AI Slides Viewer & Checker - 1

Sunghee Yun

Co-founder / CTO - AI Technology & Product Strategy

Erudio Bio, Inc.

Table of contents

About Speaker

- *Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio, CA, USA*
- *Co-founder / CTO & Chief Applied Scientist @ Gauss Labs, CA, USA – 2023*
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST
- Adjunct Professor, Electronic Engineering Department @ Sogang University
- Technology Consultant @ Gerson Lehrman Gruop (GLG)
- KFAS-Salzburg Global Leadership Initiative Fellow @ Salzburg Global Seminar
- Senior Applied Scientist @ Mobile Shopping App Org, Amazon.com, Inc. – 2020
- Principal Engineer @ Software R&D Center of DS Division, Samsung – 2017
- Principal Engineer @ Strategic Marketing & Sales Team, Samsung – 2016
- Principal Engineer @ DT Team of DRAM Development Lab, Samsung – 2015
- Senior Engineer @ CAE Team - Samsung – 2012
- M.S. & Ph.D. - Electrical Engineering @ Stanford University – 2004
- B.S. - Electrical Engineering @ Seoul National University – 1998

Highlight of career journey

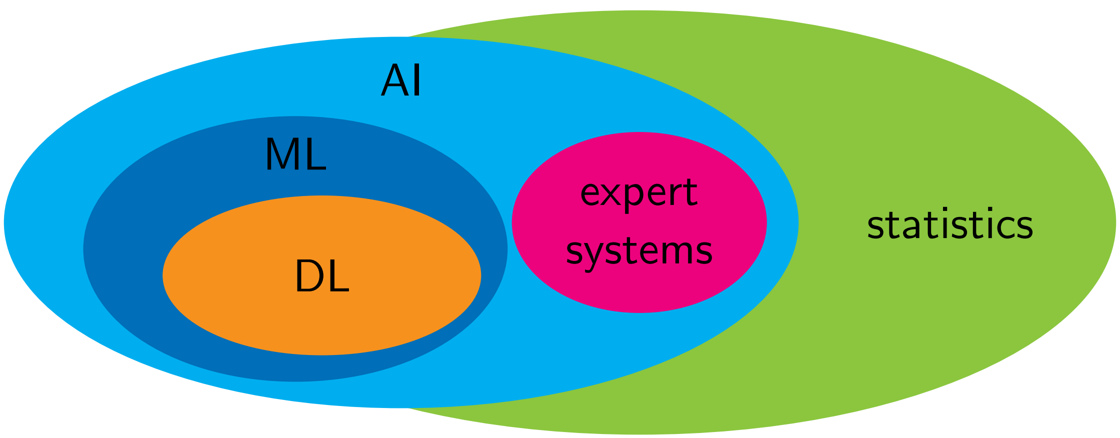
- B.S. in EE @ SNU, M.S. & Ph.D. in EE @ Stanford Univ.
  - *Convex Optimization - theory / algorithms / applications - under supervision of Prof. Stephen P. Boyd*
- Principal Engineer @ Memory Design Technology Team
  - AI & optimization partnering with *DRAM/NAND Design/Process/Test teams*
- Senior Applied Scientist @ Amazon
  - *S-Team Goal (Bezos's) project - better customer shopping experience via Amazon shopping app using AI - increased sales by $200M*
- Co-founder / CTO & Chief Applied Scientist @ Gauss Labs
  - *R&D industrial AI products & technology, market/product/investment strategies*
- Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio
  - *biotech - AI technology & product strategy*

Artificial Intelligence

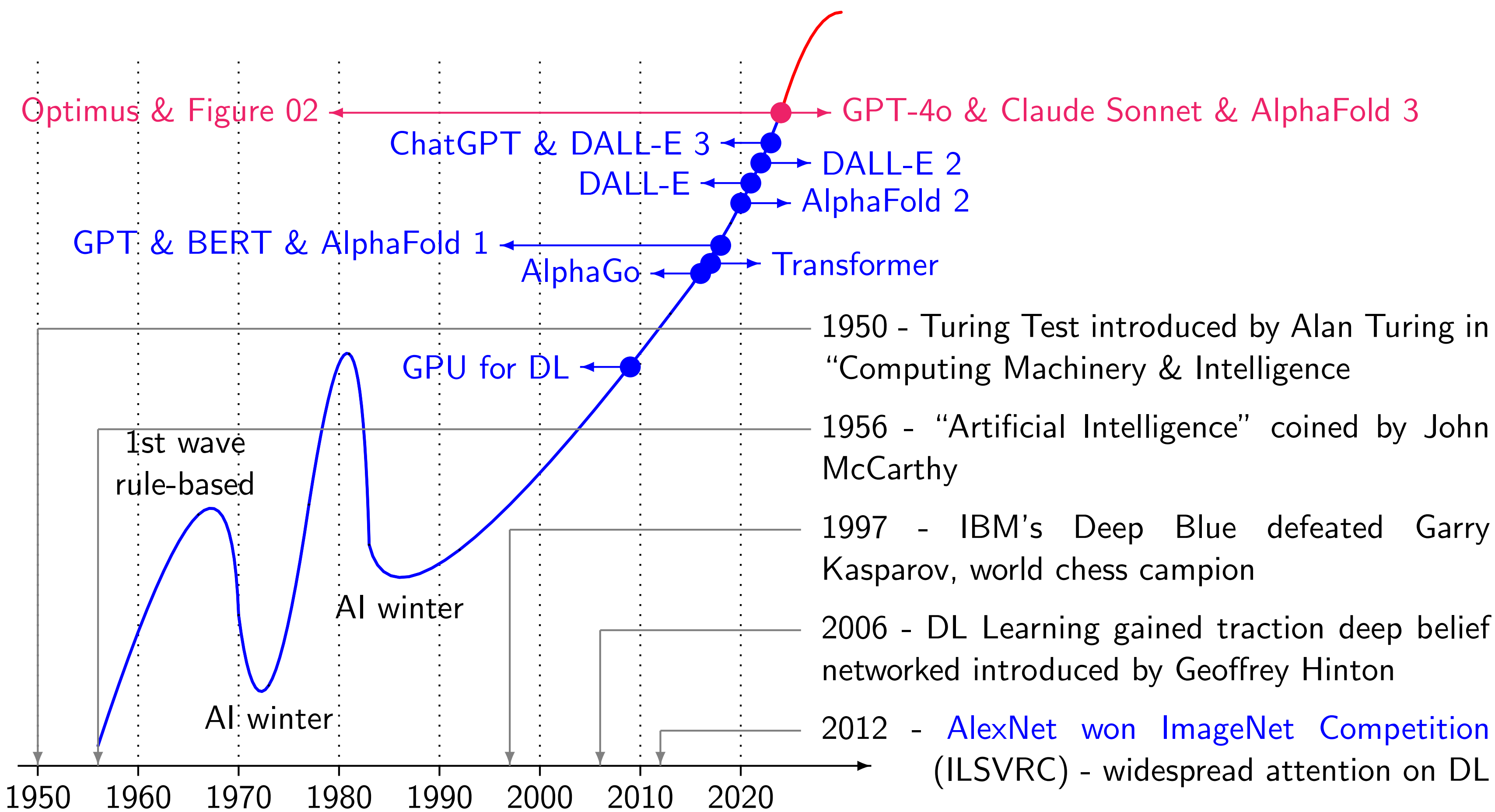
Definition and History

Definition of AI

- AI is
  - technology enabling machines to do tasks requiring human intelligence, such as learning, problem-solving, decision-making & language understanding
  - *not* one thing - encompass range of technologies, methodologies & applications
- relationship of AI, statistics, ML, DL, NN & expert system [HGH+22]



History of AI



Significant AI Achievements - 2014 – 2024

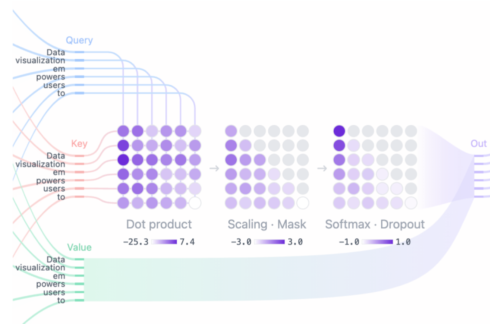
Deep learning revolution

- 2012 – 2015 - DL revolution[1]
  - CNNs demonstrated exceptional performance in image recognition, *e.g.*, *AlexNet's victory in ImageNet competition*
  - widespread adoption of DL learning in CV transforming industries
- 2016 - AlphaGo defeats human Go champion
  - DeepMind's AlphaGo defeated world champion in Go, extremely complex game *believed to be beyond AI's reach*
  - significant milestone in RL - AI's potential in solving complex & strategic problems

[1] DL: deep learning, CNN: convolutional neural network, CV: computer vision, RL: reinforcement learning

Transformer changes everything

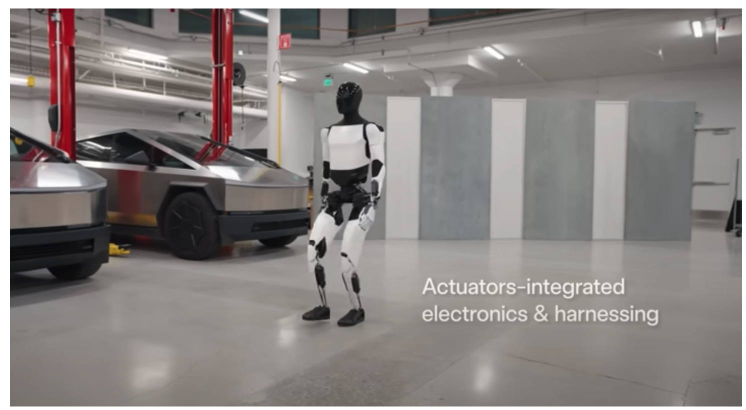
- 2017 – 2018 - Transformers & NLP breakthroughs[2]
  - *Transformer ($e.g.$, BERT & GPT) revolutionized NLP*
  - major advancements in, $e.g.$, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 3D protein structures with remarkable accuracy
  - accelerates drug discovery and personalized medicine - offering new insights into diseases and potential treatments



[2] NLP: natural language processing, GPT: generative pre-trained transformer

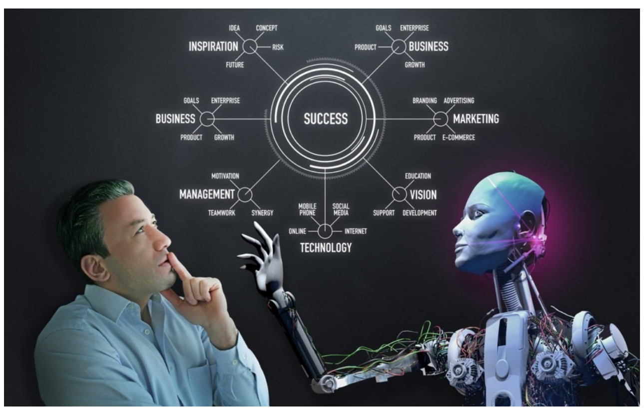
Lots of breakthroughs within 6 months in 2024

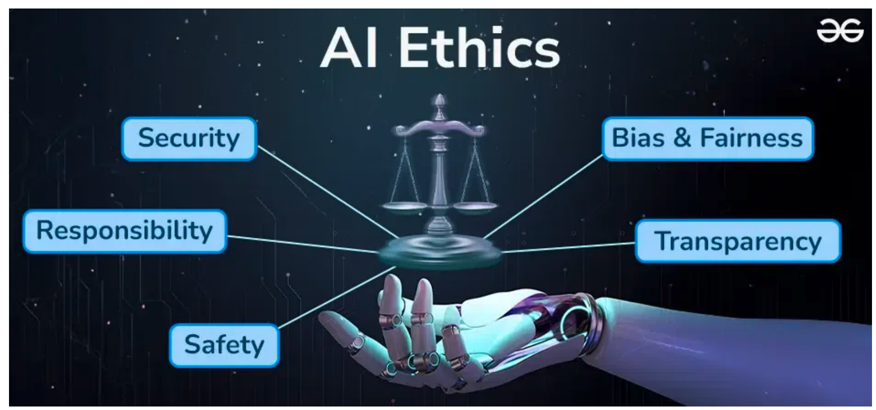
- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Llama 3, Sora
  - *transforming industries* such as content creation, customer service, education, *etc.*
- breakthroughs in specialized AI applications
  - Figure 02, Optimus, AlphaFold 3
  - driving unprecedented advancements in automation, drug discovery, scientific understanding - *profoundly affecting healthcare, manufacturing, scientific research*



Transformative impact of AI - reshaping industries, work & society

- accelerating human-AI collaboration
  - not only reshaping industries but *altering how humans interact with technology*
  - AI's role as collaborator and augmentor redefines productivity, creativity, the way we address global challenges, $e.g.$, *sustainability & healthcare*
- AI-driven automation *transforms workforce dynamics* - creating new opportunities while challenging traditional job roles
- *ethical AI considerations* becoming central not only to business strategy, but to society as a whole - *influencing regulations, corporate responsibility & public trust*

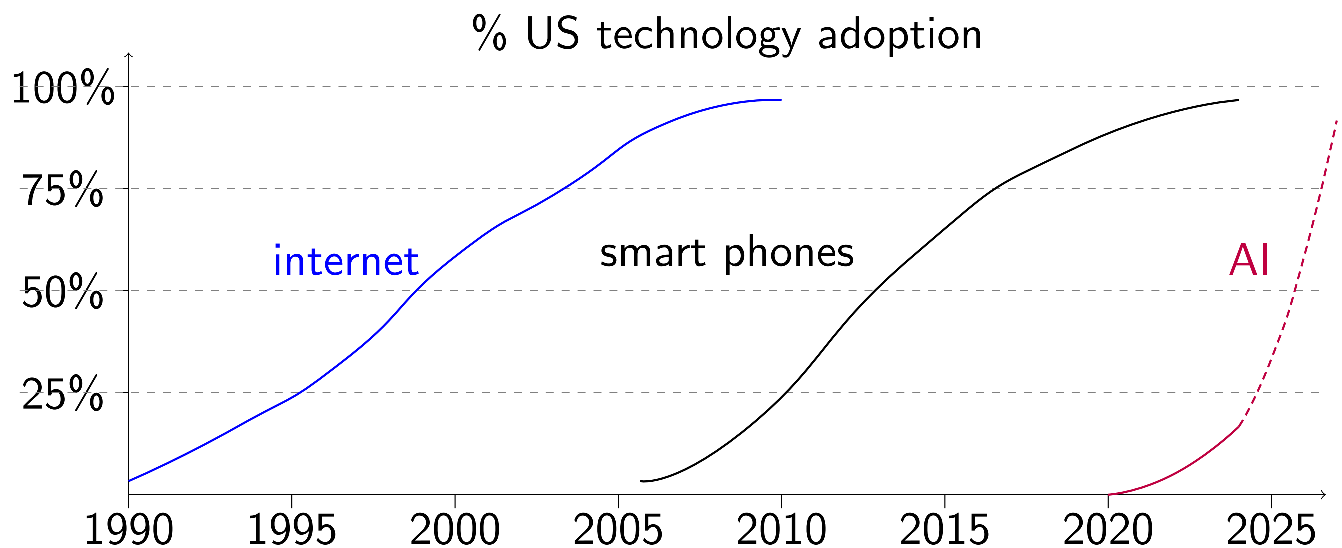




Recent Advances in AI

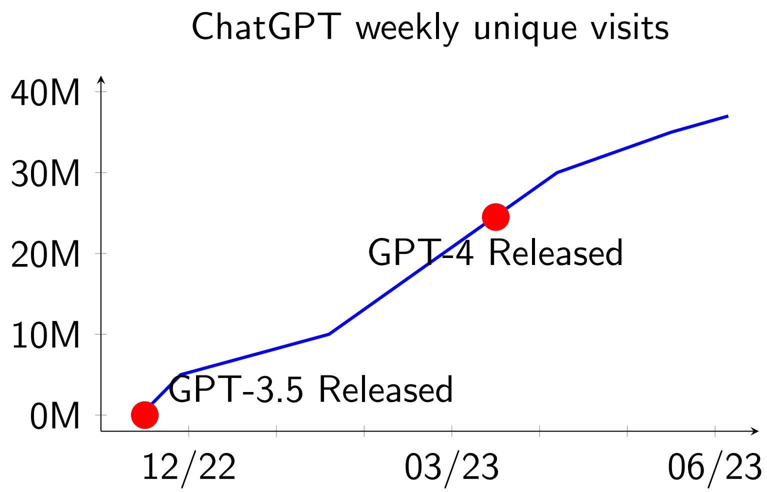
Where are we in AI today?

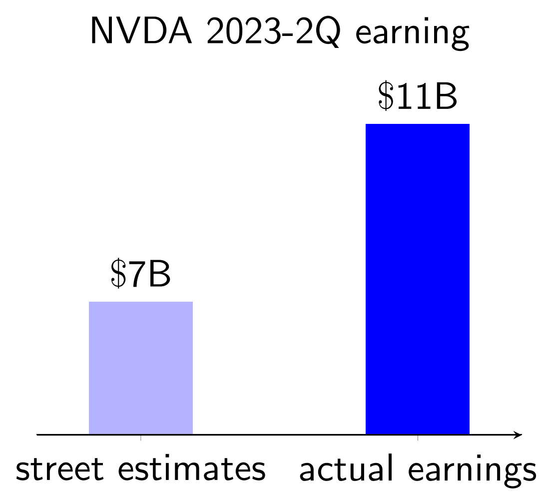
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



Explosion of AI ecosystems - ChatGPT & NVIDIA

- took only *5 months for ChatGPT users to reach 35M*
- NVDIA 2023 Q2 earning exceeds market expectation by big margin - \$7B vs \$13.5B
  - surprisingly, *101% year-to-year growth*
  - even more surprisingly *gross margin was 71.2%* - up from 43.5% in previous year[3]

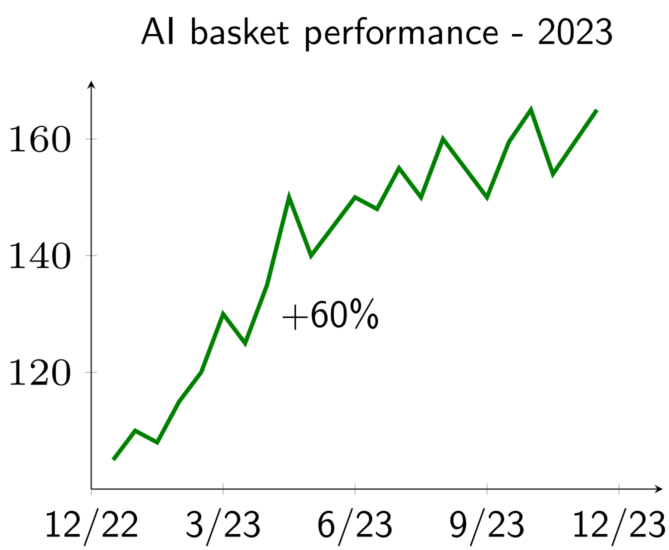


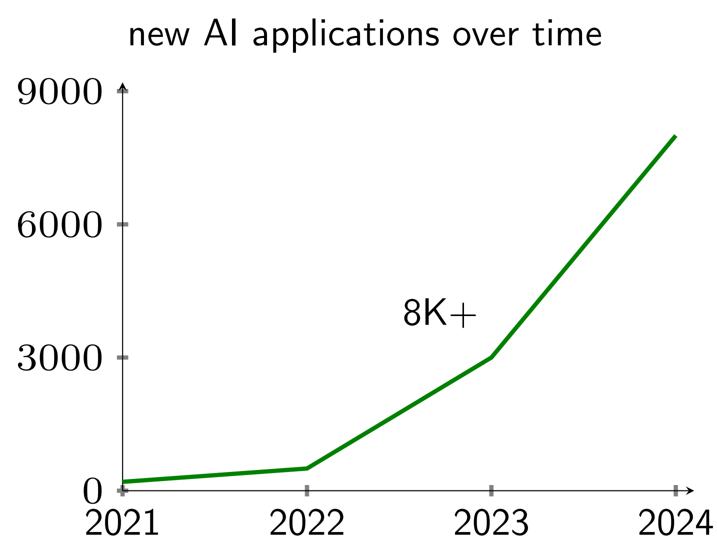


[3] source - Bloomberg

Explosion of AI ecosystems - AI stock market

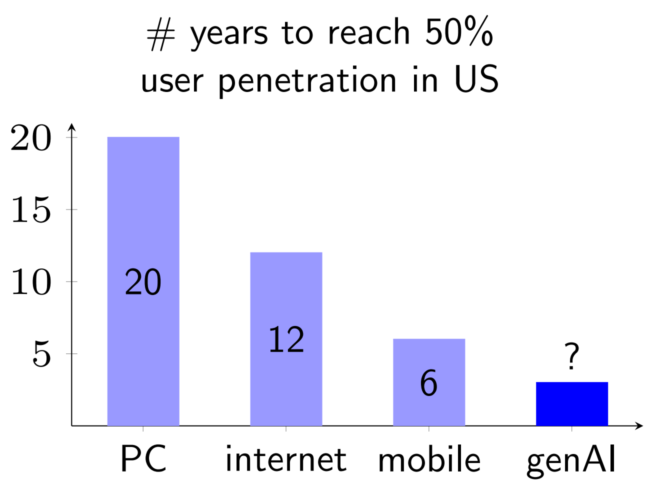
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

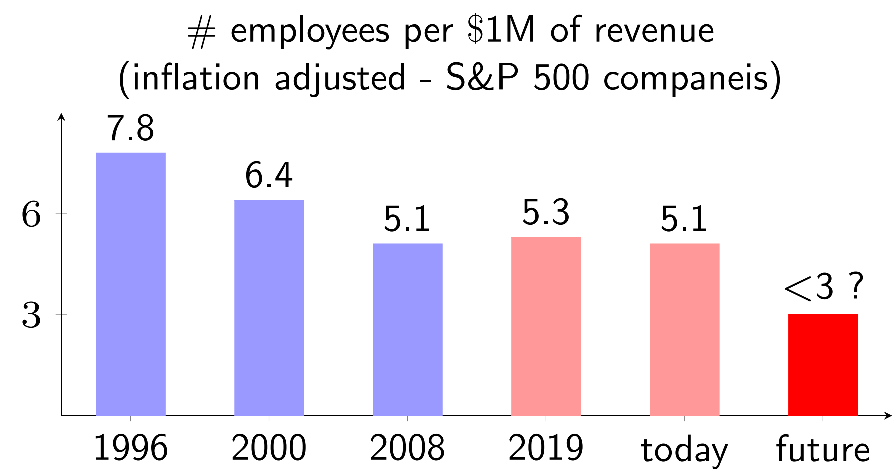




AI's transformative impact - adoption speed & economic potential

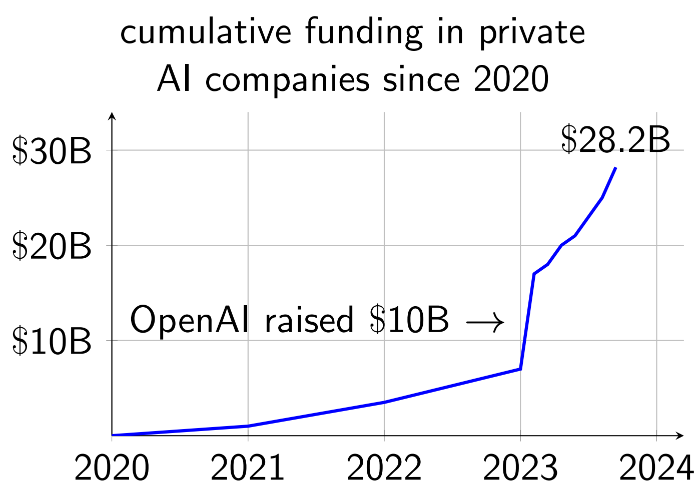
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater gains expected with AI proliferation

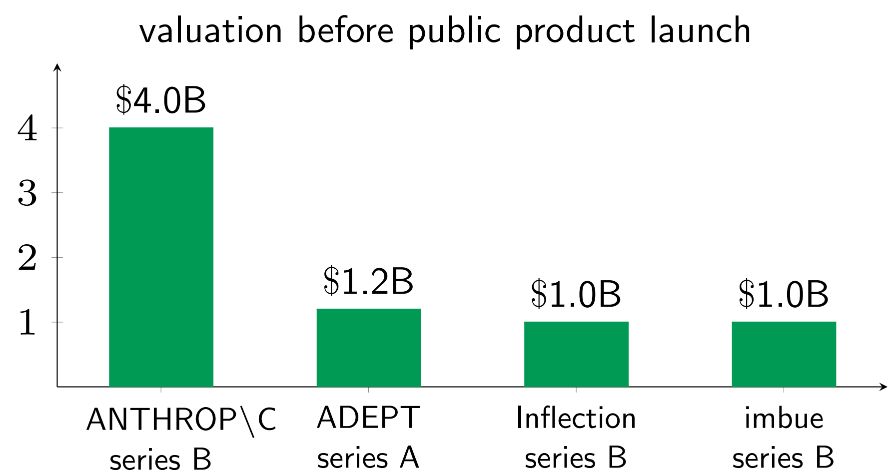




Massive investment in AI

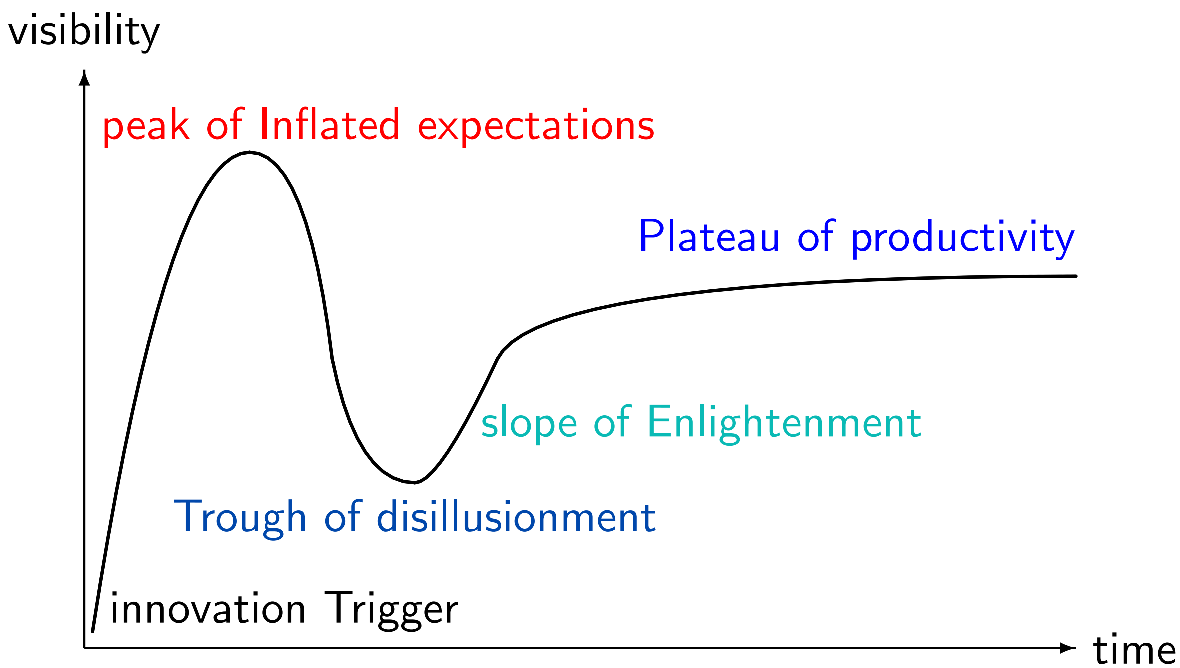
- *explosive growth* - cumulative funding skyrocketed reaching staggering $28.2B
- OpenAI - significant fundraising (=$10B) fueled rapid growth
- *valuation surge* - substantial valuations even before public products for stella companies
- *fierce competition for capital* among AI startups driving innovation & accelerating development
- massive investment indicates *strong belief in & optimistic outlook for potential of AI* to revolutionize industries & drive economic growth





Is AI hype?

Technology hype cycle



- innovation trigger - technology breakthrough kicks things off
- peak of inflated expectations - early publicity induces many successes followed by even more
- trough of disillusionment - expectations wane as technology producers shake out or fail
- slope of enlightenment - benefit enterprise, technology better understood, more enterprises fund pilots

Fiber vs cloud infrastructure

- fiber infrastructure - 1990s
  - Telco Co's raised $1.6T of equity & $600B of debt
  - bandwidth costs decreased 90% within 4 years
  - companies - Covage, NothStart, Telligent, Electric Lightwave, 360 networks, Nextlink, Broadwind, UUNET, NFS Communications, Global Crossing, Level 3 Communications
  - became *public good*
- cloud infrastructure - 2010s
  - entirely new computing paradigm
  - mostly public companeis with data centers
  - *big 4 hyperscalers generate* $150B + annual revenue

Yes & No

characteristics of hype cycles	speaker's views
value accrual misaligned with investment	• OpenAI still operating at a loss; business model *still* not clear • gradual value creation across broad range of industries and technologies (*e.g.*, CV, LLMs, RL) unlike fiber optic bubble in 1990s
overestimating timeline & capabilities of technology	• self-driving cars delayed for over 15 years, with limited hope for achieving level 5 autonomy • AI, however, has proven useful within a shorter 5-year span, with enterprises eagerly adopting
lack of widespread utility due to technology maturity	• AI already providing significant utility across various domains • vs quantum computing remains promising in theory but lacks widespread practical utility

AI Research

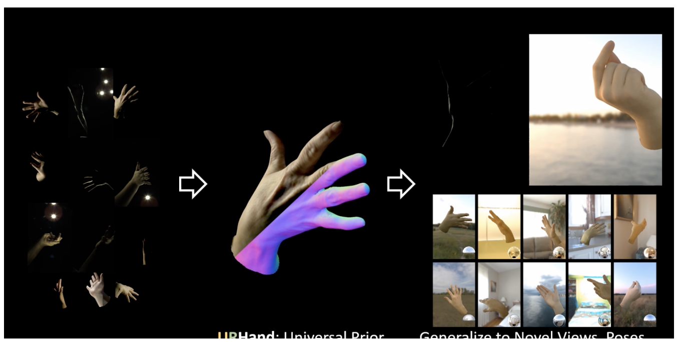
AI research race gets crazy

- practically impossible to follow all developments announced everyday
  - new announcement and publication of important work everyday!
- *industry leads research - academia lags behind*
  - trend observed even before 2015
- everyone excited to show off their work to the world - conference and `github.com`
  - biggest driving force behind unprecedented scale and speed of advancement of AI together with massive investment of capitalists

AI progress within a month - March, 2024

- UBTECH Humanoid Robot Walker S: Workstation Assistant in EV Production Line
- H1 Development of dance function
- Robot Foundation Models (Large Behavior Models) by Toyota Research Institute (TRI)
- Apple Vision Pro for Robotics
- Figure AI & OpenAI
- Human modeling
- LimX Dynamics' Biped Robot P1 Conquers the Wild Based on Reinforcement Learning
- HumanoidBench: Simulated Humanoid Benchmark for Whole-Body Locomotion and Manipulation - UC Berkeley & Yonsei Univ.
- Vision-Language-Action Generative World Model
- RFM-1 - Giving robots human-like reasoning capabilities

Papers of single company accepted by single conference



- CVPR 2024
  - PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar - MIT, Codec Avatars Lab, & Meta [KXS+24]
    - 3D reconstruction from single-view
  - Nymeria Dataset
    - large-scale multimodal egocentric dataset for full-body motion understanding
  - Relightable Gaussian Codec Avatars - Codec Avatars Lab & Meta [SSS+24]
    - build high-fidelity relightable head avatars being animated to generate novel expressions
  - Robust Human Motion Reconstruction via Diffusion (RoHM) - ETH Zürich & Reality Labs Research, Meta [ZBX+24]
    - robust 3D human motion reconstruction from monocular RGB videos

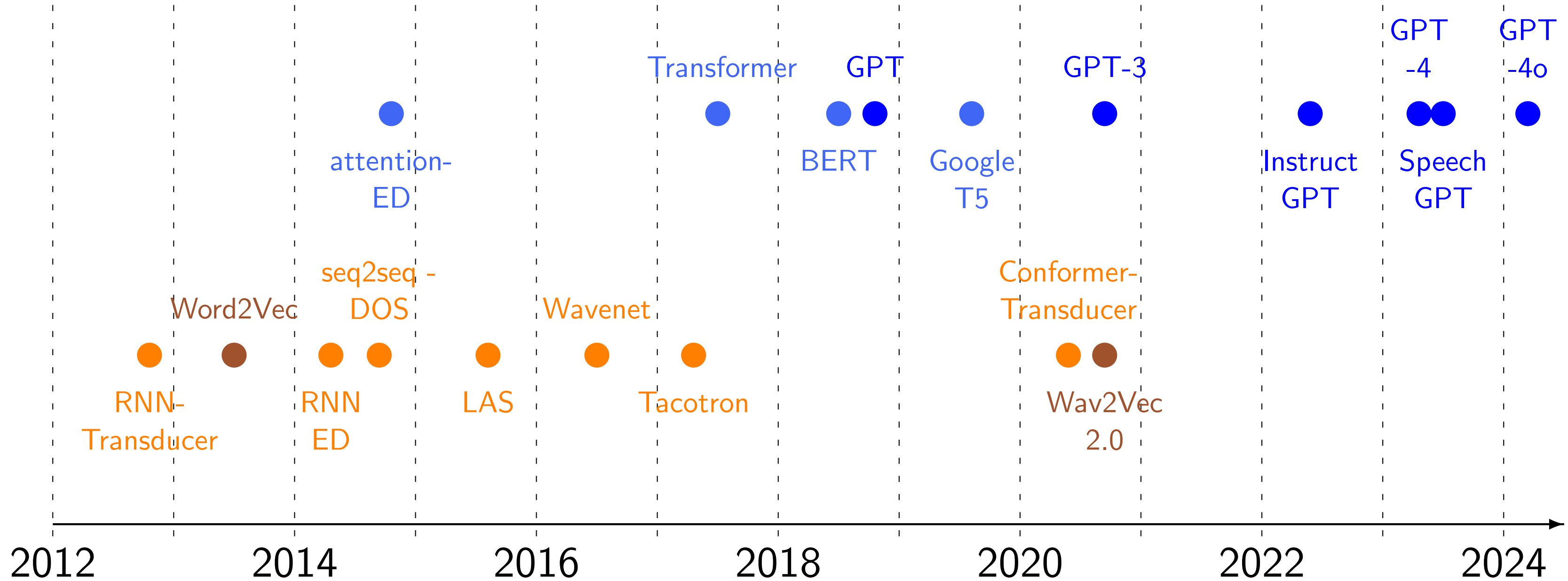
LLM

Language Models

History of language models

- bag of words - first introduced – 1954
- word embedding – 1980
- RNN based models - conceptualized by David Rumelhart – 1986
- LSTM (based on RNN) – 1997
- 380M-sized seq2seq model using LSTMs proposed – 2014
- 130M-sized seq2seq model using gated recurrent units (GRUs) – 2014
- Transformer - Attention is All You Need - A. Vaswani et al. @ Google – 2017
  - 100M-sized encoder-decoder multi-head attention model for machine translation
  - non-recurrent architecture, handle arbitrarily long dependencies
  - parallelizable, *simple* (linear-mapping-based) attention model

Recent advances in speech & language processing



- LAS: listen, attend, and spell, ED: encoder-decoder, DOS: decoder-only structure

Types of language models

- many of language models have common requirements - language representation learning
- can be learned via pre-tranining *high performing model* and fine-tuning/transfer learning/domain adaptation
- this *high performing model* learning essential language representation *is* (lanauge) foundation model
- – actually, same for other types of learning, $e.g.$, CV

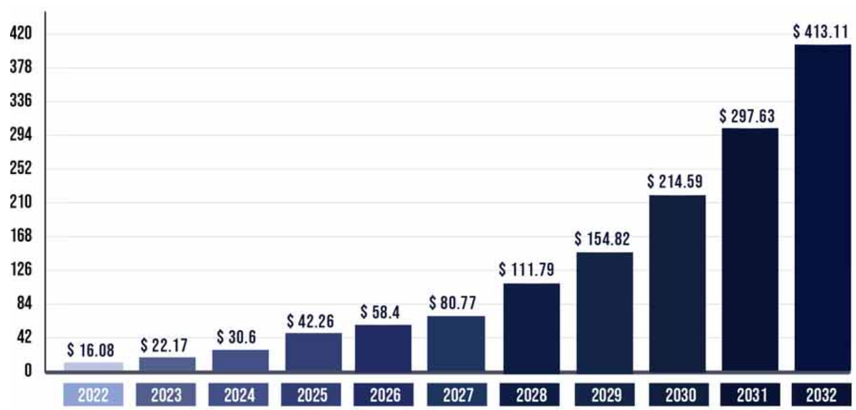
voice recognition	NL generation	summarization	spelling correction	action planning
voice synthesis	translation	topic model	topic model	
	Q&A	NLU		

NLP Market

NLP market size

- global NLP market size estimated at USD 16.08B in 2022, is expected to hit USD 413.11B by 2032 - *CAGR of 38.4%*
- in 2022
  - north america NLP market size valued at USD 8.2B
  - high tech and telecom segment accounted revenue share of over 23.1%
  - healthcare segment held a 10% market share
  - (by component) solution segment hit 76% revenue share
  - (deployment mode) on-premise segment generated 56% revenue share
  - (organizational size) large-scale segment contributed highest market share

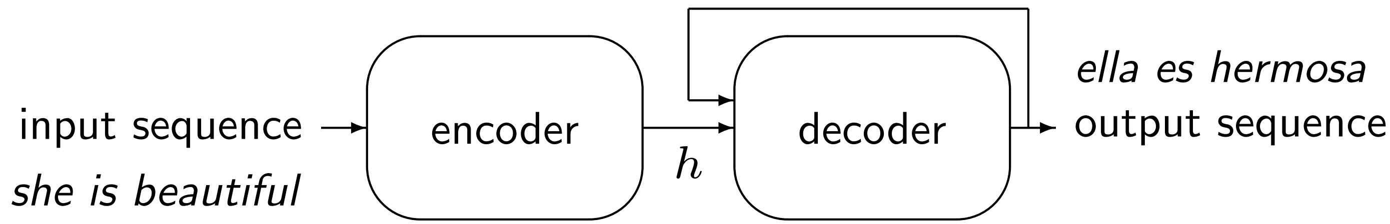
\- source - Precedence Research



Sequence-to-Sequence Models

Sequence-to-sequence (seq2seq) model

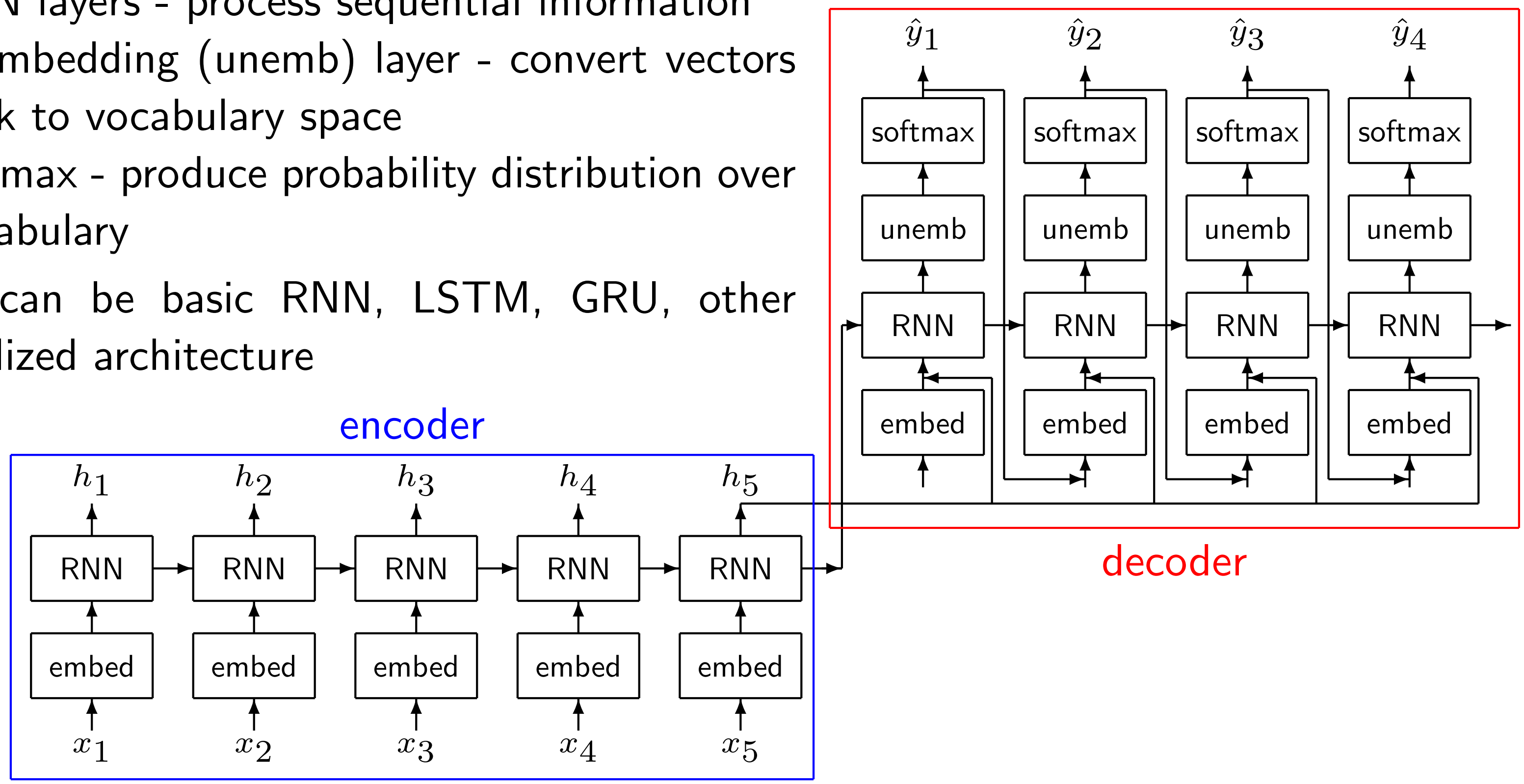
- seq2seq - take sequences as inputs and spit out sequences
- encoder-decoder architecture



  - encoder & decoder can be RNN-type models
  - $h \in \mathbf{R}^n$ - hidden state - *fixed length* vector
- (try to) condense and store information of input sequence (losslessly) in (fixed-length) hidden states
  - finite hidden state - not flexible enough, $i.e.$, cannot handle arbitrarily large information
  - memory loss for long sequences
  - LSTM was promising fix, but with (inevitable) limits

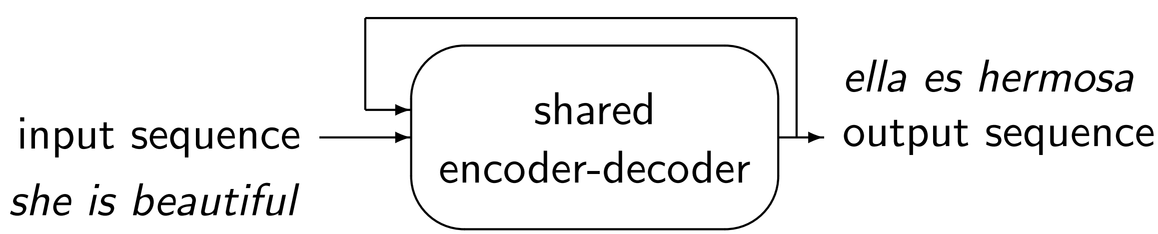
RNN-type encoder-decoder architecture

- components
  - embedding layer - convert input tokens to vector representations
  - RNN layers - process sequential information
  - unembedding (unemb) layer - convert vectors back to vocabulary space
  - softmax - produce probability distribution over vocabulary
- RNN can be basic RNN, LSTM, GRU, other specialized architecture



Shared encoder-decoder model

- single neural network structure can handle both encoding & decoding tasks
  - efficient architecture reducing model complexity
  - allow for better parameter sharing across tasks
- widely used in modern LLMs to process & generate text sequences
  - applications - machine translation, text summarization, question answering
- advantages
  - efficient use of parameters, versatile for multiple NLP tasks



Large Language Models

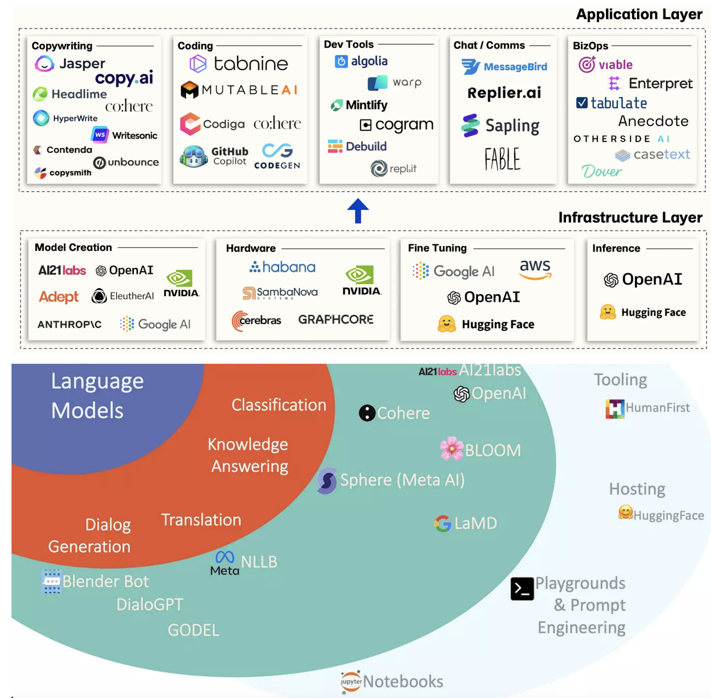
LLM

- LLM
  - type of AI aimed for NLP trained on massive corpus of texts & programming code
  - allow learn statistical relationships between words & phrases, $i.e.$, conditional probabilities
  - *amazing performance shocked everyone - unreasonable effectiveness of data (Halevry et al., 2009)*
- applications
  - conversational AI agent / virtual assistant
  - machine translation / text summarization / content creation / sentiment analysis / question answering
  - code generation
  - market research / legal service / insurance policy / triange hiring candidates

  \+ virtually infinite # of applications

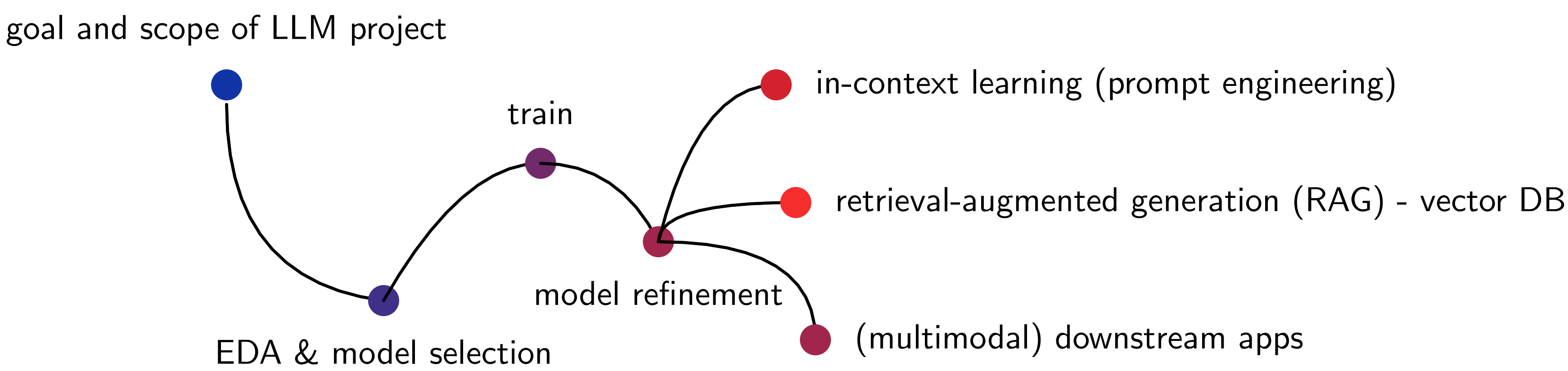
LLMs

- Foundation Models
  - GPT-x/Chat-GPT - OpenAI, Llama-x - Meta, PaLM-x (Bard) - Google
- # parameters
  - generative pre-trained transfomer (GPT) - GPT-1: 117M, GPT-2: 1.5B, GPT-3: 175B, GPT-4: 100T, GPT-4o: 200B
  - large language model Meta AI (Llama) - Llama1: 65B, Llama2: 70B, Llama3: 70B
  - scaling language modeling with pathways (PaLM) - 540B
- burns lots of cash on GPUs!
- applicable to many NLP & genAI applications



LLM building blocks

- data - trained on massive datasets of text & code
  - quality & size critical on performance
- architecture - GPT/Llama/Mistral
  - can make huge difference
- training - self-supervised/supervised learning
- inference - generates outputs
  - in-context learning, prompt engineering



Transformer

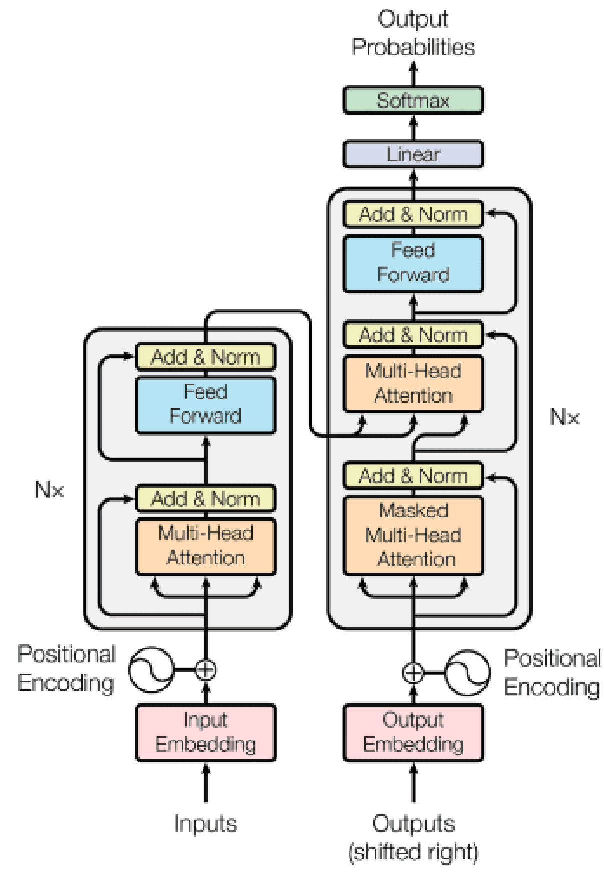
LLM architectural secret (or known) sauce

Transformer - simple parallelizable attention mechanism

A. Vaswani, et al. Attention is All You Need, 2017

Transformer architecture

- encoding-decoding architecture
  - input embedding space → multi-head & mult-layer representation space → output embedding space
- additive positional encoding - information regarding order of words @ input embedding
- multi-layer and multi-head attention followed by addition / normalization & feed forward (FF) layers
- *(relatively simple) attentions*
  - single-head (scaled dot-product) / multi-head attention
  - self attention / encoder-decoder attention
  - masked attention
- benefits
  - *evaluate dependencies between arbitrarily distant words*
  - has recurrent nature w/o recurrent architecture → parallelizable → fast w/ additional cost in computation



Single-head scaled dot-product attention

- values/keys/queries denote value/key/query *vectors*, d_k & d_v are lengths of keys/queries & vectors
- we use *standard* notions for matrices and vectors - not transposed version that (almost) all ML scientists (wrongly) use

- output: weighted-average of values where weights are attentions among tokens
- assume n queries and m key-value pairs

$$Q \in \mathbf{R}^{d_k \times n}, K \in \mathbf{R}^{d_k \times m}, V \in \mathbf{R}^{d_v \times m}$$

- attention! outputs n values (since we have n queries)

$$\mathrm{Attention}(Q, K, V) = V \mathsf{softmax}\left(K^T Q / \sqrt{d_k}\right) \in \mathbf{R}^{d_v \times n}$$

- *much simpler attention mechanism than previous work*
  - attention weights were output of complicated non-linear NN

Single-head - close look at equations

- focus on ith query, $q_i \in \mathbf{R}^{d_k}$, $Q = \begin{bmatrix} - & q_i & - \end{bmatrix} \in \mathbf{R}^{d_k \times n}$
- assume m keys and m values, $k_1, \ldots, k_m \in \mathbf{R}^{d_k}$ & $v_1, \ldots, v_m \in \mathbf{R}^{d_v}$

$$K = \begin{bmatrix} k_1 & \cdots & k_m \end{bmatrix} \in \mathbf{R}^{d_k \times m}, V = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \in \mathbf{R}^{d_v \times m}$$

- then

$$K^T Q / \sqrt{d_k} = \begin{bmatrix} & \vdots & \\ - & k_j^T q_i / \sqrt{d_k} & - \\ & \vdots & \end{bmatrix}$$

$e.g.$, dependency between ith output token and jth input token is

$$a_{ij} = \exp\left(k_j^T q_i / \sqrt{d_k}\right) / \sum_{j=1}^{m} \exp\left(k_j^T q_i / \sqrt{d_k}\right)$$

- value obtained by ith query, q_i in $\mathrm{Attention}(Q, K, V)$

$$a_{i,1} v_1 + \cdots + a_{i,m} v_m$$

Multi-head attention

- evaluate h single-head attentions (in parallel)
- d_e: dimension for embeddings
- embeddings

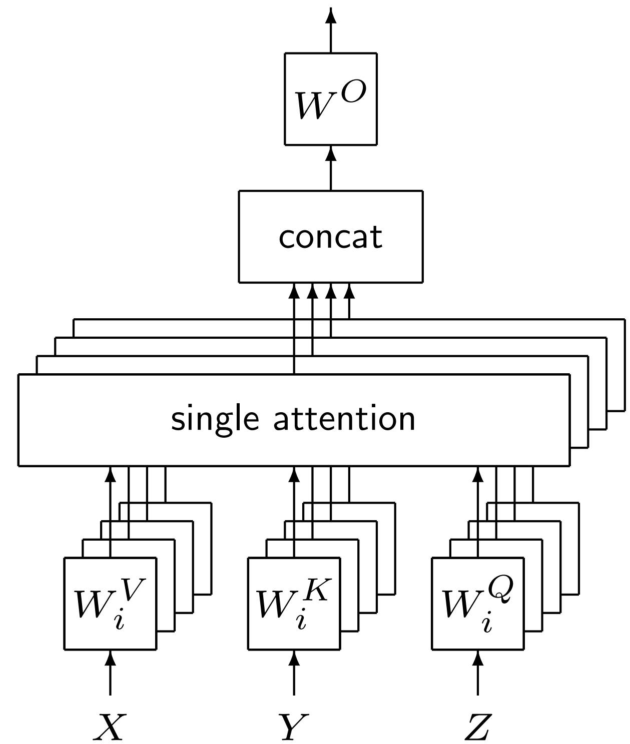
$$X \in \mathbf{R}^{de \times m},\ Y \in \mathbf{R}^{de \times m},\ Z \in \mathbf{R}^{de \times n}$$

  e.g., n: input sequence length & m: output sequence length in machine translation

- h key/query/value weight matrices: $W_i^K, W_i^Q \in \mathbf{R}^{d_k \times de}$, $W_i^V \in \mathbf{R}^{dv \times de}$ $(i = 1, \ldots, h)$
- linear output layers: $W^O \in \mathbf{R}^{de \times hdv}$
- *multi-head attention!*

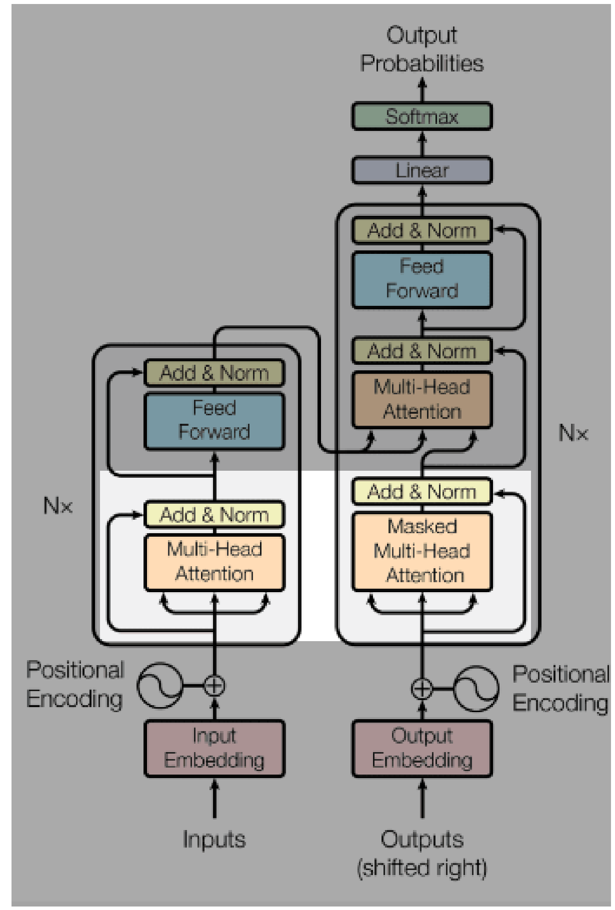
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbf{R}^{de \times n},$$

$$A_i = \mathrm{Attention}(W_i^Q Z, W_i^K Y, W_i^V X) \in \mathbf{R}^{dv \times n}$$



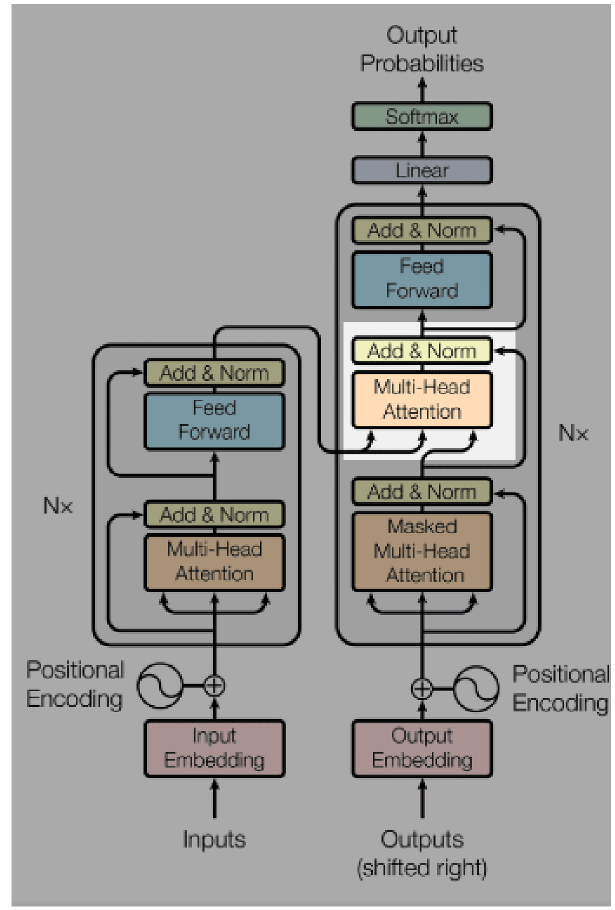
Self attention

- $m = n$
- encoder
  - keys & values & queries (K, V, Q) come from same place (from previous layer)
  - every token attends to every other token in input sequence
- decoder
  - keys & values & queries (K, V, Q) come from same place (from previous layer)
  - every token attends to other tokens up to that position
  - prevent leftward information flow to right to preserve causality
  - assign $-\infty$ for illegal connections in softmax (masking)



Encoder-decoder attention

- m: length of input sequence
- n: length of output sequence
- n queries (Q) come from previous decoder layer
- m keys / m values (K, V) come from output of encoder
- every token in output sequence attends to every token in input sequence

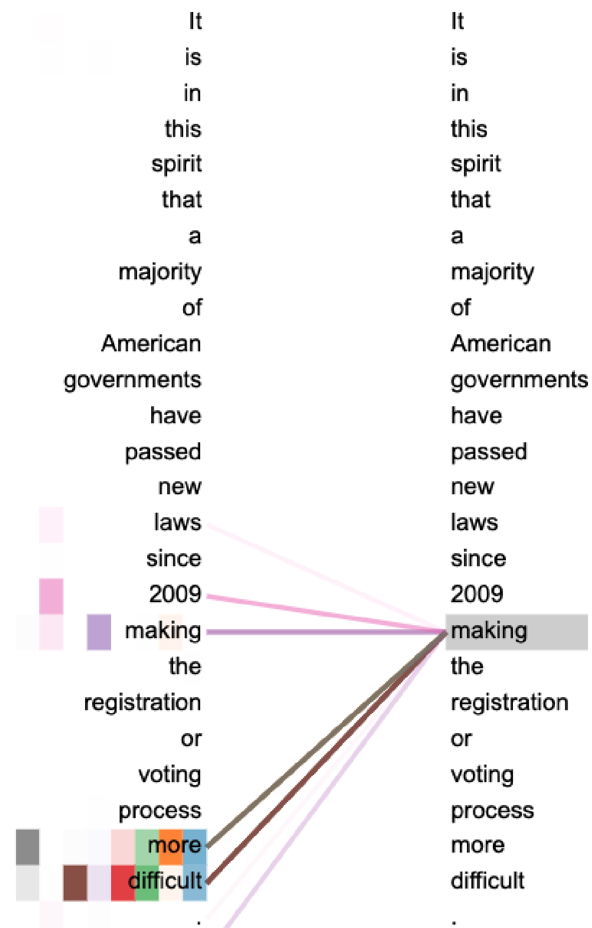


Visualization of self attentions

example sentence

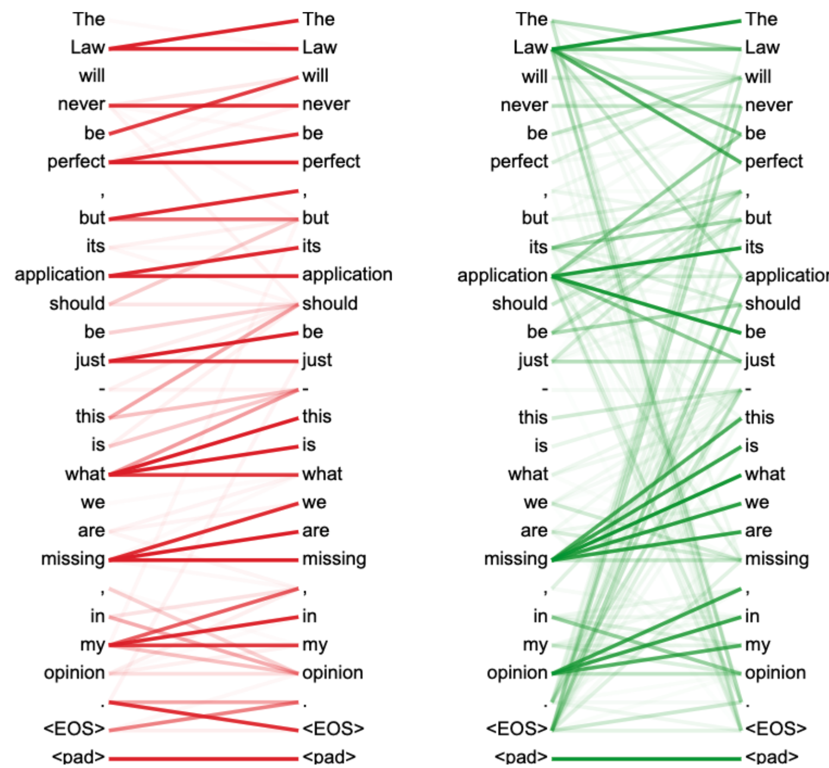
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."

- self attention of encoder (of a layer)
  - right figure
    - show dependencies between "making" and other words
    - different columns of colors represent different heads
  - "making" has strong dependency to "2009", "more", and "difficult"

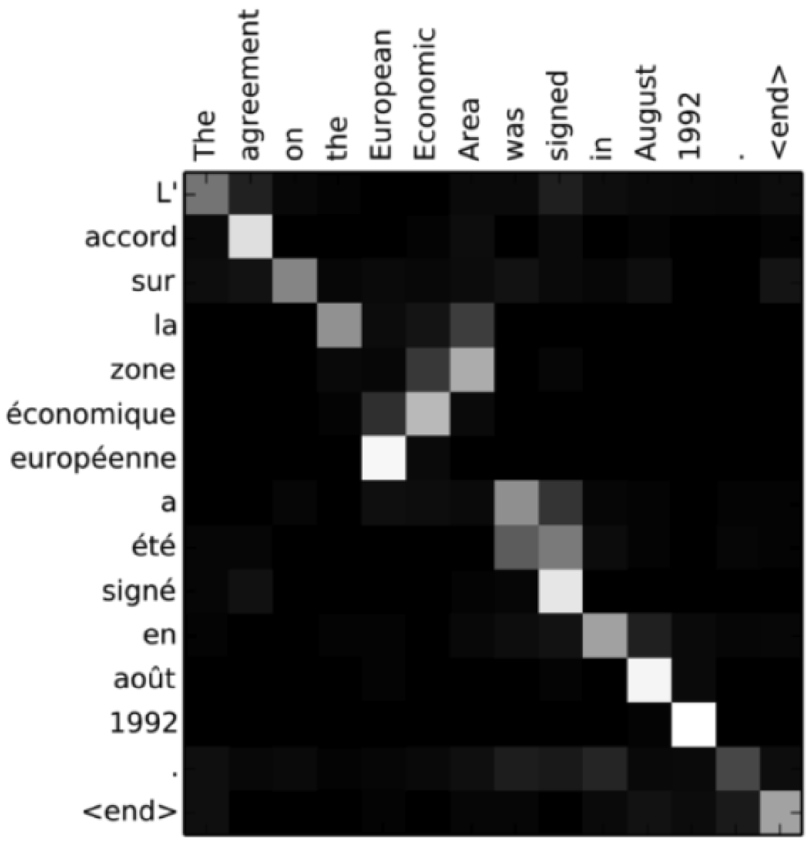


Visualization of multi-head self attentions

- self attentions of encoder for two heads (of a layer)
  - different heads represent different structures
    $\rightarrow$ advantages of multiple heads
  - multiple heads work together to colletively yield good results
  - dependencies *not* have absolute meanings (like embeddings in collaborative filtering)
  - randomness in resulting dependencies exists due to stochastic nature of ML training



Visualization of encoder-decoder attentions



- machine translation: English $\rightarrow$ French
  - input sentence: "The agreement on the European Economic Area was signed in August 1992."
  - output sentence: "L' accord sur la zone économique européenne a été signé en août 1992."
- encoder-decoder attention reveals relevance between
  - European $\leftrightarrow$ européenne
  - Economic $\leftrightarrow$ européconomique
  - Area $\leftrightarrow$ zone

Model complexity

- computational complexity
  - n: sequence length, d: embedding dimension
  - complexity per layer - self-attention: $\mathcal{O}(n^2 d)$, recurrent: $\mathcal{O}(1)$
  - sequential operations - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$
  - maximum path length - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$

- *massive parallel processing, long context windows*

  $\longrightarrow$ *makes NVidia more competitive, hence profitable!*

  $\longrightarrow$ *makes SK Hynix prevail HBM market!*

Variants of Transformer

Bidirectional encoder representations from transformers (BERT)

- Bidirectional Encoder Representations from Transformers [DCLT19]
- pre-train deep bidirectional representations from unlabeled text
- fine-tunable for multiple purposes

input	[cls]	my	dog	is	cute	[sep]	she	likes	playing	[sep]
token embeddings	$E_{\text{[cls]}}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{\text{[sep]}}$	E_{she}	E_{likes}	E_{playing}	$E_{\text{[sep]}}$
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
segment embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
poistional embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9

Challenges in LLMs

- *hallucination - can give entirely plausible outcome that is false*
- data poison attack
- unethical or illegal content generation
- huge resource necessary for both training & inference
- model size - need compact models
- outdated knowledge - can be couple of years old
- lack of reproducibility
- *biases - more on this later . . .*

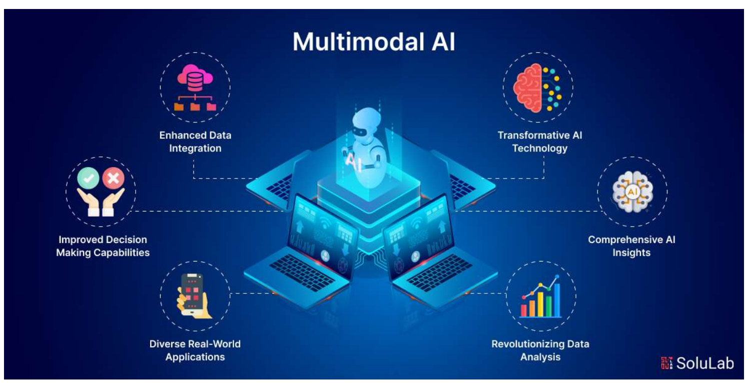
do not, though, focus on downsides but on *infinite possibilities!*

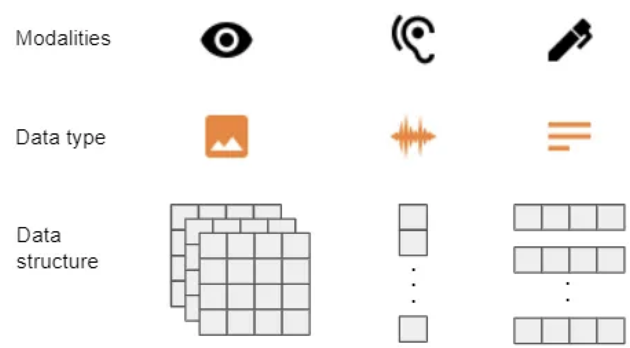
- it evolves like internet / mobile / electricity
- only "tip of the iceburg" found & releaved

Multimodal AI

Multimodal learning

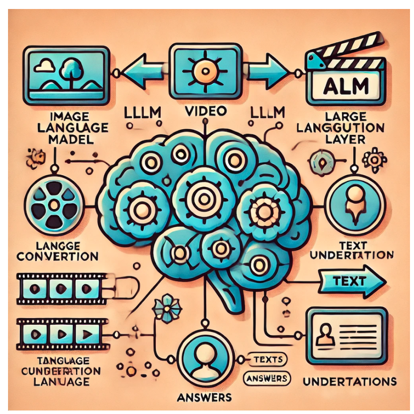
- understand information from multiple modalities, $e.g.$, text, images, audio & video
- representation learning methods
  - combine two representations, $e.g.$, language + image
  - learn multimodal representations simultaneously
- applications
  - images from text prompt, captions from images, videos with narration, musics with lyrics
- collaboration among different modalities
  - understand image world (open system) using language (closed system)





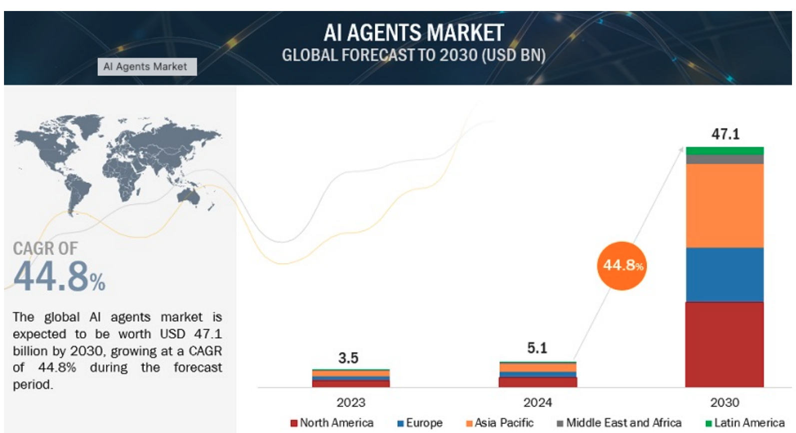
Implications of success of LLMs

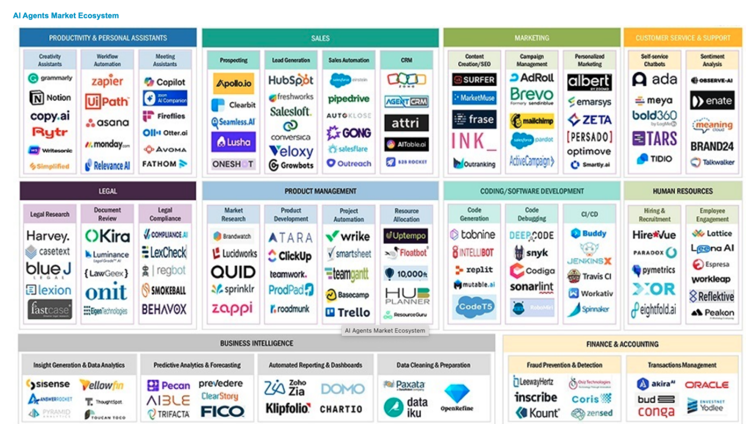
- (very) many researchers change gears towards LLM
  – from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not (only) about NLP . . .* humans have . . .
  - evolved and optimized (natural) language structures for eons
  - handed down knowledge using (this) natural languages for thousands of years
  – (internal structure or representation of) natural language optimized via evolution through *thousands of generation by evolution*
- LLM (can) *connect non-linguistic world (open system) via languages (closed system)*

Multimodal AI (mmAI) - definition & history

- definition - systems processing & integrating data from multiple sources & modalities, *e.g.*, text, images, audio, video, to generate unified response / decision
- 1990s – 2000s - early systems - initial research combining basic text & image data
- 2010s - CNNs & RNNs enabling more sophisticated handling of multimodality
- 2020s - modern multimodal models - Transformer-based architectures (*e.g.*, CLIP and GPT-4 Vision) handling complex multi-source data at highly advanced level
- mmAI *mimics human cognitive ability* to interpret and integrate information from various sources, leading to holistic decision-making





mmAI Technology

- core components
  - data preprocessing - images, text, audio & video
  - architectures - unified Transformer-based (*e.g.*, ViT) & cross-attention mechanisms / hybrid architectures (*e.g.*, CNNs + LLMs)
  - integration layers - fusion methods for combining data representations from different modalities
- technical challenges
  - data alignment - accurate alignment of multimodal data
  - computational demand - high-resource requirements for training and inferencing
  - diverse data quality - manage variations in data quality across modalities
- advancements
  - multimodal embeddings - shared feature spaces interaction between modalities
  - self-supervised learning - leverage unlabeled data to learn representations across modalities

mmAI - current and future industries

- industries
  - healthcare - enhanced diagnostic combining imaging, *e.g.*, MRI, with patient history
  - customer experience - virtual assistants understanding spoken language & visual cues
  - autonomous vehicles - integration of visual, radar & audio data
- future
  - ubiquitous mmAI - seamless integration into everyday devices
  - advances in edge computing - low-power & real-time mm processing for IoT apps
  - highly tailored personalized experience - in education, entertainment & healthcare

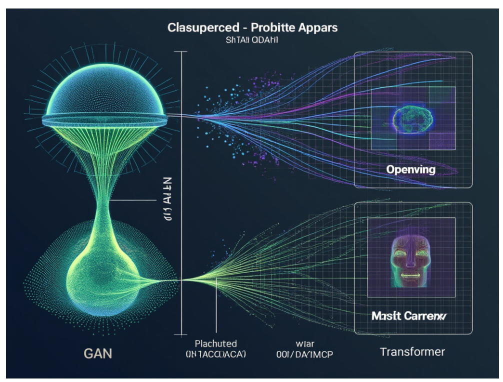
genAI

Definition of genAI

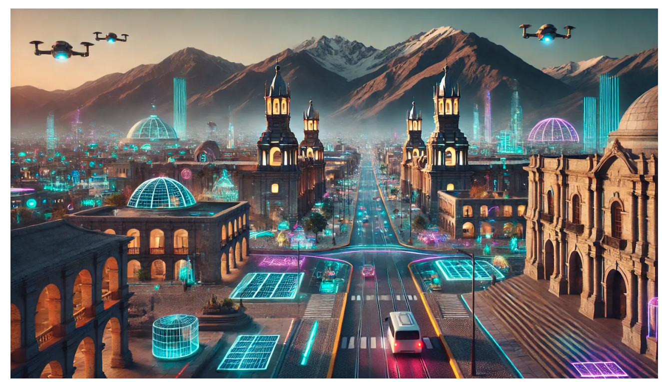
Generative AI

- genAI refers to systems capable of producing new (& original) contents based on patterns learned from training data (representation learning)
  - as opposed to discriminative models for, *e.g.*, classification, prediction & regression
  - here content can be text, images, audio, video, *etc.* - what about smell & taste?
- genAI model examples
  - generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, Transformers

by Midjourney



by Grok 2 mini



by Generative AI Lab

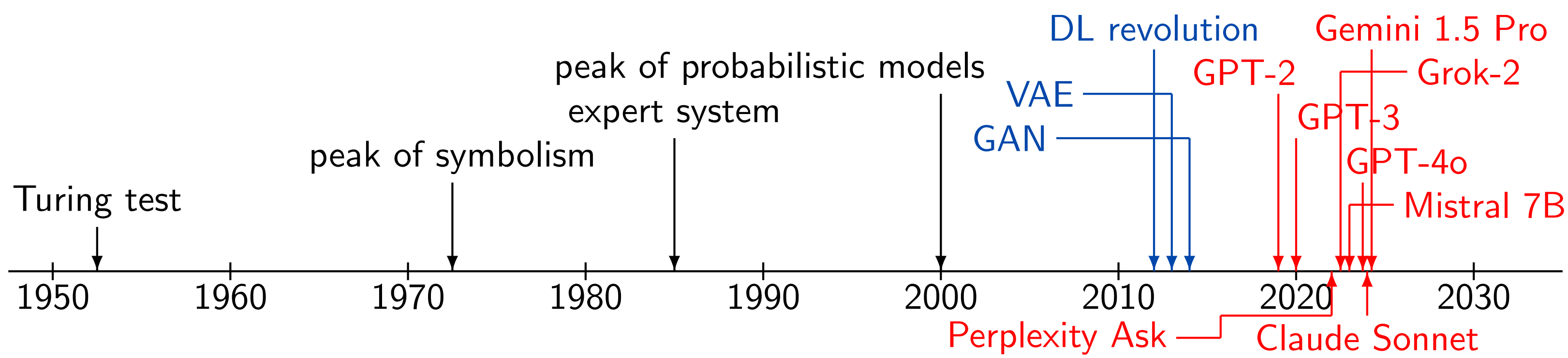
Examples of genAI in action

- text generation
  - Claude, ChatGPT, Mistral, Perplexity, Gemini, Grok
  - conversational agent writing articles, code & even poetry
- image generation
  - DALL-E - creates images based on textual descriptions
  - Stable Diffusion - uses diffusion process to generate high-quality images from text prompts (by denoising random noise)
  - MidJourney - art and visual designs generated through deep learning
- music generation
  - Amper Music - generates unique music compositions
- code generation
  - GitHub Copilot - generates code snippets based on natural language prompts

History of genAI

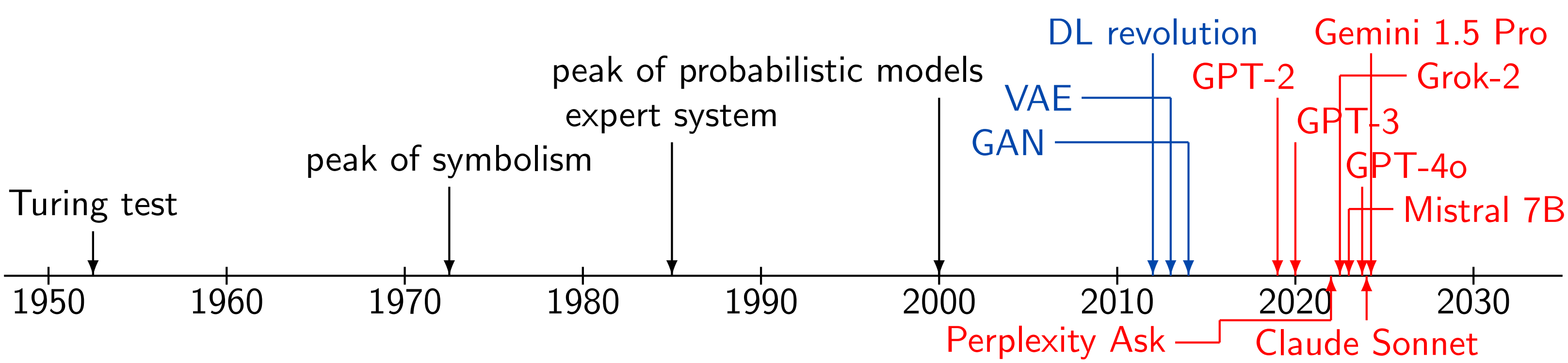
Birth of AI - early foundations & precursor technologies

- 1950s $\sim$ 1970s
  - Alan Turing - concept of *"thinking machine" & Turing test* to evaluate machine intelligence (1950s)
  - *symbolists* (as opposed to connectionists) - early AI focused on symbolic reasoning, logic & problem-solving - Dartmouth Conference in 1956 by *John McCarthy, Marvin Minsky, Allen Newell & Herbert A. Simon*
  - precursor technologies - genetic algorithms (GAs), Markov chains & *hidden Markov models (HMMs)* - laying foundation for generative processes (1970s $\sim$)



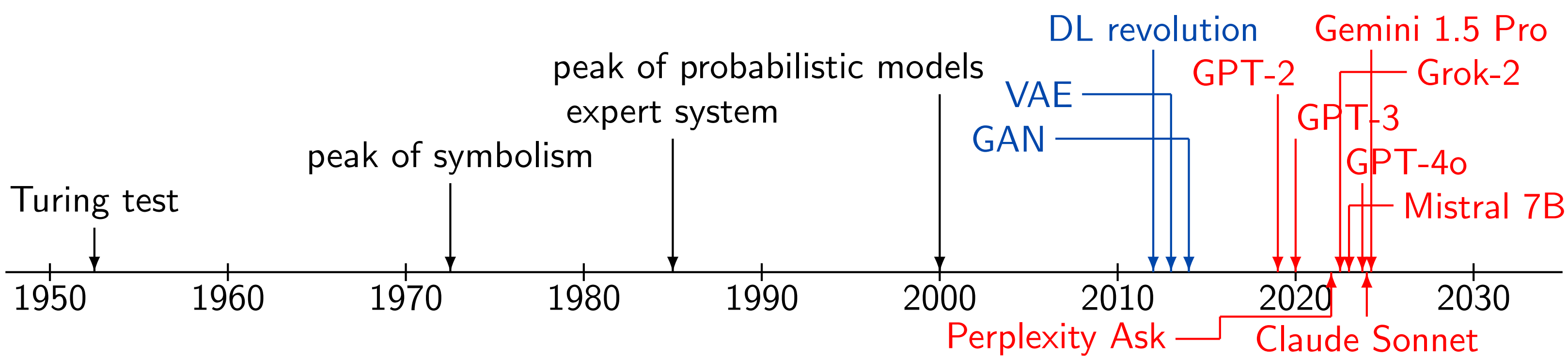
Rule-based systems & probabilistic models

- 1980s $\sim$ early 2000s
  - *expert systems* (1980s) - AI systems designed to mimic human decision-making in specific domains
  - development of neural networks (NN) w/ backpropagation *training multi-layered networks* - setting stage for way more complex generative models
  - *probabilistic models* (including network models, $i.e.$, Bayesian networks) & Markov models - laying groundwork for data generation & pattern prediction



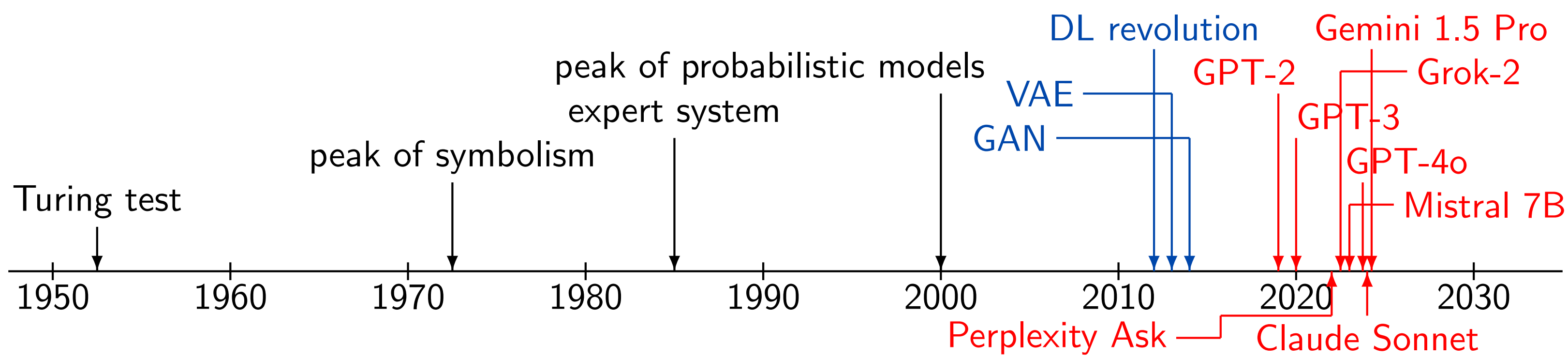
Rise of deep learning & generative models

- 2010s - breakthrough in genAI
  - *deep learning (DL) revolution* - advances in GPU computing and data availability led to the rapid development of deep neural networks.
  - *variational autoencoder (VAE)* (2013) - by Kingma and Welling - learns mappings between input and latent spaces
  - *generative adversarial network (GAN)* (2014) - by Ian Goodfellow - game-changer in generative modeling where two NNs compete each other to create realistic data
    - widely used in image generation & creative tasks



Transformer models & multimodal AI

- late 2010s $\sim$ Present
  - Transformer architecture (2017) - by Vaswani et al.
    - *revolutionized NLP,* $e.g.$, LLM & various genAI models
  - GPT series - generative pre-trained transformer
    - GPT-2 (2019) - generating human-like texts - *marking leap in language models*
    - GPT-3 (2020) - 175B params - set *new standards for LLM*
  - multimodal systems - DALL-E & CLIP (2021) - *linking text and visual data*
  - emergence of diffusion models (2020s) - new approach for generating high-quality images - progressively "denoising" random noise (DALL-E 2 & Stable Diffusion)



Mathy Views on genAI

genAI models

- definition of generative model

$$\boxed{\mathcal{Z}} \xrightarrow{g_\theta(z)} \boxed{\mathcal{X}}$$

- *generate samples in original space, $\mathcal{X}$, from samples in latent space, $\mathcal{Z}$*
- g_θ is parameterized model $e.g.$, CNN / RNN / Transformer / diffuction-based model
- training
  - finding θ that minimizes/maximizes some (statistical) loss/merit function so that $\{g_\theta(z)\}_{z\in\mathcal{Z}}$ generates plausiable point in $\mathcal{X}$
- inference
  - random samples z to generated target samples $x = g_\theta(z)$
  - $e.g.$, image, text, voice, music, video

VAE - early genAI model

- variational auto-encoder (VAE) [KW19]

$$\boxed{\mathcal{X}} \xrightarrow{q_\phi(z|x)} \boxed{\mathcal{Z}\mathrm{o}} \xrightarrow{p_\theta(x|z)} \boxed{\mathcal{X}}$$

- log-likelihood & ELBO - for any $q_\phi(z|x)$

$$\begin{aligned}
\log p_\theta(x) &= \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x) = \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p_\theta(z|x)} \\
&= \mathcal{L}(\theta,\phi;x) + D_{KL}(q_\phi(z|x)\|p_\theta(z|x)) \geq \mathcal{L}(\theta,\phi;x)
\end{aligned}$$

- (indirectly) maximize likelihood by maximizing evidence lower bound (ELBO)

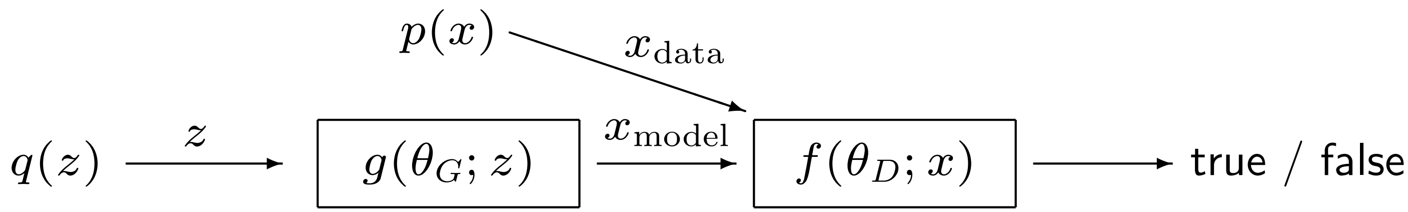
$$\mathcal{L}(\theta,\phi;x) = \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

- generative model

$$p_\theta(x|z)$$

GAN - early genAI model

- generative adversarial networks (GAN) [GPAM+14]



  - value function

$$V(\theta_D, \theta_G) = \mathop{\mathbf{E}}_{x \sim p(x)} \log f(\theta_D; x)) + \mathop{\mathbf{E}}_{z \sim q(z)} \log(1 - f(\theta_D; g(\theta_G; z)))$$

  - modeling via playing min-max game

$$\min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G)$$

  - generative model

$$g(\theta_G; z)$$

  - variants: conditional / cycle / style / Wasserstein GAN

genAI - LLM

- *maximize conditional probability*

$$\underset{\theta}{\text{maximize}} \; d\left(p_\theta(x_t|x_{t-1}, x_{t-2}, \ldots), p_{\text{data}}(x_t|x_{t-1}, x_{t-2}, \ldots)\right)$$

  where $d(\cdot, \cdot)$ distance measure between probability distributions

  – previous sequence: $x_{t-1}, x_{t-2}, \ldots$
  – next token: x_t

- p_θ represented by (extremely) complicated model
  – $e.g.$, containing multi-head & multi-layer Transformer architecture inside

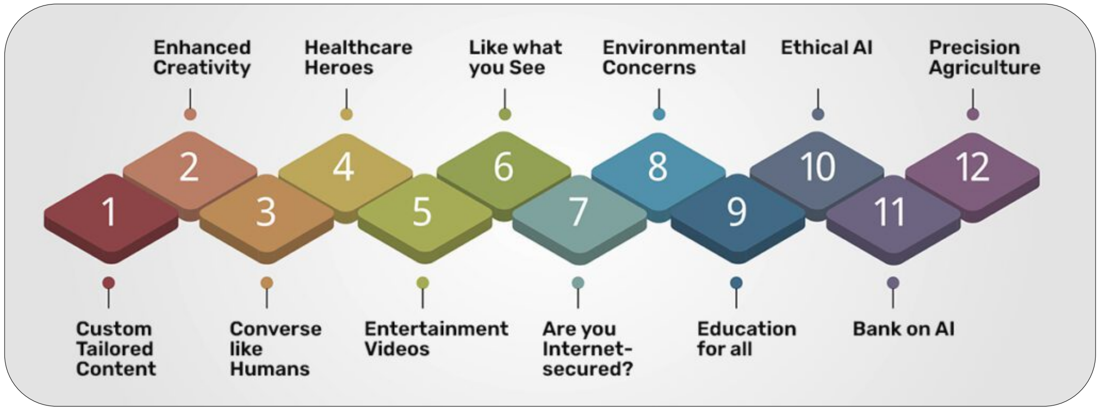
- model parameters, $e.g.$, for Llama2

$$\theta \in \mathbf{R}^{70,000,000,000}$$

Current Trend & Future Perspectives

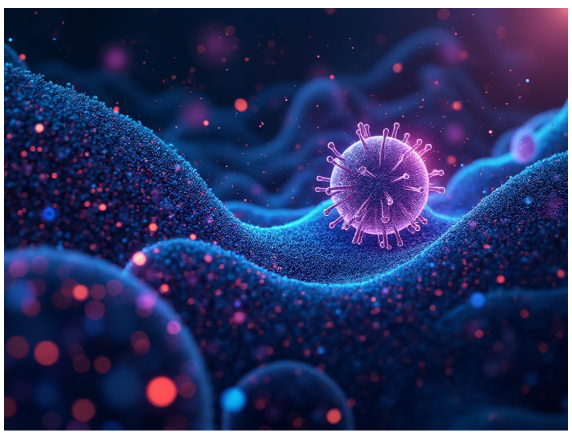
Current trend of genAI

- rapid advancement in language models & multimodal AI capabilities
- rise of AI-assisted creativity & productivity tools
- growing adoption across industries
  - creative industries - design, entertainment, marketing, software development
  - life sciences - healthcare, medical, biotech
- infrastructure & accessibility, *e.g.*, Hugging Face democratizes AI development
- integration with cloud platforms & enterprise-level tools
- increased focus on AI ethics & responsible development



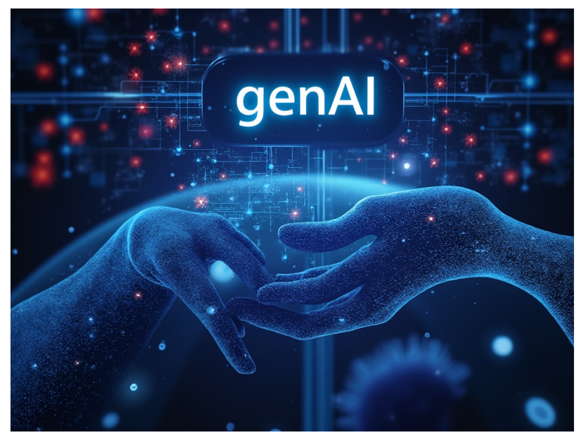
Industry & business impacts

- how genAI is transforming industries
  - creative industries - content creation - advertising, gaming, film
  - life science - enhance research, drug discovery & personalized treatments
  - finance - automating document generation, risk modeling & fraud detection
  - manufacturing & Design - rapid prototyping, 3D modeling & optimization
  - business operations - automate routine tasks to boost productivity



Future perspectives of genAI

- hyper-personalization - highly personalized content for individual users - music, products & services
- AI ethics & governance - concerns over deepfakes, misinformation & bias
- interdisciplinary synergies - integration with other fields such as quantum computing, neuroscience & robotics
- human-AI collaboration - augment human creativity rather than replace it
- energy efficiency - have to figure out how to dramatically reduce power consumption



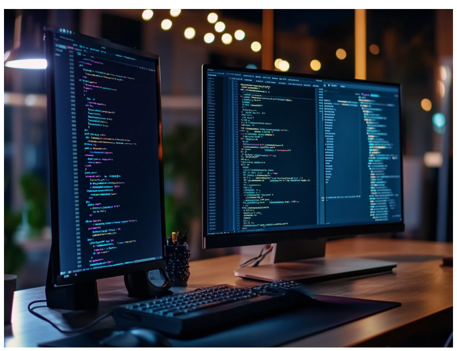
AI Products

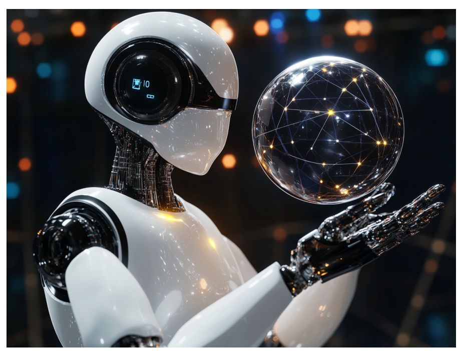
AI product development - trend and characteristics

- *rapid pace* of innovation - new AI models & products being released at unprecedented rate, improvements coming in weeks or months (rather than years)
- *LLMs dominating* - models like GPT-4 & Claude pushing boundaries in NLP & genAI
- *multimodal AI* gaining traction - models processing & generating text, images & even video becoming more common, *e.g.*, Grok, GPT-4, Gemini w/ vision capabilities
- *open-source* AI movement - growing trend of open-source AI models and tools, challenging dominance of proprietary systems
- *AI integration in everyday products* - from smartphones to home appliances, AI being integrated into wide array of consumer products

AI product development - trend and characteristics

- *ethical AI & regulatory focus* - increased attention on ethical implications of AI & calls for regulation of AI development and deployment
- AI in enterprise - businesses across industries rapidly adopting AI for various applications
- *specialized AI models* - development of AI models tailored for specific industries or tasks, $e.g.$, healthcare, biotech, financial analysis
- AI-assisted *coding and development* - help software developers write code more efficiently & tools becoming increasingly sophisticated
- *concerns about AI safety & existential risk* - growing debate about potential short & long-term risks of advanced AI





LLM products

- OpenAI - ChatGPT 4o, GPT-4 Turbo Canvas
- Anthropic - Claude 3.5 Sonnet (with Artifacts), Claude 3 Opus, Claude 3 Haiku
- Mistral AI - Mistral 7B, Mistral Large 2, Mistral Small `xx.xx`, Mistral Nemo (12B)
- Google - Gemini (w/ 1.5 Flash), Gemini Advanced (w/ 1.5 Pro)
- X - Grok [mini] [w/ Fun Mode]
- Perplexity AI - Perplexity [Pro] - combines GPT-4, Claude 3.5, and Llama 3
- Liquid AI - Liquid-40B, Liquid-3B (running on small devices)

flying cats generated by Grok, ChatGPT 4o & Gemini

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
GPT-4	OpenAI	web & books	170B	advanced reasoning & multimodal capabilities	high computational resources
LLaMA-2	Meta	public info & research articles	7∼70B	open access & good performance for different sizes	not powerful for complex tasks
Claude	Anthropic	mix of high-quality datasets	not disclosed	safety-first approach avoiding harmful responses	limited in publicly available details
PaLM 2	Google	multilingual text corpus	540B	high multilingual comprehension supporting various downstream apps	significant resources & not versatile in some contexts

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
BLOOM	BigScience Community	diverse multilingual corpus	176B	open & support multiple languages	resource-intensive & lower performance
Mistral[4]	Mistral AI	public web data	7∼13B	lower parameter count	limited scalability for specialized apps
Liquid Foundation Model (LFM)	Liquid AI	adaptive datasets	adaptive & dynamic parameters	modular & support more specialized fine-tuning for niche use-cases & adaptable in deployment	complexity in design and implementation

Multimodal genAI products

- DALL-E by OpenAI
  - *generate unique and detailed images based on textual descriptions*
  - understanding context and relationships between words
- Midjourney by Midjourney
  - let people *create imaginative artistic images*
  - can interactively guide the generative process, providing high-level directions

Multimodal genAI products

- Dream Studio by Stability AI
  - *analyze patterns in music data & generates novel compositions*
  - musicians can explore new ideas and enhance their *creative* processes
- Runway by Runway AI
  - *realistic images, manipulate photos, create 3D models & automate filmmaking*

Rise of co-pilot products

- definition - AI-powered tools designed to enhance human productivity across multiple domains including document creation, presentations & coding
- benefits
  - *efficiency* - automate repetitive tasks allowing users to focus on high-value activities
  - *error reduction* - minimize mistakes common in manual work
  - *creativity* - suggestions and prompts help users explore new ideas and approaches
  - *integration* with major productivity suites - Microsoft 365, Google Workspace
- popular products
  - GitHub Copilot, Microsoft 365 Copilot, Grammarly AI, Visual Studio Code Extensions

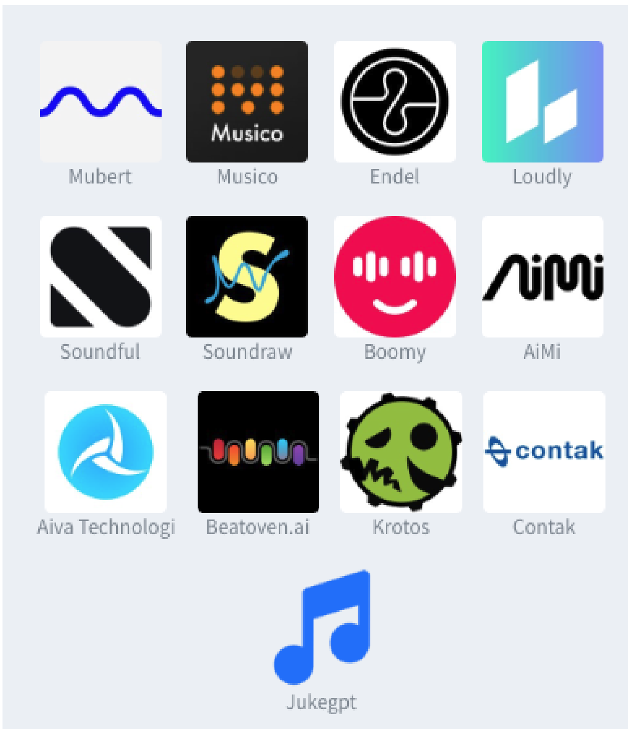
Future of co-pilot products

- potential advancements
  - wider adoption across industries and professions
  - *real-time fully automated collaboration*, *predictive content generation*, personalization
- impact on work environments & creative processes
  - *collaborative human-AI relationships* with augmented reality
  - unprecedented levels of problem-solving due to *augmented cognitive abilities*
- challenges & considerations
  - *ethical concerns around data privacy & AI decision-making*
  - potential impact on *human skills & job markets*

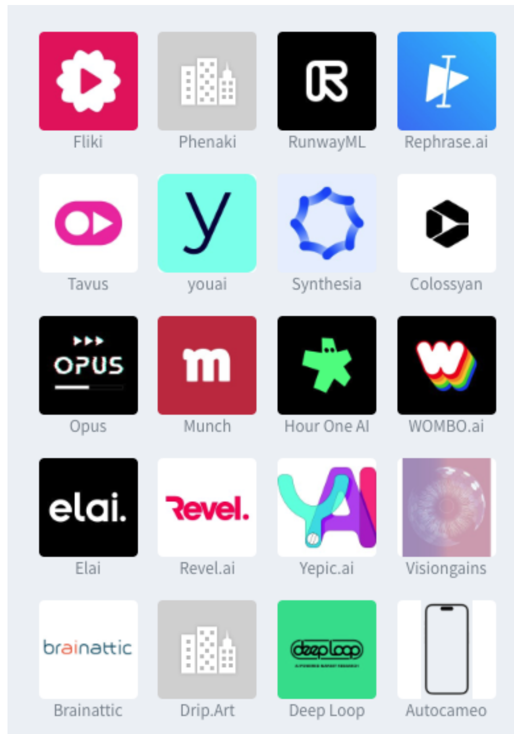


Other AI products - audio/video/text

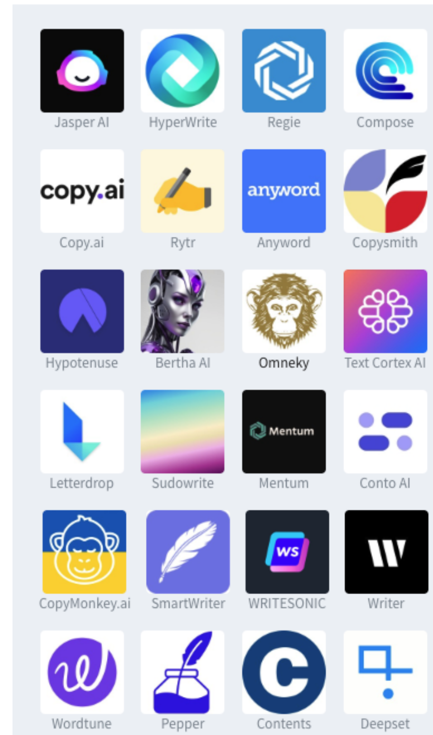
audio



vidio

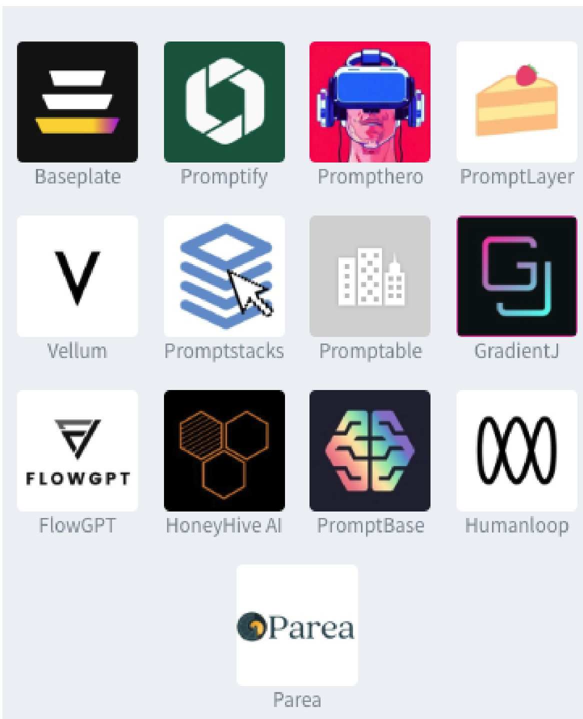


text

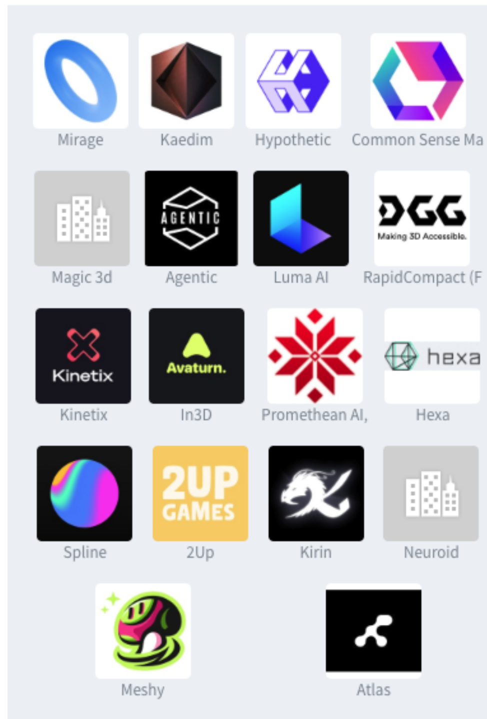


Other AI products - LLM/gaming/design/coding

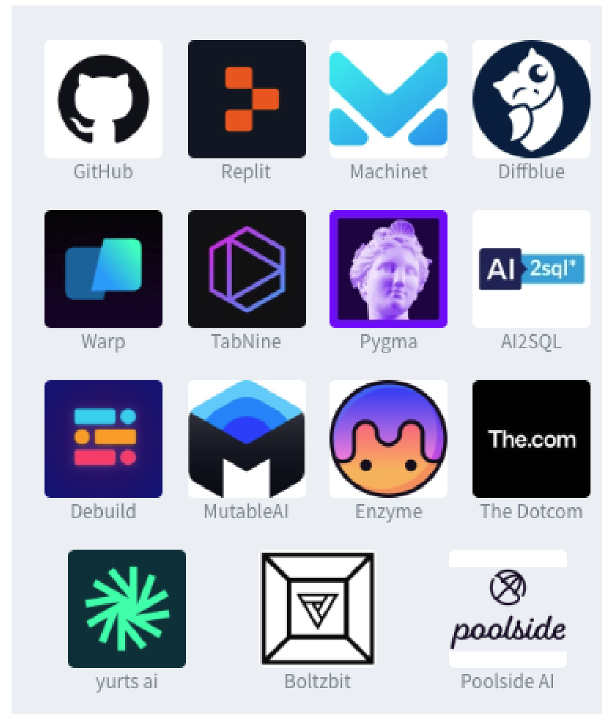
LLM



gaming & design



coding



AI Market & Values

AI market

- PwC, one of "big four" accounting firms, believes
  - *AI can add \$15.7 trillion to the global economy by 2030*

Cloud stacks

- SaaS dominates cloud stack - account for 40% of total cloud stack market with estimated TAM of $260B
- IaaS and PaaS significant players
- semi-cloud's niche presence

cloud stack	companies	estimated TAM	% total in stack
SaaS apps	Salesforce, Adobe	$260B	40%
PaaS	Confluent, snowflake	$140B	22%
IaaS	AWS, Azure, GCP	$200B	30%
cloud semis	AMD, Intel	$50B	8%

AI stacks

- AI investment landscape - AI sector witnessing significant capital inflow with total funding of approximately $29 billion across various segments
- models lead pack - AI models, particularly those developed by OpenAI and Anthropic, attracted lion's share of investments, accounting for 60% of total funding
- diverse growth - while models dominate funding, other segments like apps, AI cloud, and AI semis also experiencing substantial growth, indicating broadening AI ecosystem

AI stack	companies	total funding	% total in stack
apps	character.io, replit	∼$5B	17%
models	openAI, ANTHROP\C	∼$17B	60%
AIops	Hugging Face, Weights & Biases	∼$1B	4%
AI cloud	databricks, Lambda	∼$4B	13%
AI semis	cerebras, SambaNova	∼$2B	6%

AI model companies

- AI model companies - competing for which AI model companies will dominate 2020s
- venture funding surge - private AI model companies raised approximately $17B since 2020, indicating strong investor confidence
- growing open-source presence - becoming increasingly prevalent, adding competition and innovation to AI landscape
- key players - notable companies in AI model space include Adept, OpenAI, Anthropic, Imbue, Inflection, Cohere, and Aleph Alpha
- outcome uncertain - future success is still to be determined, reflecting dynamic and evolving nature of AI industry

AI advancing much faster

- rapid AI advancement - general AI projected to progress from basic content generation to superhuman reasoning in only 5 years
- significantly outpacing 15-year timeline for fully autonomous vehicles

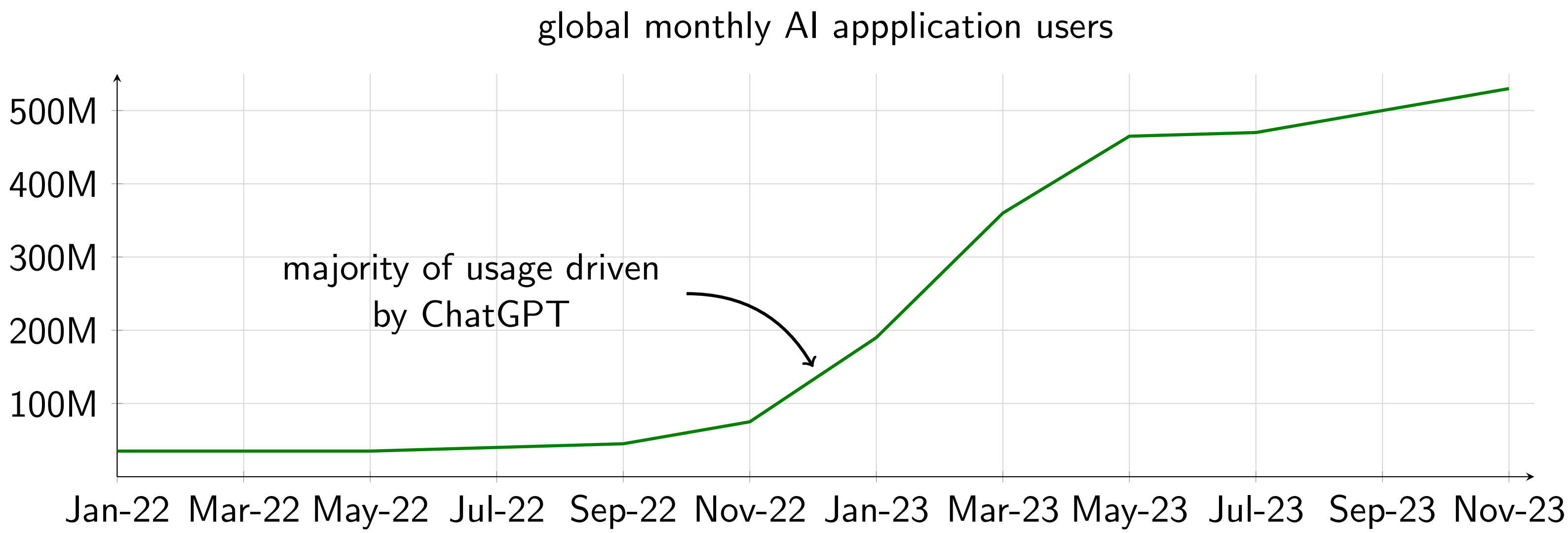
autonomy level	autonomous vehicles	genAI
L5	fullly autonomous	superhuman reasoning & perception
L4	highly autonomous	AI autopilot for complex tasks
L3	self-driving with light intervention	AI co-pilot for skilled labor
L2	Tesla autopilot	supporting humans with basic tasks
L1	cruise control	generating basic content

15 yrs

5 yrs

AI interest of users

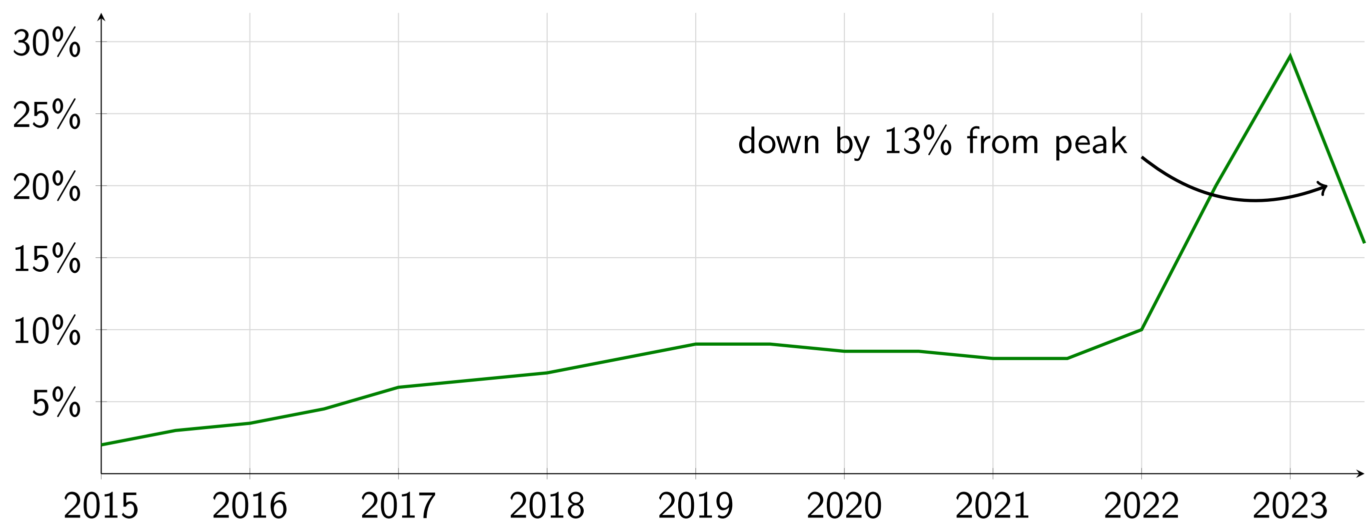
- AI adoption approaching saturation - initial wave may be nearing saturation
- future growth might come from deeper integration into professional workflows & specialized applications
- potential for market diversification - ChatGPT drove majority of early growth, but now we have other LLMs - Claude, Mistral, Gemini, Grok, Perplexity



AI interest of developers

- rising popularity - portion of new GitHub stars given to AI/ML repositories steadily increased from 2015 to 2022
- excitement waning & washing out AI "tourists" - decline of 13% from peak in 2022
- could indicate potential factors such as market saturation, economic conditions, or shifts in developer preferences

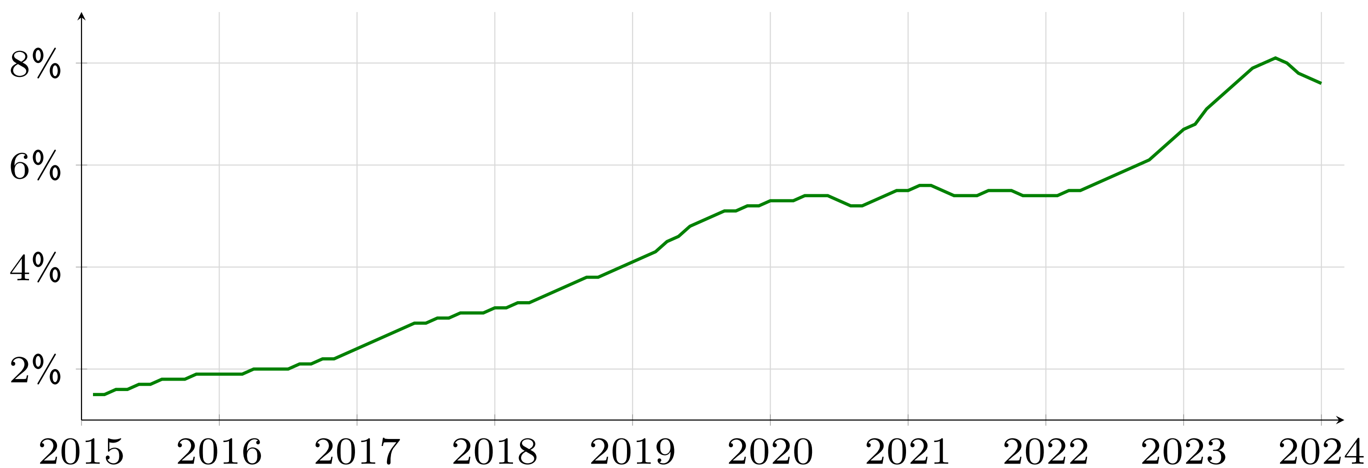
portion of new GitHub stars given to AI/ML repos



Developers' contribution to software packages

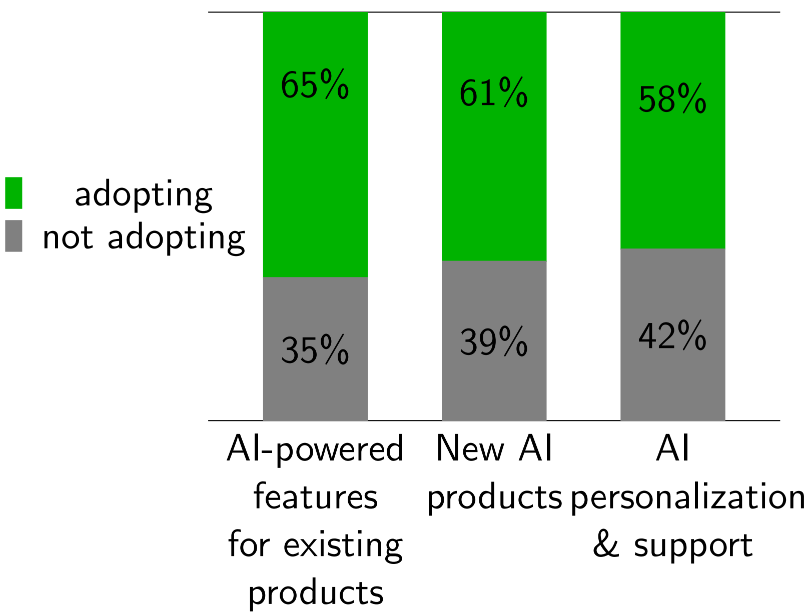
- steep acceleration from 2022 to 2024 correlates with explosion of LLMs & genAI
- suggesting transformative shift in AI landscape beyond gradual growth
- AI/ML still represents relatively small portion (less than 10%)
- indicating significant room for growth and mainstream adoption across various software domains

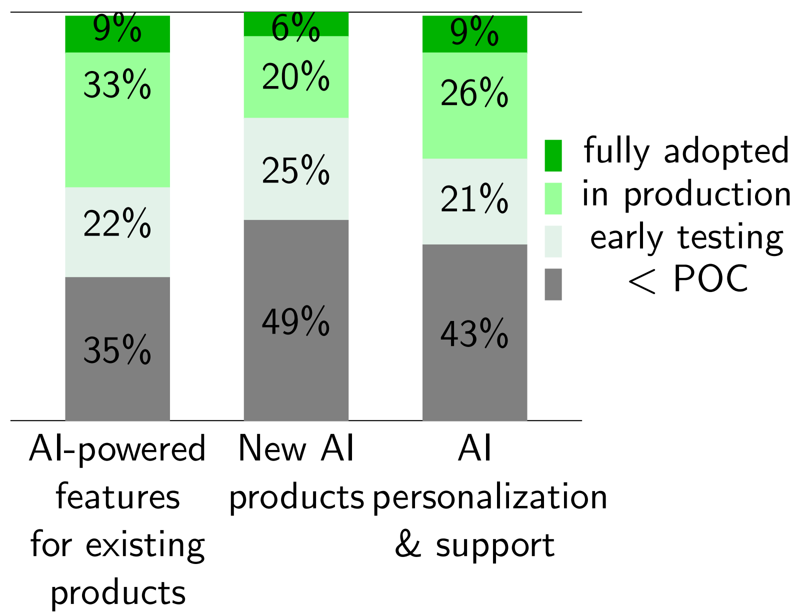
portion of AI/ML GitHub repo commits



Enterprises adoptiong AI

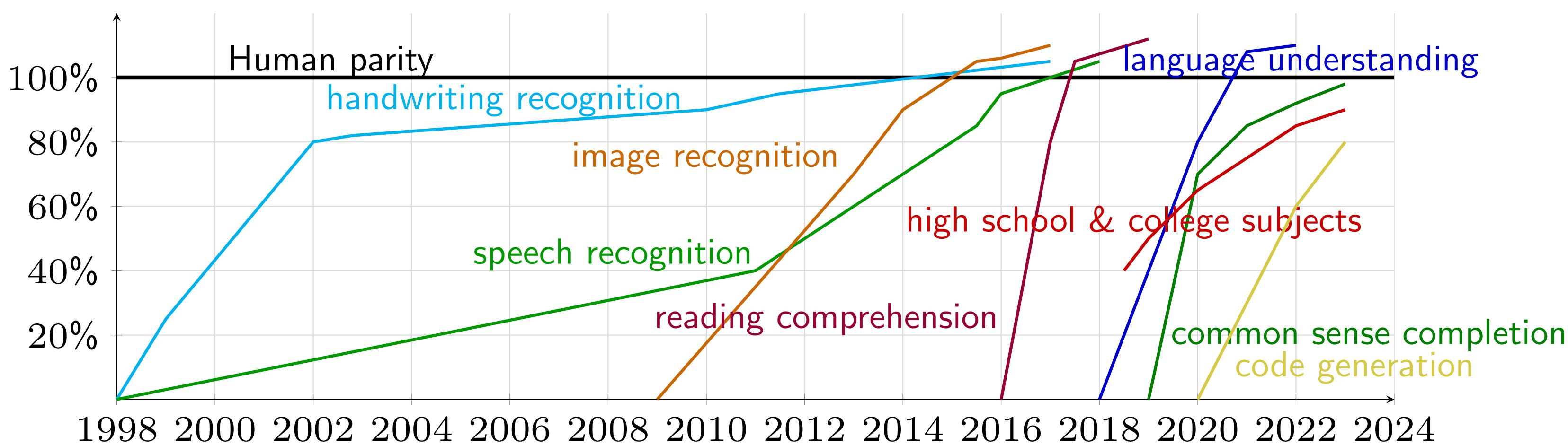
- more than 60% of enterprises planning to adopt AI
- full adoption rate is less than 10% - will take long time





AI getting better and faster

- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached human parity, while others having still way to go
  - achieving truly human-like capabilities in broad range remains a challenge



AI delivers game-changing values

- time developers save using GitHub Copilot - *55%*
  - *10M+* cumulative downloads as of 2024 & *1.3M* paid subscribers - *30%* Q2Q increase
  - improves developer productivity by *30%+*
- reduction in human-answered customer support requests - *45%*
  - cost per support interaction - *95%* save / $2.58 (human) vs $0.13 (AI)
  - median response time - *44 min* faster / 45 min (human) vs 1 min (AI)
  - median customer satisfaction - *14%* higher / 55% (human) vs 69% (AI)
- time saved from editing video in runway - *90%*
- AI chat rated higher quality compared to physician responses - *79%*

AI Industry

Heavy Lifting of LLMs

News - OpenAI's "$8.5B bills" report sparks bankruptcy speculation

- OpenAI's financial situation reflects its ambitious vision
  - projected $8.5B expenses vs $3.5–4.5B revenue in 2024 w/ massive investment in AI infrastructure and talent
- caused by Sam Altman's reckless & non-strategic commitment to AGI development
  - "Whether we burn $500M, $5B, or $50B a year, I don't care..." - prioritizing long-term impact over short-term profitability
- reflect broader AI industry trend of high burn rates
  - indicative of the resource-intensive nature of cutting-edge AI research

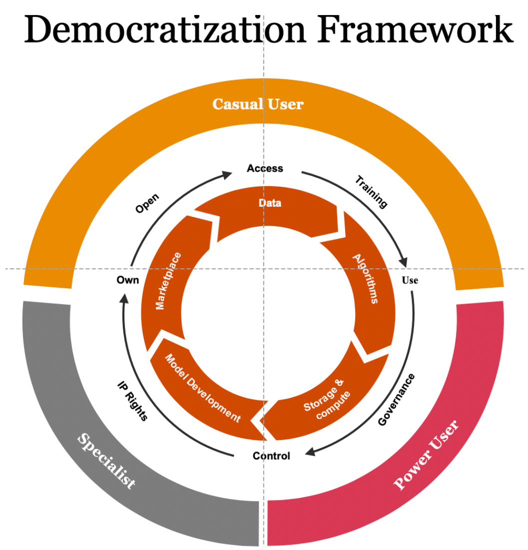
LLM - strategic challenges & industry dynamics

- evolving competitive landscape
  - threat from open-source models ($e.g.$, Meta's Llama 3.1) & potential commoditization of LLMs
- balancing act with Microsoft partnership
  - critical financial support vs maintaining independence - Microsoft's $13B investment provides both opportunity and constraint
- sustainability of current business model
  - high costs of AI development vs monetization challenges
  - need for breakthrough applications or efficiency improvements
- ethical & regulatory considerations
  - balancing rapid development with responsible AI principles
  - potential impact of future AI regulations on operations and costs

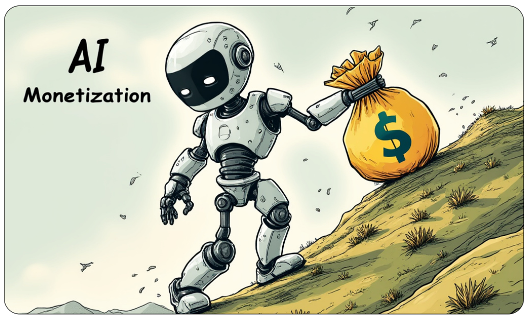
Industry disruption of open-source AI models on industry

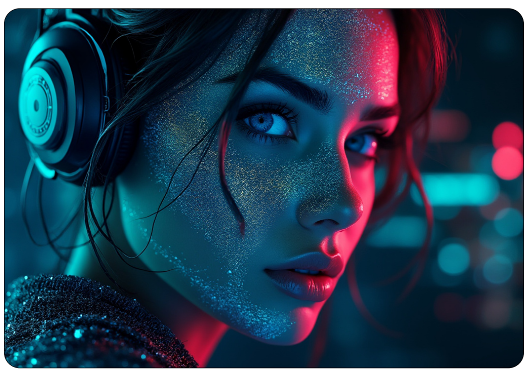
- rise of open-source models such as Meta's Llama 3.1 reshaping the AI landscape

- industry disruption
  - AI democratization - open-source making advanced AI capabilities accessible to wider range of developers and companies
  - innovation acceleratation - collaborative improvement of open-source models could lead to faster progress
  - pressure on proprietary models - companies like OpenAI may need to offer significant advantages over free alternatives to justify their costs



Impact of open-source AI models on industry





- business model challenges
  - monetization difficulties - capable models becoming freely available
  - shift to services & applications - focus may move from selling access to models to providing *specialized services* or *applications built on top of them*
- ethical & security concerns
  - responsible AI - open-source models raise questions about control and responsible use
  - dual-use potential - wider access to powerful AI models could increase risks of misuse or malicious applications, $e.g.$, *Deepfake*

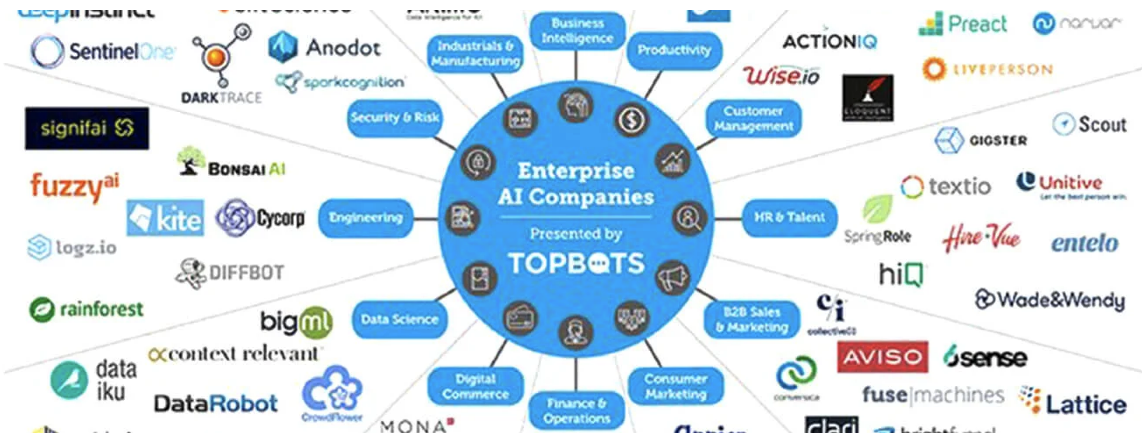
Tech Giants & AI Companies

Evolving relationship between tech giants & AI companies

- partnership between OpenAI & Microsoft exemplifies broader trend of collaboration & integration in AI industry
- symbiotic relationships
  - tech giants provide :esources & funding - AI companies research & innovation
  - provide AI companies w/ instant access to large user bases & distribution channels
- power dynamics
  - independence concerns - AI companies' risk of losing autonomy
  - tech giants' access to advanced AI potentially widening gap with smaller competitors

AI industry consolidation

- mergers & acquisitions
  - will see increased M&A activities as tech giants seek to bring AI capabilities in-house
- ecosystem development
  - tech giants creating AI-focused ecosystems, similar to cloud services, to attract and retain developers & businesses



AI Startups

Incubators

- Y Combinator
  - invests $500,000 in each startup receiving 7% equity
  - program culminates in "Demo Day" where startups present to investors
  - helped Airbnb, Dropbox, DoorDash, Reddit, Stripe, . . .
  - highly competitive - acceptance rate of 1.5% – 2% - 10k applicants per six months
  - significantly higher survival rate - around 18% valued at over $100M and 4% becoming unicorns—valued at over $1B
- XXX

We're not talking about AI chip startups enough

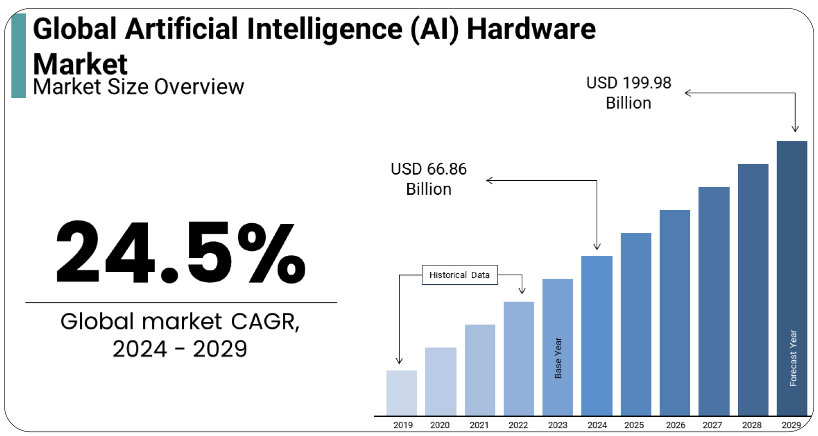
- *not* talking about AI chip startups enough
  - Etched - specializes in AI accelerator chips for LLMs - Sohu chip embeds transformer architecture into silicon
  - Tenstorrent - develops high-performance AI processors focused on efficient training and inference
  - Axelera AI - specializes in high-performance AI hardware for edge computing, excelling in CV applications
  - Recogni - develops high-performance AI chips for edge computing applications focusing on low-latency processing for autonomous vehicles
- *global AI chip market will gross \$30B in 2024 and become \$300B in next 10 years*

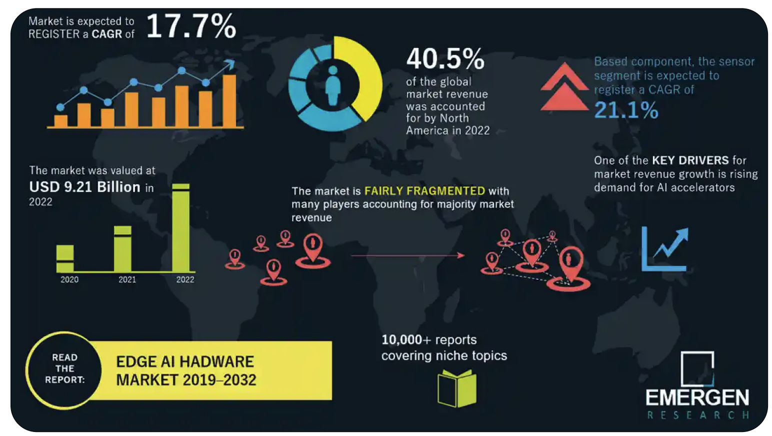
AI Hardware

AI Hardware Industry

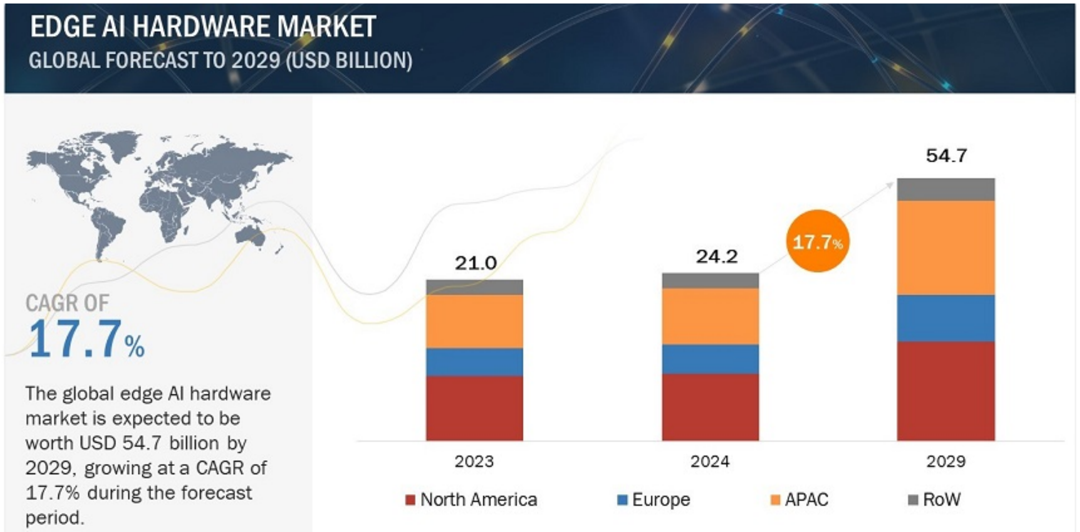
Landscape of AI hardware industry

- global AI hardware market valued at $66.96B in 2024, projected to grow significantly
- major companies - Nvidia, Intel, AMD, Qualcomm, and IBM w/ Nvidia holding substantial market share



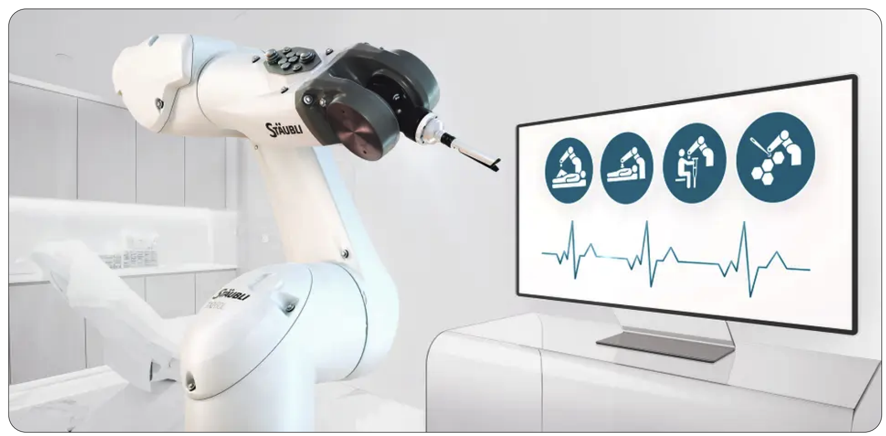


- North America leading market - high R&D investments & key industry players
- Asia Pacific rapidly expanding - strong semiconductor industries in South Korea, China & Japan
- demand for advanced processors such as GPUs, TPUs & AI accelerators rising due to complexity of AI algorithms & high computational power



Predictions for future of AI hardware market

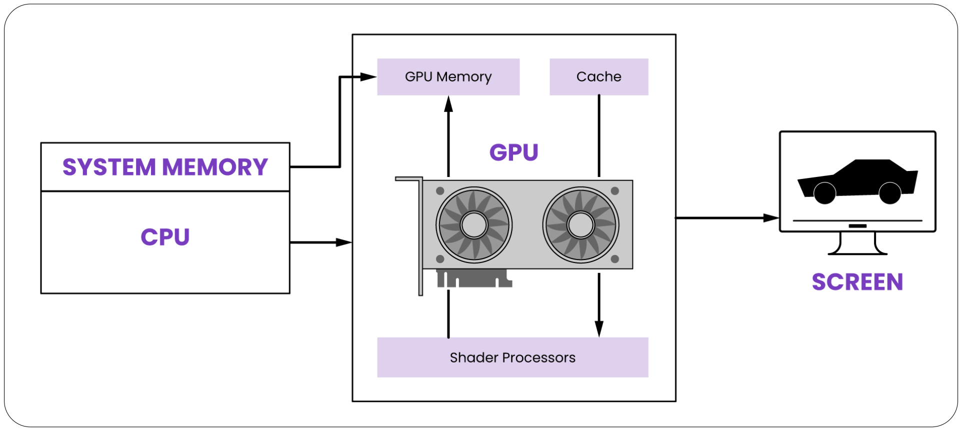
- AI hardware market expected to reach \$382B by 2032 - significant growth in data center AI chips
- integration of AI w/ 5G & increased use of AI in edge computing anticipated to drive future demand
- AI hardware becoming crucial in sectors such as autonomous vehicles, robotics & medical devices
- need to address challenges such as heat and power management along with technical complexities



GPUs and AI Accelerators

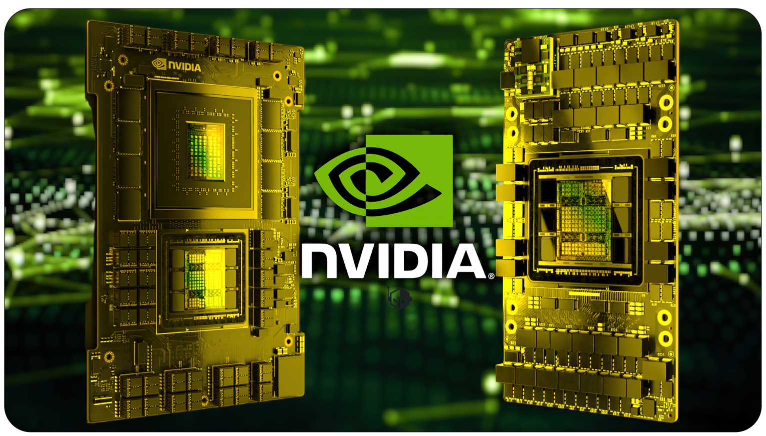
Technical challenges of GPUs & AI accelerators

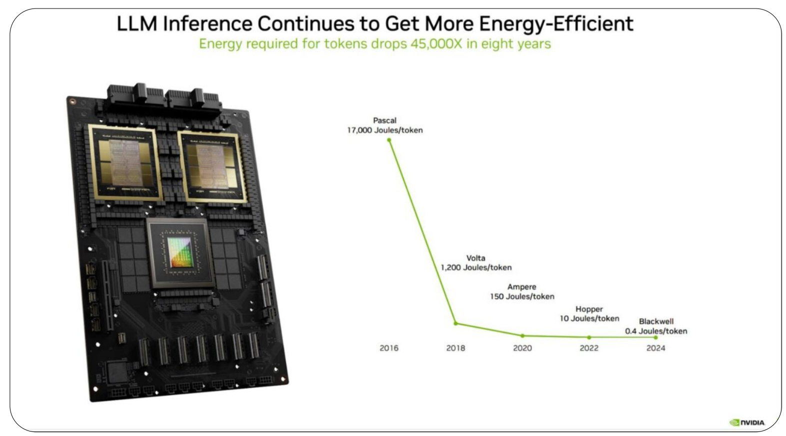
- facing challenges in scaling to handle increasingly large AI models and datasets - traditional architectures struggling w/ massive parallel processing demands of modern AI applications
- AI applications require extensive memory bandwidth often leading to bottlenecks - efficient memory management is crucial
- AI accelerators consume significant power - high operational costs and environmental concerns for both cloud-based & edge AI applications



Potential solutions for overcoming challenges

- development of AI-specific architectures such as tensor cores and custom ASICs to improve efficiency and performance - novel architectures like FPGAs for specific AI tasks, $e.g.$, for RAG & vectorDB
- implementing software optimizations to enhance hardware usability and performance - use of compilers and frameworks that maximize efficiency of existing hardware
- encouraging market competition to drive innovation and reduce monopolistic control - exploring alternative hardware solutions and improving energy efficiency standards





Big tech's in-house chip development

- shift towards in-house AI hardware - major tech companies increasingly developing their own AI chips - move to enhance AI capabilities and reduce dependence

- collaboration with specialized partners - partnering with specialized firms for manufacturing and technology blending in-house expertise with external innovation

	Microsoft	Google	Amazon	Meta
Chip	Maia 100	TPU v5e	Inferentia2	MTIA v1
Launch Date	November, 2023	August, 2023	Early 2023	2025
IP	ARM	ARM	ARM	RISC-V
Process Technology	TSMC 5nm	TSMC 5nm	TSMC 7nm	TSMC 7nm
Transistor Count	105 billion	-	-	-
INT8	-	393 TOPS	-	102.4 TOPS
FP16	-	-	-	51.2 TFLOPS
BF16	-	197 TFLOPS	-	-
Memory	-	-	-	LPDDR5
TDP	-	-	-	25W
Packaging Technology	CoWoS	CoWoS	CoWoS-S	2D
Collaborating Partners	Global Unichip Corp.	Broadcom	Alchip Technologies	Andes Technology
Application	Training/Inference	Inference	Inference	Training/Inference
LLM	GPT-3.5, GPT-4	BERT, PaLM, LaMDA	Titan FM	Llama, Llama2

AMD - Nvidia's new competitor

- key points
  - AMD launched new AI accelerator chip, *Instinct MI300X*, on Dec 6, 2023
  - CDNA 3 architecture, mix of 5nm and 6nm IPs, delivering 153B transistors
  - *outperforms Nvidia's H100 TensorRT-LLM* by 1.6X higher memory bandwidth and 1.3X FP16 TFLOPS
  - up to 40% faster vs Nvidia's Llama-2 70B model in 8x8 server configurations
- market impact
  - significant challenge to Nvidia's dominance in AI accelerator market
  - performance gains over Nvidia's offerings could drive *customer adoption and market share for AMD*
- future prediction
  - *AMD stocks soared* since launch indicating investor confidence in their competitiveness
  - Lisa Su, AMD's CEO, categorized Instinct MI300X as "next big thing" in tech industry
  - potential risks include need to *manage ROCm vs CUDA software ecosystem* & ensure rapid customer adoption and production coverage

AI Accelerator Startups

AI accelerator startups

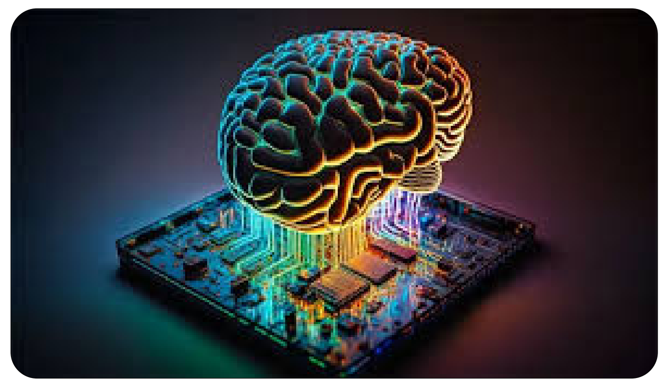
- innovative architectures - startups like Groq, SambaNova & Graphcore leading with *novel architectures designed to accelerate AI workloads*
  - *Groq* - tensor streaming processor (TSP) offering ultra-low latency & high throughput, high-performance AI inference chips enhancing speed & efficiency
  - *SambaNova* - reconfigurable dataflow architecture optimizing for various AI workloads
  - *Graphcore* - intelligence processing unit (IPU) tailored for graph-based computation excelling in sparse data processing
  - *Cerebras Systems* - develop wafer scale engine (WSE), largest chip built for AI workloads, unmatched computational power revolutionizing AI hardware capabilities
  - *Hailo* - specialize for edge devices optimizing AI processes for real-time applications, raised $120M emphasizing potential to disrupt traditional AI chip markets

Technological competitiveness

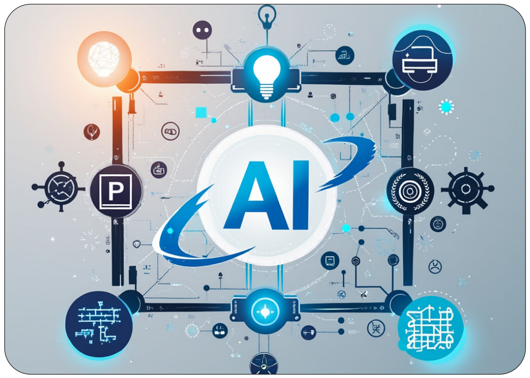
- energy efficiency
  - energy-efficient designs crucial for scalability in data centers and edge devices
  - startups developing solutions significantly reducing power consumption without compromising performance
- customization & flexibility
  - AI accelerators from startups often offer greater customization options for specific AI tasks compared to traditional GPUs
  - flexibility in hardware allows for tailored solutions that can outperform general-purpose accelerators in certain applications
- software integration
  - robust software ecosystems critical - startups investing in developing software stacks that optimize performance for their hardware
  - compatibility with existing AI frameworks is competitive advantage, *e.g.*, `TensorFlow` & `PyTorch`

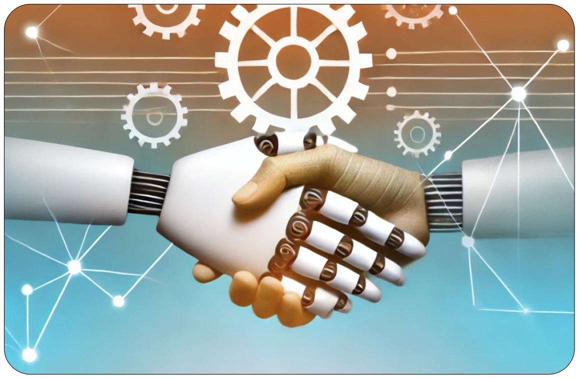
Industry and market influence

- disruption of traditional players
  - challenging dominance of established players like NVIDIA & Intel
  - unique architectures providing specialized solutions traditional GPUs and CPUs cannot efficiently handle
- driving down costs
  - offering competitive alternatives pushing down cost of AI computation
  - could lead to democratization of AI w/ more companies affording high-performance AI capabilities

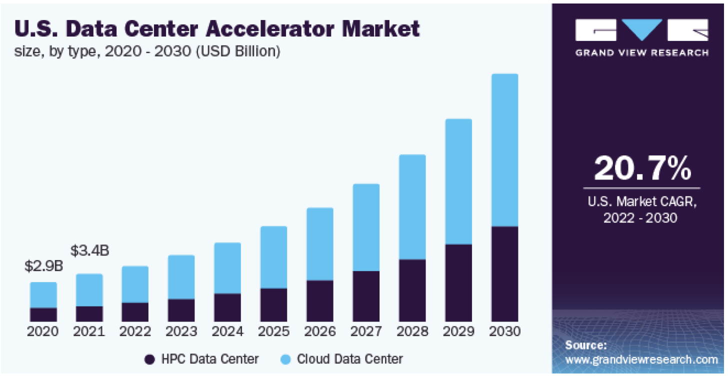


- accelerating AI innovation
  - contributing to rapid innovation providing hardware that can handle emerging AI models & workloads
  - adaptability and specialization enable advancements in AI research & faster development cycles
- strategic partnerships & acquisitions
  - big techs increasingly forming strategic partnerships or acquiring startups to stay competitive
  - collaborations can speed up integration of advanced AI hardware into mainstream products





- market growth & opportunities
  - AI accelerator market expected to grow significantly driven by demand in data centers, edge computing & autonomous systems
  - startups well-positioned to capture significant share of growing market particularly in niche applications
- future outlook
  - dependency on Asia for fabrication might lead to strategic shifts in global tech policies and investments in local manufacturing
  - increasing demand for efficient AI processing on edge devices and in data center.



Global Semiconductor Industry

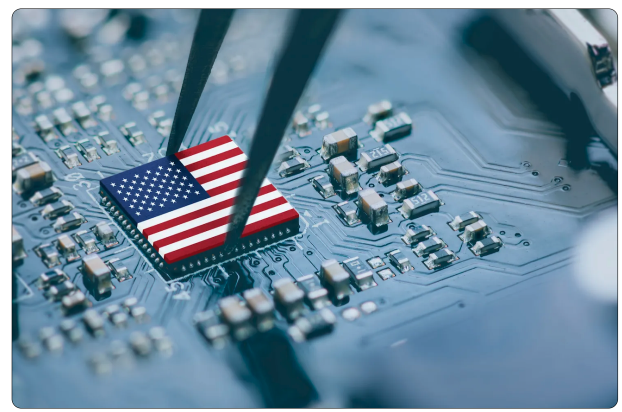
Hard-to-predict AI hardware markets

- US
  - birthplace for modern semiconductor chips driving PC market, internet, multi-media, mobile phones, and AI . . .
    - Intel, Texas Instrument (TI), Global Foundry
  - traditionally strong with design houses - NVIDIA, AMD, Broadcom, Apple, . . .
  - threatened experiencing global chip shortage & vulnerable supply chain via COVID
  - national security concerns & economic competitiveness
- China
  - strong fast followers - SMIC[5], Huawei, Hua Hong Semiconductor (foundry)
- South Korea
  - best memory chip makers - Samsung, SK hynix
  - struggling with LSI and foundry business

[5] SMIC - Semiconductor Manufacturing International Corporation

Reshoring semiconductor manufacturing industry

- trade & semiconductor WAR between US & China
  - export controls on advanced chips and equipment
- CHIPS & Science Act (Aug, 2022)
  - $52B in subsidies for domestic production, 25% investment tax credit for chip plants
  - (coerce) world-best semiconductor manufacturers build factories in US with support
    - GlobalFoundries - $1.5B @ Feb-2024
    - Intel - $8.5B @ Apr-2024 - Ohio - two fabs expandable to $100B
    - Samsung - $6.4B @ Apr-2024 - Talor, Texas
    - TSMC - $6.6B @ Apr-2024 - Phoenix, Arizona - two foundry fabs (3nm & 4nm)



Turmoils in global semiconductor business

- global context
  - EU Chips Act - €43B to boost European chip production
  - Japan & South Korea - significant investments in domestic capacity
- industry dynamics
  - Intel's foundry ambitions - targeting 50% global market share by 2030
  - TSMC expanding global footprint (US, Japan, possibly Germany)
- future outlook
  - projected shift in global semiconductor manufacturing landscape
  - increased geographical diversification of chip production

Export controls on US chip technology to China

- goal - limit China's access to advanced semiconductor tech to maintain US strategic advantage
- impacts on
  - China - advanced chips and equipment not allowed, domestic innovation increased
  - US - short-term - US lose market share and revenue in China
  - US - long-term - potential decline in US global competitiveness
- Chinese response - circumvent controls and adapt supply chains
- conclusion
  - US-China chip rivalry transforms global supply chains with deep implications for *security & industry*
  - US success hinges on better coordination and policy analysis

- reference - Balancing the Ledger - Center for Strategic & International Studies (CSIS)

China strikes back on US sanction

- Huawei's launch of Mate 60 Pro smartphone
  - these domestically produced chips represent major breakthrough against US sanctions
  - its success with *advanced 7nm Kirin 9000S chip* demonstrates significant progress in China's self-reliance in high-tech manufacturing - narrowing the technological gap with global leaders
- Huawei case highlights potential failure of US sanctions potentially leading to more aggressive US measures
  - US export controls on China's semiconductor industry are effective in the short term but insufficient to halt China's progress especially in legacy chip manufacturing
  - to maintain technological edge, US must balance further restrictions with supporting its semiconductor industry to avoid overreliance on export controls

Chinese semiconductor companies

- Chinese major semiconductor companies
  - SMIC - China's largest chip foundry, advancing 7nm technology
  - HiSilicon - Huawei's chip design arm, crucial for the Kirin processors
  - YMTC - leader in 3D NAND memory chip production
  - Huahong Group, CXMT, SMEE, GigaDevice, UnilC Semiconductors, ASMC, *etc.*
- *SMIC shows significant progress in producing 7nm chips* & YMTC leads memory chip manufacturer - both face challenges from US export controls
- industry faces internal challenges, *e.g.*, corruption & misallocation of resources
- but remains crucial to China's goal of technological self-reliance

Serendipities around AIs

Serendipity or inevitability

- What if Geoffrey Hinton had not been persistent researcher?
- What if symbolists won AI race over connectionists?
- What if attention mechanism did not perform well?
- What if Transfomer architecture did not perform super well?
- What if Jensen Hwang had not been crazy about making hardware for professional gamers?

- Is it like Alexander Fleming's Penicillin?
- Or more like Inevitability?

References

References

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[HGH+22] Sue Ellen Haupt, David John Gagne, William W. Hsieh, Vladimir Krasnopolsky, Amy McGovern, Caren Marzban, William Moninger, Valliappa Lakshmanan, Philippe Tissot, and John K. Williams. The history and practice of AI in the environmental sciences. *Bulletin of the American Meteorological Society*, 103(5):E1351 – E1370, 2022.

[KW19] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[KXS+24] Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, and Rakesh Ranjan. Platonerf: 3D reconstruction in Plato's cave via single-view two-bounce lidar, 2024.

[Mil22] Chris Miller. *Chip war: fight for the world's most critical technology*. New York: Scribner, 2022.

[MLZ22] Louis-Philippe Morency, Paul Pu Liang, and Amir Zadeh. Tutorial on multimodal machine learning. In Miguel Ballesteros, Yulia Tsvetkov, and Cecilia O. Alm, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 33–38, Seattle, United States, July 2022. Association for Computational Linguistics.

[SSS+24] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable Gaussian codec avatars, 2024.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[YFZ+24] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[ZBX+24] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion, 2024.