UNIVERSITY OF MINNESOTA TWIN CITIES

The dawn of Artificial Intelligence: Historical perspective, Challenges and Opportunities

Yousef Saad
University of Minnesota

Panel on AI / Table ronde IA

Algiers, Jan 31, 2026

The plan

1. Historical perspective [computational side]. The AI wave
2. A few details on AI methods, transformers, LLMs..
3. Hardware & impact on power consumption
4. Recent developments and trends
5. Resources

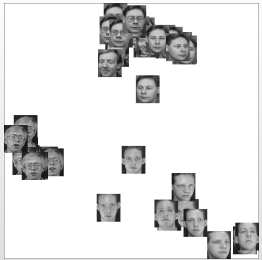
- Note: parts indicated by ϟ will be skipped or discussed very briefly

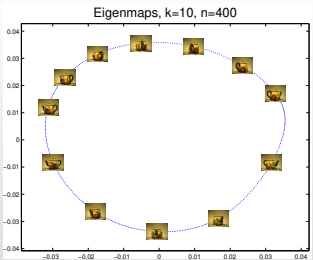
University of Minnesota

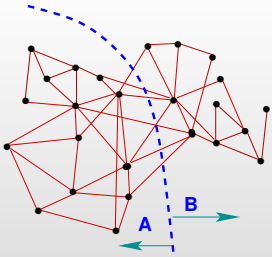
The AI wave

Progression of data-based methodologies

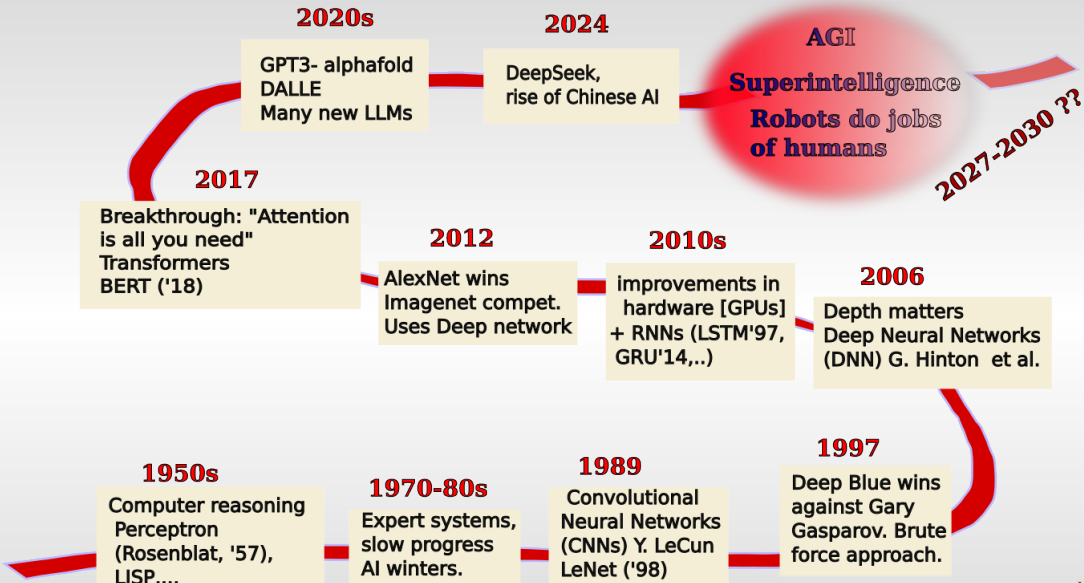
- Early days: logic programming, symbolic prog., Natural Language Processing,
- Some success [‘expert systems’] – but research hit a wall
- 1990s, 2000s: ’Data Mining’: extract information from data.
  Examples: Clustering, embedding, unsupervised learning, Network analysis, PageRank, ideas of centrality, ...

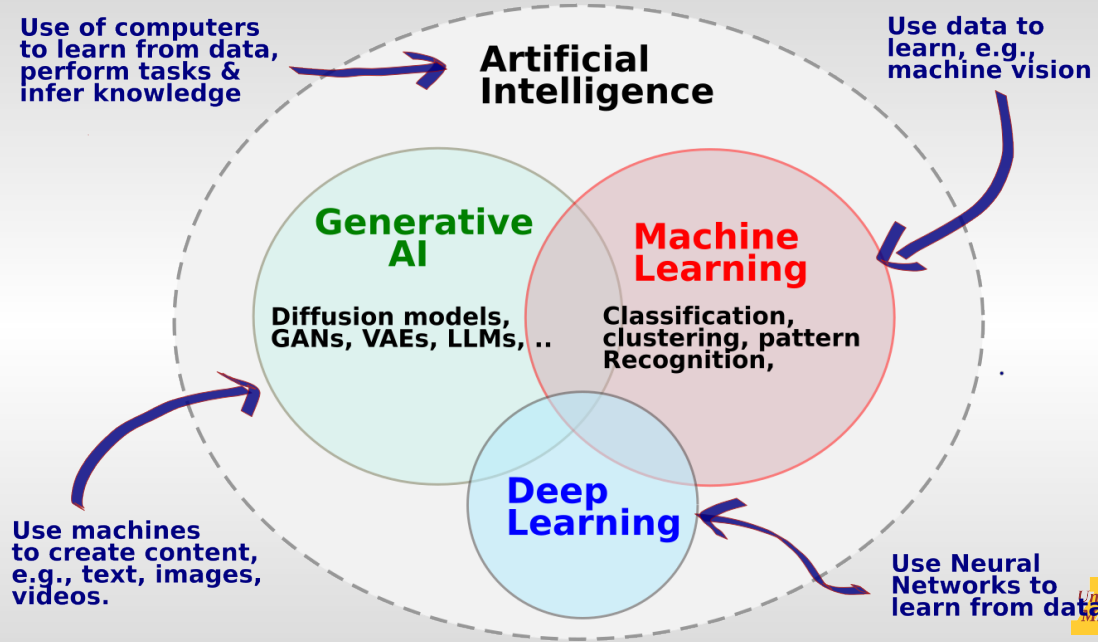






University of Minnesota

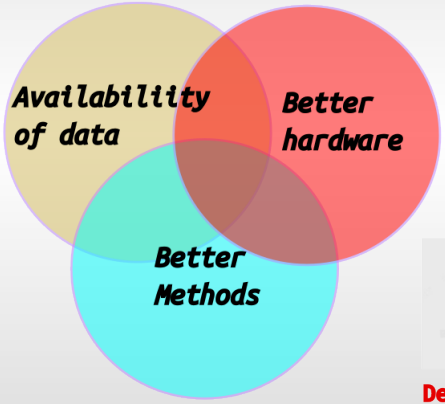

1950s
Computer reasoning
Perceptron
(Rosenblat, '57),
LISP,...
1970-80s
Expert systems,
slow progress
AI winters.
1989
Convolutional
Neural Networks
(CNNs) Y. LeCun
LeNet ('98)
1997
Deep Blue wins
against Gary
Gasparov. Brute
force approach.
2006
Depth matters
Deep Neural Networks
(DNN) G. Hinton et al.
2010s
improvements in
hardware [GPUs]
+ RNNs (LSTM'97,
GRU'14,..)
2012
AlexNet wins
Imagenet compet.
Uses Deep network
2017
Breakthrough: "Attention
is all you need"
Transformers
BERT ('18)
2020s
GPT3- alphafold
DALLE
Many new LLMs
2024
DeepSeek,
rise of Chinese AI
AGI
Superintelligence
Robots do jobs
of humans
2027-2030 ??


Use of computers to learn from data, perform tasks & infer knowledge
Artificial Intelligence
Use data to learn, e.g., machine vision
Generative AI
Diffusion models, GANs, VAEs, LLMs, ..
Machine Learning
Classification, clustering, pattern Recognition,
Deep Learning
Use machines to create content, e.g., text, images, videos.
Use Neural Networks to learn from data

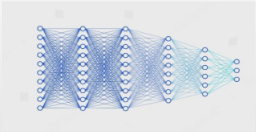
Ingredients of the AI wave

Wikipedia

NVIDIA
AMD,..



The internet



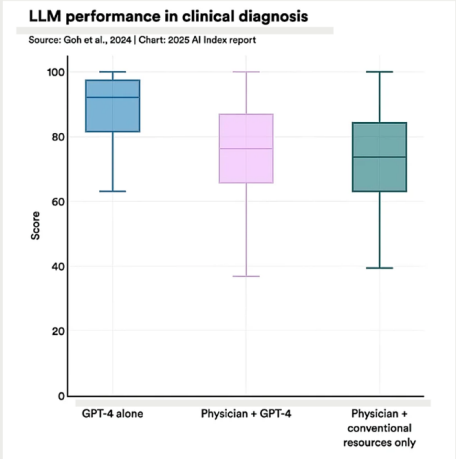
Deep Neural Networks

University of Minnesota

The BIG AI wave

- AI is making astounding progress ...
- Synergies: Hardware, openness, focus, ..
- Today: AI can solve university-level Homeworks
- Soon: will do the work of a doctor, engineer, ..

Ref: L. Aschenbrenner *'Situational awareness: The decade ahead'* (2024). Convincing argument that AGI [Artificial General Intelligence] will be with us within a few years.



University of Minnesota

A Recent Milestone: Gold at the Math Olympiad (July 25)

World

Humans triumph over AI at annual math Olympiad, but the machines are catching up

Updated on: July 22, 2025 / 12:06 PM EDT / CBS/AFP

Add CBS News on Google

Sydney — Humans beat generative AI models made by Google and OpenAI at a top international mathematics competition, but the programs reached gold-level scores for the first time, and the rate at which they are improving may be cause for some human introspection.

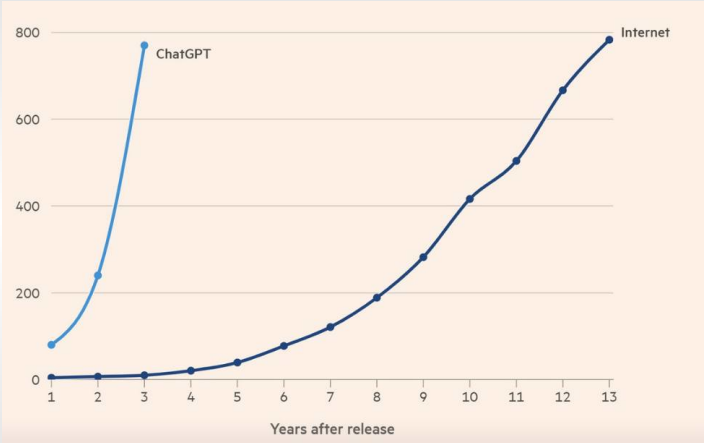
- Came 18 years earlier than predicted 5 years ago...

University of Minnesota

'The fastest technology adoption in history'

"The internet took 13 years to reach 800M users. ChatGPT just did it in 2."

source: Financial Times



University of Minnesota

Artificial General Intelligence (AGI), Artificial Super-Intelligence (ASI)

Does anything humans can do intellectually

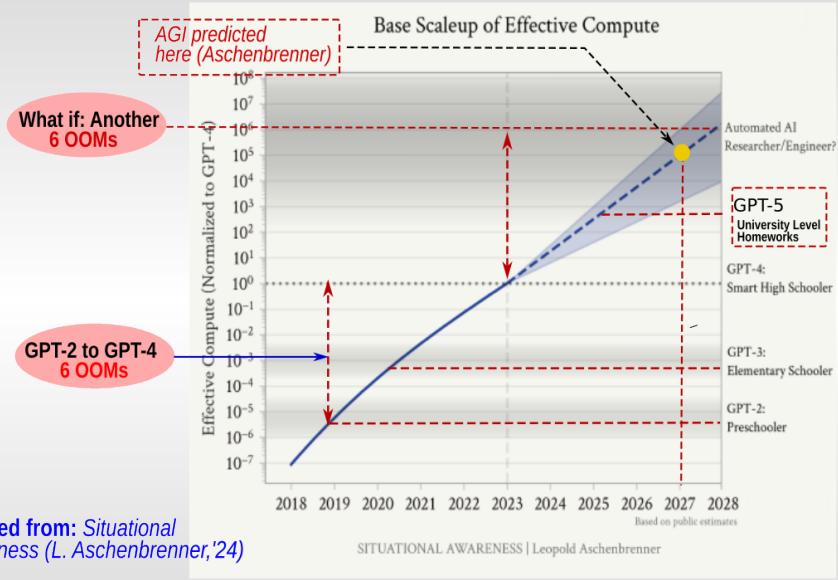
- An AGI can be a doctor, an artist,
- Can solve math problems,
- Reason, Create knowledge
- ...

Exceeds best human minds in all domains

- If AGI = 1 then ASI could be 10^4 ?
- Imagine: Scientific Breakthrough in Minutes
- Self-improvement $\rightarrow$ Exponential $\uparrow$

University of Minnesota

OOM = Order of Magnitude (compute scale, algorithms, + ...)



Adapted from: *Situational Awareness (L. Aschenbrenner,'24)*

University of Minnesota

AGI by 2027?

AGI by 2027 is strikingly plausible. GPT-2 to GPT-4 took us from pre-schooler to smart high-schooler abilities in 4 years. Tracing trend-lines in compute (0.5 orders of magnitude or OOMs/year), algorithmic efficiencies (0.5 OOMs/year), and "unhobbling" gains (from chatbot to agent), we should expect another preschooler-to-high-schooler- sized qualitative jump by 2027. (L. Aschenbrenner, 2024)

- Predicts: Superintelligence [super-human level intelligence] will follow quickly because AGIs will compress 5+OOMs into 1 year

University of Minnesota

Comparison with Moore's Law (microchips)

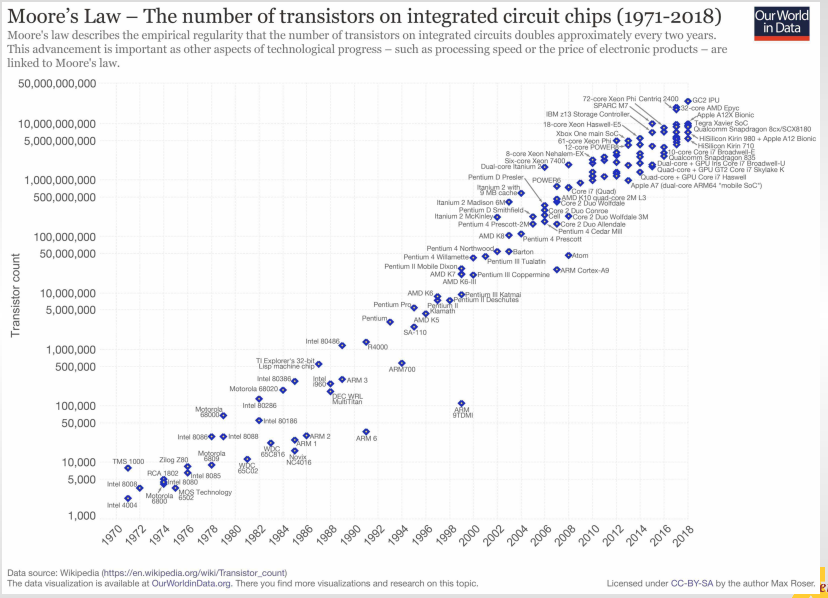
- Useful to compare with progress made in hardware
- Moore's law has been incredibly accurate in predicting advances in microchips. Note: stipulated in 1965! [actual ratio corrected in 1975.. still....]

> Moore's Law: Number of transistors placed on a chip doubles every 2 years

- One can see similar laws in progress of technology

University of Minnesota

Moore's Law: Number of transistors placed on a chip doubles every 2 years.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

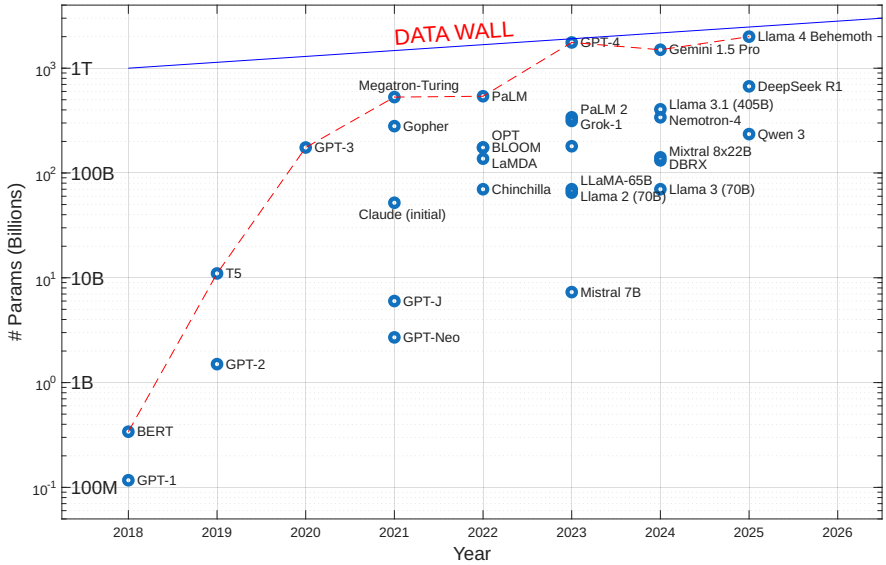
What about AI?

- Look at: Number of parameters in Large Language Models (LLMs)
- Extremely fast increase at beginning (2018–2021):

Model	GPT-1	GPT-2	GPT-3	GPT-4
Year	2018	2019	2020	2023
# Params	*177M*	*1.5B*	*175B*	*1.7 T*

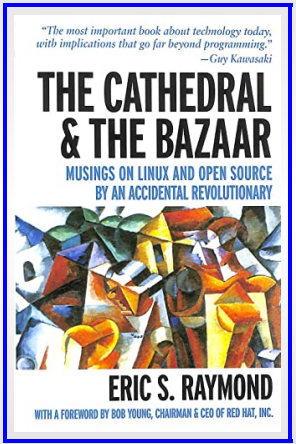
- Doubling every $\approx$ **4.6** months ($\times$**10** every $\approx$ 14 months)
- ... but number of params now stalling
- Next plot shows trend with selection of 31 recent models
- Red dashed line = max for each year

University of Minnesota


DATA WALL
Params (Billions)
1T
100B
10B
1B
100M
Year
2018
2019
2020
2021
2022
2023
2024
2025
2026
BERT
GPT-1
GPT-2
T5
GPT-3
Megatron-Turing
Gopher
GPT-J
GPT-Neo
Claude (initial)
PaLM
OPT
BLOOM
LaMDA
Chinchilla
GPT-4
PaLM 2
Grok-1
LLaMA-65B
Llama 2 (70B)
Mistral 7B
Gemini 1.5 Pro
Llama 4 Behemoth
DeepSeek R1
Llama 3.1 (405B)
Nemotron-4
Qwen 3
Mixtral 8x22B
DBRX
Llama 3 (70B)

Other factors

- Chipmaking is very competitive – Technology highly protected
- In contrast: AI is basically open.
- *Unparalleled global participation*
- The Bazaar (bottom-up) vs. The Cathedral (top-down)
- A blessing for research/science but ...
- ... also a curse: containment issues



University of Minnesota

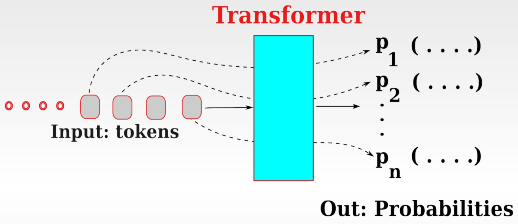
LLMs: main components

LLMs and Transformers

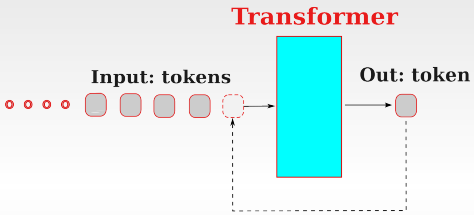
- Up-to $\approx$ 2015: MLP, CNN, RNN+LSTM, + Focus on images. Then:
- *"Attention is all you need"* paper [Vaswani al, '17] – a major breakthrough
  - Before: LLMs needed to account for sequentiality.. order in words. Difficulties: stability, ...,
  - Now: use (1) attention + (2) adding 'positional encoding' scheme to embedding.

University of Minnesota

Transformers: The big picture

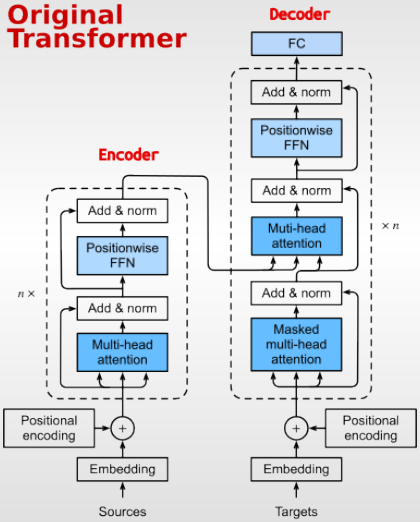


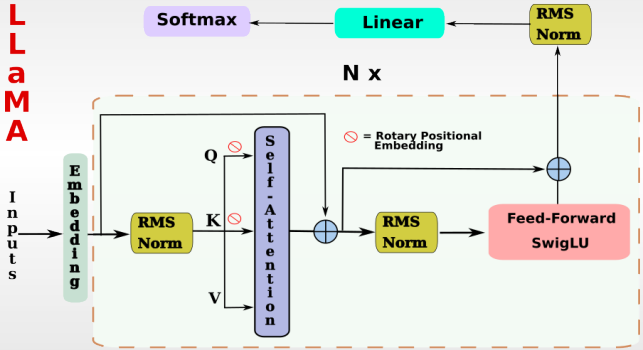
Training: get weights s.t. output probabilities for next token match target



Generation: get next token. Re-inject as input. Repeat

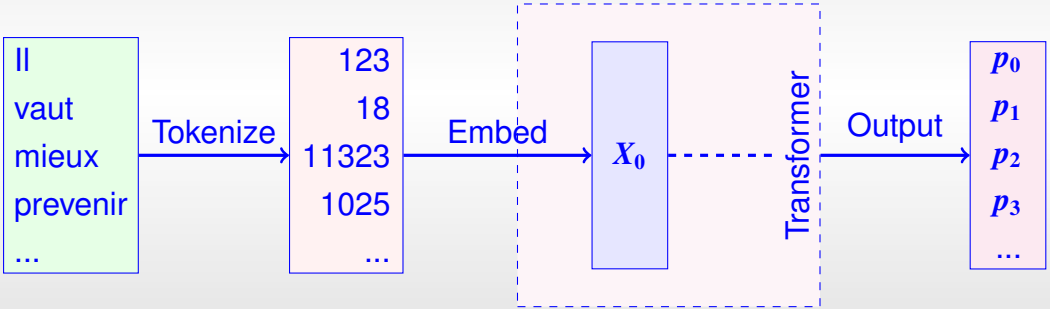
Transformer architecture: Encoder-Decoder, Decoder-Only





Tokenize, Embed, Transform

- 1st step of LLMs: transform sequence of strings (words, chars) to tokens
- Each token embedded to a vector in $\mathbb{R}^d$.



- X_0 transformed to $X_1, X_2 \cdots X_l, \cdots X_L$

University of Minnesota

- Transformer = sequence of transformations – [represented by 'boxes']
- Two types of boxes are found in every LLM:
  - *Attention layer*
  - *Feed Forward Network* (FFN)

Attention

- $X_l \rightarrow Q = X_l W_Q; \quad K = X_l W_K; \quad V = X_l W_V;$
- Attention matrix $A_l = Q^T K$
- Scores: softmax$(A_l \times V)$

A_l: a sort of distance between every pair of tokens. Scores: Which tokens are closest to those 'in V'

Multi-Layer Perceptron (MLP)- a.k.a Feed-Forward Network (FFN)

➤ Training a neural network amounts to approximating a function ϕ which is defined via sets of parameters; e.g., Multi-Layer Perceptron (MLP):

Mapping $x \rightarrow y = \phi(x)$

Input: x, **Output:** y

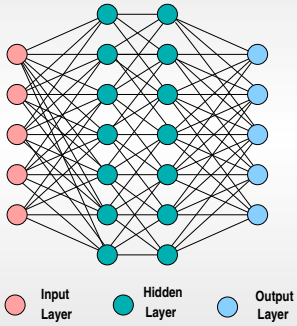
Set: $z_0 = x$

for $l = 1$: L+1 **do Do:**

$$z_l = \sigma(W_l^T z_{l-1} + b_l)$$

end for

Set: $y = \phi(x) := z_{L+1}$



- layer # 0 = input
- layer # $(L+1)$ = output
- Matrix W_l: from layer $l-1$ to layer l.

➤ Problem: Find params W_l, b_l s.t. $\phi(x_i) \approx y_i \; i = 1, \cdots, n$

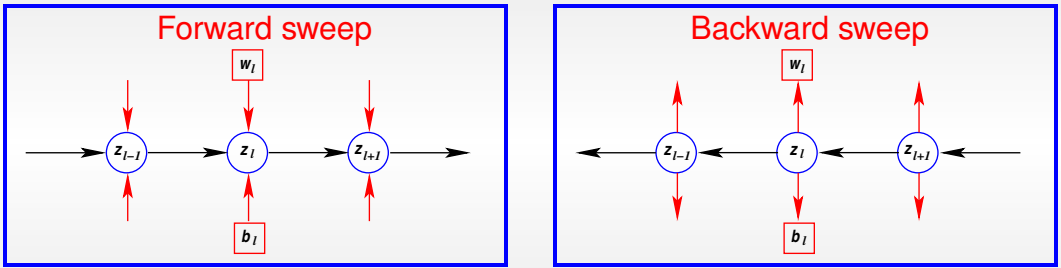
A (very quick) look at optimization in Deep Learning

- Often in ML: $f(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w) \quad \rightarrow \quad \nabla f(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w).$
- Gradient descent: $w_{j+1} = w_j - \eta \nabla f(w_j)$ $\rightarrow$ Expensive + impractical
- Stochastic Gradient Descent (SGD) approach:
  Cycle through 'mini-batches' $\mathcal{B}_j$ of functions $w_{j+1} = w_j - \eta \nabla f_{\mathcal{B}_j}(w_j)$
- Notation: $f_{\mathcal{B}_j}(w) \equiv \frac{1}{|\mathcal{B}_j|}\sum_{k\in\mathcal{B}_j} f_k(w)$
- At the heart of optimization in deep learning
- More advanced methods add momentum + various normalizations:
  Adam, RMSprop, AdaGrad, Nesterov, ..

University of Minnesota

A key ingredient: Back-propagation

Example: Basic forward propagation equation in FFN: $z_l = \sigma(W_l^T z_{l-1} + b_l)$



$$(1)\quad \frac{\partial f}{\partial z_l} = \frac{\partial f}{\partial z_{l+1}} \times \frac{\partial z_{l+1}}{\partial z_l} \qquad (2)\quad \frac{\partial f}{\partial W_l} = \frac{\partial f}{\partial z_l} \times \frac{\partial z_l}{\partial W_l} \qquad (3)\quad \frac{\partial f}{\partial b_l} = \frac{\partial f}{\partial z_l} \times \frac{\partial z_l}{\partial b_l}.$$

University of Minnesota

Hardware & impact of GPUs

AI would not be here without the hardware: The rise of GPUs

- GPUs [Graphics Processing Units] are very powerful co-processors for graphics.
- Idea in Mid- 2000s: why not use them for scientific computing?
- Difficulty: software.
- Solution: Compute Unified Device Architecture (CUDA)
- Introduced in 2006 by NVIDIA
- Idea of attached processor [or co-processor]– Not new [e.g. FPS AP-120B 'array processor' unveiled in 1981]

University of Minnesota

The CUDA environment: The big picture

- A host (CPU) and an attached device (GPU)

Typical program

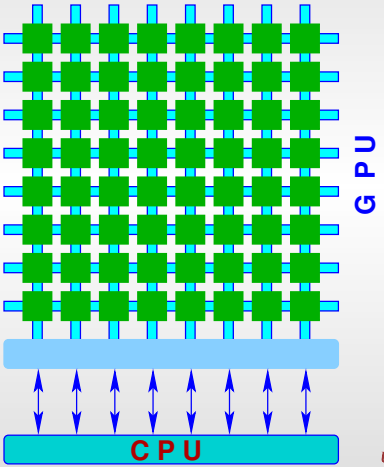
1. Generate data on CPU
2. Allocate memory on GPU

   `cudaMalloc(...)`
3. Send data Host → GPU

   `cudaMemcpy(...)`
4. Execute GPU 'kernel':

   `kernel <<<(...)>>>(..)`
5. Copy data GPU →CPU

   `cudaMemcpy(...)`



University of Minnesota

Affordable supercomputing

- [from my class notes 'Intro to parallel computing', UMN, 2021]
- Recall: 1 TFLOPS = $\mathbf{10^{12}}$ Floating-Point Operations per second
- In 2008: 1 TFLOPS for $\approx$ \$ 1,350 (Tesla C2050). Picture in 2021:

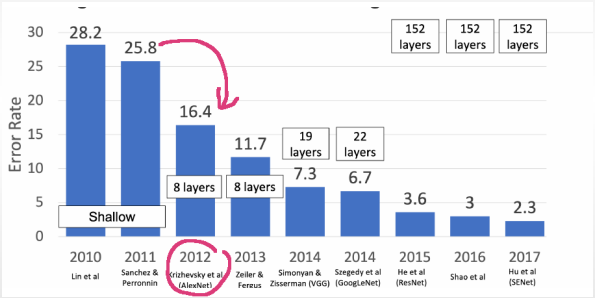
GPU model	$\approx$ Price	FP64 Perf.	\$ / TFLOPS	DL (FP32) Perf.	\$/ TensTOPS
P100 (16GB)	\$ 7,374	4.7 TFLOPS	\$ 1,569	18.7 TFLOPS	\$394.33
V100 16GB	\$10,664*	7 TFLOPS	\$1,523	112 TFLOPS	\$95.21
32GB	\$11,458*		\$1,637		\$102.3

- Note huge jump in performance for Deep Learning (DL) made in V100 vs P100

University of Minnesota

The rise of GPUs for AI

- 'A supercomputer for $\approx$ \$10K': Tesla $\rightarrow$ Fermi $\rightarrow$ Kepler $\rightarrow$ Pascal $\rightarrow$ ...



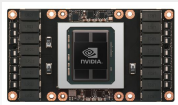
2012 AlexNet trained with NVIDIA GTX 580 GPUs (Fermi). NVIDIA saw big opportunity in AI.

$\leftarrow$ Best performance at ImageNet Competitions 2010-2017

University of Minnesota

The rise of GPUs (cont.)

2016 Introduction of FP16 in Pascal chips for AI



2017 Volta: Introduction of Tensor-cores (FP16)

2020 Ampere: Introduction of TF32 (Tensor Float 32)

2022 Big jump in performance with Hopper arch.

University of Minnesota

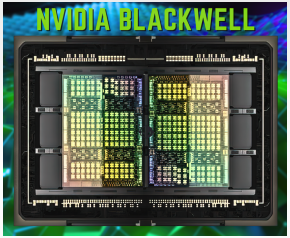
H100

$$\textit{One SM} \equiv 4 \times (16\ \textit{int32} + 32\ \textit{FP32} + 16\ \textit{FP64} + 1\ \textit{TensorCore})$$

→

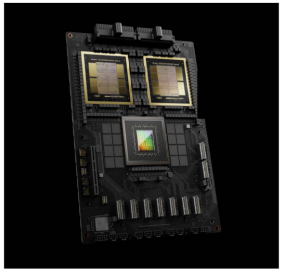
H100 chip with 144 Streaming Multiprocessors (SMs) ↓

The Latest: Blackwell



B 100 Blackwell chip

≈ $ 30-35K



GB_200 Superchip (2 GPUs + 1 Grace-CPU)

≈ $ 60-70K

The GB-200 NVL72 (36 $\times$ GB_200)

≈ $ 3M

University of Minnesota

NVIDIA Blackwell Specs (from NVIDIA)

	GB200 NVL72	GB200 Superchip
FP4 Tensor Core (Sparse|Dense)	1,440 | 720 PFLOPS	40 | 20 PFLOPS
FP8/FP6 Tensor Core (Sp)	720 PFLOPS	20 PFLOPS
INT8 Tensor Core (Sp)	720 POPS	20 POPS
FP16/BF16 Tensor Core (Sp)	360 PFLOPS	10 PFLOPS
TF32 Tensor Core (Sp)	180 PFLOPS	5 PFLOPS
FP32	5,760 TFLOPS	160 TFLOPS
FP64 / FP64 Tensor Core	2,880 TFLOPS	80 TFLOPS
GPU Memory | Bandwidth	13.4 TB | 576 TB/s	372 GB | 16 TB/s
NVLink Bandwidth	130 TB/s	3.6 TB/s
CPU Core Count	2,592 Arm V2 cores	72 Arm V2 cores

What about 'personal' or 'local' platforms?

- For training/ inference with small models. Three options:
  - NVIDIA DGX Spark (or Asus Ascent GX10). Powered by the GB10 Grace Blackwell Superchip. But: Not meant for development.
  - Workstation with AMD Ryzen-AI Max+ 395 + 128GB RAM + 2TB Disk [e.g., GMKtec, HP, ..]
  - MacBook pro with M5 processor (recent) -designed for AI. But: max RAM: 32GB

[*Platforms:* Nvidia: *CUDA*, Mac-OS: *MPS*, AMD: *ROCm*]

University of Minnesota

NVIDIA chips

Year	Architecture	Key Accelerators	Breakthrough
2006	Tesla	C870	First CUDA GPUs
2008	Tesla (2nd)	C2070	HPC adoption
2010	Fermi	M2050	ECC, DP boost
2012	Kepler	K20/K80	Efficiency leap
2016	Pascal	P100	NVLink, FP16
2017	Volta	V100	Tensor Cores
2020	Ampere	A100	TF32, MIG
2022	Hopper	H100/H200	FP8 + Transformer Engine
2024	Blackwell	B100/B200/GB200	NVFP4, 208B transistors
2025	Blackwell Refresh	GB200 NVL72	Extreme cluster scaling
2026	Rubin	R100/GR200	HBM4, NVLink 6

University of Minnesota

'Rubin' is out (production)

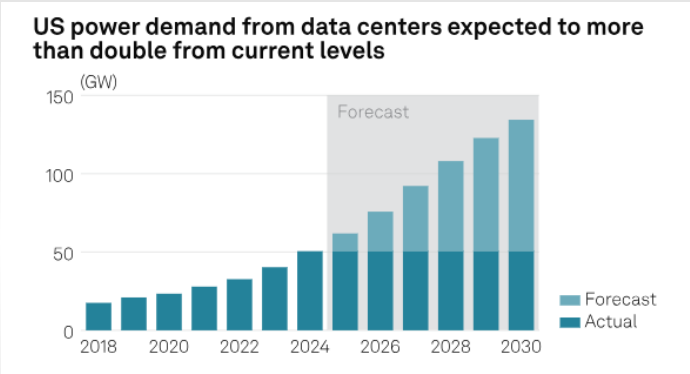
Announced at the Consumer Electronics Show (CES) 2026 (Jan. 6-9, 2026)

	B-200	Vera Rubin
Fab:	TSMC 4NP	TSMC 3nm 3NP
Mem:	192 GB HBM3e	288 GB HBM4
NVFP4 inference	10 PFLOPS	50 PFLOPS
NVFP4 training	10 PFLOPS	35 PFLOPS
Max power draw	1,200 W	1,800 W

- Claim: generational leap over Blackwell/B200,
- Compare: Ampere (A100) : 7nm, Hopper/Blackwell H200/B200 4nm)
- Compare: Performance gain for $\approx$ **50**% more energy

University of Minnesota

Big Issue: Increased Energy Consumption



Data center grid-power demand up 22% in 2025. Will triple by 2030 (Sce: 451 Research)

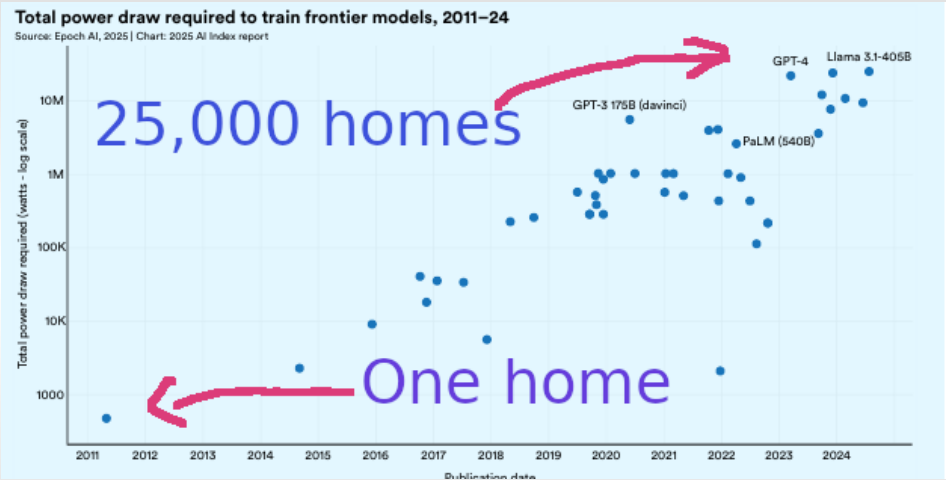
- Hopper H100: 700W, Blackwell B100: 700W, B200: 1K–1.2K W
- Between Feb 24 to Feb 26: 6M Blackwells sold $\rightarrow \approx 5GW$ power ↑
- Perfect storm at horizon: *AI* + *EVs* + *Cryptocurrencies*

University of Minnesota

From AI index report:

- Energy needed to train 1st Transformer model ('17): 4,500 Watts
- Energy needed to train LLama 3.1-405B model ('24) 25.3M Watts
- 2023: 4.4 % of electricity in US consumed by data centers
- By 2028 $\rightarrow$ 6.7% to 12%.

University of Minnesota


Total power draw required to train frontier models, 2011–24
Source: Epoch AI, 2025 | Chart: 2025 AI Index report
25,000 homes
One home
GPT-4
Llama 3.1-405B
GPT-3 175B (davinci)
PaLM (540B)
Total power draw required (watts - log scale)
10M
1M
100K
10K
1000
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Publication date

- AI hardware gets faster, cheaper, and more energy efficient.

Energy efficiency

New research suggests that machine learning hardware performance, measured in 16-bit floating-point operations, has grown 43% annually, doubling every 1.9 years. Price performance has improved, with costs dropping 30% per year, while energy efficiency has increased by 40% annually. (AI index rep.)

Recent developments and trends

Current Trends

AI is no longer just a story of what's possible—it's a story of what's happening now and how we are collectively shaping the future of humanity.
(The AI index report 2025)

University of Minnesota

- The report lists 12 “top takeaways”. Here are a few of them:
  1. AI performance on demanding benchmarks continues to improve.
  2. AI is increasingly embedded in everyday life.
  3. Business is all in on AI (investment+usage), $\rightarrow$ strong productivity impact
  4. U.S. still leads in producing top AI models - China closing performance gap
  5. AI becomes more efficient, affordable, and accessible.
  6. Industry is racing ahead in AI - but frontier is tightening.

Research & Development in AI (source: AI index report)

1. In 2024 : 40 notable new AI models in US, 15 in China, 3 in Europe
2. AI publication totals continue to grow –

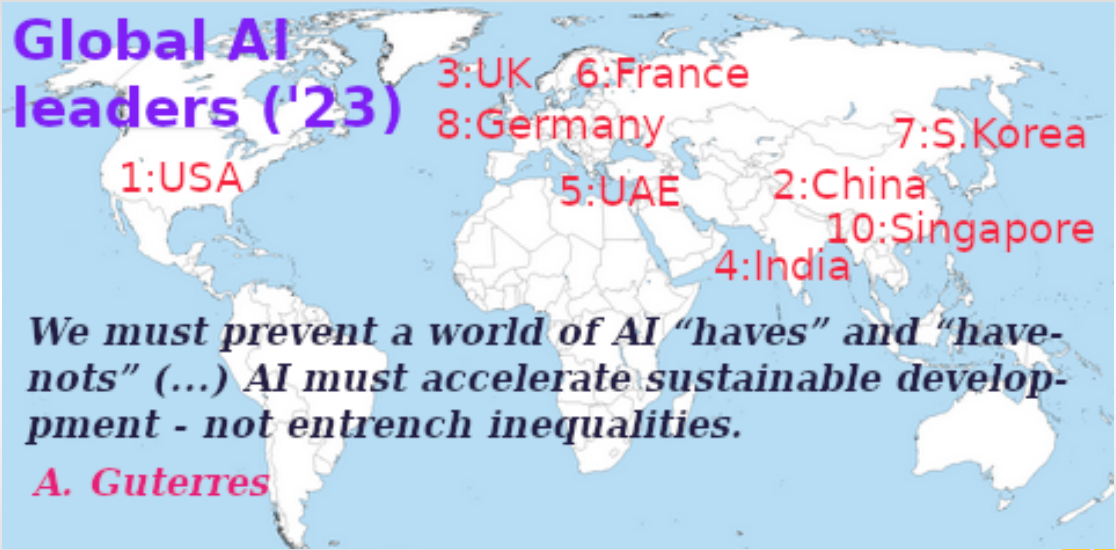
	# AI pubs	% share of CS
2013	102,000	21.6 %
2023	242,000	41.8 %

3. New models: bigger, computationally demanding, energy intensive
4. Training compute for notable AI models doubles approximately every five months

University of Minnesota

Governments caught by surprise - starting to react:

- International governing bodies created
- Independent International Scientific Panel on AI
- Global Digital Compact. Notable concern: Inequalities



University of Minnesota

Top 10 AI nations: by compute power and by # data clusters

Rank	By compute (H100 equiv.)	By num. clusters
1	USA: 39.7Mx H100 $\sim$	China: 230 c.
2	UAE: 23.1 M	USA: 187 c.
3	Saudi Arabia: 7.2 M	France: 18 c.
4	Korea: 5.1 M	Korea: 13 c.
5	France: 2.4 M	Germany: 12 c.
6	India: 1.2 M	Saudi Arabia: 9 c.
7	China: 400K	UAE: 8 c.
8	UK: 120K	India: 8 c.
9	Finland: 72K	UK: 6 c.
10	Germany: 51K	Finland: 5 c.

sce: Forbes, Sep. 2025

Issues for China:

1) Power
2) Chips Embargo

but..
China is reacting on Energy front

University of Minnesota

Resources etc..

Open weight models

- Weights (not the training data, details on architecture, or actual code)
  1. Mistral 3 (Mistral AI). Params.: 675B
  2. DeepSeek V 3.2 (DeepSeek) 2024. Params.': 671B -sparse: 37B.
  3. GPT-oss (openAI). Params.: 120B
  4. Qwen3 (Alibaba). Params 235B.
  5. LLama 3.1. Params: 8B, 70B, 405B

University of Minnesota

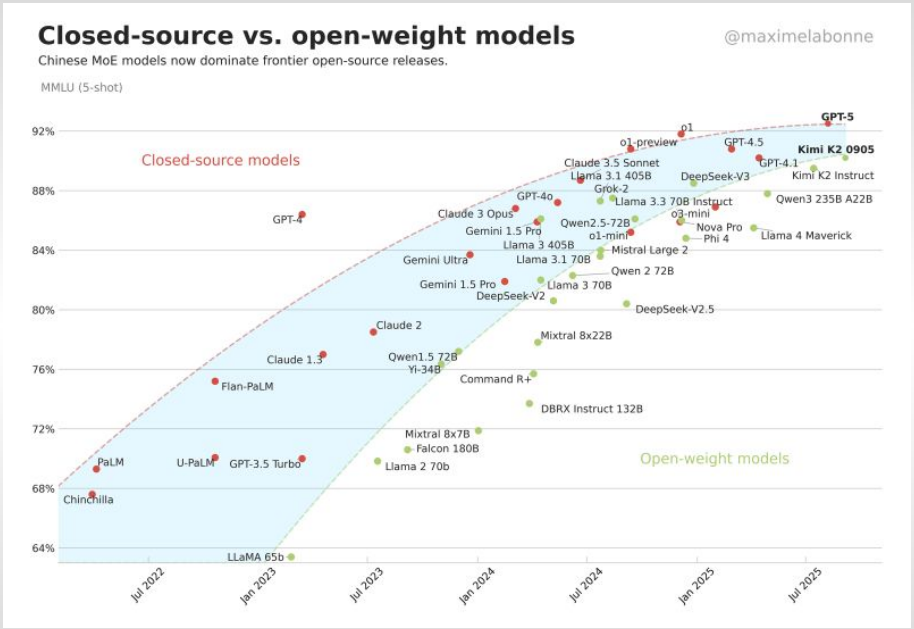
Open source models

2024 OSI: everything you need to reproduce model: weights, source code, + details about data. *Very liberal license*

1. OlMo-3 (Allen Institute) Params.: 32B
2. GPT2 (Open AI - old) * weights and source. Params: 130M
3. GPT-NeoX (EutherAI). Params: 20B
4. Falcon (TII, Abu Dhabi) Params: 11B, 180B
5. BLOOM (BigScience)
6. OpenLLaMA (openLM research)

Difficulty: Training requires resources (and huge volumes of data) which a small group cannot afford [experience of openLLaMA]

University of Minnesota



https://x.com/maximelabonne/status/1816416043511808259/photo/1

University of Minnesota

Example: the GPT-oss Open Weight from openAI

gpt-oss Released in Aug. 2025. First ever major Open Weight model by openAI

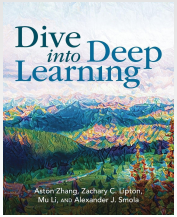
1. *gpt-oss 120B*: designed for production reasoning. Fits on a single high-end GPU
2. *gpt-oss 20B*: smaller – can run on consumer hardware (16GB RAM)

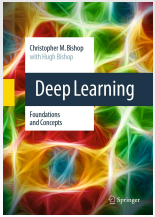
- License: Apache 2.0 license – very permissive (access to everything)
- Goal: to compete with Meta's LLaMA models

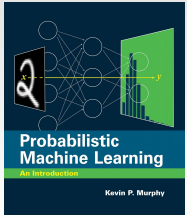
University of Minnesota

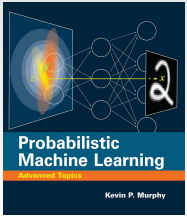
Resources

The Bazaar: *Hugging Faces*. Models, datasets, courses, codes, applications + much more
https://huggingface.co/









The Books: (Available online)

1. *Dive into Deep Learning* A. Zhang et al. d2l.ai
2. Bishop and Bishop *Deep Learning* link
3. K. Murphy. *Probabilistic Machine Learning* - Two books link-book1 - link-book2

University of Minnesota

Resources: free courses and tutorials

- Many courses/ tutorials available. Here are a few the best known:
  1. Generative AI for everyone (Andrew Ng)
  2. Intro. to Generative AI
  3. Stanford course
  4. IBM SkillBuild
  5. Elements of AI (available in French)
  6. Sebastian Raschka course

University of Minnesota

Resources: coding

- Three current major packages (all in Python)

1. Tensorflow [older (2015) - GoogleBrain]
2. PyTorch [Initiated by Meta in 2017. Then (2020) part of Linux Foundation]
3. JAX [2018, Google. Meant for replacing TensorFlow]

- Currently: PyTorch dominates
- JAX is gaining ground - especially for scientific machine learning
- It is relatively easy to write programs with, e.g., PyTorch
- All major building blocks are provided: DNN, FFN, Transformers, optimizers, ...

University of Minnesota

Concluding remarks

1. We are at the beginning of the AI wave. It is important to get involved ...
2. ... and it is easy. Our new Teachers: GPT, LLaMMa, Gemini, Mistral, ..
3. Opinion: AI == most democratic technological advancement ever
4. Opinion: *It is vital that the young talent in a place like Algeria participate*

University of Minnesota

Concluding remarks

1. We are at the beginning of the AI wave. It is important to get involved ...
2. ... and it is easy. Our new Teachers: GPT, LLaMMa, Gemini, Mistral, ..
3. Opinion: AI == most democratic technological advancement ever
4. Opinion: *It is vital that the young talent in a place like Algeria participate*

Marie Curie:

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

Louis Pasteur:

"Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world"

University of Minnesota

Thank you!