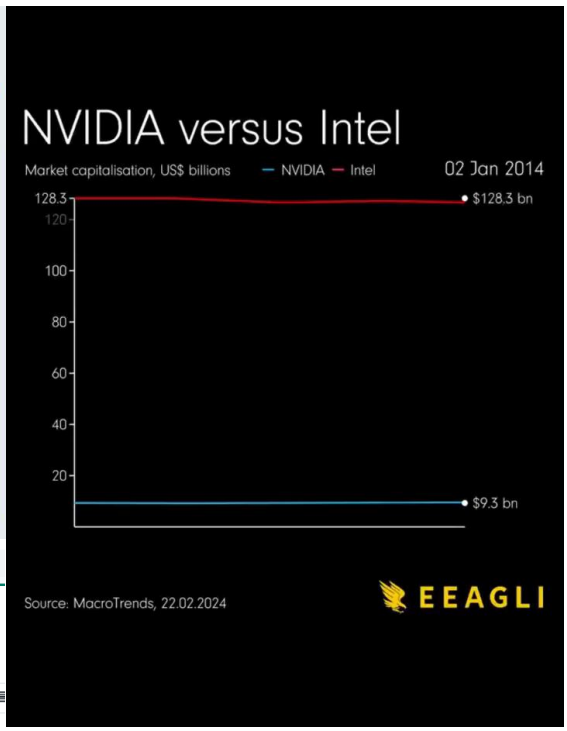


Reuters 6JUN2024 https://www.reuters.com/technology/nvidia-verge-overtaking-apple-no-2-most-valuable-company-2024-06-05/

Reuters 19JUN2024 https://www.reuters.com/technology/view-nvidia-becomes-worlds-most-valuable-company-2024-06-18/

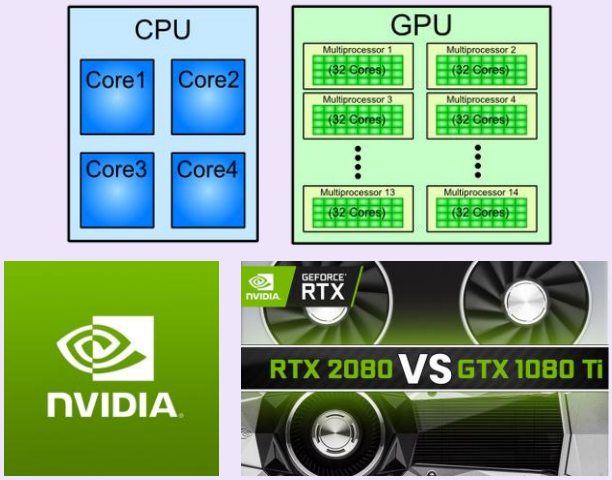
43





44

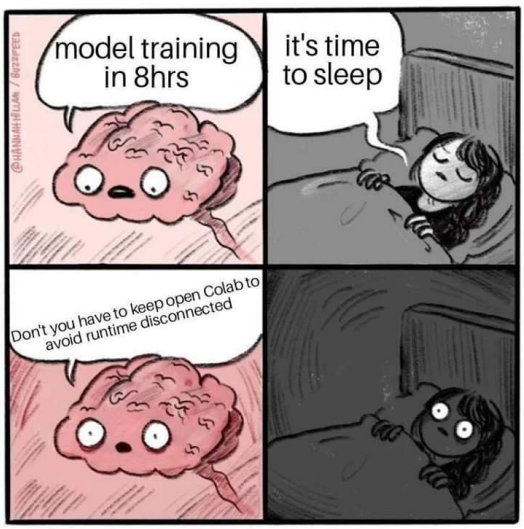
Local Machine vs. Cloud



Local Machine	Cloud

45

REVIEW: Google Colab



Version 1:

```
function ClickConnect(){
  console.log("Connnect Clicked - Start");
  document.querySelector("#top-toolbar > colab-connect-button").shadowRoot.querySele
  console.log("Connnect Clicked - End");
};
setInterval(ClickConnect, 60000)
```

Version 2: If you would like to be able to stop the function, here is the new code:

```
var startClickConnect = function startClickConnect(){
    var clickConnect = function clickConnect(){
        console.log("Connnect Clicked - Start");
        document.querySelector("#top-toolbar > colab-connect-button").shadowRoot.que
        console.log("Connnect Clicked - End");
    };

    var intervalId = setInterval(clickConnect, 60000);

    var stopClickConnectHandler = function stopClickConnect() {
        console.log("Connnect Clicked Stopped - Start");
        clearInterval(intervalId);
        console.log("Connnect Clicked Stopped - End");
    };

    return stopClickConnectHandler;
};

var stopClickConnect = startClickConnect();
```

In order to stop, call:

```
stopClickConnect();
```

UNOFFICIAL

Share Follow

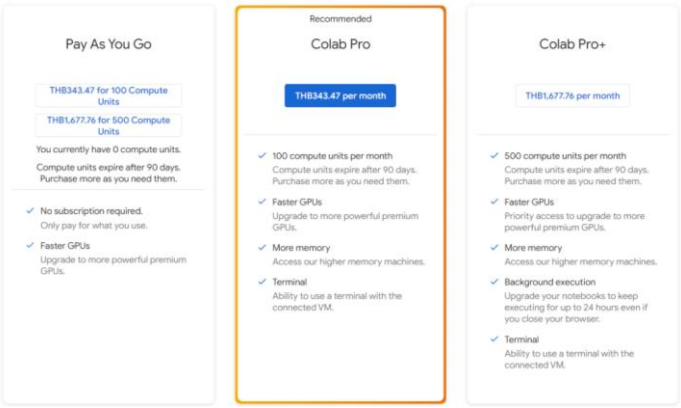
edited Jun 7, 2020 at 17:15 Christophe Prat 17 ● 5

answered Mar 7, 2020 at 15:27 barbossusus 539 ● 4 ● 4

Credit: https://stackoverflow.com/questions/57113226/how-can-i-prevent-google-colab-from-disconnecting

46

REVIEW: Google Colab Pro+



Update on AUG2023:
- Nvidia A100 GPU is available for both Colab Pro and Pro+.
- The better the GPU, the more compute units it will consume.
- When running out of compute units, we can simply buy more compute units to continue running.

- **Good:** 
  - Convenient as we can train the model while being far away from our computer or even when closing the computer
  - At least 10+ hours of continuous running without interruption
- **Bad:** 
  - After 2 continuous weeks of heavy GPU usage (e.g., 4 concurrent tabs), it banned us from further GPU usage with a suggestion to move to Google Cloud ML.

Review from a real user (using Google Colab Pro+ during January-February 2022)

47

TPU

- **TPU** is an AI accelerator hardware specially designed for neural networks and is proprietary to Google.
- **TPU** has been used to power Google's data centers internally since 2015 and in 2018 made available for 3rd party use.
- According to Google (in 2017), **TPU**-v1 was about 80x faster than CPU and 30x faster than GPU at neural network inference.
- Currently, **TPU** is mostly available as a cloud service or a smaller version of the chip (Edge TPU).

TPU v1
Launched in 2015
Inference only

TPU v2
Launched in 2017
Inference and training

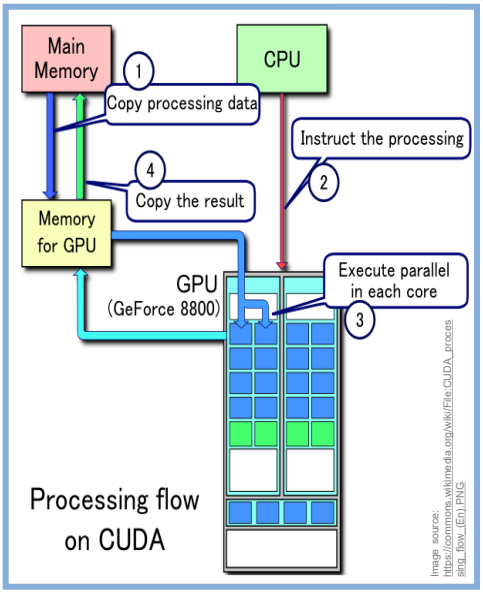
5APR2017: https://cloud.google.com/blog/products/gcp/quantifying-the-performance-of-the-tpu-our-first-machine-learning-chip

48

CPU vs. GPU vs. TPU

- **Overall, TPU** has the highest training throughput (the number of examples trained per second).
- **Deep FC networks:**
  - For a big batch size, it is **TPU**, GPU and CPU respectively.
  - For a big model, **GPU** is the best.
  - For a big batch size and model, **GPU** is better than CPU because of better parallelization.
  - For very big models, **CPU** is better than GPU, and GPU is better than TPU. This is because CPU has the highest memory per core.
- **Deep CNNs:**
  - **TPU** is better than GPU and is the best for training CNN, particularly very big CNNs. This is because TPU is specifically optimized for spatial reuse characteristics of CNN.
- **RNNs:**
  - For speed aspect, **TPU** is a lot better than GPU.
  - Right now, TPU is good at dense computation (MatMul) whereas GPU is better in non-MatMul operations. In the future, if TPU is further optimized for non-MatMul operations, it will do even better than GPU.

[31JUL2019] Harvard University, https://arxiv.org/abs/1907.10701



49

Deep Learning's Computer

Why not Mac OS? [23NOV2019] Apple and Nvidia are over https://gizmodo.com/apple-and-nvidia-are-over-1840015246
How much GPU's RAM for training DL models? [18FEB2020] Choosing the Best GPU for Deep Learning in 2020 https://lambdalabs.com/blog/choosing-a-gpu-for-deep-learning/
Why Mac OS? [15NOV2020] How is the Apple M1 going to affect Machine Learning? https://medium.com/disruptive-nerd/how-is-the-apple-m1-going-to-affect-machine-learning-2d9da1beef86

50

51

Neural Processing Unit (NPU)

NVIDIA

Home | AI | Data Center | Driving | Gaming | Pro Graphics | Robotics | Healthcare | Startups | AI Podcast | NVIDIA Life

NVIDIA RTX 500 and 1000 Professional Ada Generation Laptop GPUs Drive AI-Enhanced Workflows From Anywhere

Thin and light designs deliver advanced AI, compute and graphics horsepower for professionals on the go.

February 26, 2024 by John Della Bona

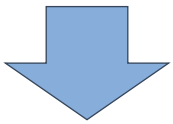
The next generation of mobile workstations with Ada Generation GPUs, including the RTX 500 and 1000 GPUs, will include both a neural processing unit (NPU), a component of the CPU, and an NVIDIA RTX GPU, which includes Tensor Cores for AI processing. The NPU helps offload light AI tasks, while the GPU provides up to an additional 682 TOPS of AI performance for more demanding day-to-day AI workflows.

26FEB2024 https://blogs.nvidia.com/blog/rtx-ada-ai-workflows/

52

Neural Processing Unit (NPU)

- **Problems of GPU:**
  - **Still rely on CPU:** Although GPU has the advantage in parallel computing capability, it does not work alone and needs the CPU's co-processing. The construction of neural network models and data streams are still carried out on the CPU.
  - **The problem of high-power consumption and large size:** The higher the performance, the larger the GPU, the higher the power consumption, and the more expensive it is, which will not be available for some small devices and mobile devices.



- We need a small size, low power consumption, high computational performance, and high computational efficiency of a dedicated chip. This is the birth of **NPU**.

Credit: (28DEC2021) https://www.utmel.com/blog/categories/integrated%20circuit/neural-processing-unit-npu-explained

53

Neural Processing Unit (NPU)

- **NPUs** are becoming increasingly common in PC, laptop, and mobile domains. It usually refers to all specialized AI processors.
  - For example: AI chip in Apple iPhone, Exynos 9 (9820) the AI-enabled NPU in Samsung mobile

- **NPUs** are designed to complement the functions of CPUs and GPUs, ensuring that no single processor gets overwhelmed, maintaining smooth operation across the system.
  - CPUs handle a broad range of tasks.
  - GPUs excel in rendering detailed graphics.
  - **NPUs** specialize in executing AI-driven tasks swiftly.

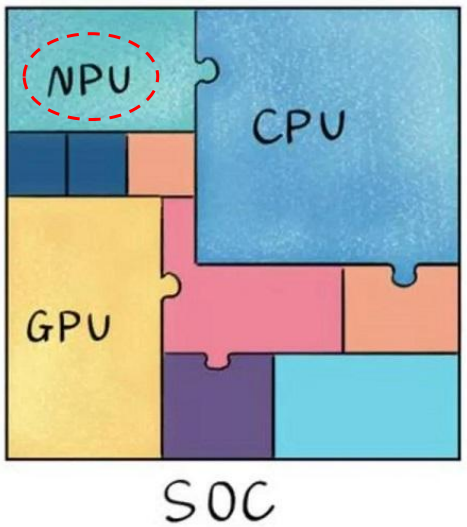


Image source: (28DEC2021) https://www.utmel.com/blog/categories/integrated%20circuit/neural-processing-unit-npu-explained

54

Neural Processing Unit (NPU)

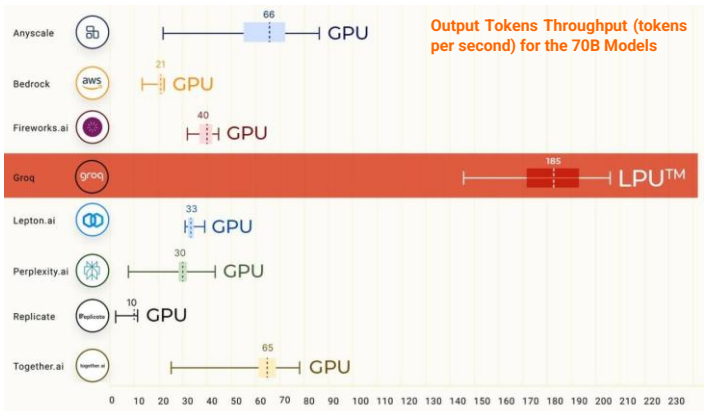
- **CPU vs. GPU vs. NPU:**
  - Unlike traditional CPUs and GPUs, an **NPU**, at its core, is a specialized processor explicitly designed/optimized for executing AI/ML/DL algorithms, particularly for ANN mathematical computation and DL acceleration.
  - This degree of specialization allows **NPUs** to deliver significantly higher performance for AI workloads compared to CPUs and even GPUs in certain scenarios.
    - **NPUs** excel in processing vast amounts of data in parallel, making them ideal for tasks like image recognition, natural language processing, and other AI-related functions.
    - For example, if you were to have an **NPU** within a GPU, the **NPU** may be responsible for a specific task like object detection or image acceleration.
  - The concept of **GPNPU** (GPU-NPU hybrid) has emerged, aiming to combine the strengths of GPUs and **NPUs**. GPNPUs leverage the parallel processing capabilities of GPUs while integrating **NPU** architecture for accelerating AI-centric tasks.

Credit: (28DEC2021) https://www.utmel.com/blog/categories/integrated%20circuit/neural-processing-unit-npu-explained

55

Language Processing Unit (LPU)

- In February 2024, a largely unknown company, **Groq** (not to be confused with Elon's Grok AI) demonstrated unprecedented speed running open-source LLMs such as Llama-2 (70 billion parameters) at more than 100 tokens per second, and Mixtral at nearly 500 tokens per second per user on a Groq's **Language Processing Unit (LPU)**.
- **LPU** is a special kind of computer brain designed to be highly specialized for language-intensive tasks, boasting optimized hardware and software. In short, it uses faster processing times and lower power consumption.
- **LPU** employs sequential processing, meticulously handling tasks step-by-step, mirroring the natural flow of language.



Source
- 26FEB2024 https://www.linkedin.com/pulse/new-ai-compute-paradigm-language-processing-unit-lpu-dheeren-v%C3%A9lu-n22bc
- 27FEB2024 https://www.analyticsvidhya.com/blog/2024/02/what-is-the-difference-between-lpu-and-gpu/

56

Forbes

Cerebras Update: The Wafer Scale Engine 3 Is A Door Opener

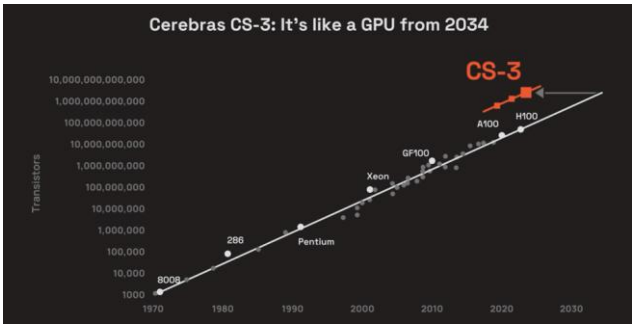
Karl Freund Contributor
Founder and Principal Analyst, Cambrian-AI Research LLC

Follow

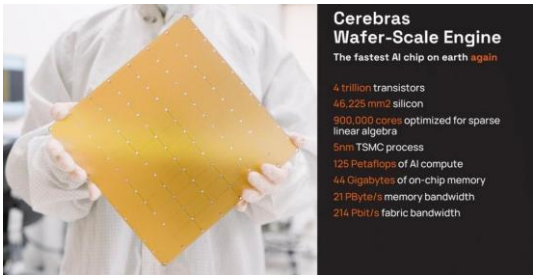
0

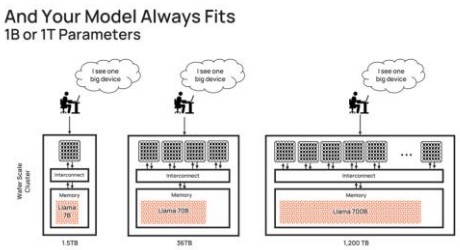
Mar 25, 2024, 03:16pm EDT

Updated Mar 25, 2024, 06:19pm EDT



AI Supercomputer





25MAR2024: https://www.forbes.com/sites/karlfreund/2024/03/25/cerebras-update-the-wafer-scale-engine-3-is-a-door-opener/

57

AI startup Cerebras debuts 'world's fastest inference' service - with a twist

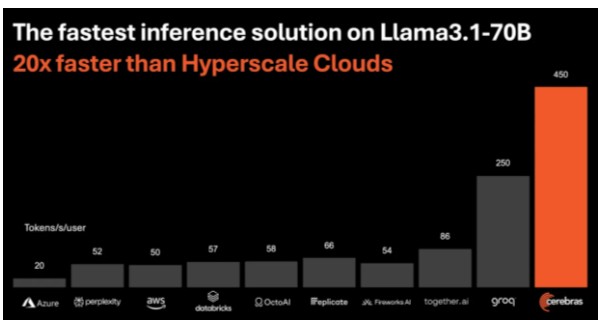
The AI computer maker claims its inference service is dramatically faster and makes new kinds of 'agentic' AI possible.

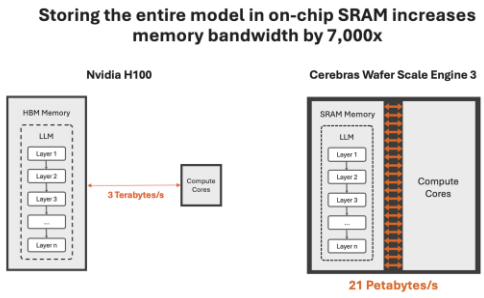
Written by **Tiernan Ray,** Senior Contributing Writer

Aug. 27, 2024 at 6:03 p.m. PT

Cerebras vs Nvidia H100
Llama3.1-70B

	Tokens/sec/user	Cost per M tokens [1]
Cerebras	450	$0.60
Hyperscale H100 Cloud	20	$2.90 [2]
X-factor	22x	1/5





27AUG2024: https://cerebras.ai/blog/introducing-cerebras-inference-ai-at-instant-speed , https://www.zdnet.com/article/ai-startup-cerebras-debuts-worlds-fastest-inference-with-a-twist/

58

CNBC MARKETS BUSINESS INVESTING TECH POLITICS VIDEO INVESTING CLUB JOIN PRO JOIN LIVESTREAM

TECH

Nvidia buying AI chip startup Groq's assets for about $20 billion in its largest deal on record

PUBLISHED WED, DEC 24 2025•3:54 PM EST | UPDATED FRI, DEC 26 2025•9:14 AM EST

David Faber
@DAVIDFABER

SHARE

KEY POINTS

- Nvidia is making its largest purchase ever, acquiring assets from 9-year-old chip startup Groq for about $20 billion.
- The company was founded by creators of Google's tensor processing unit, or TPU, which competes with Nvidia for artificial intelligence workloads.
- Groq, which was valued at $6.9 billion in a financing round in September, framed the deal as a "non-exclusive licensing agreement," with its CEO and other senior leaders joining Nvidia.

24DEC2025: https://www.cnbc.com/2025/12/24/nvidia-buying-ai-chip-startup-groq-for-about-20-billion-biggest-deal.html

59

NOTEBOOKCHECK

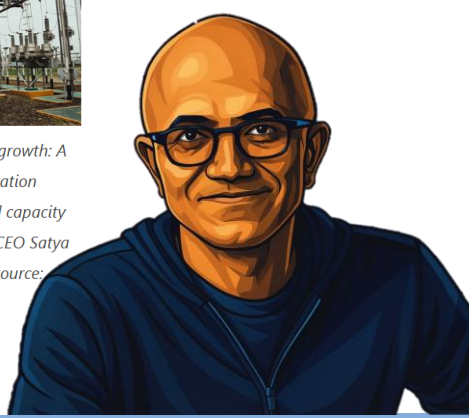
Reviews | News | Videos | Benchmarks / Tech | Buyers Guide | Magazine | Library

The GPU war is over: Nadella says AI's new ceiling is the power grid

Microsoft CEO Satya Nadella confirms the GPU shortage is over. The new AI bottleneck is power: Microsoft has chips sitting in inventory it can't plug in. Why the race for data center grid capacity and Small Modular Reactors (SMRs) is now the real AI power crunch.

Darryl Linington, Published 12/07/2025 ES PT ... AI Software

The real bottleneck for AI growth: A typical electrical substation representing the local grid capacity challenges that Microsoft CEO Satya Nadella cited. (Image source: Freepik.com)



> Microsoft CEO Satya Nadella says the AI boom has hit a new kind of ceiling... and it's not GPUs.

Speaking on the BG2 podcast alongside OpenAI CEO Sam Altman, Nadella said Microsoft is "not chip supply constrained" anymore. The real problem, he explained, is finding enough powered, fully built-out data centres — the "warm shells" close to grid capacity — actually to switch all those accelerators on.

1NOV2025: https://www.youtube.com/watch?v=Gnl833wXRz0 , https://www.notebookcheck.net/The-GPU-war-is-over-Nadella-says-AI-s-new-ceiling-is-the-power-grid.1180290.0.html

60

DL Frameworks

Which framework to use in our deep learning project

61

Deep Learning Frameworks

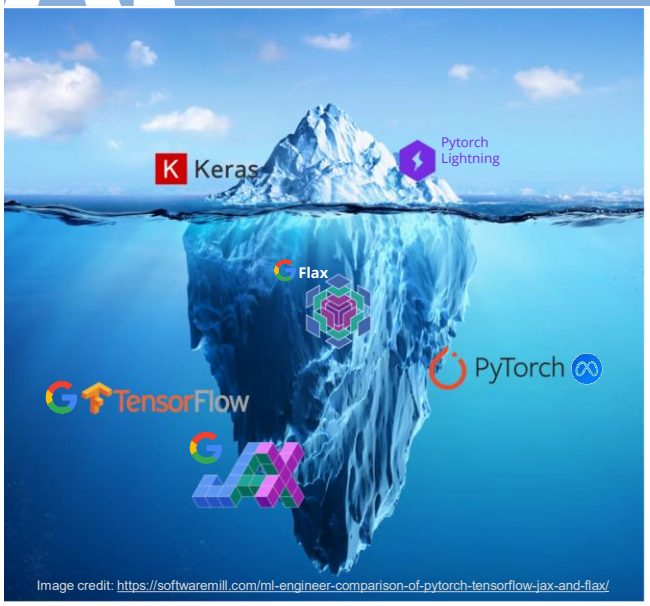
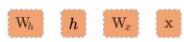


Image credit: https://softwaremill.com/ml-engineer-comparison-of-pytorch-tensorflow-jax-and-flax/

- Most deep learning frameworks can be used for two different things:
  - Replacing Numpy-like operations with GPU-accelerated operations
  - Building deep neural networks
- **TensorFlow** (before 2.0) used static computation graphs. However, **PyTorch** uses dynamic computation graph, allowing greater flexibility in building complex architectures.

A graph is created on the fly




```
from torch.autograd import Variable

x = Variable(torch.randn(1, 10))
prev_h = Variable(torch.randn(1, 20))
W_h = Variable(torch.randn(20, 20))
W_x = Variable(torch.randn(20, 10))
```

62

Static vs. Dynamic Computation Graphs

- Most static declaration paradigms use a two-step process:
  1. Define a computational architecture that is represented as a computational graph
     - For example, take a 64x64 image and perform two layers of convolution, predict the class of the image out of 100 classes, and optionally calculate the loss with respect to the correct label
  2. Execute the computation
     - For example, a user repeatedly populates the 64x64 matrix with data and the library executes the previously declared computation graph. At test time, the predictions can be used. At training time, the loss can be calculated and back-propagated through the graph to compute the gradients required for parameter updates
- Despite a number of advantages, there are many scenarios where it is difficult for simple static declaration tools to handle.
  - For example, variably sized inputs, variably structured inputs, non-trivial inference algorithms, variably structured outputs
- In the dynamic declaration paradigm, a user defines the computation graph programmatically. There are no separate steps for definition and execution: the necessary computation graph is created, on the fly, as the loss calculation is executed, and a new graph is created for each training instance.
- Recommended reading materials that comparatively explain the pros and cons of each paradigm:
  - [15JAN2017] DyNet: The Dynamic Neural Network Toolkit https://arxiv.org/abs/1701.03980
  - [11MAR2019] https://medium.com/analytics-vidhya/dynamic-vs-static-computation-graph-2579d1934ecf

63

Deep Learning Frameworks

python™

Keras

since 2015 (with several backends including TF)
Since 6/2019, officially included in TF 2.0
since 2023, Keras 3.0 includes full APIs for TF, JAX, and Torch backends

PYTORCH

since October 2016

since February 2024 (stable release)

TensorFlow

since 2015
since 6/2019 (TF 2.0), include keras's high-level APIs

Simply change from:

```
from keras.XXX import YYY
```

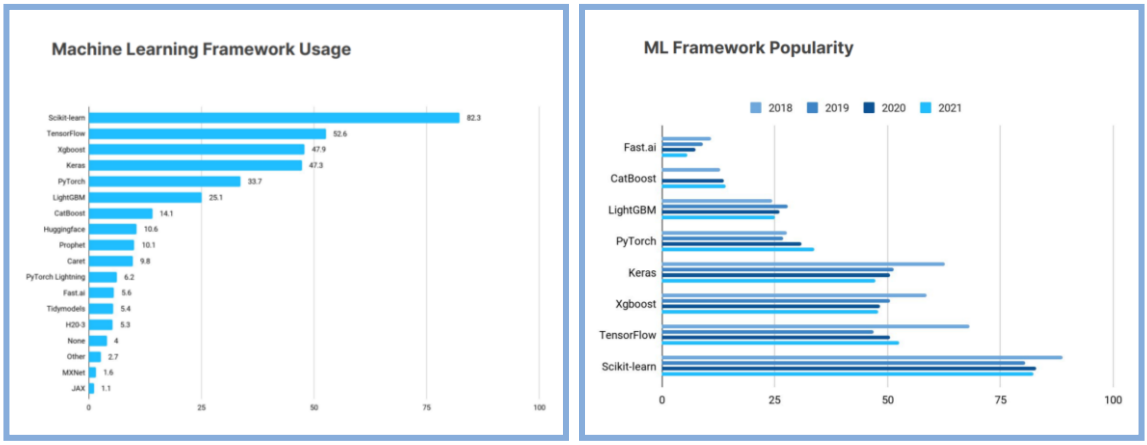
To:

```
from tensorflow.keras.XXX import YYY
```

64

Google TensorFlow vs. Meta PyTorch

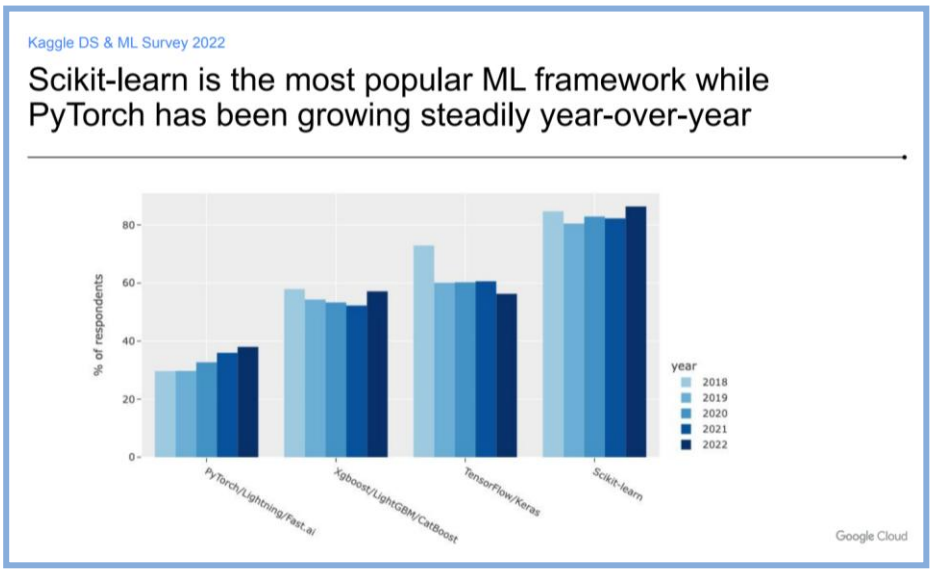
- Previously, TensorFlow took the lead when considering the number of users. It has been a go-to framework for deployment-oriented applications.



[14OCT2021] Kaggle: https://www.kaggle.com/kaggle-survey-2021

65

Google TensorFlow vs. Meta PyTorch



[11-13OCT2022] Kaggle: https://www.kaggle.com/kaggle-survey-2022

66

Google TensorFlow vs. Meta PyTorch

- PyTorch quickly became the preferred choice within research/developer communities.

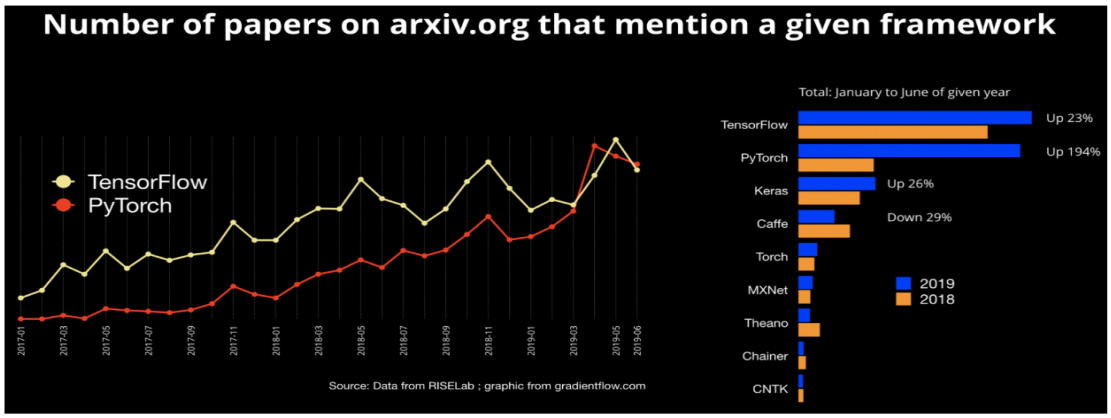


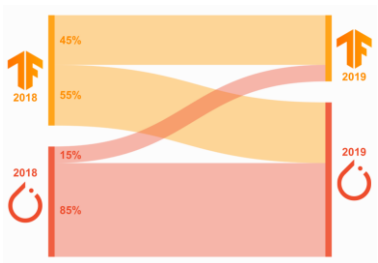
Figure 1. Number of papers posted on arXiv.org that mention each framework. Source: Data from RISELab and graphic by Ben Lorica.

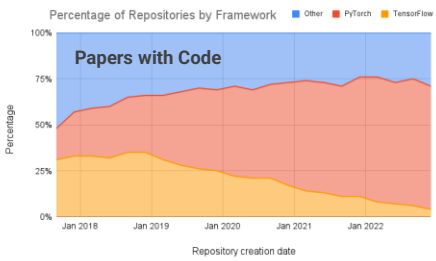
[JULY2019] https://www.oreilly.com/content/one-simple-graphic-researchers-love-pytorch-and-tensorflow/

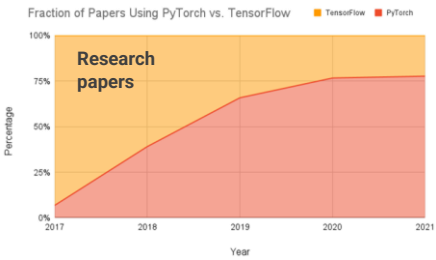
67

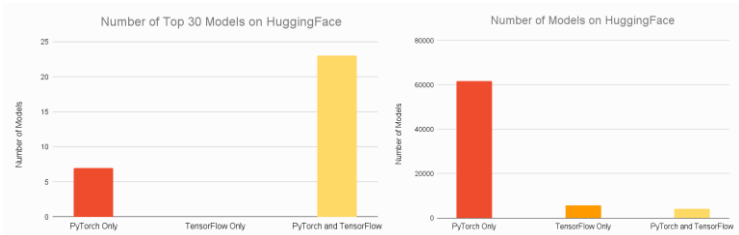
Google TensorFlow vs. Meta PyTorch

- PyTorch has steadily overshadowed TensorFlow.







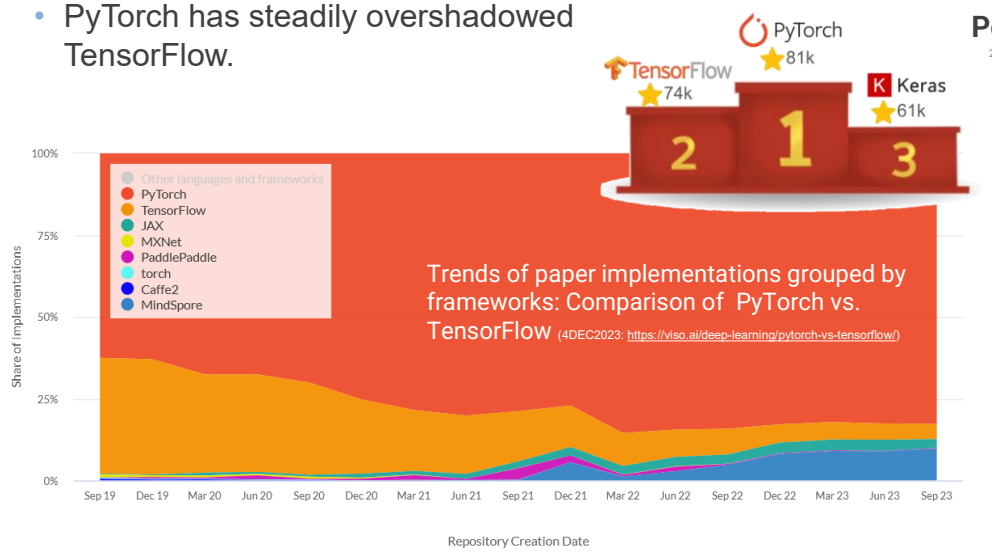


14DEC2021: https://www.assemblyai.com/blog/pytorch-vs-tensorflow-in-2023/

68

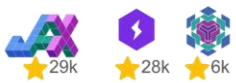
Google TensorFlow vs. Meta PyTorch

- PyTorch has steadily overshadowed TensorFlow.



Popularity on Github

24FEB2024: https://softwaremill.com/ml-engineer-comparison-of-pytorch-tensorflow-jax-and-flax/



69

Deep Learning Frameworks

python™

Keras

since 2015 (with several backends including TF)
Since 6/2019, officially included in TF 2.0
since 2023, Keras 3.0 includes full APIs for TF, JAX, and Torch backends

PYTORCH

since October 2016

since February 2024 (stable release)

since 2015
since 6/2019 (TF 2.0), include keras's high-level APIs

Simply change from:

```
from keras.XXX import YYY
```

To:

```
from tensorflow.keras.XXX import YYY
```

70

```python
import keras

model = keras.Sequential([
    keras.layers.Input(shape=(num_features,)),
    keras.layers.Dense(512, activation="relu"),
    keras.layers.Dense(512, activation="relu"),
    keras.layers.Dense(num_classes, activation="softmax"),
])
model.summary()

model.compile(
    optimizer=keras.optimizers.AdamW(learning_rate=1e-3),
    loss=keras.losses.CategoricalCrossentropy(),
    metrics=[
        keras.metrics.CategoricalAccuracy(),
        keras.metrics.AUC(),
    ],
)

history = model.fit(
    x_train, y_train, batch_size=64, epochs=8, validation_split=0.2
)
evaluation_scores = model.evaluate(x_val, y_val, return_dict=True)
predictions = model.predict(x_test)
```

```
$ python example.py
Using TensorFlow backend
```

```
$ python example.py
Using PyTorch backend
```

```
$ python example.py
Using JAX backend
```

Image credit: “Introducing Keras 3.0,” https://keras.io/keras_3/

71

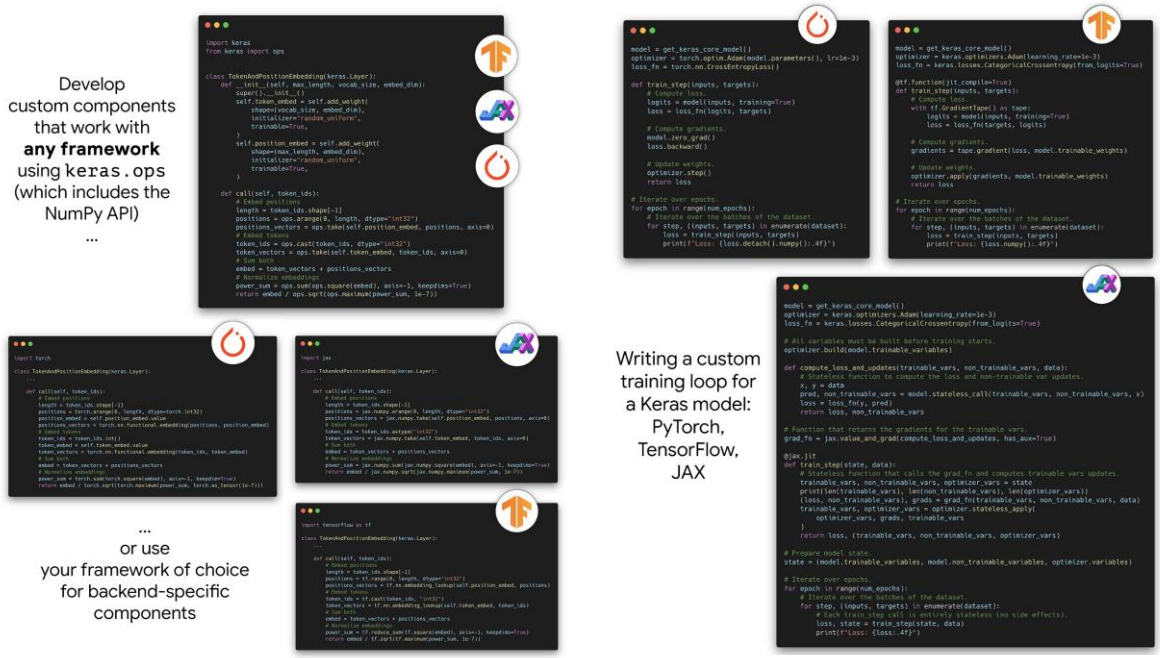
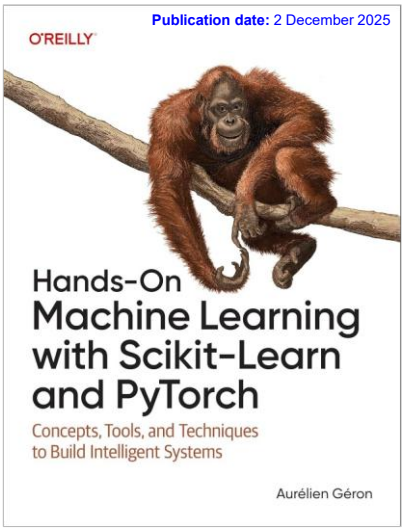
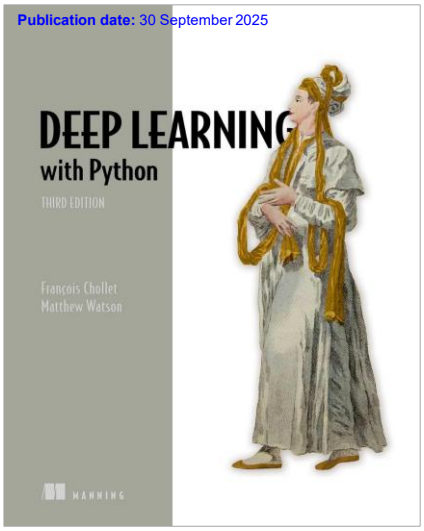


Image credit: “Introducing Keras 3.0,” https://keras.io/keras_3/

72

Google TensorFlow vs. Meta PyTorch





73

DL frameworks in this course

python™

Keras

PYTORCH

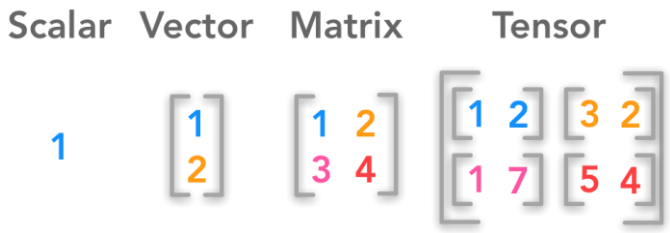
74

AI ETC.

75

Tensor in Deep Learning

- **Tensor** can be thought as a general term for nd-array where n can be 0, 1, 2, 3, … .



How many indexes required?	Computer Science	Mathematics	
0	Number	Scalar	Rank-0 tensor (0 dimension)
1	Array	Vector	Rank-1 tensor (1 dimension)
2	2D-array	Matrix	Rank-2 tensor (2 dimensions)
n	nd-array	nd-tensor	Rank-n tensor (n dimensions)

76

FLOPS vs. FLOPs

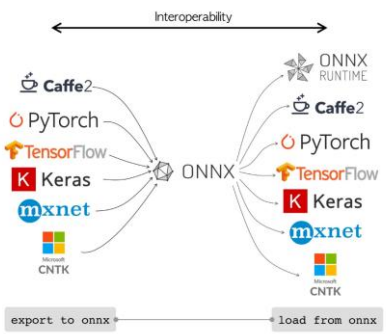
- One floating-point operation refers to one multiplication, one division, one addition, and so on.
- **FLOPS** (**FL**oating-point **OP**erations per **S**econd) is a unit of speed that often uses to measure computing hardware performance.
  - For example, NVIDIA RTX 2080 has a theoretical performance in FP32 mode of 10.07 TFLOPS. This means the theoretical maximum number of floating-point operations that the hardware might be capable of (if we are extremely lucky).
  - For modern CPUs, **FLOPS** is calculated from repeated uses of a "fused multiply then add" instruction, so that one instruction counts as two floating point operations.
- **FLOPs** (**FL**oating-point **OP**eration**s**) is a unit of amount that often uses to describe how many operations are required to run a single instance of a given model.

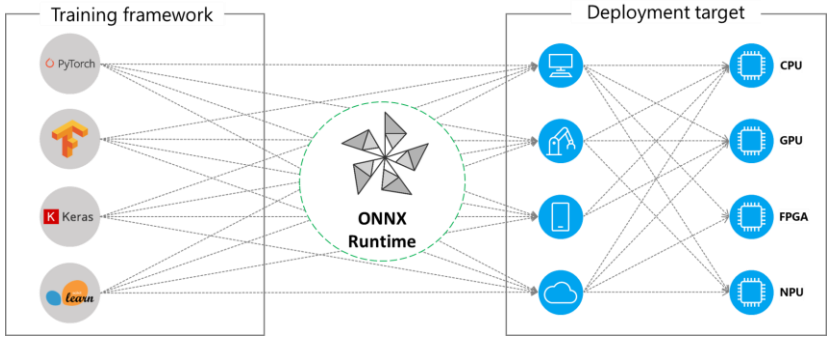
Model	Input Size	Param Size	Flops
AlexNet	227 x 227	233 MB	727 MFLOPs
CaffeNet	224 x 224	233 MB	724 MFLOPs
VGG-VD-16	224 x 224	528 MB	16 GFLOPs
VGG-VD-19	224 x 224	548 MB	20 GFLOPs
GoogleNet	224 x 224	51 MB	2 GFLOPs
ResNet-34	224 x 224	83 MB	4 GFLOPs
ResNet-152	224 x 224	230 MB	11 GFLOPs
SENet	224 x 224	440 MB	21 GFLOPs

77

ONNX , ONNX Runtime

- **O**pen **N**eural **N**etwork E**x**change **(ONNX)**

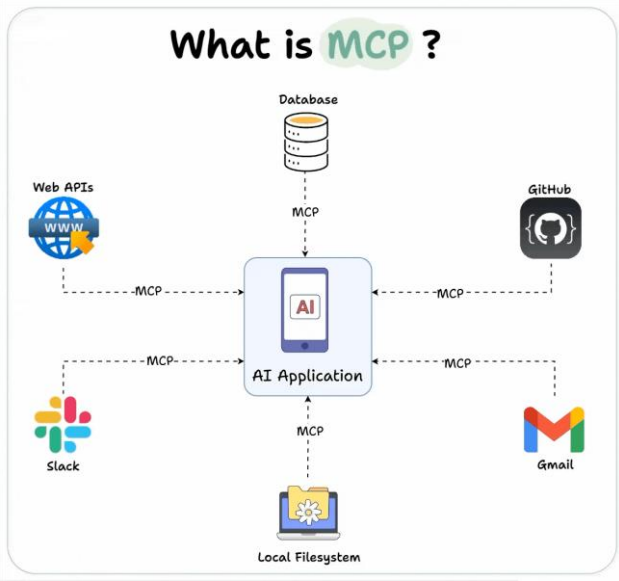




Images from: https://www.aurigait.com/blog/onnx-onnx-runtime-and-tensortrt/ , https://onnxruntime.ai/docs/execution-providers/

78

Model Context Protocol

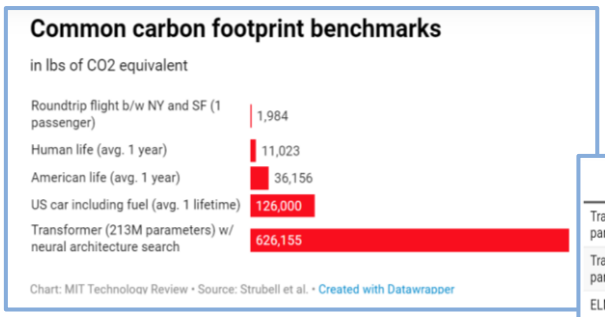


- **MCP** is like a universal connector for AI applications, allowing different AI tools and models to interact seamlessly with various data sources.
- **MCP** is an open protocol that standardizes how applications provide context to LLMs.
- **MCP** was originally built to improve Claude's ability to interact with external systems. Anthropic decided to open-source MCP in early 2024 to encourage industry-wide adoption.
- Read more https://www.ibm.com/think/topics/model-context-protocol

Image credit: (29JUL2025) https://abhishek-iiit.medium.com/model-context-protocol-mcp-assets-8cae94599875

79

Carbon Emissions Problem



	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.
Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

(6JUN2019) MIT Technology Review: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

80

Carbon Emissions Problem

- In the DL era, the computational resources needed to produce a best-in-class AI model has on average doubled every 3.4 months; this translates to a 300,000x increase between 2012 and 2018.
- A study in 2019 estimated that:
  - Training a single DL model (a particularly energy-intensive model) can generate up to 626,155 pounds of CO_2 emissions—roughly equal to the total lifetime carbon footprint of five cars. As a point of comparison, the average American generates 36,156 pounds of CO_2 emission in a year.
  - Training an average-sized model (much smaller than GPT) and examining not just the energy required to train the final version, but the total number of trial runs that went into producing the final version:
    - Over the course of six months, 4,789 different versions of the model were trained, requiring 9,998 total days' worth of GPU time (more than 27 years).
    - Taking all these runs into account, the researchers estimated that building this model generated over 78,000 pounds of CO_2 emission in total—more than the average American adult will produce in two years.

(17JUN2020): https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=7120e1176b43

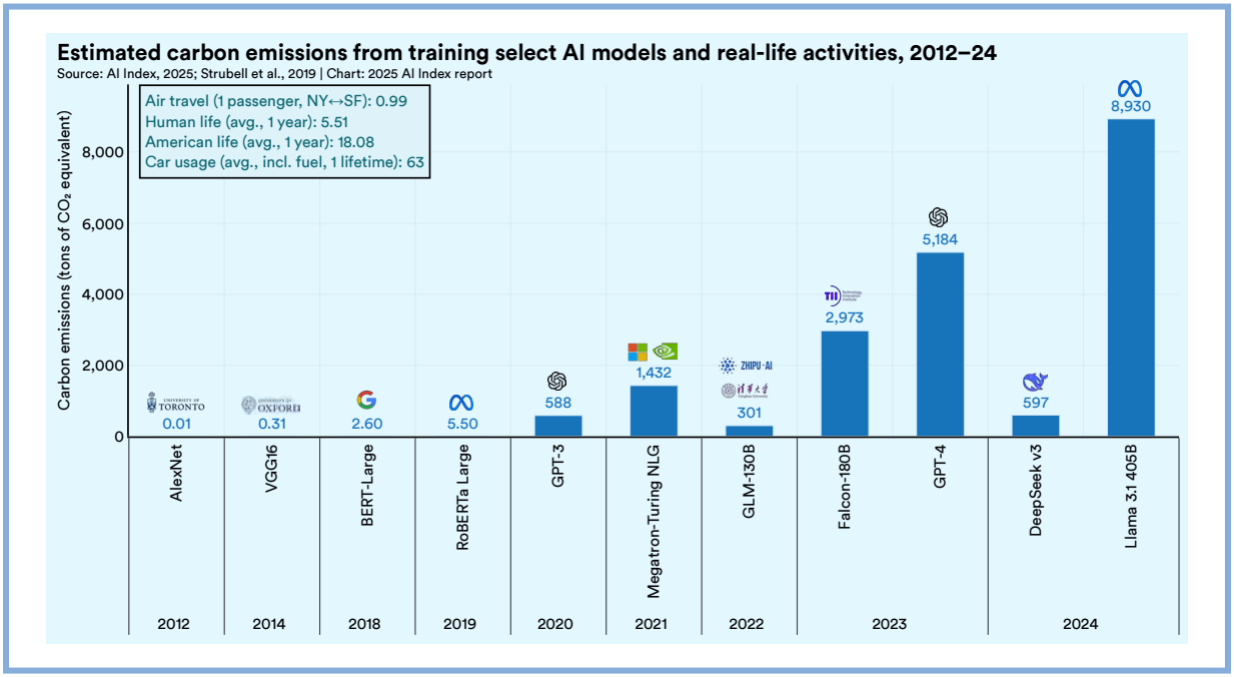
81

Carbon Emissions Problem

- Deploying AI models to take action in real-world settings—a process known as inference—comes even more energy than training does. NVIDIA estimates that 80-90% of the cost of a neural network is in inference rather than training.
- Machine Learning Emissions Calculator https://mlco2.github.io/impact/#compute

(17JUN2020): https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=7120e1176b43

82



HAI AI Index Report 2025 (April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

83

84

Next class, a ready-to-use GPU computer

- Cloud
  - Google Colab
- Local installation
  - NVIDIA driver, CUDA toolkit, cuDNN
  - https://pytorch.org/get-started/locally/
- NVIDIA's Docker containers
  - https://docs.nvidia.com/deeplearning/frameworks/user-guide/index.html

85

86