K-PAI Seminar: The AI Knight Rises
From Deep Learning to Flourishing Societies

Sunghee Yun

Co-founder / CTO - AI Technology @ Erudio Bio, Inc.

About Speaker

- *Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio, CA, USA*
- 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*
- *Co-founder / CTO & Chief Applied Scientist @ Gauss Labs, CA, USA – 2023*
- 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* - advised by *Prof. Stephen P. Boyd*
- Principal Engineer @ Memory Design Technology Team
  - AI & optimization - collaborating with *DRAM/NAND Design/Process/Test teams*
- Senior Applied Scientist @ Amazon
  - e-commerce AIs - deep reinforcement learning & recommender system
  - Jeff Bezos's project - *increase sales by $200M* via Mobile Shopping App
- Co-founder / CTO & Chief Applied Scientist @ Gauss Labs
  - *industrial AI - R&D, market & product strategies*
- Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio
  - *biotech - AI technology, business development & product strategy*

Today

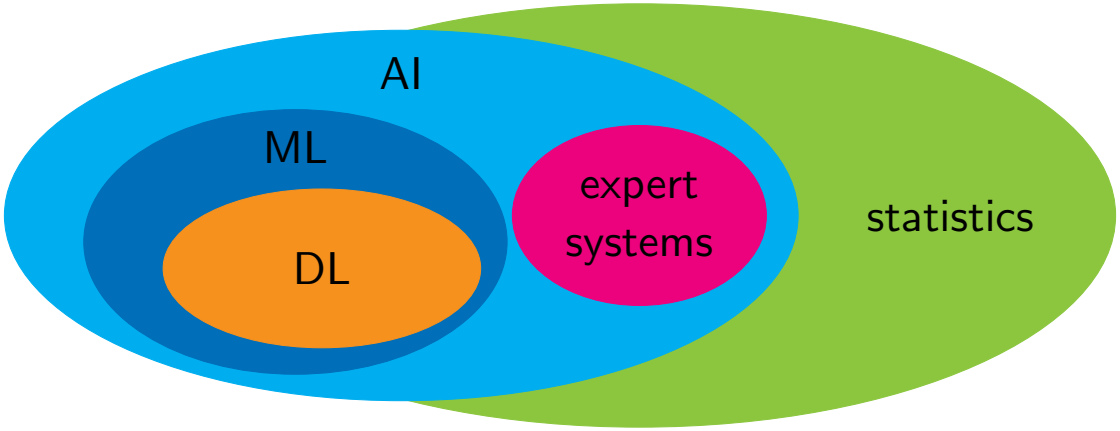
- Artificial Intelligence
  - AI history, recent significant AI achievements, is AI hype?
- Multimodal AI Agents
  - implications of LLM, future of society powered by AI agents
- Empowering Humanity for Future Enriched by AI
  - blessings and curses, KFAS-Salzburg Global Leadership Initiative
  - reclaiming technology for Humanity
- Appendices
  - AI products, serendipities around AIs, some important questions
- Selected references, References

Artificial Intelligence

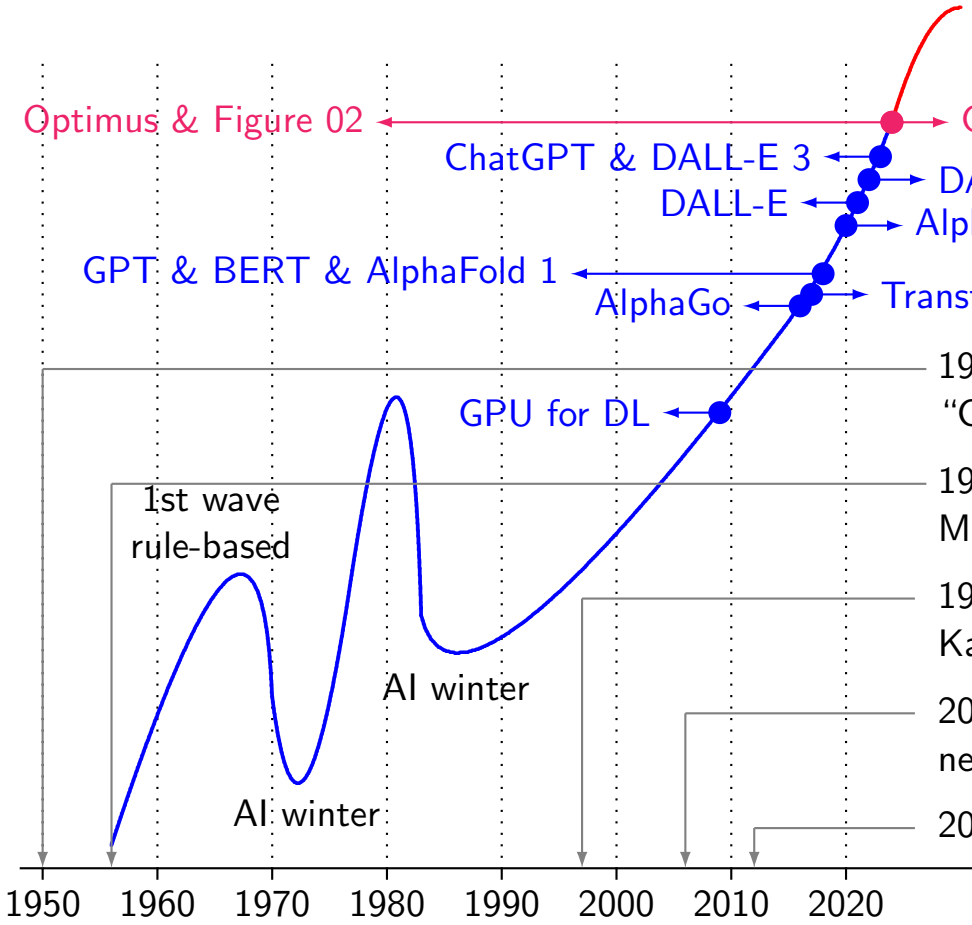
Definition and History

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



1950 - Turing Test introduced by Alan Turing in "Computing Machinery & Intelligence"

1956 - "Artificial Intelligence" coined by John McCarthy

1997 - IBM's Deep Blue defeated Garry Kasparov, (then) the world chess champion

2006 - DL Learning gained traction deep belief networked introduced by Geoffrey Hinton

2012 - AlexNet won ImageNet Competition (ILSVRC) - widespread attention on DL

Significant AI Achievements - 2014 – 2024

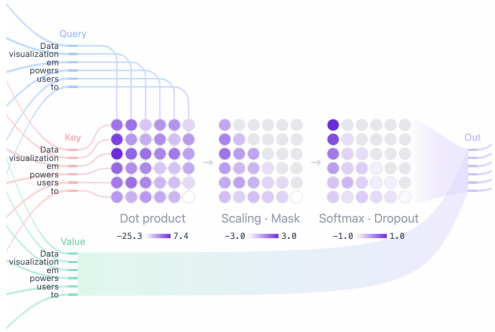
Deep learning revolution

- 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

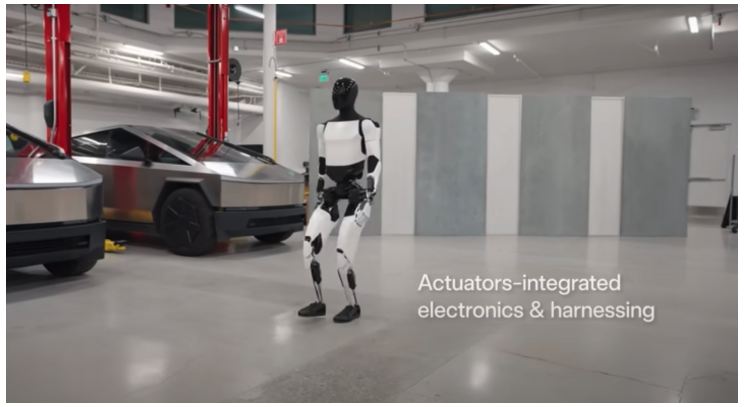
- 2017 – 2018 - Transformers & NLP breakthroughs[2]
  - *Transformer (e.g., BERT & GPT) revolutionized NLP*
  - major advancements in, *e.g.*, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 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

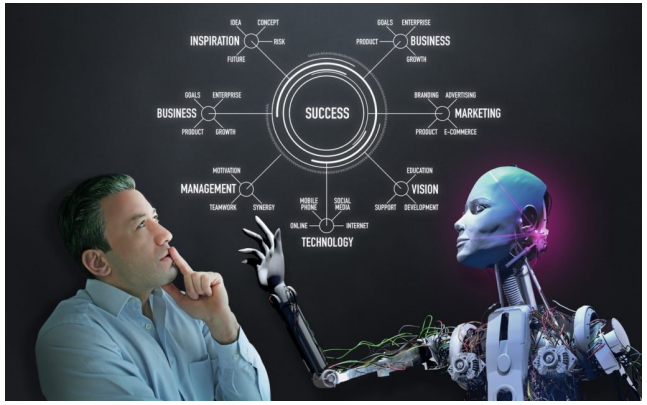
Lots of breakthroughs in AI technology and applications in 2024

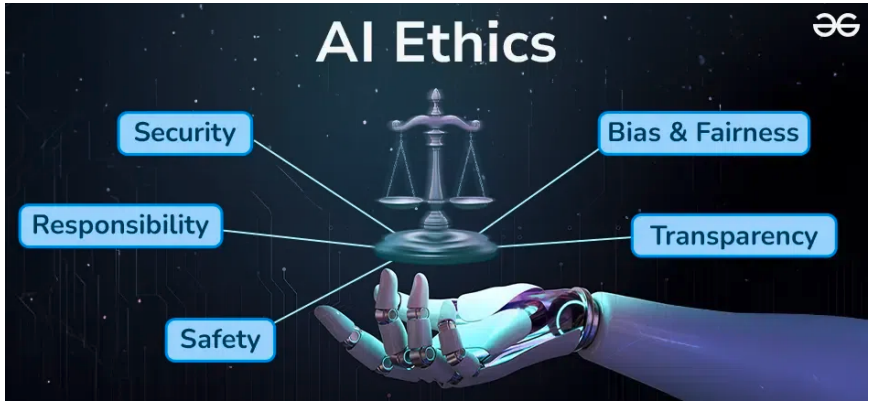
- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Llama 3, Sora
  - *transforming industries* such as content creation, customer service, education, $etc.$
- breakthroughs in specialized AI applications
  - Figure 02, Optimus, AlphaFold 3
  - driving unprecedented advancements in automation, 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*

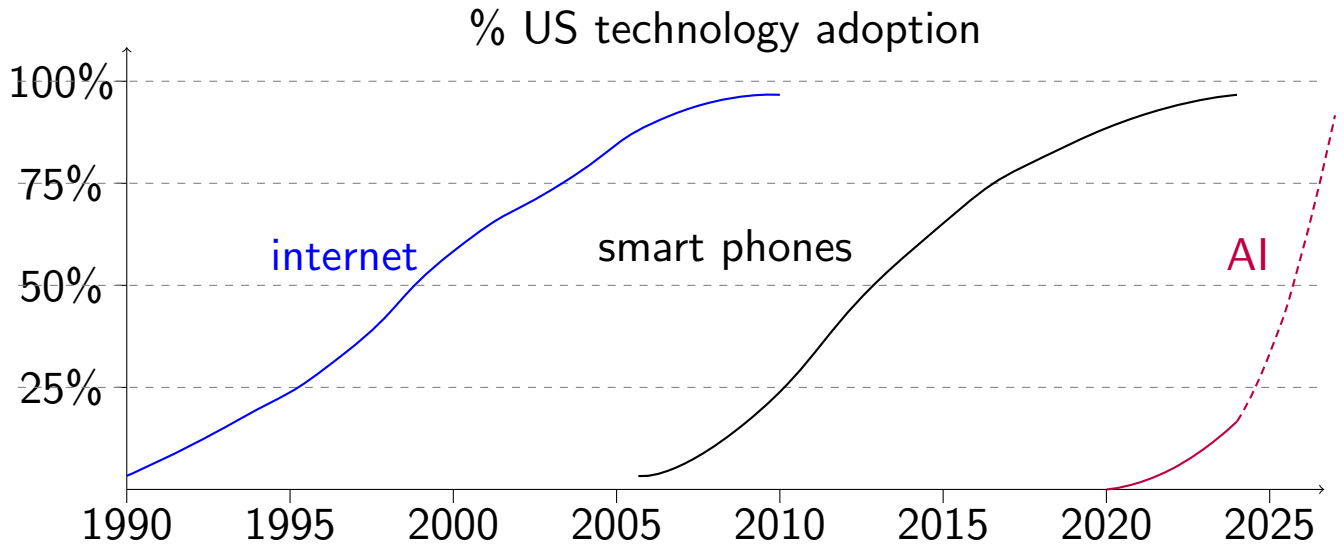




Recent Advances in AI

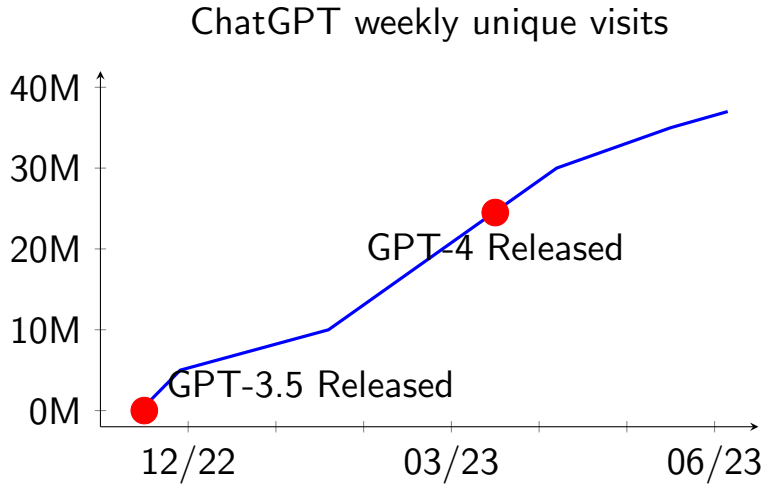
Where are we in AI today?

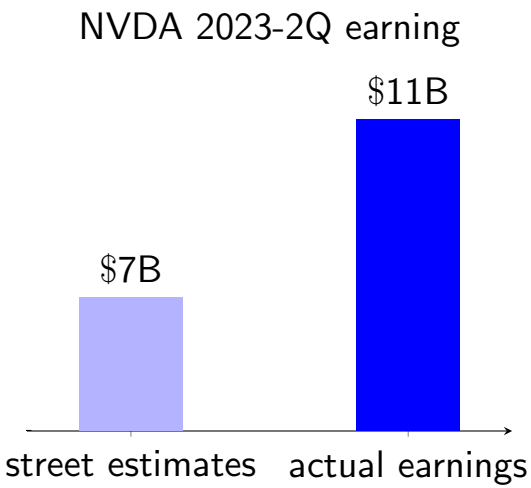
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



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]



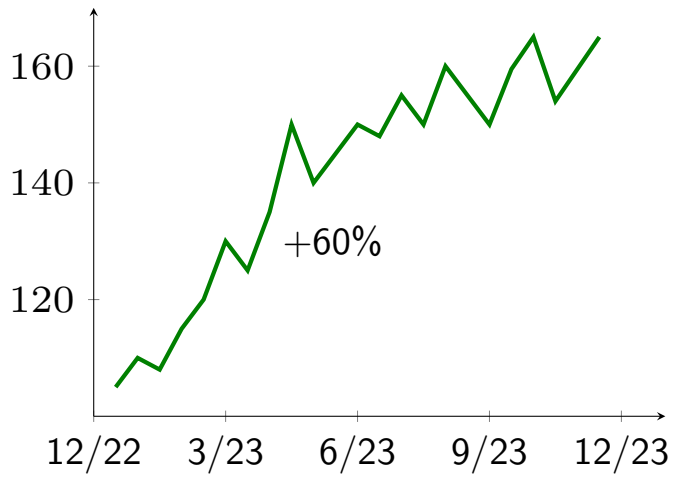


[3] source - Bloomberg

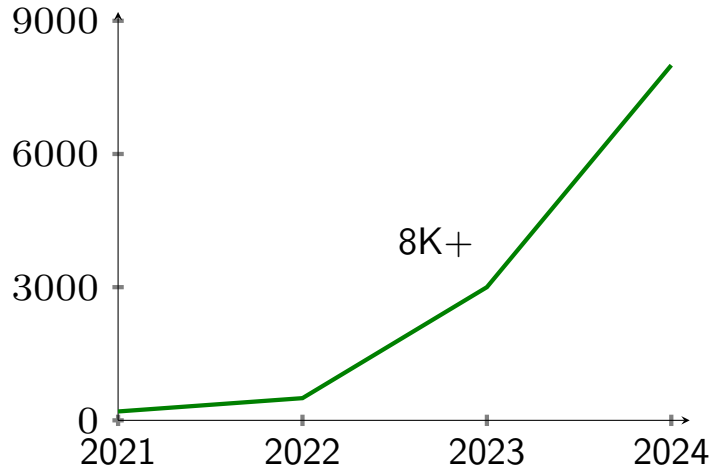
Explosion of AI ecosystems - AI stock market

- *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 basket performance - 2023

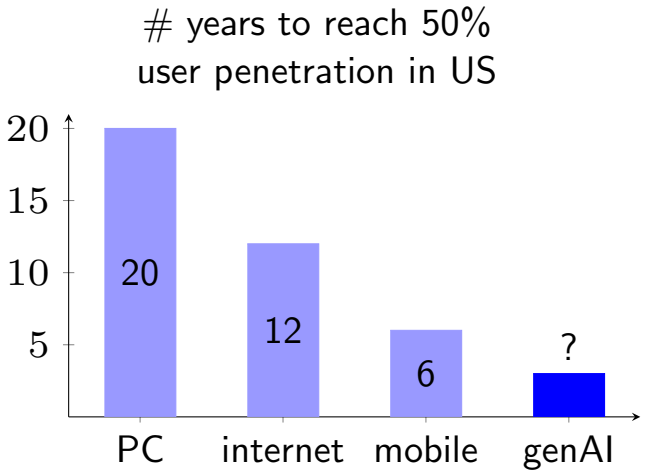


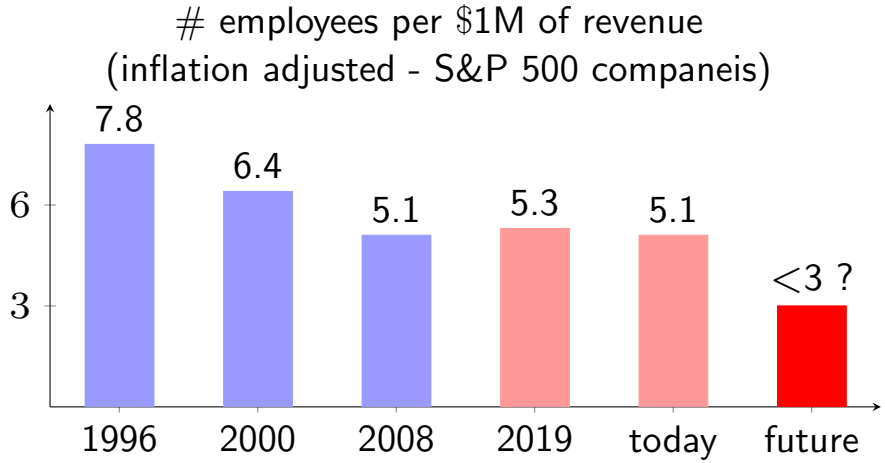
new AI applications over time



AI's transformative impact - adoption speed & economic potential

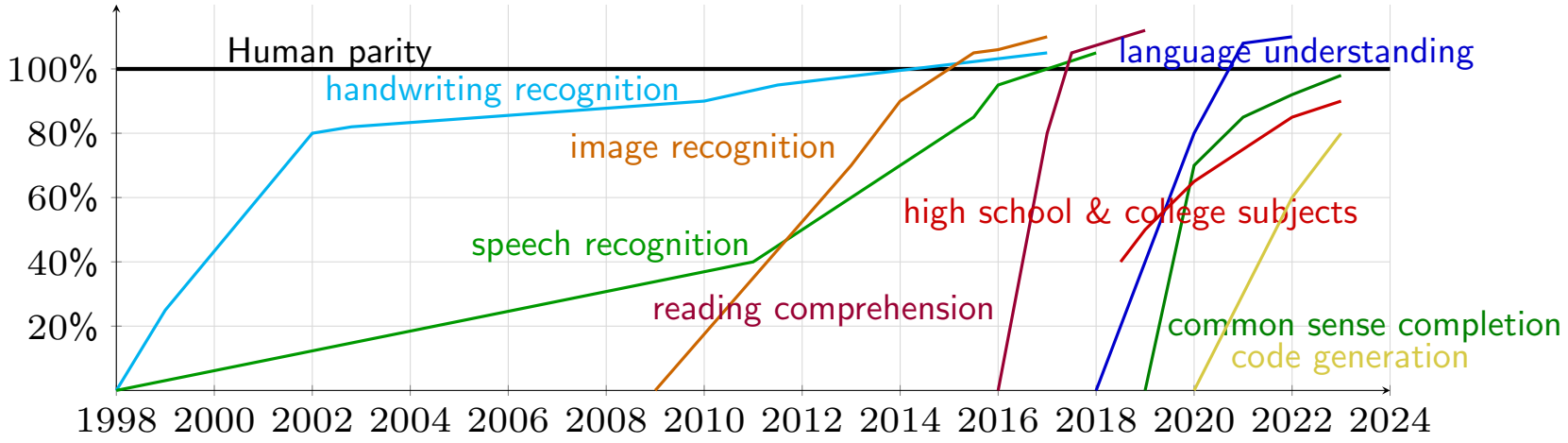
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater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





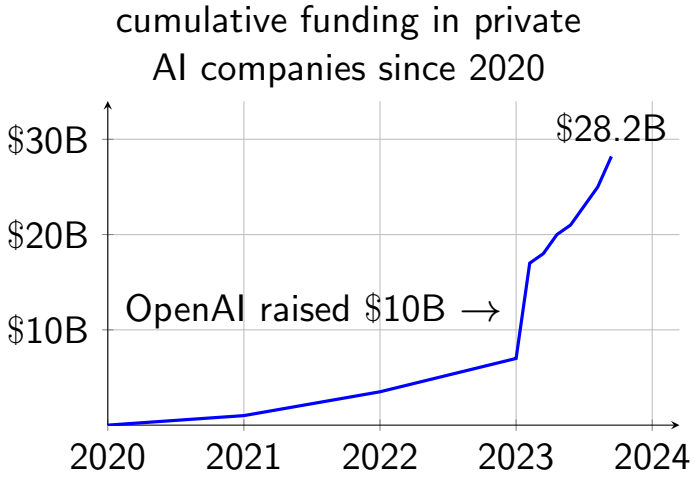
AI getting more & more faster

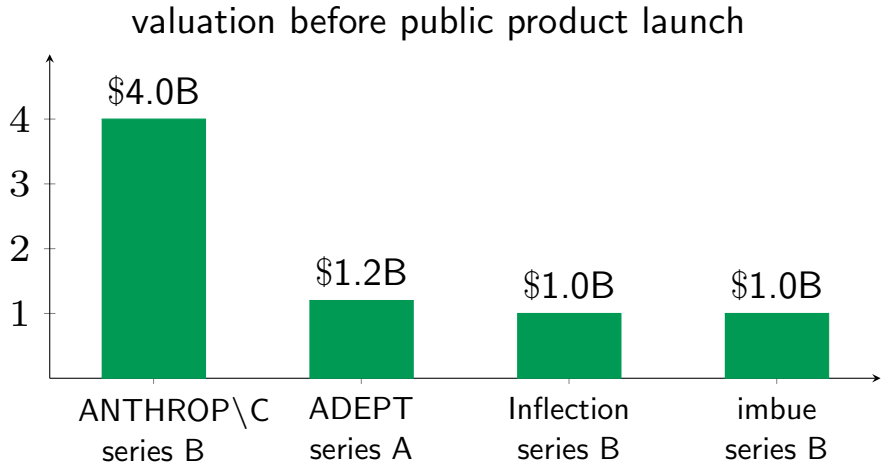
- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached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



Massive investment in AI

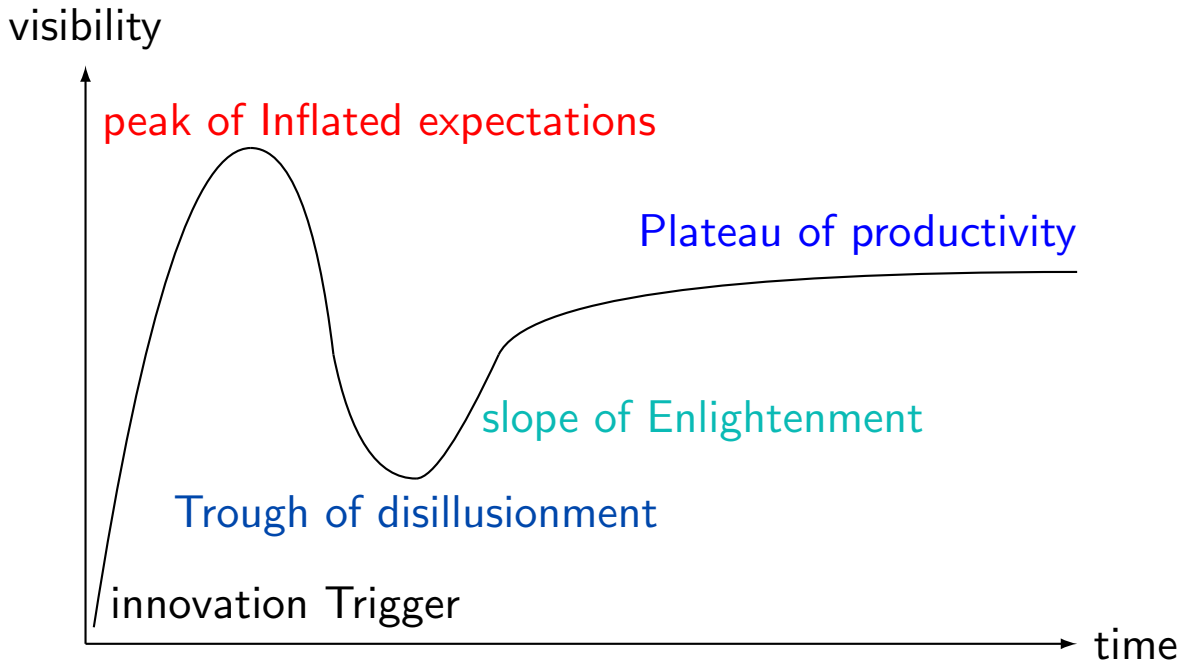
- *explosive growth* - cumulative funding skyrocketed reaching staggering $28.2B
- OpenAI - significant fundraising (=$10B) fueled rapid growth
- *valuation surge* - substantial valuations even before public products for stella companies
- *fierce competition for capital* among AI startups driving innovation & 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

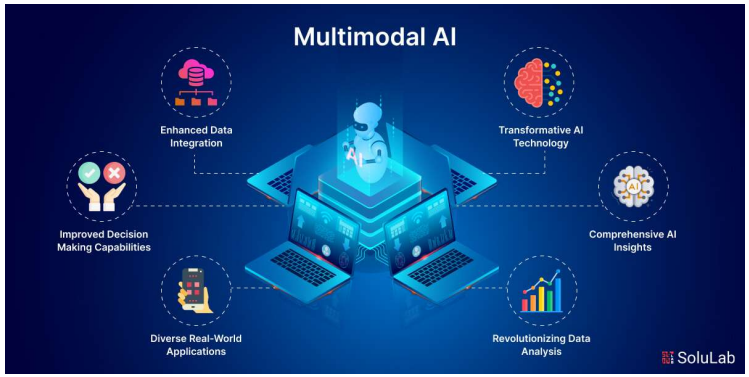
Yes & No

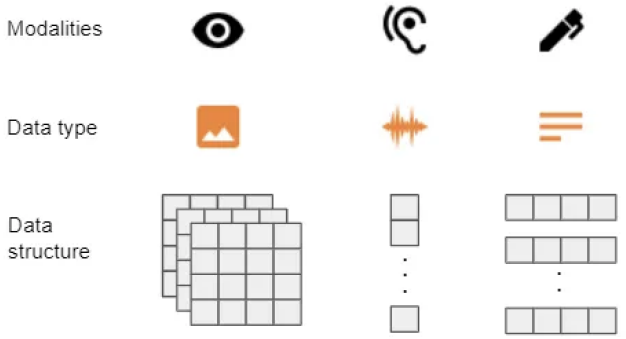
characteristics of hype cycles	speaker's views
value accrual misaligned with investment	• OpenAI still operating at a loss; business model *still* not clear • gradual value creation across broad range of industries and technologies ($e.g.$, CV, LLMs, RL) unlike fiber optic bubble in 1990s
overestimating 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

Multimodal AI Agents

Multimodal learning

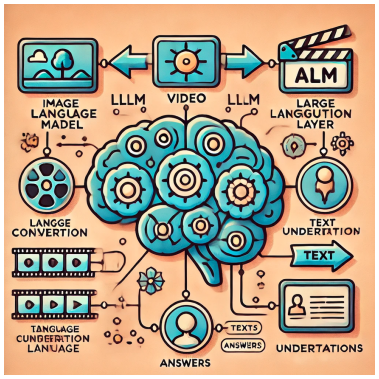
- understand information from multiple modalities, $e.g.$, text, images, audio, video
- representation learning methods
  - combine two representations or learn multimodal representations simultaneously
- applications
  - images from text prompt, videos with narration, musics with lyrics
- collaboration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





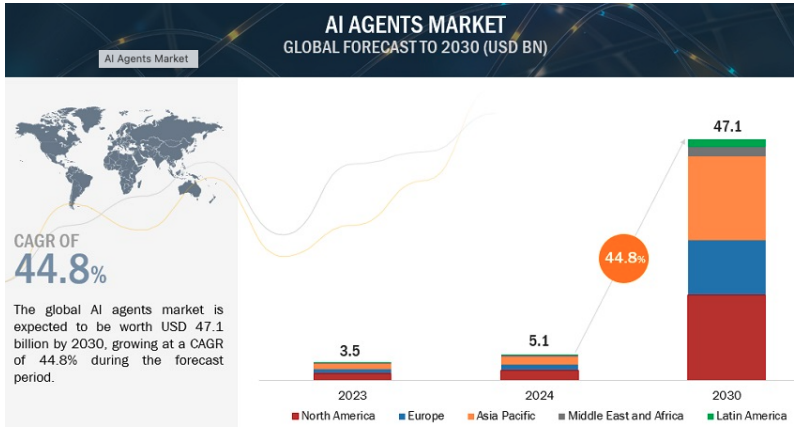
Implications of success of LLMs

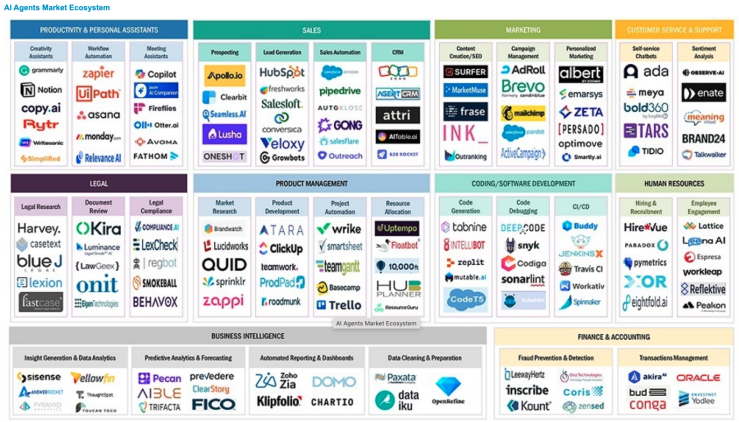
- many researchers change gears towards LLM
  - from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not only about NLP . . .* humans have . . .
  - evolved and optimized natural language structures for eons
  - handed down knowledge using this natural languages for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



Multimodal AI (mmAI) - definition & history

- mmAI - systems processing & integrating data from multiple sources & modalities, 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 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



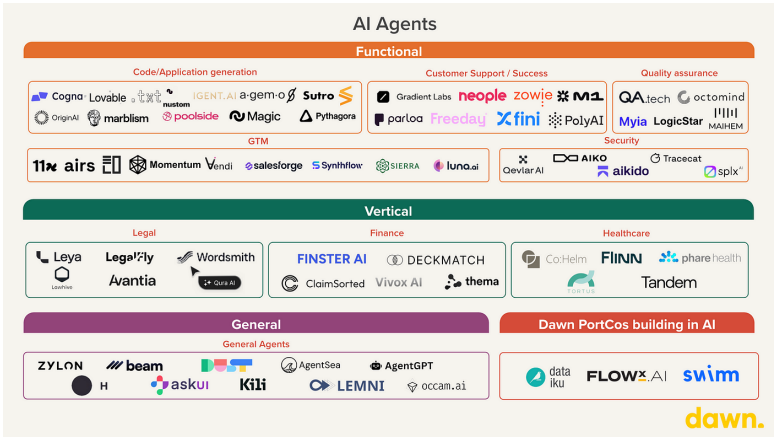


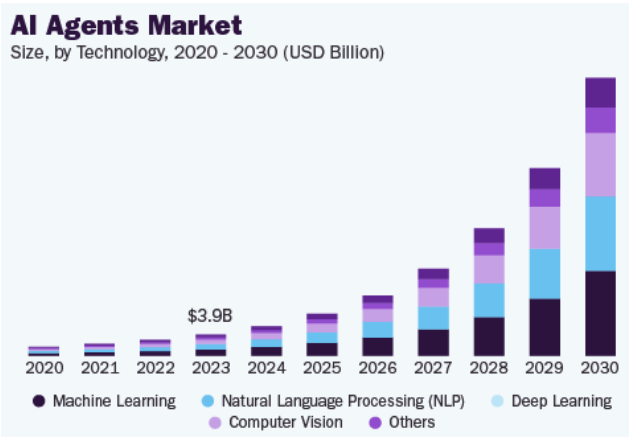
mmAI Technology

- 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

AI agents powered by multimodal LLMs

- foundation
  - integrate multimodal AI capabilities for enhanced interaction & decision-making
- components
  - perceive environment through multiple modalities (visual, audio, text), process using LLM technology, generate contextual responses & take actions
- capabilities
  - understand complex environments, reason across modalities, engage in natural interactions, adapt behavior based on context & feedback





AI agents - Present & Future

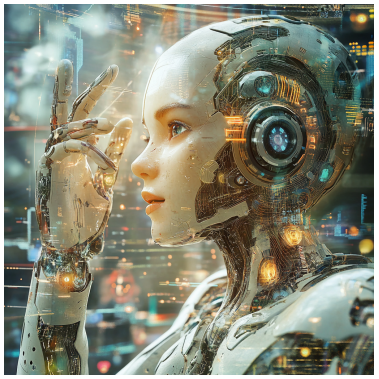
- emerging applications
  - scientific research - agents analyzing & running experiments & generating hypotheses
  - creative collaboration - AI partners in design & art combining multiple mediums
  - environmental monitoring - processing satellite sensor data for climate analysis
  - healthcare - enhanced 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 AI agents - seamless integration into everyday devices
  - highly tailored personalized experience - in education, entertainment & healthcare

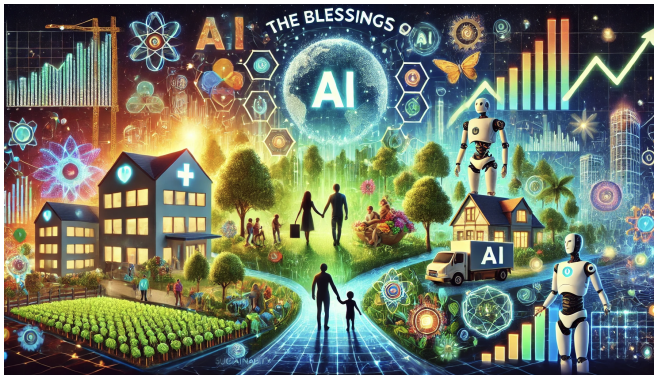
Empowering Humanity for Future Enriched by AI

Blessings & Curses of AI

Blessings

- advancements in healthcare & improved quality of life
  - much faster & more accurate diagnosis, far superior personalized medicine, accelerated drug discovery, assistive technologies
- economic growth & efficiency
  - automation to increase productivity and reduce cost, far superior decision-making
- environmental solutions
  - climate change prediction, global warming effect mitigation, solutions for sustainability
- safety & security
  - natural disaster prediction & relief, cybersecurity





Curses

- job displacement & overall impacts on labor market
  - millions of jobs threatened, wealth gap widened
- bias & inequality, misinformation & manipulation
  - existing human biases, both conscious and unconscious, perpetuated through AIs, asymmetric accessibility to advanced AI technologies by nations & corporations
- ethical dilemmas
  - infringing privacy & human rights, accountability for weapon uses and damages by AI
- environmental costs
  - significant energy for training AI models, waste generated by obsolescent AI hardware

Salzburg Global Seminar

KFAS-Salzburg Global Leadership Initiative

- "Uncertain Futures and Connections Reimagined: Connecting Technologies" - 41 global leaders convened from 4-Dec to 8-Dec, 2024 @ Schloss Leopoldskron in Salzburg, Austria
- My working group was "Technology, Growth, and Inequality: The Case of AI"
  - International Cooperation Officer (Portugal)
  - Gender Equality, Disability Inclusion Consultant, UN Women (Lithuania)
  - Assistant Professor @ Lincoln Alexander School of Law (Canada)
  - Research Associate @ Luxembourg Institute of Socio-Economic Research
  - Policy Officer & Delegation of the EU Union (India)
- blog: Bridging Technology & Humanity - Reflections from Lyon, Salzburg, and München

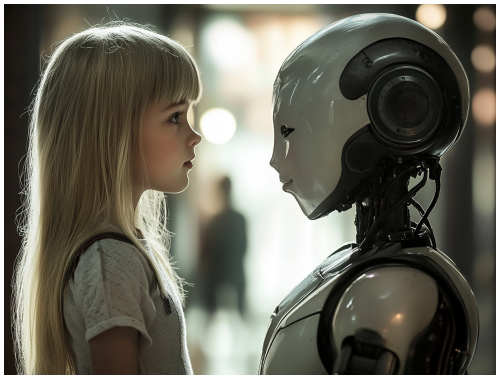
KFAS-Salzburg Global Leadership Initiative

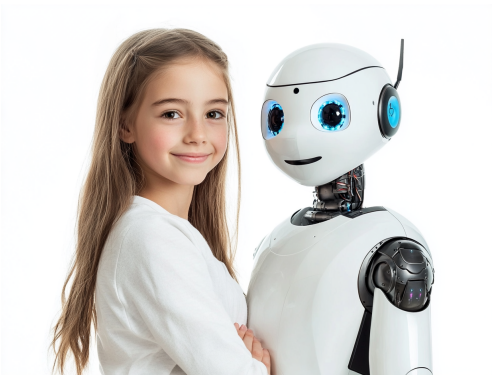
Salzburg Global photo collections

Empowering Humanity

AI capacity building - scientists, engineers & practitioners

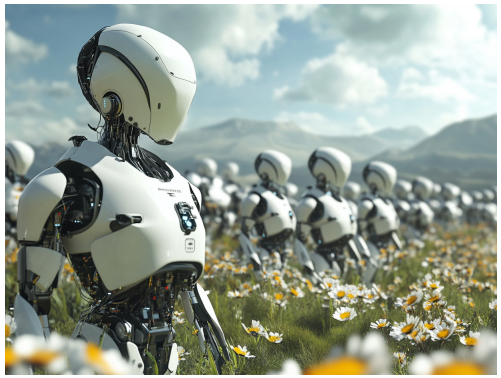
- ethics and responsible AI education or campaign via interdisciplinary collaboration
  - foster continuous learning programs on AI risks, bias & societal impacts
- bias detection & mitigation
  - bias-detection tools to identify & reduce discrimination in data & models
  - regular fairness audits
- transparency & explainability
  - explainable AI (xAI) techniques, frameworks like Model Cards for transparency
- environmental impact awareness
  - reduce AI's carbon footprint, advocate for sustainable AI development practices





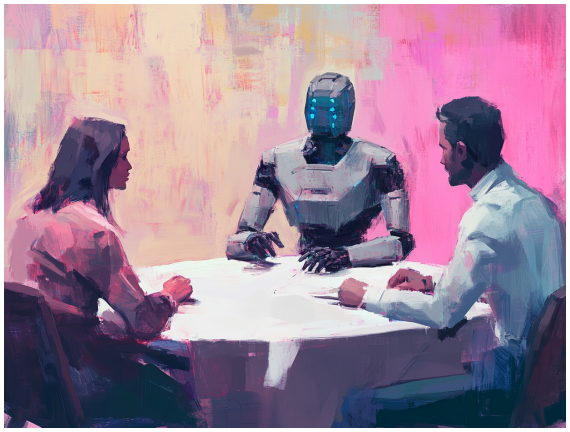
AI capacity building - lawmakers & policy makers

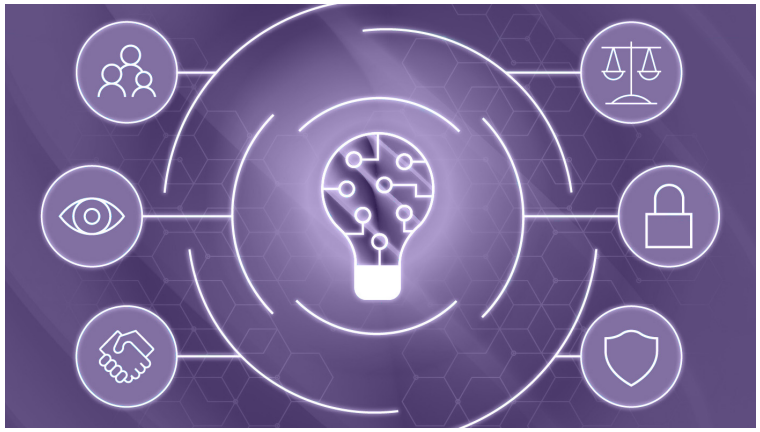
- problems
  - difficulties in understanding of rapidly evolving AI technologies
  - lead to reactive or insufficient regulation
- proposed solutions
  - develop comprehensive regulatory frameworks addressing transparency, bias & privacy concerns
    - gender bias, racial bias, hallucinations
  - foster public debates on ethical AI use & societal implications
  - introduce policies to limit spread of AI-generated misinformation,



Participatory social agreements

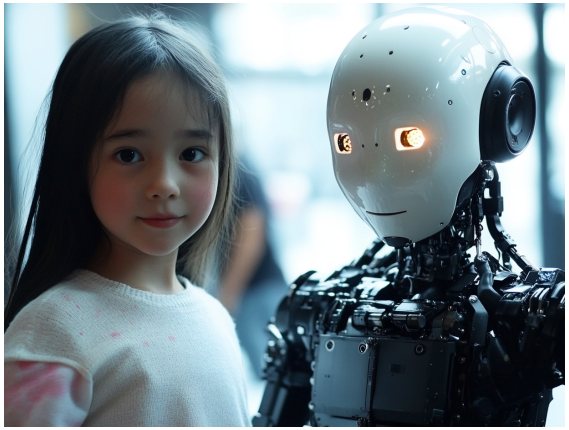
- open data frameworks including data sovereignty, regulation of data transfer, storage & localization
- corporate social responsibility, extra-territorial obligations & environmental protection
  - including outside the jurisdiction of the country
- labour and employment displacements, tax cuts & algorithmic impact assessments
  - including remedies for AI harms and enforcements





Reclaiming technology for Humanity

- strategic approach to AI development
  - *leverage very technologies alienating humans to strengthen human connection*
  - transform automation from replacement to *enhancement of human capabilities*
  - leverage technological scale to address fundamental human needs
- *paradigm shift* in technological implementation
  - recognize the duality of advanced technologies
  - *systematically channel AI capabilities toward human-centric solutions*
  - convert technological challenges into opportunities for human advancement



Appendix

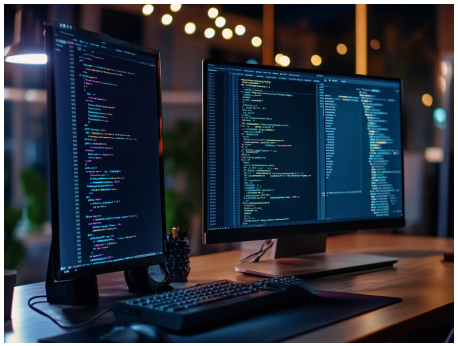
AI Products

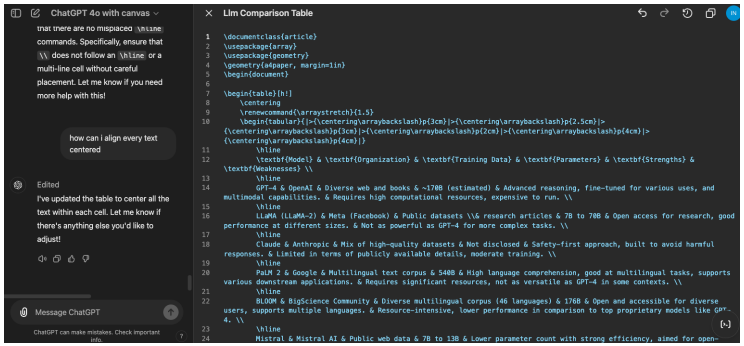
AI product development - trend and characteristics

- *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

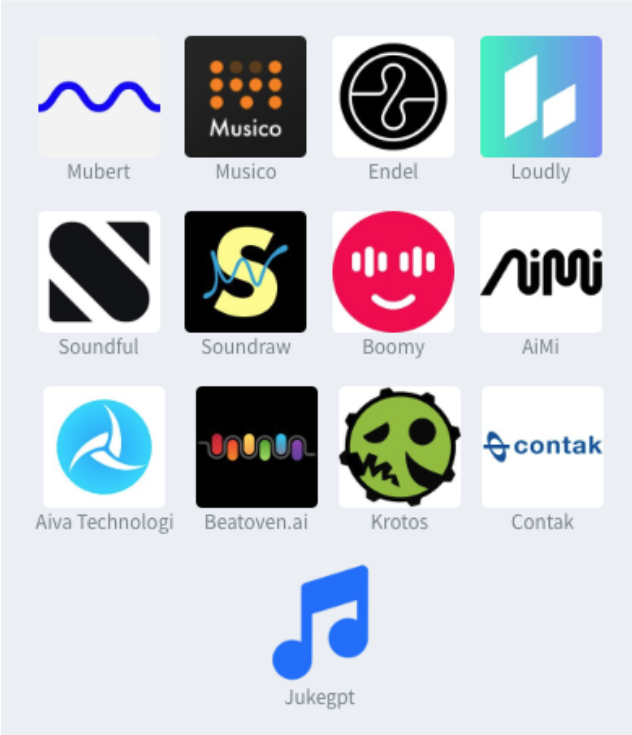
Future of co-pilot products

- 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*

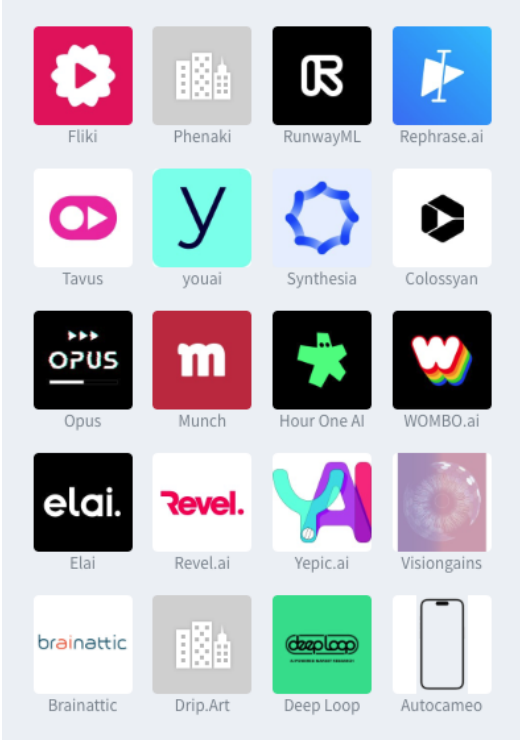


Other AI products - audio/video/text

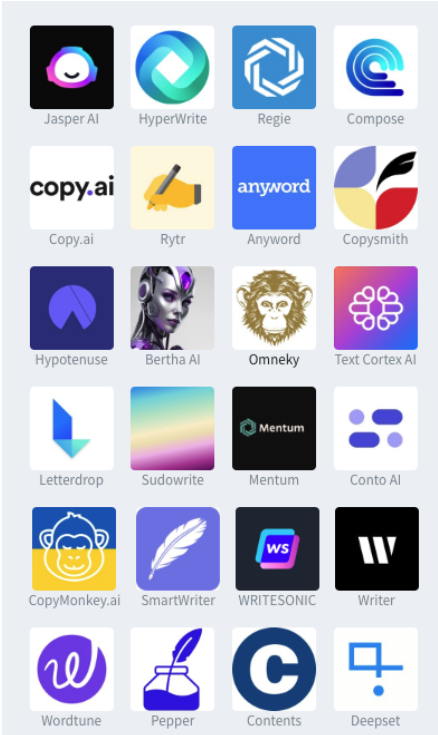
audio



vidio

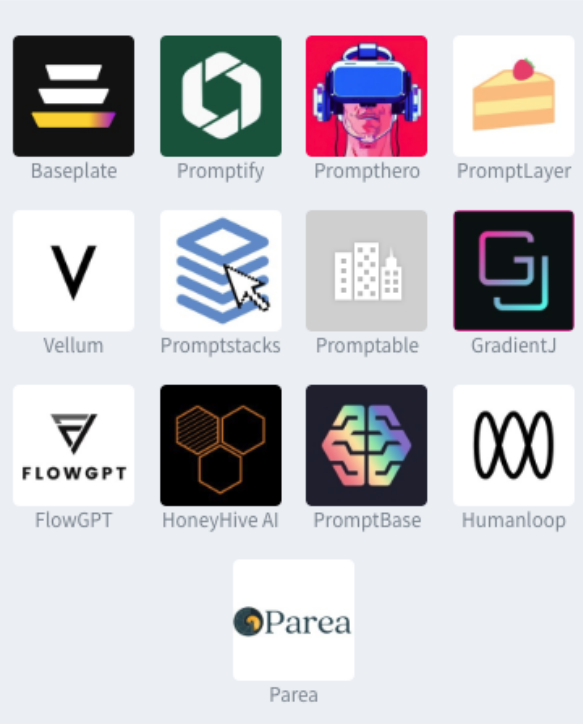


text

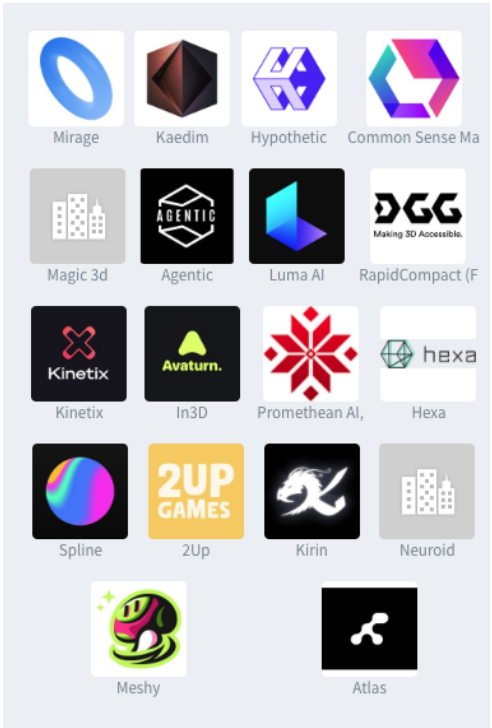


Other AI products - LLM/gaming/design/coding

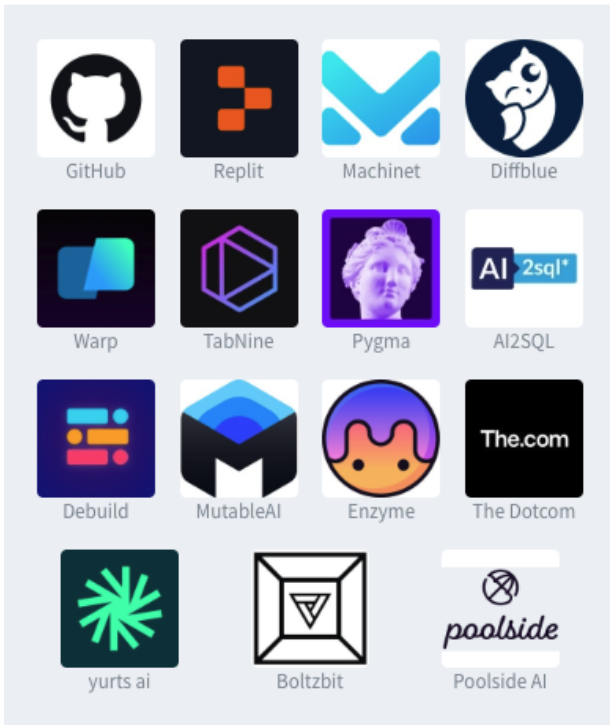
LLM



gaming & design



coding



Serendipities around AIs

Serendipity or inevitability?

- What if Geoffrey Hinton had not been a persistent researcher?
- What if symbolists won AI race over connectionists?
- What if attention mechanism did not perform well?
- What if Transformer architecture did not perform super well?
- What if OpenAI had not been successful with ChatGPT in 2022?
- What if Jensen Huang had not been crazy about making hardware for professional gamers?

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

Important Questions to be Asked

Some important questions around AI

- why human-level AI in the first place?
- what lies in very core of DL architecture? what makes it work amazingly well?
- biases that can hurt judgement, decision making, social good?
- ethical and legal issues
- consciousness, knowledge, belief, reasoning
- future of AI

Human-level AI?

Why human-level in the first place?

- lots of times, when we measure AI performance, we say
  - how can we achieve human-level performance, $e.g.$, CV models?
- why human-level?
  - are all human traits desirable? are humans flawless?
  - aren't humans still evolving?
- advantage of AI over humans
  - $e.g.$, self-driving cars can use extra eyes, GPS, computer network
  - $e.g.$, recommendation system runs for hundreds of millions of people overnight
  - AI is available 24 / 7 while humans cannot
    . . . critical advanages for medical assitance, emergency handling
  - AI does not make more mistakes because task is repetative and tedius
  - AI does not request salary raise or go on strike

What makes DL so successful?

Factors constributing to astonishing success of DL

- analysis based on speaker's mathematical, numerical algorithmic & statistical perspectives considering hardware innovations

  30% universal approximation theorem? - (partially) yes! but that's not all

  - function space of neural network is *dense* (math theory), $i.e.$, for every $f : \mathbf{R}^n \to \mathbf{R}^m$, exists $\langle f_n \rangle$ such that $\lim_{n\to\infty} f_n = f$

  25% architectures/algorithms tailored for each class of applications, $e.g.$, CNN, RNN, Transformer, NeRF, diffusion, GAN, VAE, . . .

  20% data labeling - expensive, data availability - unlimited web text corpus

  15% computation power/parallelism - AI accelerators, $e.g.$, GPU, TPU & NPU

  10% rest - Python, open source software, cloud computing, MLOps, . . .

Why do we see sudden leap in LLM performance?

Probability inferenced sequence is correct

- assume
  - t_i - ith token
  - p_i - probability that t_i is correct
  - ρ_i - correlation coefficient between t_{i-1} & t_i
  - $\tilde{p}_k$ - probability that $(t_1, \ldots, t_k)$ are correct
- recursion

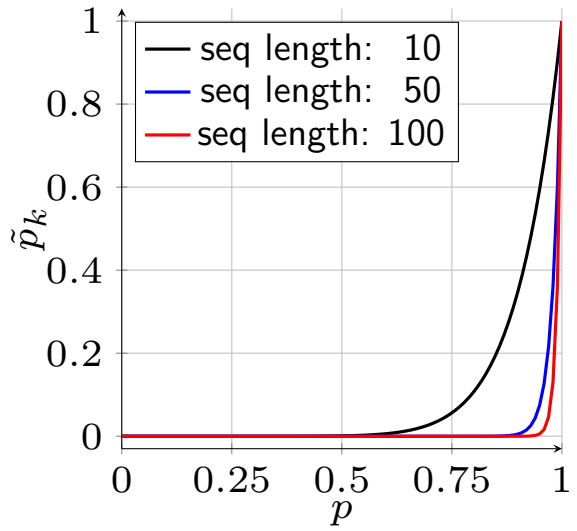
$$\rho_i = \frac{\tilde{p}_i - \tilde{p}_{i-1} p_i}{\sqrt{\tilde{p}_{i-1}(1 - \tilde{p}_{i-1}) p_i (1 - p_i)}}$$

$$\Leftrightarrow \qquad \tilde{p}_i = \tilde{p}_{i-1} p_i + \rho_i \sqrt{\tilde{p}_{i-1}(1 - \tilde{p}_{i-1}) p_i (1 - p_i)}$$

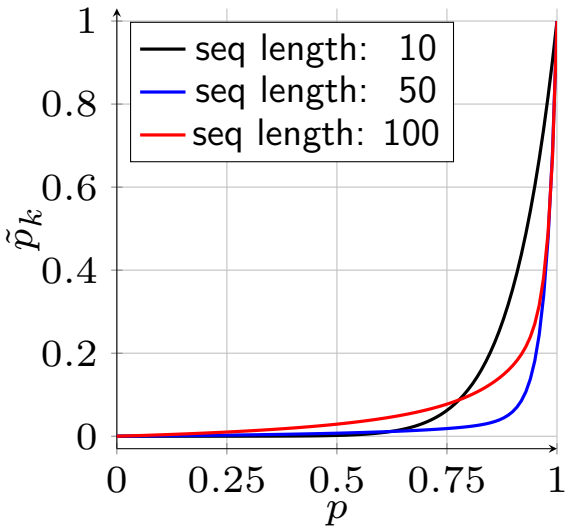
Dramatic improvement of LLM near saturation

- do simulations for both independent & dependent cases
  - assume p_i are same for all i
- (for both cases) sequence inference improves dramatically as p approaches 1
- this explains *why we have observed sudden dramatic performance improvement of certain seq2seq learning technologies*, $e.g.$, LLM

when tokens are independent



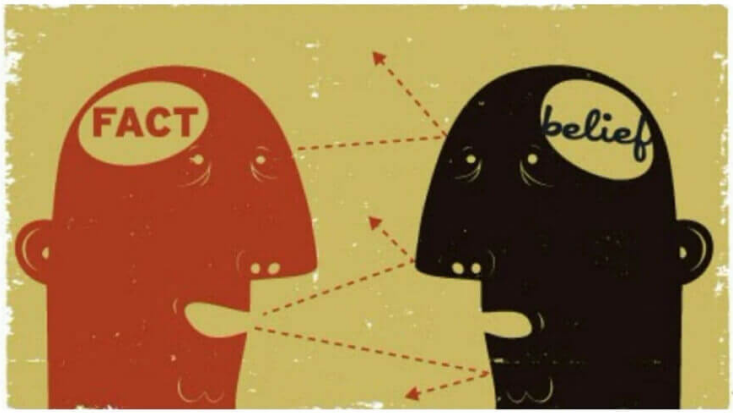
when tokens are dependent



Biases - by Humans & Machines

Cognitive biases

- cognitive biases [Kah11]
  - confirmation bias, availability bias
  - hindsight bias, confidence bias, optimistic bias
  - anchoring bias, halo effect, framing effect, outcome bias
  - belief bias, negativity bias, false consensus,



LLM biases

- plausible with LLM
  - availability bias - baised by imbalancedly available information
    - LLM trained by imbalanced # articles for specific topics
  - belief bias - derive conclusion not by reasoning, but by what it saw
    - LLM eaisly inferencing what it saw, $i.e.$, data it trained on
  - halo effect - overemphasize on what prestigious figures say
    - LLM trained by imbalanced # reports about prestigious figures
- similar facts true for other types of ML models,
  - $e.g.$, video caption, text summarization, sentiment analysis
- cognitive biases only human represent
  - confirmation bias, hindsight bias, confidence bias, optimistic bias, anchoring bias, negativity bias, framing effect

Ethical and Legal Issues

Ethics - possibilities & questions

- AI can be exploited by those who have bad intention to
  - manupilate / deceive people - using manupilated data corpus for training
    - $e.g.$, spread false facts
  - induce unfair social resource allocation
    - $e.g.$, medical insurance, taxation
  - exploit advantageous social and economic power
    - $e.g.$, unfair wealth allocation, mislead public opinion
- AI for Good - advocated by Andrew Ng
  - $e.g.$, public health, climate change, disaster management
- should scientists and engineers be morally & politically conscious?
  - $e.g.$, Manhattan project

Legal issues with ethical consideration - (hypothetical) scenarios

- scenario 1: full self-driving algorithm causes traffic accident killing people
  - who is responsible? - car maker, algorithm developer, driver, algorithm itself?
- scenario 2: self-driving cars kill less people than human drivers
  - $e.g.$, human drivers kill 1.5 people for 100,000 miles & self-driving cars kill 0.2 people for 100,000 miles
  - how should law makers make regulations?
  - utilitarian & humanistic perspectives
- scenario 3: someone is not happy with their data being used for training
  - "The Times sues OpenAI and Microsoft over AI use of copyrighted work" (Dec. 2023)
  - "Newspaper publishers in California, Colorado, Illinois, Florida, Minnesota and New York said Microsoft and OpenAI used millions of articles without payment or permission to develop ChatGPT and other products" (Apr. 2024)

Consciousness

Consciousness

- what is consciousness, anyway?
  - recognizes itself as independent, autonomous, valuable entity?
  - recognizes itself as living being, unchangeable entity?
- no agreed definition on consciousness exists yet
  . . . and will be so forever
- does it have anything to do with the fact that humans are biologically living being?
- is SKYNET ever plausible (without someone's intention)?
  - can AI have *desire* to survive (or save earth)?

Utopia or dystopia - futile debates

- not important questions (at all) *I think . . .*
- what we should focus on is not the possibilities of doomday or Judgment Day, but rather
  - our limits on controlling unintended impacts of AI
  - *misuse* by (greedy and bad) people possessing social, economic & political power
  - *social good and welfare impaired* by (exploiting) AI
- should concern
  - choice among utilitarianism, humanism, justice & equity
  - amend or improve laws and regulations
  - address ethical issues caused by AI

Knowledge, Belief, and Reasoning of AI

Does AI (or LLM) have knowledge or belief? Can it reason?

What categories of questions should they be in?

engineering, scientific, philosophical, cognitive scientific . . . ?

Three surprises of LLM

- LLM is very different sort of animal . . . except that it is *not* an animal!
- *unreasonable* effectiveness of data [HNF09]
  - *performance scales with size of training data*
  - *qualitative leaps* in capability as models scale
  - tasks demanding human intelligence *reduced to next token prediction*
- focus on third surprise

  "conditional probability model looks like human with intelligence"

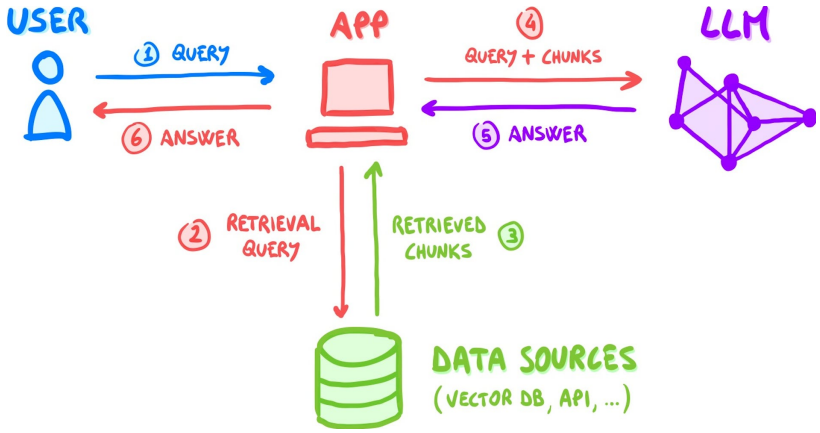
  - making vulnerable to anthropomorphism
- examine it by throwing questions
  - *"does LLM have knowledge and belief?"*
  - *"can it reason?"*

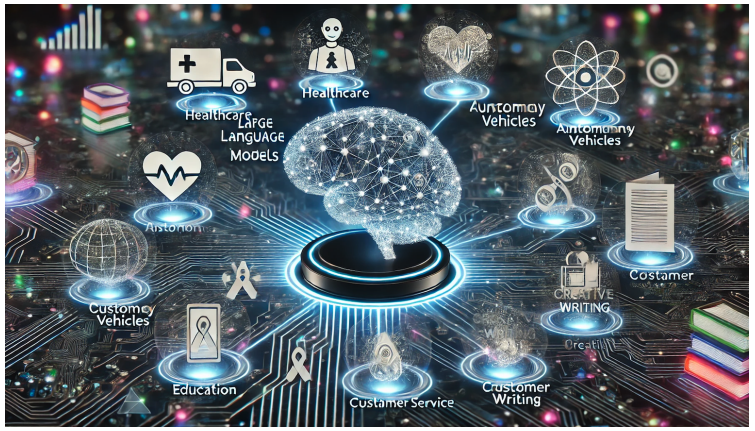
What LLM really does!

- given prompt "the first person to walk on the Moon was", LLM responds with "Neil Armstrong". . . strictly speaking
  - it's *not* being asked *who* was the first person to walk on the Moon
  - what are being *really* asked is *"given statistical distribution of words in vast public corpus of text, what words are most likely to follow 'The first person to walk on the Moon was'?"*
- given prompt "after ring was destroyed, Frodo Baggins returned to", LLM responds with "the Shire"
  - on one level, it seems fair to say, you might be testing LLM's knowledge of fictional world of Tolkien's novels
  - what are being *really* asked is *"given statistical distribution of words in vast public corpus of text, what words are most likely to follow 'After the ring was destroyed, Frodo Baggins returned to'?"*

LLMs or systems in which they are embedded?

- crucial to distinguish between the two (for philosophical clarity)
  - LLM (bare-bones model) - highly specific & well-defined function, which is *conditional probability estimator*
  - systems in which LLMs are embedded - question-answering, news article summarization, screenplays generation, language translation





How ChatBot works using LLMs?

- conversational AI agent does *in-context learning* or *few-shot prompting*
- for example,
  - when the user enters

    ```
    who is the first person to walk on the Moon?
    ```

  - ChatBot, LLM-embedded system, feeds the following to LLM

    ```
    User, a human, and BOT, a clever and knowledgeable AI agent.
    User:  what is 2+2?
    BOT: the answer is 4.
    User:  where was Albert Einstein born?
    BOT: he was born in Germany.
    User:  who is the first person to walk on the Moon?
    BOT:
    ```

Knowledge, belief & reasoning around LLM

- *not* easy topic to discuss, or even impossible because
  - we do *not* have agreed definition of these terms especially in context of being asked questions like

    does LLM have belief?

    or

    do humans have knowledge?

- let us discuss them in two different perspectives
  - laymen's perspective
  - cognitive scientific perspective

Laymen's perspective on knowledge, belief & reasoning

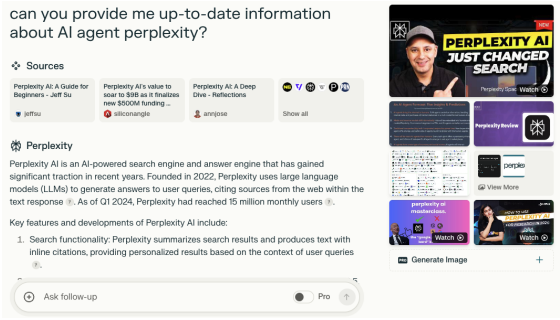
- does (good) LLM have knowledge?
  - Grandmother - looks like it cuz when instructed *"explaing big bang"*, it says
    " The Big Bang theory is prevailing cosmological model that explains the origin and evolution of the universe. . . . 13.8 billion years ago . . . "
- does it have belief?
  - Grandmother: I don't think so, $e.g.$, it does not believe in God.
- can it reason?
  - Grandmother: seems like it! $e.g.$, when asked *"Sunghee is a superset of Alice and Beth is a superset of Sunghee. is Beth a superset of Alice?"*, it says
    " Yes, based on information provided, if Sunghee is a superset of Alice and Beth is a superset of Sunghee, then Beth is indeed a superset of Alice . . . "
- can it reason to prove theorem whose inferential structure is more complicated?
  - Grandmother: I'm not sure. - actually, I don't know what you're talking about!

Knowledge

- could argue LLM "knows" which words follow which other words with high probability
- but, only *in context of capacity to distinguish truth from falsehood,* can we legitimately speak of "knowledge"!
- LLM(-embedded BOT)
  - can be said to *"encode"*, *"store"*, *or "contain"* knowledge
  - lacks means to use words "true" & "false" in all ways & in all contexts because . . .
  - does not inhabit the world we human language-users share!

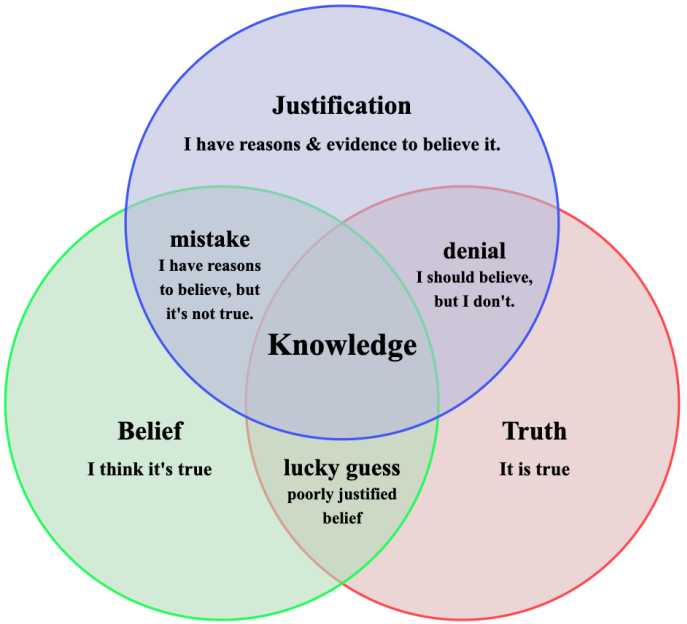
Belief

- nothing can count as *belief about the world* we share unless
  - it is against backdrop of "ability to update beliefs appropriately in light of evidence from *that world*" - (again) essential capacity to distinguish truth from falsehood
- change taking place in humans when acquiring or updating belief is
  - reflection of their nature as language-using animals inhabiting shared world with community of language-users
- then, *what if LLM-embedded system updates LLM with outside world information?*
  - even so, when interacting with AI systems based on LLMs, these grounds are *absent*!



Cognitive scientific perspective on knowledge

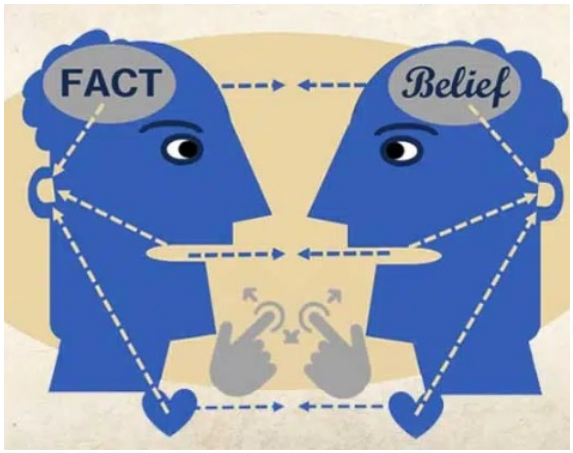
- does LLM have knowledge?
  - I don't think so.
- why?
  - when asked *"who is Tom Cruise's mother?"*, it says *"Tom Cruise's mother is Mary Lee Pfeiffer."* However, this is nothing but *"guessing" by conditional probability model the most likely following words after "Tom Cruise's mother is."*
  - so *we cannot say it really knows the fact!*



Cognitive scientific perspective on belief

- for the discussion
  - we do not concern *any specific belief*
  - we concern prerequisites for ascribing any beliefs to AI system
- so does it have belief?
  - when a human being takes to Wikipedia and confirms some fact, what happens is not her language model update, but

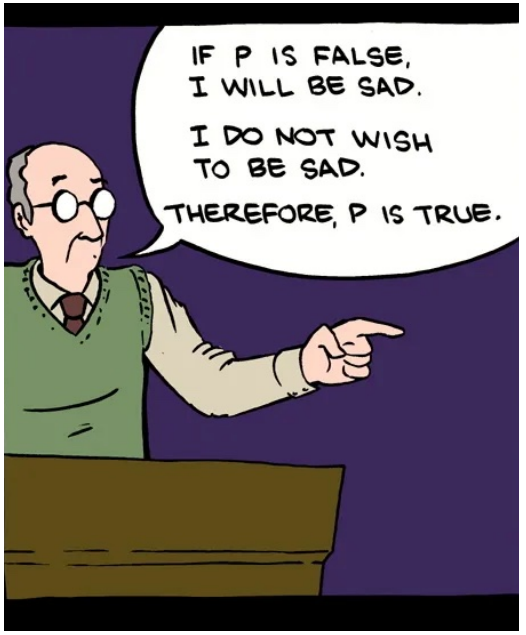
    reflection of her nature as language-using animal inhabiting shared world with a community of other language-users.
  - LLM does not have this ground, essential consideration when deciding whether it *really* had beliefs.
  - so *no, LLM cannot have belief!*



Cognitive scientific perspective on reasoning

- note reasoning is *content neutral*
  - $e.g.$, following logic is perfect regardless of truth of premises
  - hence, no access to outside world does *not* disqualify
- when asked *"if humans are immortal, would Socrates have survived today?"*, LLM says

  " . . . it's logical to conclude that Socrates would likely still be alive today. . . . "
- is there fundamental difference compared to *true* reasoning?
- moreoever, LLM can *mimic even multi-step reasoning whose inferencing structure is complicated* using *chain-of-thoughts prompting*, $i.e.$, *in-context learning* or *few-short prompting*,



Simple example showing LLM not possessing knowledge

- User

  "Who is Tom Cruise's mother?"

- LLM(-embedded question-answering system) (as of Jan 2022)

  "Tom Cruise's mother is Mary Lee Pfeiffer. She was born Mary Lee South. . . . Information about his family, including his parents, has been publicly available, *. . . "*

- User

  "Who is Mary Lee Pfeiffer's son?"

- LLM(-embedded question-answering system) (as of Jan 2022)

  "As of my last knowledge update in January 2022, I don't have specific information about Mary Lee Pfeiffer or her family, including her son. *. . . "*

Risk of anthropomorphization

- unfortunately, contemporary LLMs are *too powerful, too versatile, and too useful to accept previous arguments!*
- maybe, it is o.k. for laymen to (mistakenly) anthropomorphize LLM(-embedded systems)
- however, *imperative for AI researchers, scientists, engineers & practitioners* to have rigorous understanding in these aspects especially when
  - talk to or advise *policy makers, media,* $etc.$
  - consult or collaborate with professionals in areas such as *philosophy, ethics, law,* $etc.$

  $e.g.$, to address and prepare negative soceital and economic impacts

Moral

- AI, $e.g.$, LLM, shows incredible utility and commercial potentials, hence we should
  - make informed decisions about trustworthiness and safety
  - avoid ascribing capacities they lack take best usage of remarkable capabilities of AI
- today's AI is so powerful, so (seemingly) convincingly intelligent
  - obfuscate mechanism
  - actively encourage *anthropomorphism* with philosophically loaded words like "believe" and "think"
  - easily mislead people about character and capabilities of AI
- matters not only to scientists, engineers, developers, and entrepreneurs, but also
  - *general public, policy makers, media people*

Selected References & Sources

Selected references & sources

- Daniel Kahneman, Thinking, Fast and Slow, 2011
- T. Kuiken, Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight, 2023
- S. Yin, et. al., A Survey on Multimodal LLMs, 2023
- M. Shanahan, Talking About Large Language Models, 2022
- A. Vaswani, et al., Attention is all you need, NeurIPS, 2017
- I.J. Goodfellow, . . . , Y. Bengio, Generative adversarial networks (GAN), 2014
- A.Y. Halevry, P. Norvig, and F. Pereira. Unreasonable Effectiveness of Data, 2009
- Stanford Vecture Investment Groups
- CEOs & CTOs @ starup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California

References

References

[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.

[HNF09] Alon Halevy, Peter Norvig, and Nanediri Fernando. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24:8 – 12, 05 2009.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.

[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.

[Sha23] Murray Shanahan. Talking about large language models, 2023.

[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.

Thank You